Editors' Suggestion

Extending Shannon's ionic radii database using machine learning

Ahmer A. B. Baloch,[1] Saad M. Alqahtani,[2] Faisal Mumtaz,[3] Ali H. Muqaibel,[4] Sergey N. Rashkeev,[5]
and Fahhad H. Alharbi[2,4,*]

[1]*Research & Development Center, Dubai Electricity and Water Authority (DEWA), Dubai 564, United Arab Emirates*
[2]*Center of Research Excellence in Nanotechnology, King Fahd University of Petroleum and Minerals, Dhahran 31261, Saudi Arabia*
[3]*Open Systems International Inc., Montreal H4P2G7, Quebec, Canada*
[4]*Electrical Engineering Department, King Fahd University of Petroleum and Minerals, Dhahran 31261, Saudi Arabia*
[5]*Department of Materials Science and Engineering, University of Maryland, College Park, Maryland 20742, USA*

(Received 1 January 2021; accepted 11 March 2021; published 15 April 2021)

In computational material design, ionic radius is one of the most important physical parameters used to predict material properties. Motivated by the progress in computational materials science and material informatics, we extend the renowned Shannon's table from 475 ions to 987 ions. Accordingly, a rigorous machine learning (ML) approach is employed to extend the ionic radii table using all possible combinations of oxidation states (OS) and coordination numbers (CN) available in crystallographic repositories. An ionic-radius regression model for Shannon's database is developed as a function of the period number, the valence orbital configuration, OS, CN, and ionization potential. In the Gaussian process regression (GPR) model, the reached R^2 accuracy is 99% while the root mean square error of radii is 0.0332 Å. The optimized GPR model is then employed for predicting a new set of ionic radii for uncommon combinations of OS and CN extracted by harnessing crystal structures from materials project databases. The generated data are consolidated with the reputable Shannon's data and are made available online in a database repository.

DOI: 10.1103/PhysRevMaterials.5.043804

I. INTRODUCTION

In computational materials design, ionic radii are essential physical features for the prediction of crystal structure and material properties [1–3]. Data-driven studies have successfully employed the ionic radius to capture the physical and chemical behaviors in a variety of applications including crystallographic nature [4–6], batteries [7,8], scintillators [9], semiconductor absorbers [10–12], seawater properties [13], and mineralogy [14,15]. In this background, computing the ionic radius for arbitrary oxidation states and coordination geometries to study material properties is of a considerable theoretical and applied interest.

Ionic radius is not a fixed value for a particular ion but it changes with oxidation state, coordination environment, and orbital configurations among other properties. It is defined by the distance between the nucleus of a cation (anion) and its adjacent anion (cation) in a crystal structure. However, calculating the ionic radii is a complicated problem as the electron distribution is probabilistic and forms clouds without clear boundaries to demarcate different ions. Therefore, many scientists have worked on reasonable ways of defining the ionic radii of different chemical elements [16–18]. The most influential work was carried out by Shannon who predicted and compiled the ionic radii data for common oxidation states (OS) and coordination numbers (CN) [16]. It resulted in the most acceptable databases of ionic radii collection from Refs. [16–18]. Albeit this data has been used extensively in different fields, the Shannon's table is incomplete, as it does not cover the entire periodic table due to a lack of structural information for all common and uncommon OS with all possible coordination CN. Moreover, the absence of this data has rendered difficulties in predicting and screening novel compounds. For instance, most of the material searching for new perovskite absorbers is restricted only to cations and anions available in Shannon's ionic radii collection [11,19–23]. Researchers designing halide perovskites have been using a wide range of effective ionic radius values for tin ion (Sn^{2+}) (from 0.93 to 1.36 Å) [20–23] resulting in misleading analysis in numerous case studies. Therefore, extending Shannon's database could provide missing ion information for different materials. Fortunately, during the last decades, computational material science has evolved rapidly thanks to useful crystal prediction tools such as USPEX [24] and CALYPSO [25] and expanding databases for crystal structures [26]. Moreover, a huge amount of experimental data has been accumulated that can be employed for empirical correlations and computational materials design [27]. The ideal design or model should be able to connect any type of physical and chemical properties of a compound to its constituent parameters [28]. Driven by the growth of material informatics, we developed a robust ionic radii model for a complete periodic table and tabulated the data of the missing ions in the already existing databases for the research community [29–32]. The predicted Shannon's ionic radii can be used for classifying crystal structures, tolerance factors, geometrical properties, etc. This research is timely needed and relevant to the evolving material

[*] fahhad.alharbi@kfupm.edu.sa

informatics field; but it will also find applications in many other areas.

Theoretically, the ionic radius is a fundamental property of the atom that gains or loses an electron from its valence shell. The contribution from the valence electron modifies the ionic radius in two ways: (1) repulsions between valence electrons are changed due to the increase/decrease in the number of electrons, and (2) the effective nuclear charge experienced by the remaining core electrons is altered by adding or removing valence electrons. The seminal related works can be traced back to Goldschmidt [33], followed by Pauling [18] and Zachariasen [34]. Wasastjerna [35] originally calculated the radius of ions using their relative volumes as measured from optical spectroscopy. Pauling introduced an effective nuclear charge to consider the distance between ions as a sum of an anionic and a cationic radius with a fixed radius of 1.40 Å for O^{2-} ion [18]. A comprehensive analysis of crystal structures then subsequently led to the publication of the updated ionic radii by Shannon [16], with specific CN and different OS. To be consistent with Pauling's ionic radius values, Shannon had used an ionic radius of 1.40 Å for O^{2-} and called it an "effective" ionic radius. The effective ionic radii for nitrides was then calculated by Baur [36] and for sulfides/fluorides by Shannon [37,38]. Zachariasen [39] developed a functional form for calculating the bond length for oxygen and halogen compounds of d-orbital and f-orbital elements. The effective ionic radii of the trivalent and divalent rare-earth ions were predicted by Jia [40] who calculated unknown radii for different coordination numbers, ranging from 6 to 12, for $4f$ orbital elements. Recently, the effective ionic radii for anions have been calculated for binary alkali compounds via accessing a subset of suitable crystals from materials project [41]. Ouyang carried out a comprehensive study for designing perovskite materials using an ionic radii based ML descriptor [42].

Shannon derived the empirical ionic radius, called effective ionic radii, by systematically reproducing mean experimental cation-anion distances in crystal structures using the equation $R_{(i,\text{anion})} + R_{(i,\text{cation})} = d_{(\text{anion-cation})}$ where $R_{(i,)}$ is the ionic radii and d is the interatomic distance. The data derived by Shannon was formulated for 1000 average interatomic distances and empirical bond-length/bond-strength values [37]. Corrections to the radii were carried out for physical parameters using correlations between: (1) ionic radii and unit cell volume, (2) ionic radii and CN, (3) ionic radii and OS, and (4) ionic radii and orbital configuration [16]. However, the main limitations in Shannon's data arise from its origin in using primarily the oxide ion and hence poses a challenge when computing the cation radii for other anions. To address the issue of different anions, one can use the concept of difference in empirical ionic radii [43] as different approaches for calculating the ionic radius give relative values with a similar trend due to the geometric nature of ions. To make this comparison, one can subtract the oxide radius (O^{2-}) [16] from another anion radius (say sulfide ion, S^{2-}). Negative difference would then suggest that sulfide cation-anion distances are smaller than in the oxide of the same element [38,43].

Rationally it is noticed from the literature that ionic radii depend on many features, and their calculation requires us to take into account corrections for OS and CN. Furthermore, for better comprehension of the ionic radius and its relationship to the physical and chemical properties, interaction between ions and their electronic shells should be accurately taken into account using valence electrons in s, p, d, f orbital configurations. In addition, most of the literature work involves interpolation or extrapolation without making a generalized model based on physical descriptors [36,39–41,44]. Nonetheless, it is noticed that the differences between various methods are not random and follow particular trends. Therefore, for materials informatics, it is important to use a single standard.

Although the focus of this paper is on ionic radii, assigning suitable OS to an ion plays an important role in determining material properties. The standard definition recommended by the International Union of Pure and Applied Chemistry (IUPAC) for OS of an atom is "the charge of this atom after ionic approximation of its heteronuclear bonds" [45]. However, the practical implementation of this definition is difficult since it is not general for any ion and the rules vary for different elemental families [46]. As a consequence, it is a common practice to employ linear combination of atomic orbitals (LCAO) and electron counting for molecules. However, for inorganic crystalline compounds, electron balancing does not work very well as they are dependent on valence bond lengths and valence bond orders [47]. In these types of compounds, the oxidation states are measured using the local geometry based bond-valence (BV) analysis [48]. In this method, both metal-ligand bonds are approximated and assumed to be completely ionic. Also, the oxidation states are determined by adding up valence-bond lengths. Currently, most of the online materials' data repositories such as Cambridge Structural Database (CSD) [49] and Materials Project [32] have embedded autoprediction of OS using BV method. An alternative to the IUPAC algorithm was recently proposed by Postils *et al.* [50] to develop the effective OS (EOS) system utilizing chemical information from wave functions and not solely relying on the Lewis based approach in order to form a generic OS assignment scheme. For polyatomic ions and metal complexes, quantum chemical calculations based on wave function are also employed for electron portioning [51–53], whereas experimental spectroscopic techniques such as x-ray photoemission spectroscopy (XPS), near edge x-ray absorption fine structure (NEXAFS), and neutron spectroscopy can be used for inferring OS as well [54]. Similarly, assigning neighboring atoms and bonding types play an important role in defining CN [55]. Depending on the compound type (oxides or intermetallic), a broad range of CN evaluation algorithms exists. They are based on either local geometry or interatomic distances [56]. For example, Brunner [57] suggested a threshold value of interatomic distance for determining CN, whereas O'Keeffe and Brese [58] proposed CN prediction using bond-valence summation for assessing nearest-neighboring atoms. Interestingly, most of the state-of-the art methods for determining CN such CrystalNN [55], valence-ionic radius estimator [59], and ChemEnv [56] are based on the geometric principle of the Voronoi diagram for polyhedron [60]. Although the current ML method for the ionic radii is for arbitrary CN and OS; however, it could be useful to read about the recent benchmark study for evaluating CNs and the nuances of determining local environments [55].

Accordingly, we present a rigorous machine learning (ML) approach to extend the ionic radii table of Shannon's database

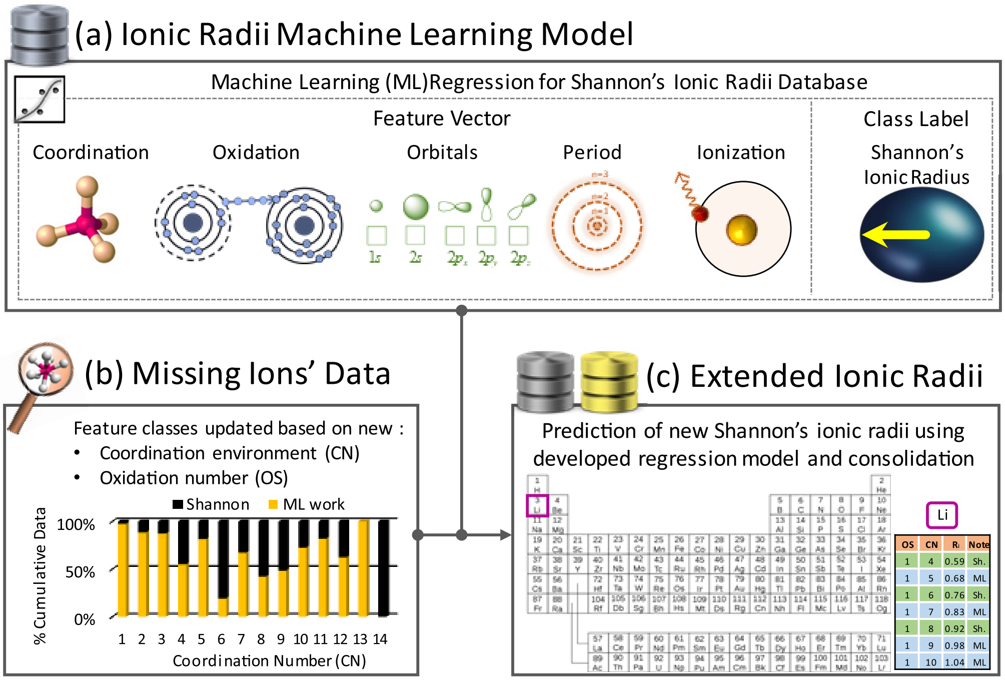

FIG. 1. Workflow for extending the Shannon's ionic radii database using machine learning and data harvesting. (a) Ionic radii machine learning model for the existing Shannon's database using physically guided features. (b) Collection of missing ions data from material repositories and its comparison with Shannon's database in terms of CN: 1–14. (c) Consolidation of both Shannon table and missing species' ionic radii databases from predicted values using missing ions' data and regression model.

using all possible combinations of OS and CN available in material informatics repositories [29,32]. Rare oxidation states and coordination environments, as well as those missing in Shannon's database, were considered. Data for different combinations of OS and CN were carefully harvested from the crystallographic database for 7969 crystal structures using material repositories with the original data stemming from experimental databases [29,30,49]. Figure 1 shows the flowchart adopted for the regression problem with selected features. Regression on the Shannon's ionic radius as a function of the period number in the periodic table of elements, OS, CN, ionization potential (E_{IP}), and the valence orbital configuration (s, p, d, f) was performed. The developed model is valid for all elemental families and not just limited to specific classes. The descriptors presented here are relatively simple and physically intuitive, relying only on eight fundamental parameters to describe an ion of any element in a periodic table. Several state-of-the-art ML-based regression models were employed including linear regression (LR), support vector machines (SVM), decision trees (DT), and Gaussian process regression (GPR). All the used methods worked satisfactory; however, the GPR model showed the best predictive accuracy. For training and testing, sevenfold cross validation was performed for the Shannon's table [16]. By optimizing the hyperparameters of the ML algorithms, Gaussian process regression (GPR) showed the minimum root mean square error (RMSE) of 0.0332 Å with a coefficient of determination (R^2) reaching 99.3%. These results illustrate effective implementation of the regression model which was then employed for predicting a new set of ionic radii for uncommon combinations of OS and CN. We extended the Shannon's ionic radii table from 475 to 987 ions by predicting ionic radii for 512 new compounds. The generated data was then consolidated with the reputable Shannon's table as shown in Fig. 2. The newly developed table should assist accurate prediction of crystal structures by considering the ionic radius value based on the exact OS/CN, rather than the common OS/CN, which translates to better prediction of material properties. The resulting data has been made available online in open database repositories for research.

II. METHODOLOGY

Prediction of Shannon's ionic radii (denoted as R_i in this paper) for missing materials and ions in Shannon's database was performed using supervised ML and data harvesting as described in Fig. 1. The regression model for all periodic table elements was evaluated on a test/train split approach with sevenfold cross validation. Almost the same results are obtained using k-fold cross validation for k ranged between 4 and 10. The obtained mean errors for all the cases are ranged between 0.0328 and 0.0391 Å. Please see Fig. S3 and Table S4 in the Supplemental Material [61]. Statistical correlation for strength and direction of the feature/target vector was evaluated using Spearman's rank coefficient (ρ). For data mining, missing ions' data in terms of coordination numbers and oxidation states were carefully collected from the ICSD, initially rooting from the experimental data [29]. The developed regression model is applied for the prediction of new ionic radii based on the recently harvested coordination environments and oxidation states from Materials Project. The predicted data are then consolidated with existing databases to

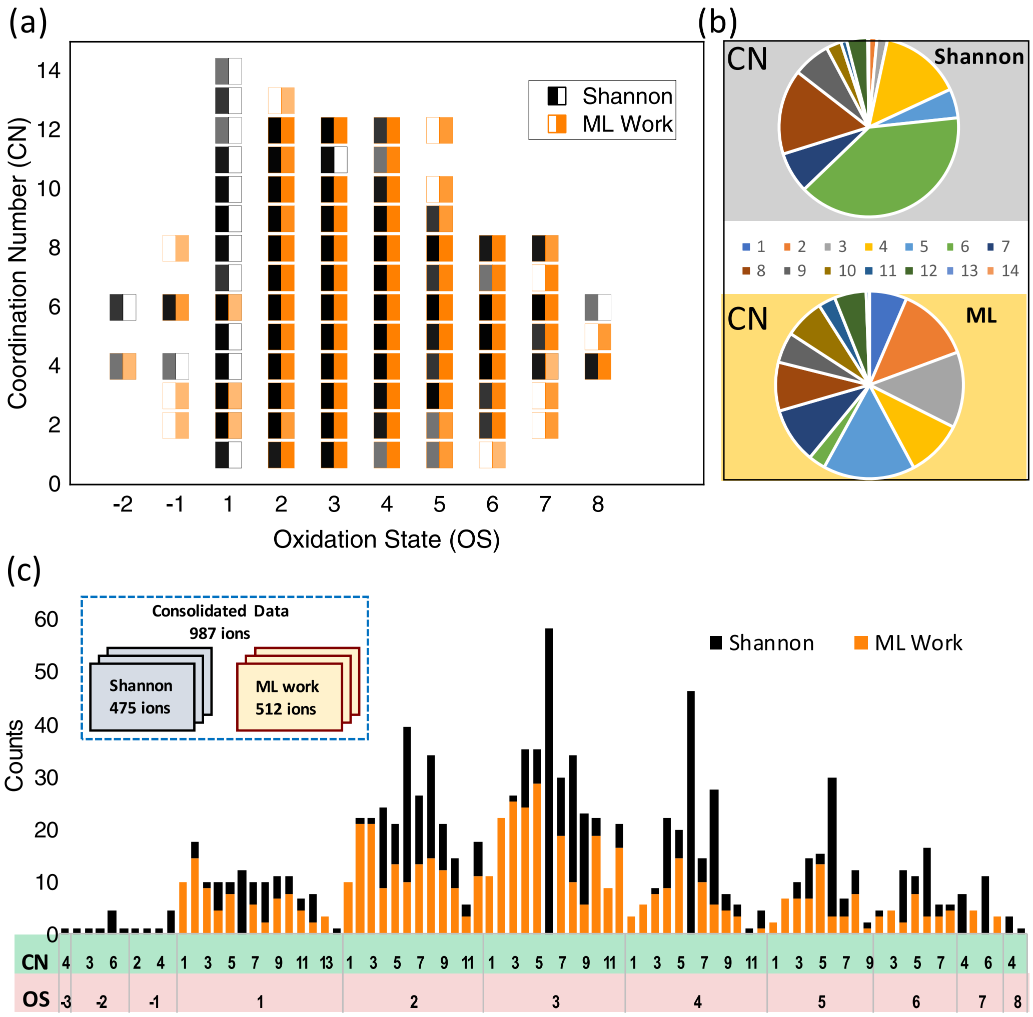

FIG. 2. Data comparison for the chemical environment in terms of oxidation state (OS) and coordination number (CN) from Shannon and the present ML work. (a) Regions for features OS and CN covered by Shannon and ML work. (b) CN: 1–14 present in both the databases considered. (c) CN present in each OS analyzed with an inset showing total Shannon and ML work data. In all these graphs, ML work consists of 512 unique ionic information harvested.

extend the ionic information in tabulated form for ease of use for the research community.

A. Data and features

Shannon's data set contains about 475 ions [16]. The table includes the ion, oxidation (formal charge), coordination, and ionic radius. We selected physically guided features for establishing a regression model for the target function of Shannon's ionic radius as shown in Fig. 1(a). The features based on the nature of the ionic radius are listed below:

(i) Atomic properties

(a) period number,

(b) *s*-orbital outer shell valence electrons,

(c) *p*-orbital outer shell valence electrons,

(d) *d*-orbital outer shell valence electrons,

(e) *f*-orbital outer shell valence electrons,

(f) $e^{-E_{\mathrm{IP}}}$, ionization potential (negative exponent).

(ii) Ionic properties

(a) oxidation state (OS),

(b) coordination number (CN).

Data for elemental properties were extracted from the web of elements [62], whereas OS, CN, and R_i were adopted from the Shannon's compilation [16] which includes Pauling [18] and Ahrens [17] as well. The descriptors presented here are relatively simple, having only eight parameters to describe an ion of any element in the periodic table. However, it has not been benchmarked for high or low spin materials as the number of counts in the original Shannon's table is low to make any statistical significance for such a model. Orbital outer shell valence electrons as a feature can be explained by considering an example of scandium with orbital configuration, Sc:$1s^2\ 2s^2\ 2p^6\ 3s^2\ 3p^6\ 4s^2\ 3d^1$. When condensed [Sc] uses a noble gas configuration it becomes [Ar] $4s^2\ 3d^1$ and accordingly our input feature vector for [$s\ p\ d\ f$] would be [2 0 1 0].

For a detailed clarification of the materials' data acquisition methodology and harnessing data for 7969 oxides, the Materials Project Representational State Transfer (REST) Application Programming Interface (API) was interfaced using the Python Materials Genomics (pymatgen) library [32,59]. The raw computed data were acquired based on user-defined criteria and then utilized to perform post-processing analysis to derive further properties of the materials via the pymatgen library. The pymatgen is an open-source library that has several packages such as core, electronic_structure, entries, io, etc. In this work, OS has been estimated using the BV module available in Python Materials Genomics (pymatgen) code [59], whereas CN is extracted from the ChemEnv module employed by Waroquiers *et al.* [56]. For OS, BV sum is calculated for all unique symmetrical points in the structure using elemental parameters tabulated by O'Keefe and Brese [58] and $\mathrm{BV} = \exp[(L_0 - L)/B]$ relation. Here L is bond length between two atoms and BV indicates bond strength. L_0 is the single bond length and it is dependent on OS and CN implicitly, whereas B is a constant factor usually kept at 0.37 [63]. Then the maximum *a posteriori* probability (MAP) estimation is carried out to find an OS combination that should result in a charged balance cell. For CN, materials space of the considered 7982 oxides in the work of Waroquiers *et al.* [56] were used as an input criterion to obtain the matching existed data in the Materials Project database [32] using query class, which is based on MongoDB-like syntax. Details of the CN extraction method are provided in the recent work of Waroquiers *et al.* [56] where they have compared the distorted local environment in structures with perfect polyhedral geometries to find symmetry measures and predict CN accordingly. Their data were curated using low pressure stable phases and energy above hull smaller than 100 meV/atom. Moreover, ions with partial vacancies were neglected and only the oxygen anion was considered (for example, no oxisulfide). After verifying their existence in ICSD as either experimental or theoretical data, we were left with 7969 oxides which represented an approximately 80% of the structures.

B. Regression procedures

State-of-the-art supervised learning models were used to develop the ionic radii ML model for Shannon's data. The model was trained with an objective function of RMSE minimization. First, we performed sevenfold cross validation to learn the hyperparameters and avoid overfitting. In the second stage, the best hyperparameter model for which k-fold reports the lowest error was then selected to test the model for prediction. Using this method, we chose the ML model with the best average prediction error. Hyperparameters in fitting the model are automatically determined internally for each regression algorithm using MATLAB [64]. The following machine learning models and their subclasses applied for this work are mentioned below:

(i) Linear regression (LR): linear, interactions linear, robust linear, and stepwise linear.

(ii) Support vector machines (SVM): linear, quadratic, cubic, fine Gaussian, medium Gaussian, and coarse Gaussian.

(iii) Decision trees (DT): fine, medium, and coarse.

(iv) Ensemble of decision trees: boosted trees and bagged trees.

(v) Gaussian process regression (GPR): squared exponential, exponential, rational quadratic, Matern 5/2 and Matern 3/2 kernel.

C. Extension methodology for ionic radius

We analyzed the chemical information (OS and CN) for a total of 7969 crystal structures from available databases to extend the ionic radii database, as highlighted in Fig. 1(b). Ionic information for 7969 oxides was extracted from the Materials Project Database [32] for ICSD [29]. They are summarized (excluding duplicates) in the Supplemental Material Table S1 [61]. This resource provided oxidation states and coordination environments for species, which were necessary for the correct prediction of missing ions. A total of 512 unique ions (in terms of CN and OS) were found after removing the duplicates from the Shannon's database. Accordingly, these unique ionic features consisting of a new oxidation state and coordination environment along with their elemental properties were then curated for the prediction of R_i.

GPR model was selected based on the lowest RMSE and highest R^2 achieved among other regression models. The technical details of the GPR model are provided in the Supplemental Material [61]. This ML model was then supplied with missing ionic properties (OS and CN) shown in Fig. 1(b) along with their respective elemental properties vector (period, E_{IP}, outer shell valence electrons in s, p, d, f orbitals) to extend the ionic radii database. The predicted values were then consolidated with Shannon's original data to build an up-to-date comprehensive table of 987 species and their respective ionic radii. It should be noted that in the case of an overlap, we kept the original Shannon's empirical values, i.e., no value of Shannon is altered in the proposed improved table. Figure 1(c) displays the web interface [65] in a periodic table style that was created for dissemination of the results, which can be valuable to many natural sciences.

III. RESULTS AND DISCUSSION

Prediction of the ionic radii for missing ions in Shannon's database was performed using supervised machine learning and data harvesting from materials project. The developed GPR regression model was primarily used to predict ionic radii of rare oxidation states and coordination numbers not considered in the current ionic radii databases. Regression algorithms as function $g(x)$, physical features of the period number OS, CN, $\exp(-E_{\mathrm{IP}})$, and outer shell valence electrons in s, p, d, f orbitals were employed to Shannon's ionic radii database in the form of

$$R_i = g[\mathrm{OS}, \mathrm{CN}, \mathrm{Period}, s, p, d, f, \exp(-E_{\mathrm{IP}})]. \quad (1)$$

It is important to highlight that the ionic radii predicted from this model are extensions of Shannon's ionic radii.

A. Data analysis

We have analyzed the ionic radii data for the common and uncommon oxidation states and coordination numbers by comparing Shannon's data to current online material databases. Rare oxidation states and coordination environments, as well as those missing in Shannon's database, are considered for extending the ionic radii database. To highlight the gaps in Shannon's table and this study's contribution, called "ML work" hereon, a visualization for OS and CN parameter space is provided in Fig. 2(a). The scatter plot

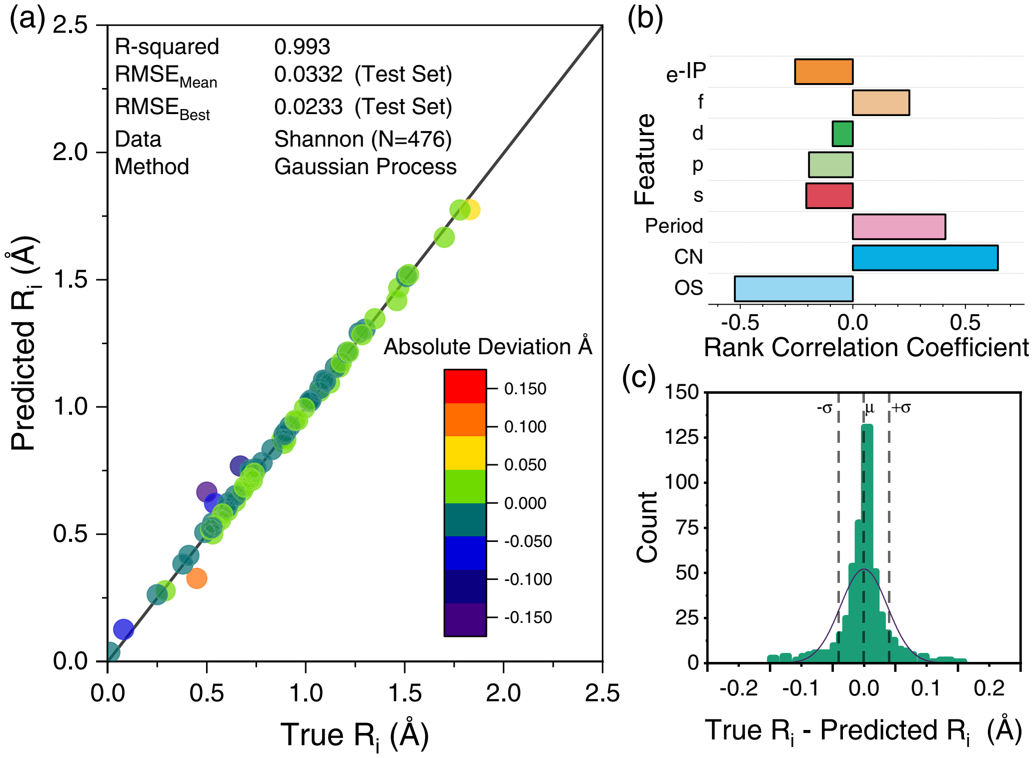

FIG. 3. Results from regression analysis for Shannon's data (testing set) using a feature vector containing period number, CN, OS, outer shell valence electrons in orbitals s, p, d, f, and $\exp(-E_{\mathrm{IP}})$ for ionic radius prediction. (a) Best performing Gaussian process regression (GPR) model for predicting R_i with an inset showing absolute deviation error (true R_i predicted R_i), (b) rank correlation coefficient for R_i with the feature vector, and (c) distribution for absolute deviation and corresponding standard deviation (σ) of 0.0398 Å.

shows the regions of OS and CN by Shannon and ML work with the color showing the density of the occurrences.

Naturally the majority of the data points in Shannon's region covers common OS and CN, whereas ML work was able to identify uncommon unreported regions of CN as shown in the pie charts of Fig. 2(b). Shannon's data (total 475 ions considered) comprises primarily the following common CN geometries: CN = 6 with octahedral and trigonal prism has 187 occurrences, CN = 4 with tetrahedral and square planar geometry has 70 ions, whereas CN = 8 has a frequency of 73. These three covers 69.5% of the total CN space reported by Shannon [16]. On the other hand, data harvested for ML work and missing ions showed primarily rare CN and OS. For instance, out of 512 new unique ions, 15.8% were found in CN = 5 with coordination geometry of trigonal bipyramidal and square pyramidal. Shannon's data, on the contrary, had only 5.2% of CN = 5. Similarly, CN = 2 for ML work had a 12.8% occurrence, whereas Shannon's table had 1.47% respectively in their corresponding data sets. In terms of formal charge on the ion, i.e., oxidation state, most of the data for Shannon's table is cation as the database itself was developed using O^{2-} anion with a sixfold coordination number and the ionic radius of 1.40 Å. Figure 2(c) shows the summary for a CN connecting to a particular OS covered by the present work and Shannon. Both these sources provide a total of 987 species, with the majority of the ions found in OS = 2 (23.7%) and OS = 3 (30.1%) for both sets together. A total of 512 unique ions in terms of CN and OS were found after removing the duplicate information in the data sets as shown in Fig. 2(c). These new unique ionic features consisting of OS, CN, and elemental properties were then employed to extend the ionic radii database using GPR.

B. Regression analysis for Shannon's database

Regression on the ionic radius as a function of the period number, oxidation state, coordination environment, electron affinity, ionization potential, and orbital configuration was performed for Shannon's data. For evaluating the model accuracy and features, these measures were used:

Root mean square error (RMSE):

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(R_i - \tilde{R}_i)^2}, \tag{2}$$

R-square:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(R_i - \tilde{R}_i)^2}{\sum_{i=1}^{n}(R_i - \bar{R}_i)^2}, \tag{3}$$

Spearman's rank correlation coefficient:

$$\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n^3 - n}. \tag{4}$$

Here n is the number of observations, $\tilde{R}_i$ is the predicted Shannon's ionic radius, and $\bar{R}_i$ is the mean of Shannon's ionic radii. $d_i = \mathrm{rank}(x_i) - \mathrm{rank}(y_i)$ is the difference between the two ranks of each observation in different variables.

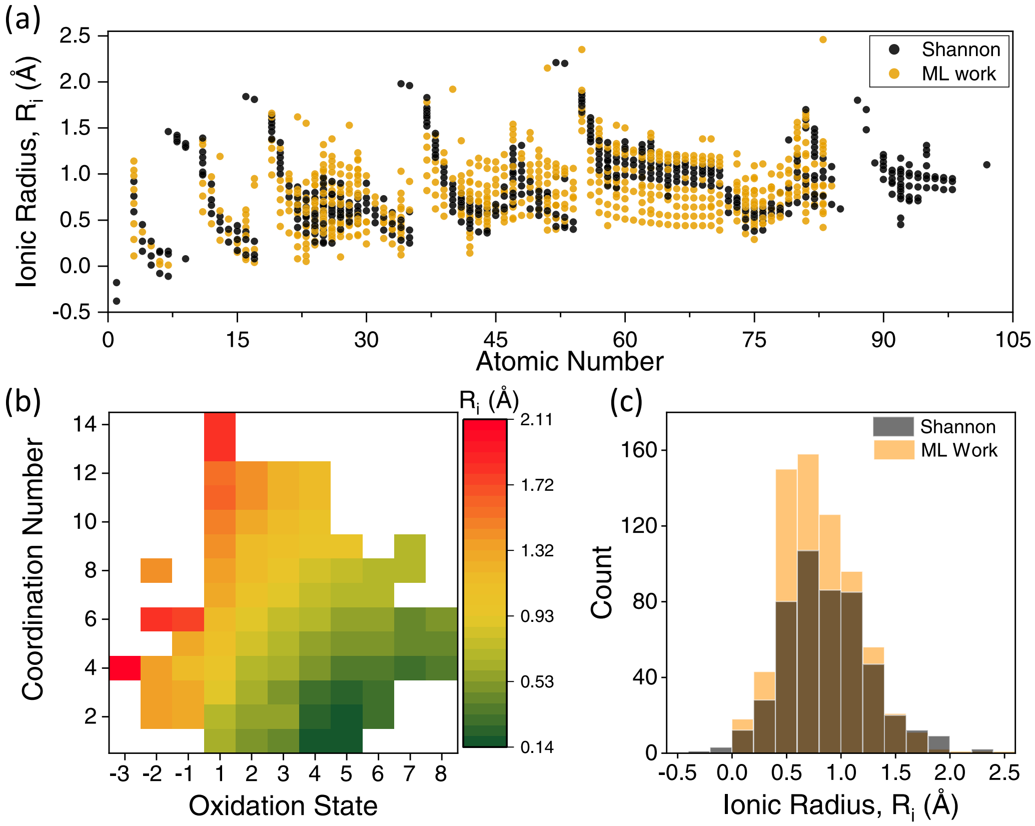

FIG. 4. Extension to new ions with results from the prediction and consolidated database. (a) Ionic radii predicted from ML work (GPR model) merged with Shannon's database as a function of atomic number. (b) General behavior of ionic radius as a function of coordination number and oxidation state for $N = 987$ consolidated species. (c) Histogram of Shannon's empirical database and current ML work showing the similarity of the ionic radii distribution.

Twenty different regression models were developed as highlighted in the Supplemental Material [61]. Among all the developed models, the GPR model with the Matern 3/2 kernel function (details in the Supplemental Material [61]) was the most accurate in predicting the ionic radii with an R^2 of 99.3% and RMSE mean of 0.0332 Å. Figure 3(a) shows the results for testing set error (sevenfold validation) using the fixed optimal hyperparameter model. Using the selected features, Fig. 3(a) shows the regression results of the GPR model where it was able to achieve the minimum RMSE mean of 0.0332 Å over sevenfold. The results are promising as this is a first general-purpose model for all periodic elements, whereas previous attempts have separately dealt with transition metals, nitrides, sulfides, etc. [36,38,40]. Moreover, the main advantage of GPR is that it directly captures the model uncertainty. To assess the robustness of the model, we performed sevenfold cross validation on Shannon's data of 475 ions where the hyperparameters of the model were optimized using the quasi-Newton approach with a function tolerance of 10^{-6} for Matern 3/2 kernel. The color bar in Fig. 3(a) shows the absolute deviation $|R_i - \tilde{R}_i|$ for each point where the majority of the data points lie in the standard deviation of ±0.0398 Å as shown in the inset figure. To validate the significance of the eight features—period number, OS, CN, $\exp(-E_{\mathrm{IP}})$, and orbital configuration s, p, d, f—we assess the Spearman's rank correlation coefficient (ρ) between individual feature and ionic radii as shown in Fig. 3(b). In the rank correlation, the maximum correlation happens at a value of 1, whereas the direction is shown by the positive or negative sign. Interestingly, the maximum correlation $\rho = +0.64$ was found for CN. It shows that with an increase in CN, R_i also increases. OS was found to have $\rho = -0.52$ showing a negative correlation with R_i. This is because as the OS increases, atoms lose electrons hence the overall effective nuclear charge increases resulting in reduced R_i. Period number resulted in a value of +0.41 due to the addition of outermost shell which causes the R_i to increase. This was followed by $\exp(-E_{\mathrm{IP}})$ with $\rho = -0.25$. Figure 3(c) shows that the absolute deviation from the model follows normal distribution, signifying that there is no systematic error present in the developed regression method.

C. Extension to missing ions

We performed an extension of the new ions using the GPR prediction model on a set of 512 ions as depicted in Fig. 4(a). The completed consolidated table is shown in the Supplemental Material Table S2 [61] with the feature vector and ionic radii values. Here we kept the original Shannon's empirical values where we found an overlap, i.e., no ionic radius from Shannon was changed in the proposed table. Figure 4(a) shows that the predicted elements were inclusive in the range of atomic numbers by Shannon (from 1 to 102). In Fig. 4(b) we see the effect of the oxidation state and coordination number on the total consolidated data set. The radii decrease with an increase in the oxidation state due to the additional effective nuclear charge by losing an electron. The ionic radius was found to increase with higher CN because the electron field is stretched out by the existence of additional surrounding ions for higher dimensional polygons.

Figure 4(c) shows the distribution of ionic radius predicted and Shannon's original database. The shape of these histograms shows that the predicted R_i is comparable to the data set used for Shannon's ionic radii. The difference is primarily due to the number of counts and rare OS/CN in our feature vectors. The descriptors are relatively simple and physically intuitive, having only features highlighting its robustness for extending to new ions as they appear in online databases. The completed consolidated table is uploaded at the open materials database website. A comparison of ionic radii for cations calculated using Brown and Shannon (BS) [66], Shannon [16], Ouyang [42], and current ML work is presented in the Supplemental Material Fig. S2 [61]. It should be noted that as the code can predict ionic radii for any arbitrary OS/CN, care must be taken when selecting chemical environment information for realistic ionic radii calculation.

D. Further analysis

Our predictive model was extended to new ions based on compounds that were primarily experimentally observed. According to our statistical analysis, about 96% of these compounds (7658 of 7969) are experimentally associated with the multiple Inorganic Crystal Structure Database (ICSD) IDs [56] (the complete data are provided in the Supplemental Material Table S4 [61]). Hereunder, arbitrarily selected ions from our predicted ionic radii data set are discussed to illustrate the importance of our predictive model. For crystal structure classification, revised ionic radii have been adopted for accurate tolerance factor predictions [20]. Searching for Sm^{2+} based perovskites, Travis *et al.* [20] had to apply Sm^{2+} (CN-7) ionic radius as Shannon's table provides no 6 coordinate Sm^{2+} ion [20]. An experimental investigation of $BiCoO_3$ with pyramidal polar coordination of Co^{3+} (CN-5) under high pressure was implemented and spin transition was observed due to the change of Co^{3+} (CN-5) in the atmospheric pressure phase to the approximately isotropic octahedral coordination (CN-6) in the high-pressure phase [67]; this is not covered in the original Shannon table. Also, the structural and magnetic properties of $LiRO_2$ (R = rare earth) were experimentally studied by Hashimoto *et al.* [68]. The x-ray diffraction measurements have shown that the $LiErO_2$ compound was found to form β-type (space group: $P21/c$) with Li^+ (CN-3) below room temperature [68]; this is not covered in the original Shannon table as well. For octahedral coordination of Sn^{2+} with (CN-6), there are efforts to stabilize the $CsSnCl_3$ perovskite structure, which prefers to form pyramidal coordination instead of octahedral coordination, by ionic substitution of Sn^{2+} with smaller octahedral cations [69]. The octahedral coordination Sn^{2+} was used as a possible candidate to replace Pb^{2+} in perovskite structure due to its toxicity [69]. This experimentally observed OS of Sn is not included in the original Shannon's table. Table I lists the considered oxidation states that are not tabulated in Shannon's table. For a concise reference, all of them are among the listed oxidation state in the seminal book by Greenwood and Earnshaw [70]. Many other references can be found for each one of them. Nonetheless, they are certainly not among the common oxidation states but they cannot be ignored.

TABLE I. The considered oxidation states that are not tabulated in Shannon's table

C: +2, +3	P: +4	S: +2, +3, +5
Cl: +1, +4	Sc: +2	Fe: +5
Co: +1	Ni: +1	Ge: +3
Se: +2	Y: +2	Nb: +2
Ru: +6	In: +1, +2	Sn: +2
Te: +5	La: +2	Ce: +2
Pr: +2	Gd: +2	Ir: +6
Pt: +6	Au: +2	Th: +3

IV. CONCLUSION

A very rigorous and highly accurate machine learning approach is employed to extend the renowned Shannon's table from 475 ions to 987 ions. In ML implementation, the original Shannon's table is used to develop the ionic-radius regression model as a function of the period number, the valence orbital configuration, OS, CN, and ionization potential. The model is then implemented to extend the ionic radii table for all possible combinations of OS and CN available in crystallographic repositories. Many ML methods are considered and a comparison was carried out. In the Gaussian process regression (GPR) model, the reached R^2 accuracy is 99% while the root mean square error of radii is 0.0332 Å.

The generated data are consolidated with the reputable Shannon's data and are made available online in a database repository [65].

ACKNOWLEDGMENT

S.M.A. would like to thank the Ministry of Education in the Kingdom of Saudi Arabia for the financial support by its Research and Development Office (RDO).

[1] W. Lu, R. Xiao, J. Yang, H. Li, and W. Zhang, Data mining-aided materials discovery and optimization, J. Materiomics **3**, 191 (2017).

[2] D. Xue, P. Balachandran, J. Hogden, J. Theiler, D. Xue, and T. Lookman, Accelerated search for materials with targeted properties by adaptive design, Nat. Commun. **7**, 11241 (2016).

[3] P. V. Balachandran, A. A. Emery, J. E. Gubernatis, T. Lookman, C. Wolverton, and A. Zunger, Predictions of new ABO_3 perovskite compounds by combining machine learning and density functional theory, Phys. Rev. Mater. **2**, 043802 (2018).

[4] M. G. Brik, A. Suchocki, and A. Kaminska, Lattice parameters and stability of the spinel compounds in relation to the ionic

radii and electronegativities of constituting chemical elements, Inorg. Chem. **53**, 5088 (2014).

[5] A. O. Oliynyk, L. A. Adutwum, J. J. Harynuk, and A. Mar, Classifying crystal structures of binary compounds ab through cluster resolution feature selection and support vector machine analysis, Chem. Mater. **28**, 6672 (2016).

[6] A. Seko, A. Togo, and I. Tanaka, Descriptors for machine learning of materials data, in *Nanoinformatics* (Springer, Singapore, 2018), pp. 3–23.

[7] J. C. Bachman, S. Muy, A. Grimaud, H.-H. Chang, N. Pour, S. F. Lux, O. Paschos, F. Maglia, S. Lupart, P. Lamp *et al.*, Inorganic solid-state electrolytes for lithium batteries: Mechanisms and properties governing ion conduction, Chem. Rev. **116**, 140 (2016).

[8] A. Ishikawa, K. Sodeyama, Y. Igarashi, T. Nakayama, Y. Tateyama, and M. Okada, Machine learning prediction of coordination energies for alkali group elements in battery electrolyte solvents, Phys. Chem. Chem. Phys. **21**, 26399 (2019).

[9] L. D. Williams and P. Ghanshyam, Machine Learning using local environment descriptors to predict new scintillator materials, Tech. Rep. (Los Alamos National Lab., Los Alamos, NM, 2018).

[10] G. Kieslich, S. Sun, and A. K. Cheetham, An extended tolerance factor approach for organic–inorganic perovskites, Chem. Sci. **6**, 3430 (2015).

[11] C. J. Bartel, C. Sutton, B. R. Goldsmith, R. Ouyang, C. B. Musgrave, L. M. Ghiringhelli, and M. Scheffler, New tolerance factor to predict the stability of perovskite oxides and halides, Sci. Adv. **5**, eaav0693 (2019).

[12] S. Lu, Q. Zhou, L. Ma, Y. Guo, and J. Wang, Rapid discovery of ferroelectric photovoltaic perovskites and material descriptors via machine learning, Small Methods **3**, 1900360 (2019).

[13] L. F. Mendes, L. Zambotti-Villela, N. S. Yokoya, E. L. Bastos, C. V. Stevani, and P. Colepicolo, Prediction of mono-, bi-, and trivalent metal cation relative toxicity to the seaweed *Gracilaria domingensis* (gracilariales, rhodophyta) in synthetic seawater, Environ. Toxicol. Chem. **32**, 2571 (2013).

[14] J. Kyziol-Komosinska *et al.*, Influence of properties of selected metal ions on their sorption onto neogene clays, Fresenius Environ. Bull. **18**, 1080 (2009), https://www.cabdirect.org/cabdirect/abstract/20093257109.

[15] G. D. Price and N. L. Ross, *The Stability of Minerals* (Springer Science & Business Media, New York, 1992), Vol. 3.

[16] R. D. Shannon, Revised effective ionic radii and systematic studies of interatomic distances in halides and chalcogenides, Acta Crystallogr. Sect. A **32**, 751 (1976).

[17] L. H. Ahrens, The use of ionization potentials part 1. Ionic radii of the elements, Geochim. Cosmochim. Acta **2**, 155 (1952).

[18] L. Pauling, The sizes of ions and the structure of ionic crystals, J. Am. Chem. Soc. **49**, 765 (1927).

[19] M. R. Filip and F. Giustino, The geometric blueprint of perovskites, Proc. Natl. Acad. Sci. USA **115**, 5397 (2018).

[20] W. Travis, E. Glover, H. Bronstein, D. Scanlon, and R. Palgrave, On the application of the tolerance factor to inorganic and hybrid halide perovskites: A revised system, Chem. Sci. **7**, 4548 (2016).

[21] M. Pazoki and T. Edvinsson, Metal replacement in perovskite solar cell materials: Chemical bonding effects and optoelectronic properties, Sustainable Energy Fuels **2**, 1430 (2018).

[22] Q. Chen, N. De Marco, Y. M. Yang, T.-B. Song, C.-C. Chen, H. Zhao, Z. Hong, H. Zhou, and Y. Yang, Under the spotlight: The organic–inorganic hybrid halide perovskite for optoelectronic applications, Nano Today **10**, 355 (2015).

[23] F. Liu, C. Ding, Y. Zhang, T. S. Ripolles, T. Kamisaka, T. Toyoda, S. Hayase, T. Minemoto, K. Yoshino, S. Dai *et al.*, Colloidal synthesis of air-stable alloyed $CsSn_{1-x}Pb_xI_3$ perovskite nanocrystals for use in solar cells, J. Am. Chem. Soc. **139**, 16708 (2017).

[24] C. W. Glass, A. R. Oganov, and N. Hansen, Uspex'evolutionary crystal structure prediction, Comput. Phys. Commun. **175**, 713 (2006).

[25] Y. Wang, J. Lv, L. Zhu, and Y. Ma, Calypso: A method for crystal structure prediction, Comput. Phys. Commun. **183**, 2063 (2012).

[26] S. Curtarolo, G. L. Hart, M. B. Nardelli, N. Mingo, S. Sanvito, and O. Levy, The high-throughput highway to computational materials design, Nat. Mater. **12**, 191 (2013).

[27] L. Ward, A. Agrawal, A. Choudhary, and C. Wolverton, A general-purpose machine learning framework for predicting properties of inorganic materials, npj Comput. Mater. **2**, 1 (2016).

[28] A. Jain, G. Hautier, S. P. Ong, and K. Persson, New opportunities for materials informatics: Resources and data mining techniques for uncovering hidden relationships, J. Mater. Res. **31**, 977 (2016).

[29] R. Allmann and R. Hinek, The introduction of structure types into the inorganic crystal structure database ICSD, Acta Crystallogr. Sect. A **63**, 412 (2007).

[30] M. Hellenbrandt, The inorganic crystal structure database (ICSD)—present and future, Crystallogr. Rev. **10**, 17 (2004).

[31] S. Trepalin, Y. E. Bessonov, B. Fel'dman, E. Kochetova, N. Churakova, and L. Koroleva, The structural chemical database of the all-Russian institute for scientific and technical information, Russian academy of sciences. An autonomous system for structural searches, Autom. Doc. Math. Ling. **52**, 297 (2018).

[32] A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder *et al.*, Commentary: The materials project: A materials genome approach to accelerating materials innovation, APL Mater. **1**, 011002 (2013).

[33] V. M. Goldschmidt, Die gesetze der krystallochemie, Naturwissenschaften **14**, 477 (1926).

[34] W. Zachariasen, A set of empirical crystal radii for ions with inert gas configuration, Z. Kristallogr. - Cryst. Mater. **80**, 137 (1931).

[35] J. A. Wasastjerna, *On the Radii of Ions* (Soc. Scientiarum Fennica, Helsinki, Finland, 1923).

[36] W. H. Baur, Effective ionic radii in nitrides, Crystallogr. Rev. **1**, 59 (1987).

[37] R. T. Shannon and C. T. Prewitt, Effective ionic radii in oxides and fluorides, Acta Crystallogr. Sect. B **25**, 925 (1969).

[38] R. Shannon, Bond distances in sulfides and a preliminary table of sulfide crystal radii, Struct. Bonding Cryst. **2**, 53 (1981).

[39] W. Zachariasen, Bond lengths in oxygen and halogen compounds of d and f elements, J. Less-Common Met. **62**, 1 (1978).

[40] Y. Jia, Crystal radii and effective ionic radii of the rare earth ions, J. Solid State Chem. **95**, 184 (1991).

[41] J. Gebhardt and A. M. Rappe, Big data approach for effective ionic radii, Comput. Phys. Commun. **237**, 238 (2019).

[42] R. Ouyang, Exploiting ionic radii for rational design of halide perovskites, Chem. Mater. **32**, 595 (2019).

[43] M. O'Keeffe, *Structure and Bonding in Crystals* (Elsevier, Amsterdam, 2012).

[44] V. Sidey, On the effective ionic radii for ammonium, Acta Crystallogr. Sect. B **72**, 626 (2016).

[45] P. Karen, P. McArdle, and J. Takats, Comprehensive definition of oxidation state (IUPAC recommendations 2016), Pure Appl. Chem. **88**, 831 (2016).

[46] D. Steinborn, The concept of oxidation states in metal complexes, J. Chem. Educ. **81**, 1148 (2004).

[47] A. Walsh, A. A. Sokol, J. Buckeridge, D. O. Scanlon, and C. R. A. Catlow, Electron counting in solids: Oxidation states, partial charges, and ionicity, J. Phys. Chem. Lett. **8**, 2074 (2017).

[48] I. D. Brown, Recent developments in the methods and applications of the bond valence model, Chem. Rev. **109**, 6858 (2009).

[49] C. R. Groom, I. J. Bruno, M. P. Lightfoot, and S. C. Ward, The Cambridge structural database, Acta Crystallogr. Sect. B **72**, 171 (2016).

[50] V. Postils, C. Delgado-Alonso, J. M. Luis, and P. Salvador, An objective alternative to IUPAC's approach to assign oxidation states, Angew. Chem. **130**, 10685 (2018).

[51] K. M. Jablonka, D. Ongari, S. M. Moosavi, and B. Smit, Using collective knowledge to assign oxidation states, ChemRxiv preprint (2020), doi: 10.26434/chemrxiv.11604129.v1.

[52] L. Jiang, S. V. Levchenko, and A. M. Rappe, Rigorous Definition of Oxidation States of Ions in Solids, Phys. Rev. Lett. **108**, 166403 (2012).

[53] H. Raebiger, S. Lany, and A. Zunger, Charge self-regulation upon changing the oxidation state of transition metals in insulators, Nature (London) **453**, 763 (2008).

[54] A. Walsh, A. A. Sokol, J. Buckeridge, D. O. Scanlon, and C. R. A. Catlow, Oxidation states and ionicity, Nat. Mater. **17**, 958 (2018).

[55] H. Pan, A. M. Ganose, M. Horton, M. Aykol, K. A. Persson, N. E. Zimmermann, and A. Jain, Benchmarking coordination number prediction algorithms on inorganic crystal structures, Inorg. Chem. **60**, 1590 (2021).

[56] D. Waroquiers, X. Gonze, G.-M. Rignanese, C. Welker-Nieuwoudt, F. Rosowski, M. G obel, S. Schenk, P. Degelmann, R. Andre, R. Glaum *et al.*, Statistical analysis of coordination environments in oxides, Chem. Mater. **29**, 8346 (2017).

[57] G. Brunner, A definition of coordination and its relevance in the structure types AlB_2 and NiAs, Acta Crystallogr. Sect. A **33**, 226 (1977).

[58] M. O'Keefe and N. Brese, Atom sizes and bond lengths in molecules and crystals, J. Am. Chem. Soc. **113**, 3226 (1991).

[59] S. P. Ong, W. D. Richards, A. Jain, G. Hautier, M. Kocher, S. Cholia, D. Gunter, V. L. Chevrier, K. A. Persson, and G. Ceder, Python materials genomics (pymatgen): A robust, open-source python library for materials analysis, Comput. Mater. Sci. **68**, 314 (2013).

[60] M. O'Keeffe, A proposed rigorous definition of coordination number, Acta Crystallogr. Sect. A **35**, 772 (1979).

[61] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevMaterials.5.043804 for machine learning model details, code and data repository links.

[62] The periodic table of the elements by webelements (2020).

[63] D. W. Davies, K. T. Butler, O. Isayev, and A. Walsh, Materials discovery by chemical analogy: Role of oxidation states in structure prediction, Faraday Discuss. **211**, 553 (2018).

[64] *MATLAB and Statistics Toolbox Release 2019a* (The MathWorks Inc., Natick, MA, 2019).

[65] The database of ionic radii is made open through https://cmd-ml.github.io/.

[66] I. Brown and R. Shannon, Empirical bond-strength–bond-length curves for oxides, Acta Crystallogr. Sect. A **29**, 266 (1973).

[67] K. Oka, M. Azuma, W.-t. Chen, H. Yusa, A. A. Belik, E. Takayama-Muromachi, M. Mizumaki, N. Ishimatsu, N. Hiraoka, M. Tsujimoto *et al.*, Pressure-induced spin-state transition in $BiCoO_3$, J. Am. Chem. Soc. **132**, 9438 (2010).

[68] Y. Hashimoto, M. Wakeshima, K. Matsuhira, Y. Hinatsu, and Y. Ishii, Structures and magnetic properties of ternary lithium oxides $LiRO_2$ (R = rare earths), Chem. Mater. **14**, 3245 (2002).

[69] Z. Wu, Q. Zhang, B. Li, Z. Shi, K. Xu, Y. Chen, Z. Ning, and Q. Mi, Stabilizing the $CsSnCl_3$ perovskite lattice by b-site substitution for enhanced light emission, Chem. Mater. **31**, 4999 (2019).

[70] N. Greenwood and A. Earnshaw, Chemical periodicity and the periodic table, Chem. Elem. **1**, 24 (1984).