Chapter 2

Prediction of Mohs Hardness with Machine Learning Methods Using Compositional Features

Joy C. Garnett*

Fisk University, Department of Life and Physical Sciences, Nashville, Tennessee 37208, United States

Vanderbilt University, Department of Physics and Astronomy, Nashville, Tennessee 37212, United States

***E-mail: jgarnett@fisk.edu, joy.garnett@vanderbilt.edu**

Hardness, or the quantitative value of resistance to permanent or plastic deformation, plays a crucial role in materials design for many applications, such as ceramic coatings and abrasives. Hardness testing is an especially useful method because it is nondestructive and simple to implement and gauge the plastic properties of a material. In this study, I proposed a machine, or statistical, learning approach to predict hardness in naturally occurring ceramic materials, which integrates atomic and electronic features from composition directly across a wide variety of mineral compositions and crystal systems. First, atomic and electronic features, such as van der Waals, covalent radii, and the number of valence electrons, were extracted from composition. The results showed that this proposed method is very promising for predicting Mohs hardness with F1-scores >0.85. The dataset in this study included modeling across a larger set of materials and hardness values, which have never been predicted in previous studies. Next, feature importances were used to identify the strongest contributions of these compositional features across multiple regimes of hardness. Finally, the models that were trained on naturally occurring ceramic minerals were applied to synthetic, artificially grown single crystal ceramics.

Introduction

Hardness plays a key role in materials design for many industrial applications, such as drilling (*1*, *2*), boring (*3*, *4*), abrasives (*5*–*7*), medical implants (*8*–*10*), and protective coatings (*11*–*13*). Increased manufacturing demand fuels the drive for new materials of varying hardnesses, which makes the fundamental understanding of the physical origin of this property necessary. Hardness testing is a nondestructive measurement of a material's resistance to permanent or plastic deformation. One such hardness test is the Mohs scratch test, in which one material is scratched with

another of a specified hardness number between 1 and 10. Materials that are easily scratched, such as talc, are given a low Mohs number (talc's is 1) while materials that are highly resistant to plastic deformation and difficult to scratch, such as diamond, are given a high Mohs number (diamond's is 10).

In the 1950s, Tabor established that Mohs scratch hardness is associated with deformation during the plastic indentation process and found that indentation hardness rises monotonically about 60% for each increment of the Mohs scale (*14*, *15*). With this correlation, Tabor identified a relationship between Mohs hardness and Vickers and Knoop indentation hardness. Tabor then correlated the stress-strain characteristics of a material to the stress that produces plastic flow (*16*). The relationship between indentation hardness (H) and Mohs scratch-hardness number (M) is

$$\ln H = kM$$

where k=*1.6,* based on experimental data comparing the indentation hardness numbers found by Vickers or Knoop measurements to the Mohs hardness value. It is unclear which atomic, electronic, or structural factors contribute to k or hardness as a whole. So, identifying the key features of a material that are involved in hardness can broaden our understanding of the mechanism of plastic deformation, and therefore guide the design of novel materials.

The Mohs hardness of a material is influenced by many factors. Material hardness for single-crystal brittle materials like minerals can depend on the type of chemical bonding, which can affect a material's ability to start dislocations under stress (*17–19*). Materials low on the Mohs scale, such as talc (M = 1) and gypsum (M = 2), exhibit van der Waals bonding between molecular chains or sheets. Materials with ionic or electrostatic bonding have a larger Mohs hardness. Materials at the top of the Mohs scale, such as boron nitride (M = 9) and diamond (M = 10), have large covalent components. Covalent bonding restricts the start of dislocations under stress, producing a resistance to plastic deformation. Hardness is also related to the correlation of composition and bond strength (*20–24*). Light elements have extremely short and strong bonds, as do transition metals which have a high number of valence bonds. Higher Mohs hardness is correlated to high average bond length, high number of bonds per unit volume, and a higher average number of valence electrons per atom.

Typically, calculations of hardness include multiple length scales to account for atomic interactions that contribute to intrinsic hardness and microstructure which in turn contribute to extrinsic hardness. Computational methodologies combined with high-performance computing methods have been utilized to see deformation and compositional factors for hardness on multiple length scales (*25*). Specific methodologies include molecular dynamics (MD), density functional theory, and machine learning (ML). MD to compute indentation hardness involve the evolution of atomic configuration over an atomistic system of millions of atoms. By calculating the evolution of the system to observe the nucleation of dislocations, MD allow the investigation of the relation between mechanical properties and microstructure. Recently, deformation processes are commonly due to dislocation generation with respect to grain sizes (*26*, *27*) . This gives great insight into how atomic interactions and microstructure contribute to deformation, however there is a major challenge with respect to how that translates into indentation hardness. Specifically, atomic relaxations calculated using MD happen faster than experimental indentation rates. This is due to a lack of force fields, which produce unreliable interatomic potentials. To address these issues, MD approaches have previously been combined with ML to compute more accurate atomic forces and interatomic potentials (*28–31*). Even so, there would still be the major challenge of the

computational cost to implement a molecular dynamic methodology across hundreds of materials in a varied chemical and structural space.

Modeling hardness has been proven difficult since hardness is not a fundamental property of a material and cannot be directly evaluated from quantum mechanics across all crystal systems with one model. Complications previously found in energy-based calculations of other properties include sensitivity to bond length, overestimation in density functionals (*20*), finite-size supercell effects (*32*, *33*), and choice of exchange-correlation function (*32*, *33*), which leads to multiple issues in large-scale implementation across different crystal structures and varied chemical spaces. Issues include large computational costs when considering several materials at a time, especially when considering the costs of optimization and of energy determination of multiple deformations across one material class.

There have been multiple computational approaches to connect hardness to bond behavior. For instance, Gao et al. introduced the concept that hardness of polar covalent crystals is related to three factors (*23*): bond density or electronic density, bond length, and degree of covalent bonding. This approach utilized first principles calculations to uncover a link between hardness and electronic structure with multiple semiempirical relationships depending on the type of bonding in the material. The advantage of this approach is that this link that was demonstrated for 29 polar and nonpolar covalent solids could be extended across a broad class of materials. One disadvantage of this approach is that there are different semiempirical relationships of microhardness depending on if the material is a pure covalent, polar covalent, or a multicomponent solid.

Šimůnek and Vackář extends this concept of expressing the hardness of covalent and ionic crystals through the bond strength (*24*), which is determined by the number of valence electron in the component atoms, crystal valence electron density, the number of bonds, and bond length between pairs of atoms. They predicted the hardness of 30 covalent and ionic crystals including binary A^{III}-B^{V} and A^{II}-B^{VI} compounds and nitride spinel crystals (C_3N_4 and Si_3N_4). While their results were close to experimental values for nitride spinels and A^{III}-B^{V} materials, there was deviation from experiment for the A^{II}-B^{VI} materials reported. One drawback of both methods is that they depend on first principles calculations, which can become computationally expensive when expanded to calculate all the bonds for hundreds of materials.

Mukhanov et al. circumvented *ab initio* calculations by utilizing thermodynamic properties to find a simple quantitative dependence of hardness and compressibility of 9 materials (*34*). They employ the standard Gibbs energy of formation and the Gibbs energy of atomization of elements in the material. In addition, they introduce the factors of bond rigidity and bond covalency, which are based on the electronegativites of the elements, as well as the ratio between the mean number of valence electrons per atom and the number of bonds with neighboring atoms. One advantage is that this method can be applied to a large number of compounds with various types of chemical bonding and structures. The hardness predictions for refractory crystalline compounds agree within 7% of experimental values. Another advantage is the flexibility of this method to calculate hardness as a function of temperature. However, there are factors that are estimated from experimental values of hardness of other materials. For instance, the coefficient of relative plasticity varies for elementary substances, compounds of second period elements, and compounds for period elements greater than 3. This coefficient is attributed to reflect the difference in bond strength depending on the elements of different periods, but it is unclear as to how directly atomic radii relates to this coefficient. While these relationships hold for superhard high-pressure phases, it may not hold true for softer materials.

Li et al. also circumvented *ab initio* calculations by using chemophysical parameters to predict the hardness of ionic materials (*35*, *36*). The chemophysical parameters of electronegativity values of elements in different valence states were used to relate the stiffness of the atoms, the electron-holding energy of an atom, and bond ionicity indicators to hardness in 8 superhard materials. The calculated hardness values are in good agreement with experimental data. However, it remains unclear if this relationship of electronegativity and hardness is only applicable for superhard materials or if it can be expanded to understand softer materials as well.

The thrust of this study is to combine all of these factors that have been theoretically connected to hardness and understand how they may interact with each other and contribute to the hardness of crystalline ceramic materials. Previously, these factors have been used to explain hardness across a small range of crystal structures, bonding frameworks, and hardness values. In this study, I look to expand these concepts to a large number of compounds with various types of chemical bonding types, structures, and compositions. These chemophysical parameters may interact with each other to predict a range of hardness values. These factors, specific to superhard bonding, may or may not equally apply to other bonding frameworks, which are either noncubic or not purely covalent.

To circumvent the issues found in solely energy-based calculations, machine or statistical learning offers a less computationally expensive method to improve predictions of material properties and accelerate the design and development of new materials. Recently, ML methods applied to existing data have been proven effective for predicting hard-to-compute material properties at reduced time, cost, and effort (*37*–*43*). Predictive models based on experimental data have proven to be extremely powerful in materials research. Examples include the prediction of the intrinsic electrical breakdown field of insulators in extreme electric fields (*44*, *45*), the crystal structure classification of binary *sp*-block transition metal compounds (*46*–*48*), the prediction of Heusler structures based on compositional features (*49*), and the prediction of band gaps of insulators (*50*–*54*). A major advantage of ML methods for rapid material property prediction on past data is their power to uncover quantitative structure-property relationships across varied compositional spaces.

Previous studies predicting various properties of materials with ML have used a broad range of chemo-structural descriptor fingerprints. Typically, the approach is to map a unique set of descriptors that act as fingerprints connecting a material to a property of interest. These descriptors range from composition to quantities based on quantum mechanical calculations. A set of materials informatics methods built strictly on compositional descriptors and experimental hardness data may be more effective at determining relationships concerning the hardness of materials than previous approaches. This study implements an approach to establish a set of ML algorithms to uncover connections between calculable atomic parameters and the Mohs hardness of single crystalline ceramic materials.

The application of ML requires a dataset of feature descriptors that relate the chemical composition of diverse crystals to their physical properties. In this study, the database is based on compositional quantities. The aim of this study is to predict mechanical properties from compositional features without the need for computationally heavy energy-based modeling. Along this line, I wanted to test how well ML models can be utilized to predict a comparative material property, such as Mohs hardness. Can one improve the prediction of hardness at a less computational expense and with greater fidelity than current methods? Are there identifiable atomic factors that contribute to plastic deformation that can be applied across a variety of crystal structures and

compositions? Is there a simple formalism that allows one to track the importance of atomic mechanisms as a function of hardness irrespective of the chemical complexity of the material?

In this study, ML is used to predict the hardness-related plastic properties of naturally occurring ceramic minerals. Using compositional-based features, the ML approach is able to predict Mohs hardness across a broad structural and chemical space. Specifically, 622 naturally occurring ceramic minerals were screened using the random forests (RFs) ensemble-based ML method, as well as support vector machines (SVMs). The results show that ML based purely on compositional features of crystalline ceramics gives better results across a more varied chemical space than previous methods. Moreover, the influence of atomic and electronic compositional features on the resulting Mohs hardness prediction is evaluated. Finally, to demonstrate the efficiency of this model, it was used to predict the Mohs hardness of 52 synthetic crystals with similar atomic and structural characteristics. The resulting classification models accurately differentiate regimes of hardness by identifying relevant and significant features that affect hardness, suggesting a connection between the existence of a common panel of compositional markers and material hardness or resistance to plastic deformation.

Methods

Datasets

In this study, the author trained a set of classifiers to understand whether compositional features can be used to predict the Mohs hardness of minerals with different chemical compositions, crystal structures, and crystal classes. The dataset for training and testing the classification models used in this study originated from experimental Mohs hardness data, their crystal classes, and chemical compositions of naturally occurring cermamic minerals reported in the Physical and Optical Properties of Minerals CRC Handbook of Chemistry and Physics and the American Mineralogist Crystal Structure Database (*55*, *56*). The database is composed of 369 uniquely named minerals. Due to the presence of multiple composition combinations for minerals referred to by the same name, the first step was to perform compositional permutations on these minerals. This produced a database of 622 minerals of unique compositions, comprising 210 monoclinic, 96 rhombohedral, 89 hexagonal, 80 tetragonal, 73 cubic, 50 orthorhombic, 22 triclinic, 1 trigonal, and 1 amorphous structure. An independent dataset was compiled to validate the model performance. The validation dataset contains the composition, crystal structure, and Mohs hardness values of 52 synthetic single crystals reported in the literature. The validation dataset includes 15 monoclinic, 8 tetragonal, 7 hexagonal, 6 orthorhombic, 4 cubic, and 3 rhombohedral crystal structures. Both datasets were processed by in-house Python scripts. The datasets for model development, evaluation, and validation have been uploaded as a dataset onto Mendeley Data (*57*). Histograms of the distributions indicating hardness values for both datasets are presented in Figures 1a and 1b.

Classes

The classification bins used in this study are based on relationships previously seen in the literature from the studies of Gao et al. and Šimůnek and Vackář calculations of Vickers hardness (*58*, *59*). Gao et al. showed a correlation of calculated bond lengths to calculated Vickers hardness values for binary and multicomponent oxides (*58*). The multicomponent oxides were broken down into systems of pseudobinary compounds to reflect the nature of bonding in the material. These calculations contain three groupings of hardness and bond length. For materials with bond lengths

greater than 2.5 Å, the Vickers hardness values were calculated to be under 5 GPa (Mohs value (0.991, 4]). For materials with bond lengths between 2 and 2.5 Å, the Vickers hardness values were calculated to be between 5 GPa and 12 GPa (Mohs value (4, 7]). For materials with bond lengths less than 2 Å, the Vickers hardness values were calculated to be between 12 GPa and 40 GPa (Mohs value (7, 10]). Similarly, Šimůnek and Vackář showed a correlation between bond length and calculated Vickers hardness (*59*). However, it was more binarized. For materials with bond lengths greater than 2.4 Å, the Vickers hardness values were calculated to be less than 6.8 GPa (Mohs value (0.991, 5.5]). For materials with bond lengths less than 2.4 Å, the Vickers hardness values were calculated to be greater than 6.8 GPa (Mohs value (5.5, 10]).

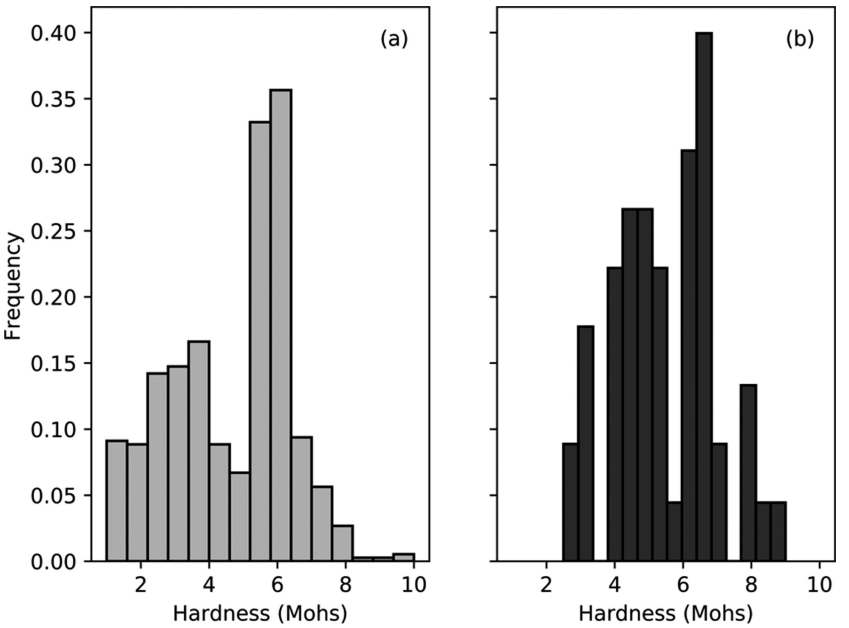

Figure 1. Histogram of the Mohs hardnesses of the datasets of (a) 622 naturally occurring mineral and (b) 52 artificially grown single crystals.

Based on these groupings, the calculated Vickers hardness values from both studies were converted to approximate Mohs hardness values and used as bins in this study. Minerals were grouped according to their Mohs hardness values as shown in Table 1. Separate binary and ternary classification groups were established as follows: Binary 0 (0.991, 5.5], Binary 1 (5.5, 10.0], Ternary 0 (0.991, 4.0], Ternary 1 (4.0, 7.0], and Ternary 2 (7.0, 10.0]. Thus, minerals of Mohs hardness between 0.991 and 5.5 were assigned to the Binary 0 group, minerals with hardness between 5.5 and 10 were assigned to the Binary 1 group, and so on.

Features

In this study, the author constructed a database of compositional feature descriptors that characterize naturally occurring materials, which were obtained directly from the Physical and Optical Properties of Minerals CRC Handbook (*55*). This comprehensive compositional-based dataset allows us to train models that are able to predict hardness across a wide variety of mineral compositions and crystal classes. Each material in both the naturally occurring mineral and artificial single crystal datasets was represented by 11 atomic descriptors, numbered 0 to 10, listed in Table 2 below. The elemental features are number of electrons, number of valence electrons, atomic number,

Pauling electronegativity of the most common oxidation state, covalent atomic radii, van der Waals radii, ionization energy (IE) of neutral atoms in the ground state (also known as the first IE), the atomic number (Z) to mass number (A) ratio, and density. These features were collected for all elements from the NIST X-ray Mass Attenuation Coefficients Database and the CRC Handbook of Chemistry and Physics (55, *60*).

Table 1. Binary and Ternary Classes in This Study Based on Mohs Hardness Values

Classes	
Binary (2-Class) Classification	
	Mohs Hardness
0	(0.991, 5.5]
1	(5.5, 10.0]
Ternary (3-Class) Classification	
0	(0.991, 4.0]
1	(4.0, 7.0]
2	(7.0, 10.0]

Table 2. List of All Primary Features

ID	*Name*	*Feature Description*
0	allelectrons_Total	Total number of electrons
1	density_Total	Total elemental density
2	allelectrons_Average	Atomic average number of electrons
3	val_e_Average	Atomic average number of valence electrons
4	atomicweight_Average	Atomic average atomic weight
5	ionenergy_Average	Atomic average first IE
6	el_neg_chi_Average	Atomic average Pauling electronegativity of the most common oxidation state
7	R_vdw_element_Average	Atomic average van der Waals atomic radius
8	R_cov_element_Average	Atomic average covalent atomic radius
9	zaratio_Average	Atomic average atomic number to mass number ratio
10	density_Average	Atomic average elemental density

The atomic averages of nine features were calculated for each mineral. The atomic average is the sum of the compositional feature (f_i) divided by the number of atoms (n) present in the mineral's empirical chemical formula, or

$$AA = \frac{1}{n}\Sigma_{i=1}^{n} f_i$$

Two additional feature descriptors were added based on the total number of electrons and the total of the elemental densities for each compound, for a total of 11 features listed in Table 2 below.

The features for this study were chosen based on factors implemented in previous methods to predict material hardness. The related factors from these studies were included as features that are easily calculated from the number of atoms in the empirical formula and elemental characteristics. The number of valence electrons per bond was included as a factor in Gao et al. (*23*), Šimůnek et al. (*24*), and Mukhanov et al. (*34*). In this study, the effect of valence electrons on hardness is considered by a simplified feature of atomic average of valence electrons. Atomic weight was included in this study since it is used to calculate molar volume, which was a factor in the Mukhanov et al. study as well (*34*). Atomic radii (covalent and van der Waals) were included as features in this study since they are related to the bond length factor in Gao et al. and the molar volume in Mukhanov et al. (*23*, *34*). Electronegativity was included in the feature set as the atomic average of Pauling electronegativity for all elements in a material's empirical formula. This atomic average is a simplified version of the electonegativity-derived factors of bond electronegativity, stiffness of atoms, and bond ionicity factors in Li et al. used to predict hardness (*35*, *36*).

In addition to features based on characteristics previously utilized in hardness calculations, three more features are also included: the first IE, the total number of all electrons, and the atomic number to mass ratio for each compound. Each of these have a connection to either the atomic radii or the strength of bonds of these materials. The first IE, or the amount of energy to remove the most loosely bound valence electron, is directly related to the nature of bonding in a material (*61*, *62*). According to Dimitrov and Komatsu (*61*), bond strength can be modeled as an electron binding force related to first IE through a Hooke's law potential energy relationship,

$$k = IE/2r_{eff}^2$$

where r_{eff} is the effective ionic radii. This has not been previously connected to hardness, so it is included as a novel feature in this study. Since hardness has been previously connected to bond strength, it makes sense that this could also be a related factor to mechanical properties like hardness.

The total number of electrons (both bonding and nonbonding) are also included in this study as a feature due to their contribution to atomic radii. As the number of electrons in inner shells increases, the repulsive force acting on the outermost shell electrons in a process known as shielding. This repulsive force increases the atomic radius, which could directly affect the bond length of a material. The atomic number to mass number ratio (Z/A) is directly related to the total electron cross-section, or the effective electronic energy-absorption cross section of an individual element. While it is commonly used to describe X-ray attenuation, it may also help in this case to describe an effective area of electronic activity that can contribute in a different context.

ML Models

In this study, nine supervised learning models were built and trained to classify hardness values in naturally occurring minerals and artificial single crystals. Specifically, I implemented RF and SVMs to predict Mohs hardness. This section reviews the models, optimization schema, feature importance calculations, and evaluation criteria utilized in this study.

RFs and Gini Feature Importance

Decision trees use decision-making rules implemented on features or attributes of the data to predict target properties. Major issues with decision trees are that they can be highly variable and sensitive to overfitting. To resolve this issue, one can employ ensemble methods, which implement multiple decision trees. Such methods include extremely randomized trees (extra trees) and RFs, which is one of the methods employed in this study. In a RFs ensemble, each tree is built on a bootstrap sample from the training dataset. At each node, a set of attributes is randomly selected from among all possible attributes to test the best split. For RF classifiers, the best split is chosen by minimizing variance, which leads to a reduction in misclassification. In the end, the value of a target property is the average of all of the predictions for that property.

This study not only predicts Mohs hardness based on feature descriptors, but also identifies which of these descriptors are most important to making the predictions for several RF models. To do this, the variable importance metric called Gini importance is employed to find the relative importances of a set of predictors based on the Gini index. The Gini index is commonly used as the splitting criterion in tree-based classifiers, as a function to measure the quality of a split. The reduction of the Gini index brought on by a feature is called the Gini importance or the mean decrease impurity. This decrease in node impurity is given by the equation

$$\hat{I}(t) = \sum_{j=1}^{J} \widehat{\phi_j}(t)(1 - \widehat{\phi_j}(t))$$

where $\hat{\phi}(t)$ is the class frequency for class j in node t (*63*, *64*).

To summarize, the Gini importance for a feature indicates that feature's overall discriminative value during the classification. If the decrease is low, then the feature is not important. An irrelevant variable has an importance of zero. The sum of the importances across all features is equal to 1. In this study, Gini feature importance is used to gauge the relative importance of a set of compositional-based features on binary and ternary RF classifications of Mohs hardness values.

SVMs

A SVM is a supervised ML method that fits a boundary or separating hyperplane around elements with similar features. The input features are mapped to a high-dimension feature space to produce a linear decision surface or hyperplane. This decision surface is based on a core set of points or support vectors. An SVM finds the linear hyperplane with the maximum margin in the higher-dimensional feature space.

Considering a training dataset of n input feature-label pairs (x_i, y_i), $i = 1 \ldots n$, the SVMs require the solution of the following optimization problem:

$$\min_{w,b,\zeta} \frac{1}{2} w^T w + C \sum_{i=1}^{l} \zeta_i$$

subject to

$$y_i(w^T\phi(x_i) + b) \geq 1 - \zeta_i,$$

$$\zeta_i \geq 0$$

where w is the set of weights for each feature or the weight vector, ϕ is the radial function that maps the data into a higher-dimensional space, b is bias, ζ_i is a slack variable to allow some misclassifications while maximizing the margin and minimizing the total error, and C is the soft margin cost function that controls the influence of each support vector (*65*, *66*). Due to the construction of the hyperplane on these support vectors, it is not sensitive to small changes to the data. This robustness to data variation means the SVM can generalize quite well. Also, the construction of the hyperplane results in complex boundaries, which resists overfitting. The SVM algorithms utilized in this study were implemented by the Scikit-Learn Python package (*67*).

Feature spaces are not always readily linearly separable. In order to improve separability within the feature space, one can map the feature space onto a higher dimensional space. By applying a nonlinear kernel mapping function, SVMs are more easily able to be applied to classification problems with this type of feature space. One common kernel function K that is utilized for feature space transformation is the radial basis function (RBF) shown in the equation below. This study uses the radial basis kernel function k_{RBF} given in the equation

$$k_{RBF}(x_i, x_j) = \exp\left(-\gamma\left|x_i - x_j\right|^2\right), \gamma < 0$$

where γ is the parameter for the Gaussian RBF (*65*).

However, not all feature spaces are smooth. For data that may have discontinuities in the feature space, a better option is to include a variable to adjust for the possible lack of smoothness of the data. A generalization of the RBF called Matérn includes a variable ν to adapt to data roughness. This variable increases the flexibility of the kernel by allowing adaptation of the kernel to properties of the true underlying functional relation that may not be smooth. In addition to the RBF kernel, this study employs the Matérn kernel function given in the equation

$$k_M(x_i, x_j) = \frac{2^{1-\nu}}{\Gamma(\nu)2^{\nu-1}}\left(\frac{\sqrt{2\nu}(x_i - x_j)}{l}\right)^{\nu} K_\nu\left(\frac{\sqrt{2\nu}(x_i - x_j)}{l}\right)$$

where Γ is the gamma function (*68*, *69*), l is a scalar length-scale parameter for amplitude, ν is the roughness factor, and K_ν is the modified Bessel function of the second kind. As ν approaches infinity, the Matérn kernel converges with the RBF kernel, which is smooth due to its infinite differentiability. This lack of smoothness is relevant to the feature space in this study. Mohs is not a linear scale, but it is an ordinal that jumps in Vickers hardness. It is entirely possible that there are jumps or discontinuities in feature contributions to the physical phenomena between adjacent Mohs hardness values. To account for this possibility, RBF and Matérn kernels are employed separately as kernels in SVMs to observe if there is a difference in performance and the underlying nature of the continuity of the feature space.

Classifiers built with a SVM are referred to in this work as support vector classifiers (SVCs). The support vector machine model in this study that was applied to the ternary classification problem was constructed with a One-vs-One (OVO) decision-making function. An OVO classifier develops multiple binary classifiers, one for each pair of classes. Specifically, it trains *N(N-1)/2* binary

classifiers for an *N*-way classification problem. In the case of the ternary (3-way) classification problem, three binary classifiers are generated: Class 0 versus 1, 1 versus 2, and 0 versus 2, each of which would only distinguish between two classes at a time. To classify a given material, the material is presented to all three classifiers and majority voting is used to assign a label to the material.

Each feature was standardized by individually centering to the mean and scaling to unit variance or standard deviation. While RFs are less sensitive to absolute values, SVMs are sensitive to feature scaling. This is due to the construction of the hyperplane on the distance between the nearest data points with different classification labels, or support vectors. If one of the dimensions have a drastically larger value range, it may influence this distance and thereby affect the hyperplane. For consistency, all models in this study used this standardized feature space.

Study Models

Nine ML models were implemented, which are listed in Table 3 below. Models 1 and 2 are RBF-kernel SVMs applied to the binary and ternary classifications outlined in Table 1, respectively. Models 3 and 4 are RFs applied to the binary and ternary classifications, respectively. Model 5 is a binary RF in which Class 0 (0.991, 4.0] is classified against a combined superclass of Classes 1 (4.0, 7.0] and 2 (7.0, 10.0]. This model is employed to separate materials with low hardness values from the rest of the dataset. Model 6 is similarly constructed in that it is a binary RF that separates the medium hardness Class 1 (4.0, 7.0] against the superclass of Classes 0 and 2, (i.e., low and high hardness values). Model 7 is similarly constructed to classify materials with high Mohs hardness values from the combined superclass of low and medium Mohs hardness values. Models 8 and 9 are Matérn-kernel SVMs applied to the binary and ternary classifications, respectively. The implementations from Scikit-learn were used for all models.

Table 3. The Nine ML Models Utilized in This Study. The Acronyms Following a Class Type Indicate the Type of Model Used

ID	*Model*
1	Binary RBF SVC
2	Ternary RBF SVC – OVO
3	Binary RF
4	Ternary RF – multiclass
5	Ternary RF – OVR: 0 versus 1, 2
6	Ternary RF – OVR: 1 versus 0, 2
7	Ternary RF – OVR: 2 versus 0, 1
8	Binary Matérn SVC
9	Ternary Matérn SVC – OVO

These nine models allow several comparisons. First, we can compare the effectiveness of the OVO nature of RBF-kernel SVCs (RBF SVCs) and Matérn-kernel SVCs (Matérn SVCs) to the inherently multiclass decision-making scheme of RF ensemble methods for binary (Models 1, 3, and 8) and ternary classifications (Models 2, 4, and 9). Second, we can compare the effectiveness of an inherently multiclass RF scheme to the One-vs-Rest (OVR) ternary RF classification scheme for Models 4–7 and determine the best way to classify materials as having low, medium, and high

Mohs hardness values. Finally, comparing feature importances for Models 5–7 tells us which features contribute most to the classification between low, medium, and high Mohs hardness. This information can highlight which material properties are most important for low, medium, and high resistance to plastic deformation.

Grid Optimization for Binary and Ternary SVC: Models 1, 2, 8, and 9

The hyperparameters C and γ for RBF-based SVMs can drastically affect the performance of a classifier. The hyperparameter of C is the cost function that implements a penalty for misclassification. If C is too low for the dataset, then a simpler decision function is applied to the model, but it may underfit the data. This represents a soft margin, which may allow training points to either be ignored or misclassified. If C is too high for the dataset, then a more complicated decision function is applied to the model, but it may overfit to the training data.

The hyperparameter γ in the RBF is connected to the spread of influence of a single training point. If γ is too low, the decision boundary is too broad and does not separate the data well. If γ is too high, the RBF overfits to the training data by forming decision boundary islands. For any dataset to be classified, a balance exists between these two hyperparameters. The hyperparameters C and ν for Matérn-kernel SVM parameters can drastically affect the performance of a classifier. The hyperparameter C performs similarly for SVCs with a Matérn kernel or RBF kernel. The hyperparameter ν in the Matérn kernel is connected to the smoothness of the function to the data. If ν is too high, the kernel may be too smooth to capture any underlying discontinuities in the feature space.

For any dataset to be classified, a balance exists between these hyperparameters. One approach to find optimal values for these hyperparameters is a grid search method. In a grid search, each possible hyperparameter combination is applied to the dataset and the accuracy is reported to find the combination that produces the highest accuracy without overfitting. To optimize the parameters for the four SVC models in this study, a grid search was performed exploring the effects that various hyperparameters have on model performance. For the binary and ternary SVCs built with RBF kernel, combinations of the hyperparameters C and γ were tested to observe their effects on accuracy. For the binary and ternary SVCs built with the Matérn kernel, combinations of the hyperparameters C and ν were tested to observe their effects on accuracy. The ranges for the RBF SVC used in this paper is C between 10^{-2} and 10^{13} and γ between 10^{-9} and 10^{13}. The ranges for the Matérn SVC hyperparameters used in this paper is C between 10^{-2} and 10^{13} and ν between 0.5 to 6.0. These should be adequate to find a suitable combination to prevent both under- and over-fitting.

For each classifier undergoing grid optimization, the mineral dataset of 622 minerals (each having 11 feature descriptors) was split into two shuffled stratified subsets: a development set (66.7%) and an evaluation set (33.3%). These datasets were rearranged to ensure each subset was representative of the whole with respect to the distribution of Mohs hardness values. The development subset was used for training while the evaluation subset was used to test each classifier. This process was repeated for all hyperparameter combinations to perform the two-dimensional grid optimization.

Evaluation Criteria

In this study, all nine ML models are trained to predict Mohs hardness through binary or ternary classification methods. Their performance is evaluated with four metrics based on the true positives (T_p), true negatives (T_n), false positives (F_p), and false negatives (F_n) predicted by a given

classification model. The metrics used in this study are accuracy, specificity, precision, recall, and F1-scores. Accuracy (*A*) gives the proportion of true positive results in a population. Precision (*P*) describes how many of true positive predictions are actually positive. Specificity (*S*) is the probability that a classification model will identify true negative results. The higher the specificity, the lower the probability of false negative results. Recall (*R*) or sensitivity indicates the proportion of actual positives that were predicted as positive. *R* is the probability that a classification model will identify true positive results. The higher the recall, the lower the probability of false positive results. Typically, precision and recall are considered together through the F1-score (*F1*). *F1* is the harmonic average of precision and recall and gives equal importance to both. It is an important metric for datasets with uneven class distribution. The closer *F1* is to 1, the closer the model comes to perfect recall and precision. Overall these five metrics give great insight into the performance of the classification models, and their equations are as follows:

$$A = \frac{T_p + T_n}{T_p + T_n + F_p + F_n}$$

$$S = \frac{T_n}{T_n + F_p}$$

$$P = \frac{T_p}{T_p + F_p}$$

$$R = \frac{T_p}{T_p + F_n}$$

$$F1 = \frac{2(P * R)}{P + R}$$

Results and Discussion

In this study, several machine or statistical learning approaches are presented to quantitatively study the relationship between material composition and Mohs hardness values, which is a complex property relating to elasticity and plastic deformation of a material.

Grid Optimization Results for Binary and Ternary SVCs: Models 1, 2, 8, and 9

In this study, grid search optimization on the binary and ternary Matérn and RBF SVC models was performed. For each classifier undergoing the grid optimization scheme, the mineral dataset of 622 minerals was split into two stratified subsets: a development set (66.7%) and an evaluation set (33.3%). For the RBF SVC models (Models 1 and 2), grid search optimization was performed by

methodically building and evaluating a model for each hyper-parameter combination of C between 10^{-2} and 10^{13} and γ between 10^{-9} and 10^{13}. For the Matérn kernel SVC models (Models 8 and 9), grid search optimization was performed by methodically building and evaluating a model for each hyper-parameter combination of C between 10^{-2} and 10^{13} and ν between 0.5 to 6.0.

The binary RBF SVC classifier Model 1 achieved an accuracy of 86.4% with $C = 10$ and $\gamma = 1$, as shown in Figure 2a. The ternary RBF SVC Model 2 achieved an accuracy of 85.0% with $C = 10$ and $\gamma = 1$, as shown in Figure 2b. The binary Matérn SVC Model 8 achieved an accuracy of 86.4% with hyperparameters $C = 10$ and $\nu = 2.5$, as shown in Figure 2c. The ternary Matérn SVC Model 9 achieved an accuracy of 87.4% with hyperparameters $C = 1$ and $\nu = 1$, as shown in Figure 2d. There is a moderate gain in accuracy in classifiers employing the Matérn kernel compared to the RBF kernel, but they are both close. This may suggest that discontinuities that may exist in the feature space may be small enough to approximate with a smooth function like RBF. For the remainder of the study, these grid-optimized hyperparameters were utilized for Models 1, 2, 8, and 9. These prediction accuracies suggest that binary and ternary SVCs built on either RBF or Matérn kernels to classify the Mohs hardness of ceramics can be helpful in materials discovery and development.

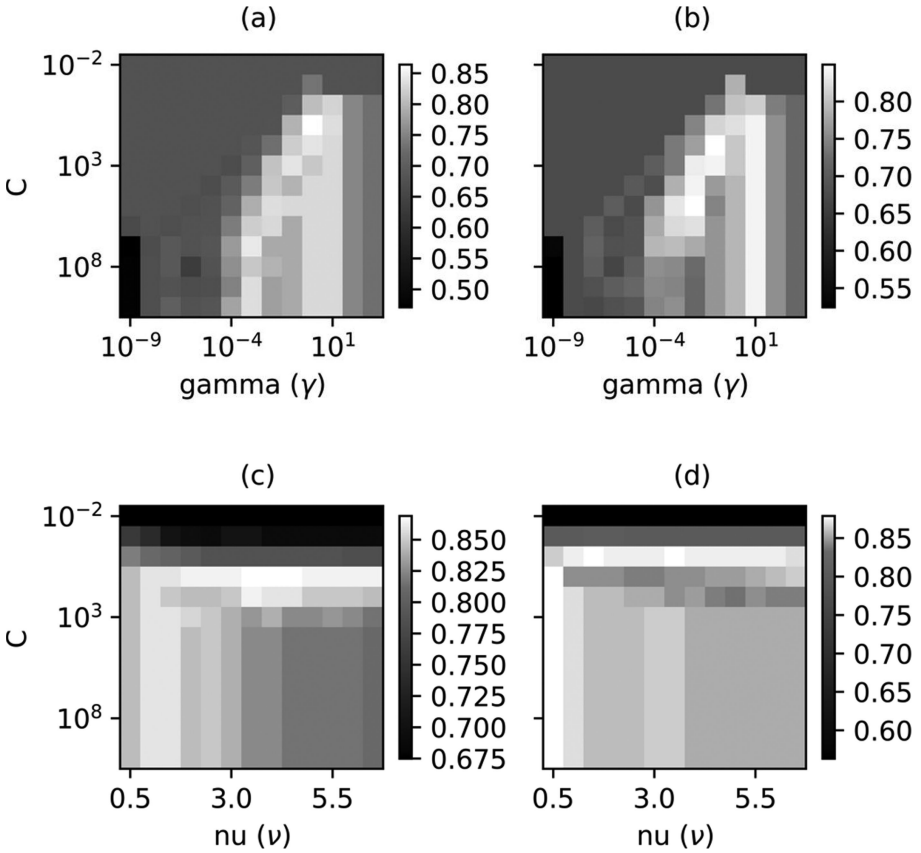

Figure 2. The cross-validation grid optimization accuracies for the (a) binary and (b) ternary RBF SVCs, Models 1 and 2, respectively. The y axis is the value of C, or the soft margin cost function. The x axis is the value of γ, or the parameter for the Gaussian RBF. The cross-validation grid optimization accuracies for the (c) binary and (d) ternary Matérn SVCs, Models 8 and 9, respectively. The color represents the model accuracy.

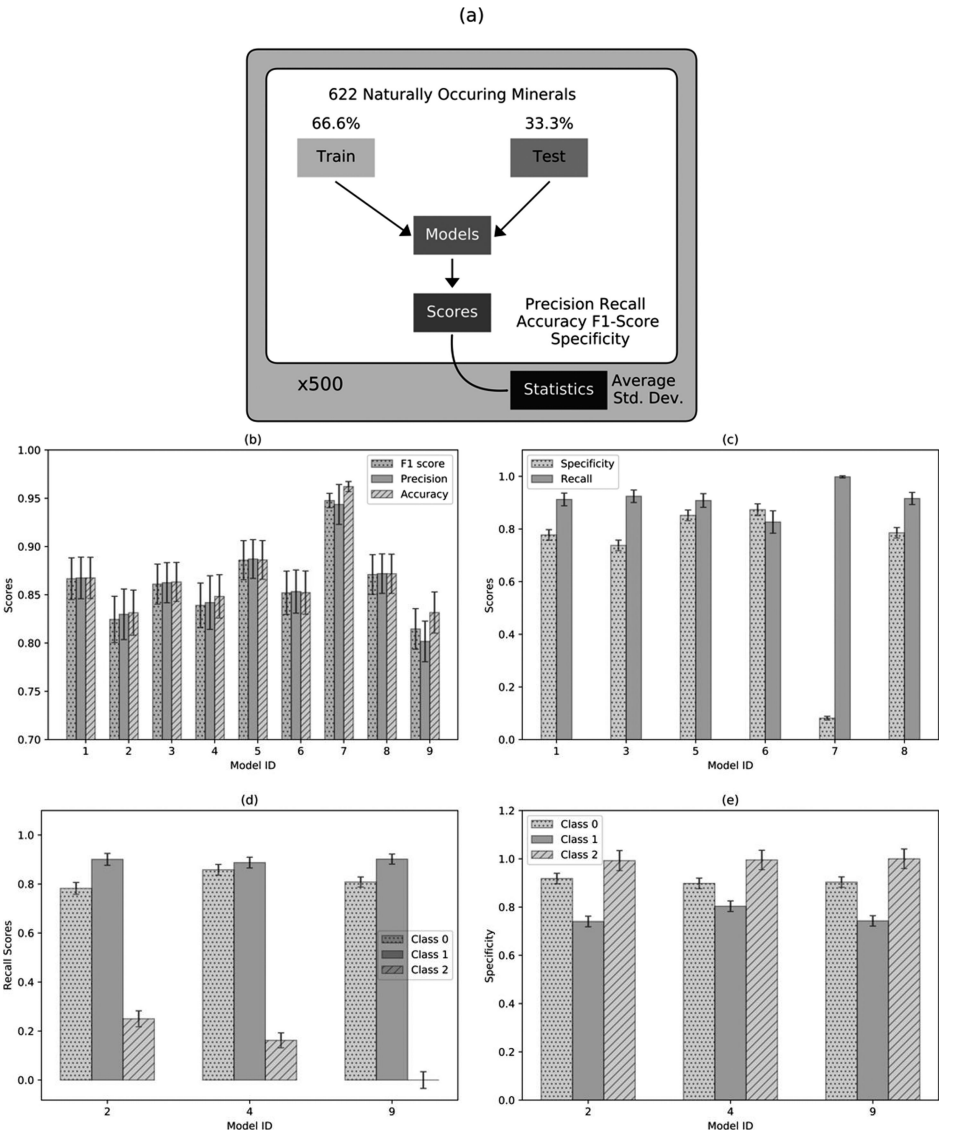

Figure 3. Performance from models trained under 500 stratified train-test splits (a) workflow of model performance. (b) The specificity and recall scores from for all binary models. (c) The recall scores for ternary models. (d) The specificity scores for ternary models. The bar height corresponds to the average respective score. The black error bars correspond to the standard deviation of the respective metric.

Model Performance on Naturally Occurring Minerals Dataset

To determine the performance of the models utilized in this study, all models were constructed with the naturally occurring mineral dataset, which was split 500 times into three-fold training and test subsets. The workflow is shown in Figure 3a. Upon completion of all 500 splits, the F1-score, precision, recall, specificity, and accuracy were calculated based on the predicted and known values of the test subsets. Figure 3b shows the weighted F1, precision, and accuracy scores for all nine models over 500 splits, along with their standard deviations. For Figure 3b and 3c, the hashed bars represent the same performance metric across each model. For Figure 3d and 3e, the hashed bars

represent the same class across each model. The height of the bars is the magnitude of the respective metric for that attribute. The x-axis labels for Figure 3b–3e correspond to the Model ID in Table 3.

All of the classifiers were able to classify the vast majority of Mohs hardness values with weighted F1, precision, and accuracy scores of 0.79 or higher. The ternary Matérn SVC (Model 9) underperforms its ternary RBF SVC (Model 2) and ternary multiclass RF (Model 4) counterparts. Also, the ternary OVR RF models (Models 5–7) appear perform similarly to or better than the ternary multiclass RF model (Model 4) with scores >0.82. For the specificity and recall for the binary models shown in Figure 3c, the scores are good (>0.7) in all models except 7. In Model 7, there is a drastic decrease in specificity to 0.2. This suggests that the model overclassified false positives. This is unlikely due to the model predicting based on the features themselves but to the model predicting based on the bias in the training data. Materials in Class 2 (7.0, 10.0] are underrepresented, only counting toward 6% of the entire dataset. This is due to the natural rarity of materials in that range.

Overall, Models 5 and 6 had the strongest prediction performance across all 5 metrics. Model 5 is a binary RF in which Class 0 (0.991, 4.0] is classified against a combined superclass of Classes 1 (4.0, 7.0] and 2 (7.0, 10.0]. This model is employed to separate materials with low hardness values from the rest of the dataset. Model 6 is similarly constructed in that it is a binary RF that separates the medium hardness Class 1 (4.0, 7.0] against the superclass of Classes 0 and 2 (i.e., low and high hardness values). These one-vs-rest binary classification of ternary bins best captured the underlying patterns of material hardness for naturally occurring ceramic minerals. However, given the small size of the population in the Class 2 classification bins, both of these models could condense into a pseudobinary classification task of Class 0 (0.991, 4.0] and Class 1 (4.0, 10.0], where the separation is at Mohs value 4.0 instead of 5.5 as in the true binary bins in this study. The effect of data bias t plagued Model 7 is less pronounced in Models 5 and 6, due to their population sizes closer to parity.

Models 5 and 6 may correspond more to the grouping presented in the correlation between the Gao calculated bond lengths and calculated hardness values than Šimůnek and Vackář's more binarized grouping found in Model 3 on this dataset of naturally occurring minerals. Model 3 is a binary RF in which Class 0 (0.991, 5.5] is classified against a combined superclass of Class 1 (5.5, 10.0]. This model is also employed to separate materials with low hardness values from the rest of the dataset. Even with a specificity is closer to 0.74, Model 3 still performs well with a recall score $\pm$ std. dev. of 0.8633 ± 0.02024. Overall, these models yield insight into the connection between bond characteristics and hardness. According to the Gao et al. study (*23*), the three factors connecting the hardness of polar covalent crystals to bond behavior are the bond density or electronic density, bond length, and degree of covalent bonding. According to the Šimůnek and Vackář study (*59*), the factors connecting the hardness of binary covalently bonded solids are crystal valence electron density, number of bonds, and bond length between pairs of atoms. From the performance of Models 3, 5, and 6, all of these factors are closely related to the hardness of naturally occurring ceramic minerals. To further understand the impact of these factors in these RF models, feature importances can be used to gauge their relative importances and increase understanding of the physical basis in the hardness regimes of these ceramic minerals.

Next, the effectiveness of several binary classifiers was evaluated using the quantitative variables of true positive rate, which represents the total number of correctly classified Mohs hardness values in the positive class, and the false positive rate, which represents the total number of incorrectly classified Mohs hardness values assigned to the positive class. With these variables, the receiver operating characteristic (ROC) curves were calculated. ROC curves plot the true positive rate for a binary classifier as a function of its false positive rate to gauge model performance. The area under the curve (AUC) is a quality measure of the classifier's ability to correctly classify a given material. The

ideal AUC is unity, or 1. To compare the effectiveness of the binary (Models 1, 3, and 8) and OVR (Models 5, 6, 7) superclass classifiers used in this study, the author implemented ROC curves and calculated the areas under the curves. These curves, are given in Figure 4 below.

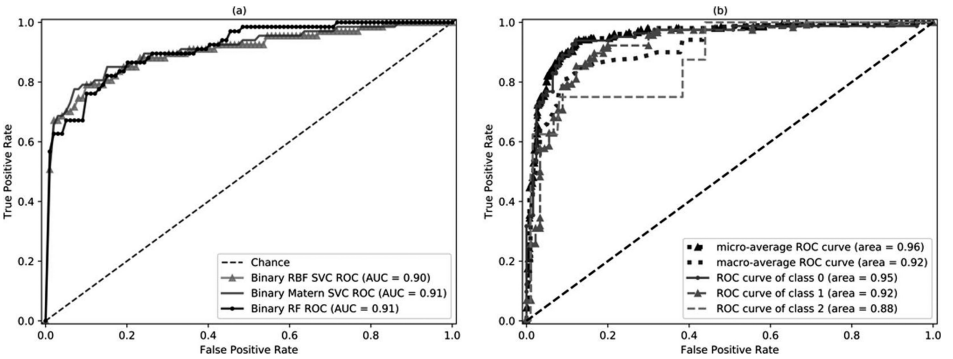

Figure 4. ROC plots using the false positive ratio and false negative ratio are used to evaluate the ability of classifiers to predict Mohs hardness values. (a) Comparison of binary RBF SVC, RF, Matérn SVC, Models 1, 3, and 8, respectively. (b) Comparison of the ternary OVR classifiers, Models 5, 6, and 7, which predict low, medium, and high hardness values, respectively.

Both the binary and ternary OVR superclass classifiers were able to discriminate the vast majority of naturally occurring minerals with an AUC of 0.88 or greater. The ROC plots in Figure 4b illustrate the similar performance of ternary OVR superclass classifiers when applied to the same set of compositional features. This appears to reflect the This suggests that compositional predictors developed for these materials can be generally applied with reasonable reliability to other single crystalline materials across a wide-ranging compositional and structural space.

Feature Importances

To determine feature importance for the RF-based models utilized in this study (Models 3–7), 10,000 trees were constructed for a single forest. Upon completion of each forest, the list of input features with their representative Gini importances was returned. Figure 5a shows the relative importance of different atomic features for binary (Model 3) and ternary (Model 4) RFs constructed as inherently multiclass. Figure 5b shows the relative importances of different atomic features for ternary RFs constructed as OVR classifiers (Models 5–7). The *x*-axis labels correspond to the feature ID outlined in Table 3. The heights of the color bars in Figures 5a and 5b indicate the Gini importance of each feature. The colors represent the different models.

The most important features vary on three points: (1) whether the model is binary or ternary, (2) whether the model is constructed as multiclass or binary OVR, and (3) the regime of hardness classified. For the binary RF model (Model 3) in Figure 5a, the four most important features are Features 5, 0, 3, and 6 with feature importances of 0.124, 0.129, 0.120, and 0.111, respectively. These features correspond to the atomic average of IE, the total of the number of electrons in the empirical formula, the atomic average of the valence electrons, and the atomic average of the Pauling electronegativities, respectively. For the ternary multiclass model (Model 4) in Figure 5a, the four most important features are Features 7, 8, 3, and 5 with feature importances of 0.129, 0.113, 0.112, and 0.111, respectively. These features correspond to the atomic average of the van der Waals atomic radii, the atomic average of the covalent atomic radii, the atomic average of the valence electrons, and the atomic average of IE, respectively.

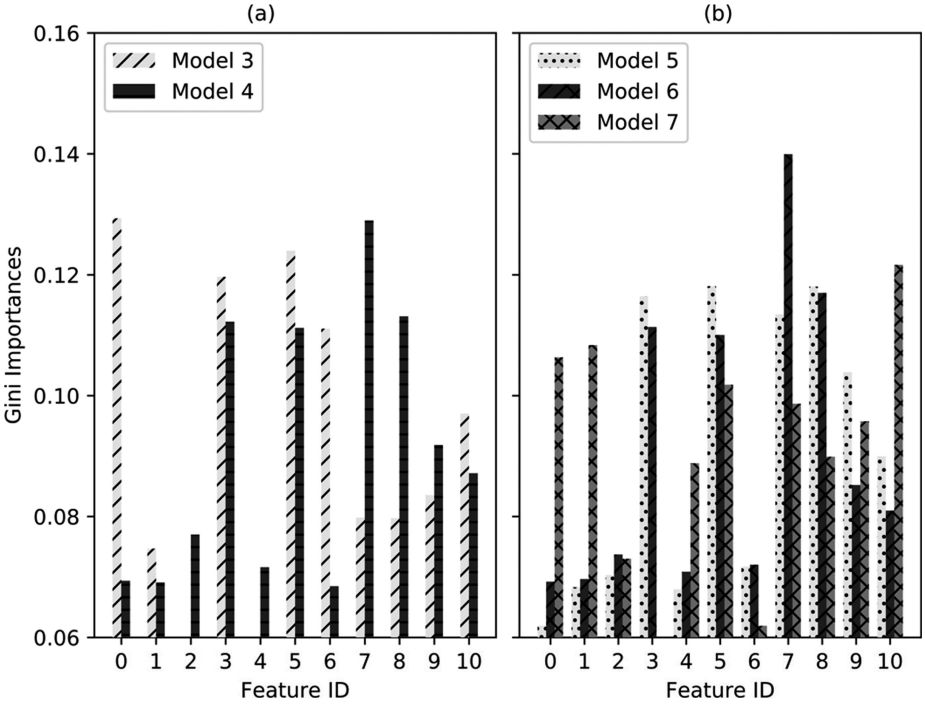

Figure 5. The Gini feature importances of a 10,000 tree RF for (a) binary (Model 3) and ternary multiclass (Model 4) models and (b) OVR binary RF classifiers (Models 5–7) with low, medium, and high hardness ternary class as a positive class, respectively.

For the ternary-bin, OVR binary RF in which Class 0 (0.991, 4.0] is classified against a combined superclass of Classes 1 (4.0, 7.0] and 2 (7.0, 10.0] (Model 5) in Figure 5b, the four most important features are Features 5, 8, 3, and 7 with feature importances of 0.118, 0.118, 0.116, and 0.113, respectively. These features correspond to the atomic average of IE, the atomic average of the covalent atomic radii, the atomic average of the valence electrons, and the atomic average of the van der Waals atomic radii, respectively. For the ternary-bin, OVR binary RF in which the medium hardness Class 1 (4.0, 7.0] against the superclass of Classes 0 and 2 (Model 6) in Figure 5b, the four most important features are Features 7, 8, 3, and 5 with feature importances of 0.140, 0.117, 0.111, and 0.110, respectively. These features correspond to the atomic average of the van der Waals atomic radii, the atomic average of the covalent atomic radii, the atomic average of the valence electrons, and the atomic average of IE, respectively. For the ternary-bin, one-vs-rest binary RF in which the medium hardness Class 2 (7.0, 10.0] against the superclass of Classes 0 and 1 (Model 7) in Figure 5b, the four most important features are Features 10, 1, 0, and 5 with feature importances of 0.121, 0.108, 0.106, and 0.101, respectively. These features correspond to the atomic average of the densities of the elements found in the empirical formula, the total densities, and the atomic average of IE, respectively.

Earlier in Figures 3b–3e, it was shown that Models 5 and 6 are similar and perform well on the dataset of naturally occurring crystalline ceramic minerals. Here in Figure 5b, it can be found that these two models share the top 4 features: 7, 8, 3, and 5. These features correspond to the atomic average of the van der Waals atomic radii, the atomic average of the covalent atomic radii, the atomic average of the valence electrons, and the atomic average of IE, respectively. The related factors from these studies directly correspond to material characteristics previously attributed as contributors to material hardness. The number of valence electrons per bond was included as a factor in Gao et al.

(*23*), Šimůnek et al. (*24*), and Mukhanov et al. (*34*). Atomic radii (both covalent and van der Waals) are related to the bond length factor in Gao et al. and the molar volume in Mukhanov et al. (*23*, *34*). The first IE is related to the bond strength of the material (*61*, *62*), which Šimůnek and Vackář attribute as a major factor in hardness (*59*). Three of the four importances (Features 8, 3, and 5) are the same between Models 5 and 6. However, Feature 7, or the atomic average of the van der Waals atomic radii, varies greatly between the two models. For Model 5, the importance of Feature 7 is 0.118. For Model 6, the importance of Feature 7 is 0.140. From Model 5 to Model 6, the importance of this feature drops 15.7%. Therefore, it may then follow that a major difference in performance between these models on a validation dataset would likely depend on this feature.

Model Validation with Validation Set

To determine the generalizability of the models to artificial ceramic crystals, all models were trained with the naturally occurring mineral dataset and tested on a validation dataset. The validation dataset consists of 52 artificial single crystals, with 15 monoclinic, 8 tetragonal, 7 hexagonal, 6 orthorhombic, 4 cubic, and 3 rhombohedral crystal structures. The F1, precision, recall, specificity and accuracy scores were calculated based on the validation set. This workflow is diagrammed in Figure 6a. Figure 6b shows the weighted F1, precision, and accuracy scores for all nine models. For Figure 6b and 6c, the hashed bars represent the same performance metric across each model. For Figure 6d and 6e, the hashed bars represent the same class across each model. The height of the bars is the magnitude of the respective metric for that attribute. The x-axis labels for Figure 6b–6e correspond to the Model IDs in Table 3.

All models in Figures 6b–6e shown have similar trends as in the naturally occurring dataset but with lower performance values. Such a dip in performance can be expected. In the growth of artificial crystals, different crystal phases can be created that may produce crystals with higher or lower Mohs hardness values than would occur naturally. Growth conditions were not factored into the feature set in this study. To address this experimental intervention, it may be useful to include crystal growth conditions as a feature in future ML models.

According to Figure 6b, all models have F1, precision, and accuracy scores greater than 0.70. Model 5 is the strongest performing model with F1, precision, and accuracy scores around 92%. Model 7 also shows strong performance metrics. However, this model is affected by data bias due to the small number in the positive class. This is also seen by the low recall score in Figure 6c. Model 5 is a stronger than Model 7 due to the reduced data bias from the more evenly matched populations of the positive and negative classes. Therefore, the author has strong confidence that the models are predicting on the feature set instead of data bias of imbalanced classes.

As discussed earlier, Models 5 and 6 are basically pseudo-inverses of each other due to the underrepresentation of naturally occurring ceramic minerals with Mohs hardness values greater than 7.0. In this case, both of these models could condense into a pseudobinary classification task of Class 0 (0.991, 4.0] and Class 1 (4.0, 10.0], where the separation is at Mohs value 4.0 instead of 5.5 as in the true binary bins in this study. However, on the validation dataset, Model 6 performs much better than Model 5. A major reason for this could reflect back to the feature importance analysis. Feature 7, or the atomic average of the van der Waals atomic radii, decreases around 15.7% between Models 5 and 6. Model 5 appears to have enough importance on that factor to produce reasonable predictions. For Model 6, there appears to be an important overreliance on the atomic average van der Waals atomic radius in predicting hardness values for man-made artificial materials. This may be

due to the sensitivity of van der Waals bonding interactions at the solid-liquid interface to growth conditions during the crystallization process, such as temperature and solvents (*70*, *71*).

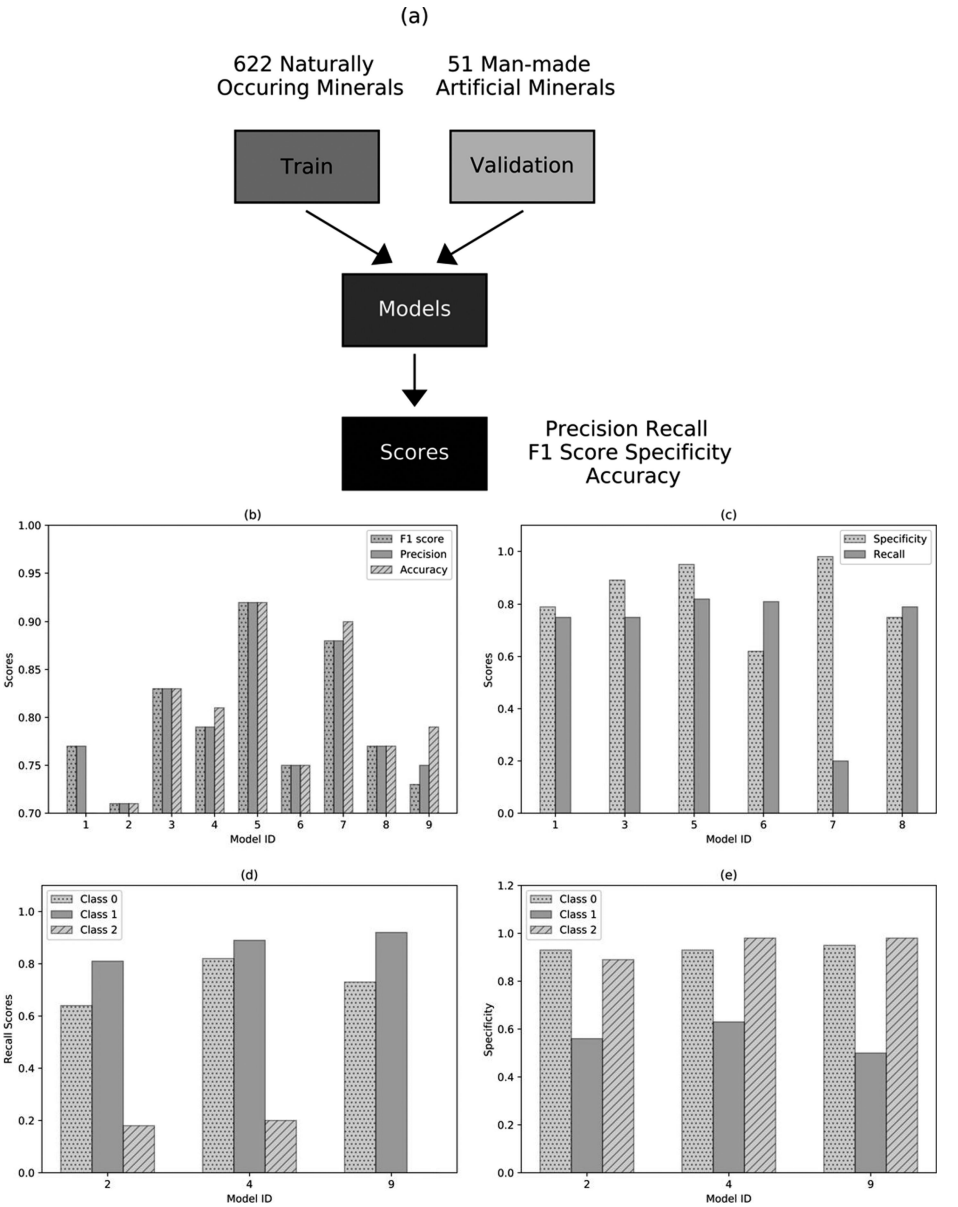

Figure 6. (a) Workflow of performance testing for model validation. (b) The specificity and recall scores for all binary models. (c) The recall scores for ternary models. (d) The specificity scores for ternary models. The bar height corresponds to the average respective score. All models were trained on the naturally occurring mineral dataset and validated with the artificial single crystal dataset.

Considerations

Note that the prediction models in this study have only considered a small number of composition-based factors and other easily accessible attributes for an extremely varied chemical space. This is a reasonable first screening step that allows us to efficiently gauge important factors that

may contribute to material hardness. However, to make larger generalizations about the nature of hardness, more features would have to be considered and then narrowed down with feature selection methods. Specifically, feature selection methods such as principal component analysis, univariate selection, and recursive feature elimination across a larger set of features would yield a deeper understanding about the nature of factors that contribute to material hardness.

In addition, predicting the hardness of superhard ($M > 9$) materials would be particularly problematic with the current dataset. Materials with a Mohs hardness less than 2 and greater than 9 are not equally represented, which leads to an imbalanced dataset. The implementation of data handling methods specifically constructed for handling imbalanced datasets, such as oversampling, undersampling, or synthetic minority oversampling technique, may allow more accurate prediction of the Mohs hardness in those regimes. This effect extends to ceramic materials in the 7–10 Mohs range. While there are more minerals included in this range, there is still a great imbalance that produces a data bias in the training of statistical and ML models. Therefore, to extend this application to predict minerals in the 7–10 Mohs range in the future, more artificial materials would need to be included. These approaches would allow possible design and prediction of novel superhard crystal ceramics.

Please note that our prediction models have only considered single crystalline materials. For other types of materials, different factors affect hardness. For metals, hardness is affected by structural factors like dislocation entanglements (*16*). Also in metals, there is a connection between bulk modulus, shear modulus, hardness, and ductility. This connection has previously been referenced by Chen (*72*), Tabor (*73*), and Pugh among others (*74*). Due to the delocalized nature of the bonding in metals, plastic deformations locally accumulate before fracture, resulting in ductility and reduced hardness. To explore this effect, the inclusion of metals that stretch across the ductile-to-brittle transition into the feature set could offer insight into the connection due to the nature of bonding strength in these materials.

For plastics, elastic and plastic properties depend on chain length, degree of cross-linking, and the degree of crystallinity of the material. Inclusion of nanomaterials into single-crystalline matrices has also been shown to increase hardness. Those were ignored in this study but may also be an avenue to consider in future studies. The continued growth of data repositories based on experimental characterization of materials is expected to enable the development of models for mechanical and microstructural material properties not covered in this study, specifically fracture toughness, thermal stability, bulk modulus, shear modulus, and work hardening, among others.

Conclusions

This study shows that comparative material properties like Mohs hardness can be modeled with ML algorithms using features based solely on material composition. The results show that RFs and SVMs are able to produce reasonable predictions of materials property. They also show that different features are relatively important for predicting Mohs hardness values. These features include the atomic average of the van der Waals atomic radii, the atomic average of the covalent atomic radii, the atomic average of the valence electrons, and the atomic average of IE among others. These features were previously included in separate studies but were combined into this study to further understand their interrelated physical contributions to materials hardness (*23*, *34*, *59*). In conclusion, I have demonstrated that a ML model can be useful in classifying comparative material properties. The methodology described in this study could be applied to other types of materials for accelerated design and materials science discovery of novel materials.

Acknowledgments

The author acknowledges the support provided by the National Science Foundation under grant numbers HRD-1547757 (CREST-BioSS Center) and HRD-1647013. I would also like to acknowledge and show my appreciation to the Vanderbilt Advanced Computing Center for Research and Education (ACCRE) as well as Professor Kelly Holley-Bocklemann and Dr. Caleb Wheeler.

References

1. Plinninger, R. J.; Spaun, G.; Thuro, K. Prediction and Classification of Tool Wear in Drill and Blast Tunnelling. In *Proceedings of 9th Congress of the International Association for Engineering Geology and the Environment* [Online]; Engineering Geology for Developing Countries: Durban, South Africa, 2002; pp 16–20. http://www.geo.tum.de/people/thuro/pubs/2002_iaeg_durban_pli.pdf (accessed Jan 26, 2019).
2. Hoseinie, S. H.; Ataei, M.; Mikaiel, R. Comparison of Some Rock Hardness Scales Applied in Drillability Studies. *Arab. J. Sci. Eng.* **2012**, *37*, 1451–1458.
3. Thuro, K.; Plinninger, R. J. Hard Rock Tunnel Boring, Cutting, Drilling and Blasting: Rock Parameters for Excavatability. In *10th ISRM Congress*; International Society for Rock Mechanics and Rock Engineering: Sandton, South Africa, 2003.
4. Ellecosta, P.; Schneider, S.; Kasling, H.; Thuro, K. Hardness–A New Method for Characterising the Interaction of TBM Disc Cutters and Rocks? In *13th ISRM International Congress of Rock Mechanics*; International Society for Rock Mechanics and Rock Engineering: Montreal, Canada, 2015.
5. Moore, M. A. The Relationship between the Abrasive Wear Resistance, Hardness and Microstructure of Ferritic Materials. *Wear* **1974**, *28*, 59–68.
6. Axén, N.; Jacobson, S.; Hogmark, S. Influence of Hardness of the Counterbody in Three-Body Abrasive Wear — an Overlooked Hardness Effect. *Tribol. Int.* **1994**, *27*, 233–241.
7. Jefferies, S. R. Abrasive Finishing and Polishing in Restorative Dentistry: A State-of-the-Art Review. *Dent. Clin. North Am.* **2007**, *51*, 379–397.
8. Balaceanu, M.; Petreus, T.; Braic, V.; Zoita, C. N.; Vladescu, A.; Cotrutz, C. E.; Braic, M. Characterization of Zr-Based Hard Coatings for Medical Implant Applications. *Surf. Coatings Technol.* **2010**, *204*, 2046–2050.
9. Parsons, J. R.; Lee, C. K.; Langrana, N. A.; Clemow, A. J.; Chen, E. H. *Functional and Biocompatible Intervertebral Disc Spacer Containing Elastomeric Material of Varying Hardness*. U.S. Patent 5,545,229, December 15, 1992.
10. Okazaki, Y.; Ito, Y.; Ito, A.; Tateishi, T. Effect of Alloying Elements on Mechanical Properties of Titanium Alloys for Medical Implants. *Mater. Trans. JIM* **1993**, *34*, 1217–1222.
11. Kanyanta, V. Hard, Superhard and Ultrahard Materials: An Overview. In *Microstructure-Property Correlations for Hard, Superhard, and Ultrahard Materials*; Springer International Publishing: Cham, Switzerland, 2016; pp 1–23.
12. Hwang, D. K.; Moon, J. H.; Shul, Y. G.; Jung, K. T.; Kim, D. H.; Lee, D. W. Scratch Resistant and Transparent UV-Protective Coating on Polycarbonate. *J. Sol-Gel Sci. Technol.* **2003**, *26*, 783–787.

13. Luber, J. R.; Bunick, F. J. Protective Coating for Tablet. Official Gazette of the United States Patent & Trademark Office Patents 1249(3), August 21, 2001.
14. Tabor, D. Mohs's Hardness Scale - A Physical Interpretation. *Proc. Phys. Soc. Sect. B* **1954**, *67*, 249–257.
15. Tabor, D. The Physical Meaning of Indentation and Scratch Hardness. *Br. J. Appl. Phys.* **1956**, *7*, 159–166.
16. Tabor, D. The Hardness of Solids. *Rev. Phys. Technol.* **1970**, *1*, 145–179.
17. Li, K.; Yang, P.; Niu, L.; Xue, D. Group Electronegativity for Prediction of Materials Hardness. *J. Phys. Chem. A* **2012**, *116*, 6911–6916.
18. Broz, M. E.; Cook, R. F.; Whitney, D. L. Microhardness, Toughness, and Modulus of Mohs Scale Minerals. *Am. Mineral.* **2006**, *91*, 135–142.
19. Gilman, J. J. *Chemistry and Physics of Mechanical Hardness*; John Wiley & Sons: Hoboken, NJ, 2009; Vol. 5.
20. Oganov, A. R.; Lyakhov, A. O. Towards the Theory of Hardness of Materials. *Orig. Russ. Text © A.R. Oganov, A.O. Lyakhov, J. Superhard Mater.* **2010**, *32*, 3–8.
21. Li, K.; Yang, P.; Niu, L.; Xue, D. Hardness of Inorganic Functional Materials. *Rev. Adv. Sci. Eng.* **2012**, *1*, 265–279.
22. Cohen, M. L. Predicting Useful Materials. *Science* **1993**, *261*, 307–309.
23. Gao, F.; He, J.; Wu, E.; Liu, S.; Yu, D.; Li, D.; Zhang, S.; Tian, Y. Hardness of Covalent Crystals. *Phys. Rev. Lett.* **2003**, *91*, 015502.
24. Šimůnek, A.; Vackář, J. Hardness of Covalent and Ionic Crystals: First-Principle Calculations. *Phys. Rev. Lett.* **2006**, *96*, 085501.
25. Inal, K.; Neale, K. W. High Performance Computational Modelling of Microstructural Phenomena in Polycrystalline Metals. In *Advances in Engineering Structures, Mechanics & Construction*; Springer Netherlands: Dordrecht, 2006; pp 583–593.
26. Vo, N. Q.; Averback, R. S.; Bellon, P.; Caro, A. Limits of Hardness at the Nanoscale: Molecular Dynamics Simulations. *Phys. Rev. B* **2008**, *78*, 241402.
27. Van Swygenhoven, H. Grain Boundaries and Dislocations. *Science* **2002**, *296*, 66–67.
28. Botu, V.; Ramprasad, R. Adaptive Machine Learning Framework to Accelerate *Ab Initio* Molecular Dynamics. *Int. J. Quantum Chem.* **2015**, *115*, 1074–1083.
29. Hansen, K.; Biegler, F.; Ramakrishnan, R.; Pronobis, W.; von Lilienfeld, O. A.; Müller, K.-R.; Tkatchenko, A. Machine Learning Predictions of Molecular Properties: Accurate Many-Body Potentials and Nonlocality in Chemical Space. *J. Phys. Chem. Lett.* **2015**, *6*, 2326–2331.
30. Behler, J. Perspective: Machine Learning Potentials for Atomistic Simulations. *J. Chem. Phys.* **2016**, *145*, 170901.
31. Li, Z.; Kermode, J. R.; De Vita, A. Molecular Dynamics with On-the-Fly Machine Learning of Quantum-Mechanical Forces. *Phys. Rev. Lett.* **2015**, *114*, 096405.
32. Zeng, Y.; Li, Q.; Bai, K. Prediction of Interstitial Diffusion Activation Energies of Nitrogen, Oxygen, Boron and Carbon in Bcc, Fcc, and Hcp Metals Using Machine Learning. *Comput. Mater. Sci.* **2018**, *144*, 232–247.
33. Wu, H.; Mayeshiba, T.; Morgan, D. High-Throughput Ab-Initio Dilute Solute Diffusion Database. *Sci. Data* **2016**, *3*, 160054.

34. Mukhanov, V. A.; Kurakevych, O. O.; Solozhenko, V. L. Thermodynamic Aspects of Materials' Hardness: Prediction of Novel Superhard High-Pressure Phases. *High Press. Res.* **2008**, *28*, 531–537.
35. Li, K.; Xue, D. Estimation of Electronegativity Values of Elements in Different Valence States. *J. Phys. Chem. A* **2006**, *110*, 11332–11337.
36. Li, K.; Wang, X.; Zhang, F.; Xue, D. Electronegativity Identification of Novel Superhard Materials. *Phys. Rev. Lett.* **2008**, *100*, 235504.
37. Rajan, K. Materials Informatics: The Materials "t; Gene "t; and Big Data. *Annu. Rev. Mater. Res* **2015**, *45*, 153–169.
38. Butler, K. T.; Davies, D. W.; Cartwright, H.; Isayev, O.; Walsh, A. Machine Learning for Molecular and Materials Science. *Nature* **2018**, *559*, 547.
39. Ward, L.; Agrawal, A.; Choudhary, A.; Wolverton, C. A General-Purpose Machine Learning Framework for Predicting Properties of Inorganic Materials. *npj Comput. Mater.* **2016**, *2*, 16028.
40. Ward, L.; Wolverton, C. Atomistic Calculations and Materials Informatics: A Review. *Curr. Opin. Solid State Mater. Sci.* **2017**, *21*, 167–176.
41. Mueller, T.; Kusne, A. G.; Ramprasad, R. Machine Learning in Materials Science: Recent Progress and Emerging Applications. *Rev. Comput. Chem.* **2016**, *29*, 186–273.
42. Liu, Y.; Zhao, T.; Ju, W.; Shi, S. Materials Discovery and Design Using Machine Learning. *J. Mater.* **2017**, *3*, 159–177.
43. Curtarolo, S.; W Hart, G. L.; Buongiorno Nardelli, M.; Mingo, N.; Sanvito, S.; Levy, O. The High-Throughput Highway to Computational Materials Design. *Nat. Mater.* **2013**, *12*.
44. Kim, C.; Pilania, G.; Ramprasad, R. From Organized High-Throughput Data to Phenomenological Theory Using Machine Learning: The Example of Dielectric Breakdown. *Chem. Mater.* **2016**, *28*, 1304–1311.
45. Kim, C.; Pilania, G.; Ramprasad, R. Machine Learning Assisted Predictions of Intrinsic Dielectric Breakdown Strength of ABX_3 Perovskites. *J. Phys. Chem. C* **2016**, *120*, 14575–14580.
46. Ghiringhelli, L. M.; Vybiral, J.; Levchenko, S. V.; Draxl, C.; Scheffler, M. Big Data of Materials Science: Critical Role of the Descriptor. *Phys. Rev. Lett.* **2015**, *114*, 105503.
47. Goldsmith, B. R.; Boley, M.; Vreeken, J.; Scheffler, M.; Ghiringhelli, L. M. Uncovering Structure-Property Relationships of Materials by Subgroup Discovery. *New J. Phys.* **2017**, *19*, 013031.
48. Ghiringhelli, L. M.; Vybiral, J.; Ahmetcik, E.; Ouyang, R.; Levchenko, S. V; Draxl, C.; Scheffler, M. Learning Physical Descriptors for Materials Science by Compressed Sensing. *New J. Phys.* **2017**, *19*, 023017.
49. Oliynyk, A. O.; Antono, E.; Sparks, T. D.; Ghadbeigi, L.; Gaultois, M. W.; Meredig, B.; Mar, A. High-Throughput Machine-Learning-Driven Synthesis of Full-Heusler Compounds. *Chem. Mater.* **2016**, *28*, 7324–7331.
50. Dey, P.; Bible, J.; Datta, S.; Broderick, S.; Jasinski, J.; Sunkara, M.; Menon, M.; Rajan, K. Informatics-Aided Bandgap Engineering for Solar Materials. *Comput. Mater. Sci.* **2014**, *83*, 185–195.

51. Ward, L.; Agrawal, A.; Choudhary, A.; Wolverton, C. A General-Purpose Machine Learning Framework for Predicting Properties of Inorganic Materials. *npj Comput. Mater.* **2016**[Online]. https://www.nature.com/articles/npjcompumats201628
52. Lee, J.; Seko, A.; Shitara, K.; Tanaka, I. *Prediction Model of Band-Gap for AX Binary Compounds by Combination of Density Functional Theory Calculations and Machine Learning Techniques*. 2015, arXiv:1509.0097. arXiv.org e-Print archive. https://arxiv.org/abs/1509.00973 (accessed June 27, 2019).
53. Pilania, G.; Mannodi-Kanakkithodi, A.; Uberuaga, B. P.; Ramprasad, R.; Gubernatis, J. E.; Lookman, T. Machine Learning Bandgaps of Double Perovskites. *Nature Sci. Rep.* **2015**.
54. Pilania, G.; Gubernatis, J. E.; Lookman, T. Multi-Fidelity Machine Learning Models for Accurate Bandgap Predictions of Solids. *Comput. Mater. Sci.* **2017**, *129*, 156–163.
55. CRC. *CRC Handbook of Chemistry and Physics*, 98th ed.; Rumble, J. R., Ed.; CRC Press/Taylor & Francis: Boca Raton, FL, 2018.
56. Downs, R. T.; Hall-Wallace, M. The American Mineralogist Crystal Structure Database. *Am. Mineral.* **2003**, *88*, 247–250.
57. Garnett, J. *Prediction of Mohs hardness with machine learning methods using compositional features*. 2019**.** https://data.mendeley.com/datasets/jm79zfps6b/1 (accessed Jan 26, 2019)
58. Gao, F. Hardness Estimation of Complex Oxide Materials. *Phys. Rev. B* **2004**, *69*, 094113.
59. Šimůnek, A.; Vackář, J. Hardness of Covalent and Ionic Crystals: First-Principle Calculations. *Phys. Rev. Lett.* **2006**, *96*, 085501.
60. Berger, M. J.; Hubbell, J. H. *NIST X-Ray and Gamma-Ray Attenuation Coefficients and Cross Sections Database*; U.S. Department of Commerce: Gaithersburg, MD 1990.
61. Dimitrov, V.; Komatsu, T. Correlation among Electronegativity, Cation Polarizability, Optical Basicity and Single Bond Strength of Simple Oxides. *J. Solid State Chem.* **2012**, *196*, 574–578.
62. Plenge, J.; Kühl, S.; Vogel, B.; Müller, R.; Stroh, F.; von Hobe, M.; Flesch, R.; Rühl, E. Bond Strength of Chlorine Peroxide. *J. Phys. Chem. A* **2005**, *109*, 6730–6734.
63. Nembrini, S.; König, I. R.; Wright, M. N. The Revival of the Gini Importance? *Bioinformatics* **2018**, *34*, 3711–3718.
64. Ishwaran, H. The Effect of Splitting on Random Forests. *Mach. Learn.* **2015**, *99*, 75–118.
65. Boser, B. E.; Guyon, I. M.; Vapnik, V. N. A Training Algorithm for Optimal Margin Classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory - COLT '92*; ACM Press: New York, NY, 1992; pp 144–152.
66. Yang, C.; Fernandez, C. J.; Nichols, R. L.; Hsu, C.-W.; Chang, C.-C.; Lin, C.-J. *A Practical Guide to Support Vector Classification*.
67. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
68. Rasmussen, C. E.; Williams, C. K. I. *Gaussian Processes for Machine Learning*; MIT Press: Cambridge, MA; 2006.
69. Matérn, B. Spatial Variation. In *Lecture Notes in Statistics*; Springer-Verlag New York: New York, NY, 2013; Vol. 36.

70. Stoica, C.; Verwer, P.; Meekes, H.; van Hoof, P. J. C. M.; Kaspersen, F. M.; Vlieg, E. Understanding the Effect of a Solvent on the Crystal Habit. *Cryst. Growth Des.* **2004**, *4*, 765–768.
71. Liu, Y.; Lai, W.; Yu, T.; Ma, Y.; Kang, Y.; Ge, Z. Understanding the Growth Morphology of Explosive Crystals in Solution: Insights from Solvent Behavior at the Crystal Surface. *RSC Adv.* **2017**, *7*, 1305–1312.
72. Chen, X.-Q.; Niu, H.; Li, D.; Li, Y. Modeling Hardness of Polycrystalline Materials and Bulk Metallic Glasses. *Intermetallics* **2011**, *19*, 1275–1281.
73. Tabor, D. *The Hardness of Metals*; Oxford University Press: Oxford, United Kingdom, 2000.
74. Pugh, S. F. XCII. Relations between the Elastic Moduli and the Plastic Properties of Polycrystalline Pure Metals. *London, Edinburgh, Dublin Philos. Mag. J. Sci.* **1954**, *45*, 823–843.