JOM, Vol. 73, No. 1, 2021
https://doi.org/10.1007/s11837-020-04411-1

AUGMENTING PHYSICS-BASED MODELS IN ICME WITH MACHINE LEARNING AND UNCERTAINTY QUANTIFICATION

Expanding Materials Selection Via Transfer Learning for High-Temperature Oxide Selection

ZACHARY D. MCCLURE [1] and ALEJANDRO STRACHAN[1,2]

1.—School of Materials Engineering and Birck Nanotechnology Center, Purdue University, West Lafayette, IN 47907, USA. 2.—e-mail: strachan@purdue.edu

Materials with higher operating temperatures than today's state of the art can improve system performance in several applications and enable new technologies. Under most scenarios, a protective oxide scale with high melting temperatures and thermodynamic stability as well as low ionic diffusivity is required. Thus, the design of high-temperature systems would benefit from knowledge of these properties and related ones for a large number of oxides. While some properties of interest are available for many oxides (e.g., elastic constants exist for > 1000 oxides), the melting temperature is known for a relatively small subset. The determination of melting temperatures is time consuming and costly, both experimentally and computationally; thus, we use data science tools to develop predictive models from the existing data. Since the relatively small number of available melting temperature values precludes the use of standard tools, we use a multi-step approach based on transfer learning where surrogate data from first principles calculations are leveraged to develop models using small datasets. We use these models to predict the desired properties for nearly 11,000 oxides and quantify uncertainties in the space.

INTRODUCTION

Materials capable of operating at high temperatures are critical for applications ranging from aerospace to energy,[1] and increasing their operating envelope over the current state of the art is highly desirable. For example, increasing the operating temperature of land-based turbines by 30°C would result in an approximately 1% efficiency increase and can translate into sector-wide fuel savings of $66 billion with significant environmental impact over a 15-year period.[2] In addition, high temperature metallic alloys can enable rotation detonation engines for hypersonic vehicles.[3] In all of these applications, high-temperature mechanical integrity or high strength is required, and so is oxidation resistance. The latter can be achieved either by the formation of a protective oxide scale during operation[4] or by the incorporation of a protective oxide (often sacrificial) during fabrication.[5,6] This article combines existing experimental, first principles data and physics-based models with data science tools, including uncertainty quantification, to create a comprehensive dataset of potential oxides and the physical properties relevant for materials selection.

In recent years, complex concentrated alloys (CCAs, multi-principal component alloys that lack a single dominant component) and the closely related high-entropy alloys (HEAs)[7,8] have attracted significant attention as they have been shown to exhibit properties not possible with traditional metallic alloys.[9] Particularly interesting for high-temperature applications are refractory CCAs (RCCAs),[10] which have emerged as an attractive alternative to current superalloys. While RCCAs exhibit high-temperature strength surpassing the state of the art, their oxidation resistance is far from ideal. For example, the mass gain at T = 1000°C for TiZrNbHfTa during 1 h in air is 65 mg/cm^2, almost an order of magnitude higher than the Cr_2O_3-forming wrought Ni-based superalloys.[11,12] Thus, efforts are underway to design RCCAs capable of growing effective oxide scales at temperatures > 1000°C.[11,13] Beyond RCCAs, high-temperature

(Received February 20, 2020; accepted September 4, 2020; published online November 10, 2020)

protective oxides are required in a range of applications. Carefully engineered oxide scales can be used to prevent further oxidation and embrittlement of alloys in high temperature applications,[11,12] corrosion resistance in adverse environments,[12] or as a protective coating during aerospace re-entry applications.[5]

Desirable properties in these oxides include high-melting temperatures, good thermodynamic and mechanical stability to facilitate their formation over competing oxides, and low oxygen ion and cation mobility to slow down oxidation kinetics. Other properties are also desirable: a coefficient of thermal expansion (CTE) matching that of the substrate, a Pilling-Bedworth ratio (defined as $\frac{V_{oxide}}{V_{metal}}$ with V the molar volume) near one, and good adhesion to the substrate.[14] Designing RCCAs with desirable oxide scales presents additional challenges since the large number of metallic elements results in various possible competing oxides and complex multi-layer scales.[11] The design of RCCAs with appropriate high-temperature oxidation resistance and the selection of oxide coatings that can be added to structures would benefit enormously from an extensive database of all possible high-temperature oxides and their properties of interest.

Unfortunately, the required information is not available for the majority of the tens of thousands of stable oxides known. To date, < 60,000 metastable oxides have been studied by the Materials Project (MP) via first principles calculations.[15] Of these, about 11,000 are either the ground state or low-energy metastable structures at zero temperature. Elastic constants are known for a small subset, totaling roughly 1000 oxides in the MP database.[15] However, melting temperatures are known for an even smaller subset. In this article we use data science tools including machine learning to generate materials property information that can be used for materials selection for the majority of known oxides. We build on the fact that some of these properties are correlated to each other because of the similar underlying physics and can be used to create physics-based surrogate models of the quantities of interest to address the challenge of small datasets.

Cyber-infrastructure for materials data Motivated by the need for faster and less expensive materials discovery and deployment cycles,[16] great strides have been made in the development of cyberinfrastructure for materials science and engineering over the last decade. Examples of this infrastructure include open and queryable repositories with first principles data, such as MP or the Open Quantum Materials Database (OQMD),[15,17] open repositories of materials properties such as Citrination,[18] and even published interatomic potential models for atomistic simulations.[19] In addition to data and models, platforms for online simulations and data analysis such as nanoHUB[20] and Google Colab[21] lower the barrier of access to simulation and data science tools for research and education.[22] These repositories are making strides towards making data findable, accessible, interoperable, and reproducible (FAIR).[23] Data can be queried through online user interfaces or via application programming interfaces (APIs) for rapid querying and analysis of data.

Transfer learning for materials selection Materials selection requires access to data and often involves a multi-objective optimization.[24,25] This was traditionally done with existing experimental data, sometimes combined with simple models.[26] More recently, *ab initio* electronic structure calculations have been incorporated to such efforts,[27] and progress in multiscale modeling is providing additional tools to materials design and optimization.[28] In addition to such data, machine learning tools are being used to assess the current state of knowledge and make decisions. In our application, it would be tempting to use machine learning to develop models to predict our quantities of interest (QoI), such as melting temperature, from compositions using the available data; these models could then be used to explore the properties of a wide range of oxides. Unfortunately, the limited set of known melting temperatures precludes such an approach as standard ML methods require vast amounts of data, on the order of 10^3-10^6 datapoints. The lack of data is common in materials applications, and several approaches have been developed to address it.

Many of the most relevant studies have successfully been able to use traditional machine learning approaches, albeit on minimal datasets ranging from 10^1-10^3, through the use of careful descriptors and transfer learning approaches. This has been used, for example, to screen billions of compositions for Li-ion conductivity,[29] using preliminary *ab initio* data to extend 101 compounds into a space of 54,779[30] and showing that by distinguishing between descriptors such as energy difference between phases accurate predictions can be made on datasets of as few as 82 samples.[31] These methods compensate the lack of large amounts of data with domain expertise, physics, and chemistry. These approaches are not unique to the field of materials. In fact, they have been extensively used in chemistry for polymer selection[32] and design of chemical compounds[33] for decades.

One such method is to enhance the information fed to the model by adding surrogate properties as inputs. These surrogate properties should be both easy to obtain and expected to correlate to the quantity of interest. In this article, we use the oxide stiffness (easily computable via *ab initio* simulations) and melting temperature estimates using Lindemann's law[34] as additional inputs to the model to predict melting temperatures. Lindemann's estimates can be easily obtained from available

properties and can be expected to serve as good surrogates based on prior studies in oxides[35] and minerals.[36] We note that stiffness and melting temperature are both governed by the strength of the inter-atomic interactions; there is a correlation between these properties, and adding stiffness as an input to the models results better accuracy.

CURRENTLY AVAILABLE DATA

The design or selection of protective oxide scales would benefit from access to materials properties for all possible oxides that are either stable or metastable at the operating temperatures. As discussed above, a large number of oxides structures are known, but high-temperature data, including melting temperatures, are known for a small subset. Thus, we start from all known oxides and combine existing data with machine learning to provide information about structures for which we lack experimental data. This section explores the relevant data available in online repositories, and "Extending Oxide Data Via Data-Driven Transfer Learning" section discusses the use, combination, and extension of the data.

As discussed above, several materials data repositories focusing on various types of data and materials classes are available today. We leveraged the MP database, Citrination, and WolframAlpha.[37] The MP is a database with density functional theory (DFT) results including crystal structure data, relative stability to the ground state, elastic constants for select materials, calculated x-ray diffraction (XRD) and x-ray photoelectron spectoscopy (XPS) spectra, and even T = 0 K phase diagrams for compounds. MP has information about a majority of known oxides, and we start our search within this list. The properties in the MP can be accurately calculated from first principles calculations; however, properties such as melting temperature are computationally too intensive for high-throughput DFT calculations. Therefore we turn to repositories with extended datasets for additional information such as melting temperature and oxygen vacancy formation energy (VFE).

The Citrination database[18] is an open repository where researchers can upload their own data and share it with the community at large. At the time of writing, citrination contains 454 public databases curated by public users and Citrine staff. Databases previously curated through research efforts have been published in their database and are freely available for download and use. For our efforts we turned to Citrination for databases of oxygen VFE.

We were unable to find an electronic database with melting temperatures for oxides. Most reported melting points exist within individual papers, collected handbooks, or commercial databases. However, we were able to find some of these properties in WolframAlpha, a general purpose, queryable, compute engine. Using WolframAlpha, we generated a list of melting temperatures for a subset of the oxides queried from MP with elasticity data.

The Materials Project: Basic Oxide Data

We accessed the MP database using the Pymatgen API[38] and analyzed every oxide available. MP contains information about 60,000 distinct oxides (differing either by composition or crystal structure). All these structures were obtained by energy minimization using DFT within the generalized gradient approximation, and additional details of the calculations can be found in work by Jain et al.[39] Confirming the metastability of these structures would require positive phonon frequencies and elastic constants to discard local energy maxima; these quantities are not available for all these oxides. To address this challenge we first filtered the data to retain only structures that are 1 meV above the convex hull (i.e., the predicted ground state for that composition). We note that the energies resulting from energy minimization correspond to a temperature of T = 0 K (minus zero point energy), and phases with free energy higher that the ground state at 0 K can be stabilized at higher temperatures because of entropic contributions to the free energy. Furthermore, many metastable structures are long-lived and used in applications. After this stability constraint, we are left with $\sim$ 11,000 possible oxides. However, elastic constants are documented in MP (from DFT calculations) for a subset of 855 oxides. To illustrate the available data, Fig. 1 compares two properties of the available oxides after filtering by energy stability and elastic constants. We plot the ionic packing fraction (defined as the total atomic volume assuming hard spheres with the corresponding ionic radius in the unit cell divided by the cell volume) vs. density obtained from the crystal structure data. Red points indicate oxides with at least one element that is found in RCCAs: we select Ti, V, Cr, Zr, Nb, Mo, Ru, Hf, Ta, W, and Re as well as Al, Cr, and Si since they are useful additives. Of the 855 oxides with elasticity data, 235 contain an element pertaining to an RCCA or additive compound. The figure also highlights common protective oxides Al_2O_3 and Cr_2O_3; as expected, these oxides have high packing fractions (which correlate in low ionic diffusivity). Interestingly, there are a number of potential compounds with comparable properties to both these common oxides.

With this basic information at hand, we now focus on the remaining properties: melting temperature and ionic mobility. While these properties can be obtained, in theory, from first principles, they are computationally very intensive and are not included in the MP. Therefore we turn to other repositories for additional DFT and experimental data.

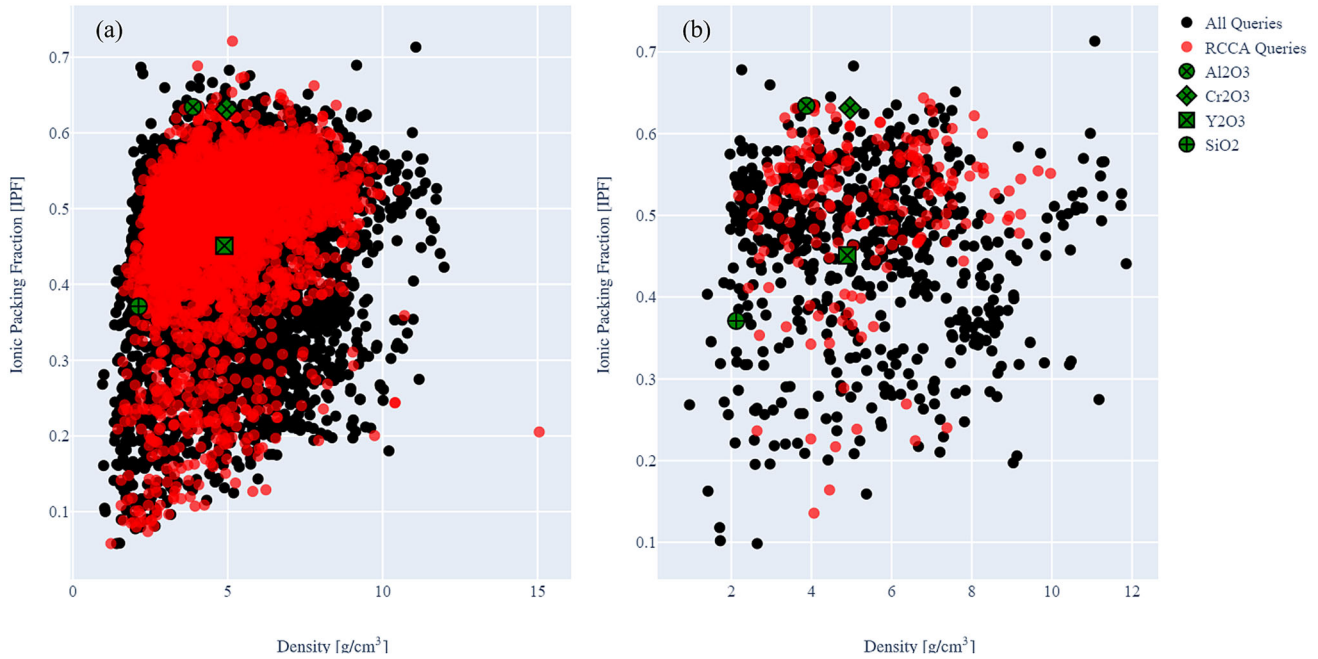

Fig. 1. Calculated ionic packing fraction of oxides and their queried densities. (a) Database curated post-energy stability filtering. (b) Database curated post-elasticity filtering.

WolframAlpha: Melting Temperatures

At the time of writing, melting temperatures of the oxides of interest were not available in materials-specific online repositories. A single curated inorganic melting point database on Citrination exists, but many of the values are not oxides and do not overlap where we have existing elasticity data. Fortunately, WolframAlpha provides an API for data exploration. Through a series of string queries we obtained and curated melting points of 158 oxides into our database. Since these data are significantly less abundant than the elasticity data from Materials Project, we will consider the melting point to be our harder to acquire, or more expensive, set of data. Improvements to this dataset can be made through literature searches, or analyzing phase diagram textbooks, but our goal is to illustrate a rapid acquisition of data rather than the traditional task of searching through physical copies of information. Figure 2 shows the results of the melting point query with respect to density and IPF properties. RCCAs contaning oxides are highlighted to guide the eye, and as expected a number of them have comparable properties to common oxides such as Al_2O_3 and Cr_2O_3.

Citrination: Vacancy Formation Energies and Thermal Expansion

Ionic mobility is another critical material property in the design of protective oxides; unfortunately, ionic mobility (oxygen or cation) is not widely available. However, since oxygen mobility is mediated by vacancies, the vacancy formation energy is a good surrogate for ionic transport: the higher the vacancy formation, the lower the vacancy concentration and oxygen ion mobility. Citrination includes a database of nearly 2000 charge-neutral vacancy formation energies of oxides based on first principles approaches originally published by Deml et al.[40] Of this dataset, 1200 were unique oxide compounds.

Another database of importance is available based on work by Shick et al.[41] containing 69 average coefficient of thermal expansion (CTE) values obtained from anharmonic phonon calculations. We will consider the use of this database in future work, but at this time the limited dataset provides a great challenge to accurate predictions outside the selected compounds via machine learning methods. A transfer learning approach could be used, but our focus here will be the melting temperature.

EXTENDING OXIDE DATA VIA DATA-DRIVEN TRANSFER LEARNING

In summary, from online accessible databases we were able to extract 11,000 stable and metastable oxides from an initial 60,000 queries on MP. Of these 11,000 possible oxides roughly 1000 have existing elasticity data. From the list of 1000 oxides with elasticity information, only 162 melting points were obtained through queries. In addition, we have VFE values for 1200 cases and CTE for 69. Ultimately, our goal is be to build models with each of these properties and use the information leveraged from each to extend a materials search into the original 11,000 stable and metastable oxides.

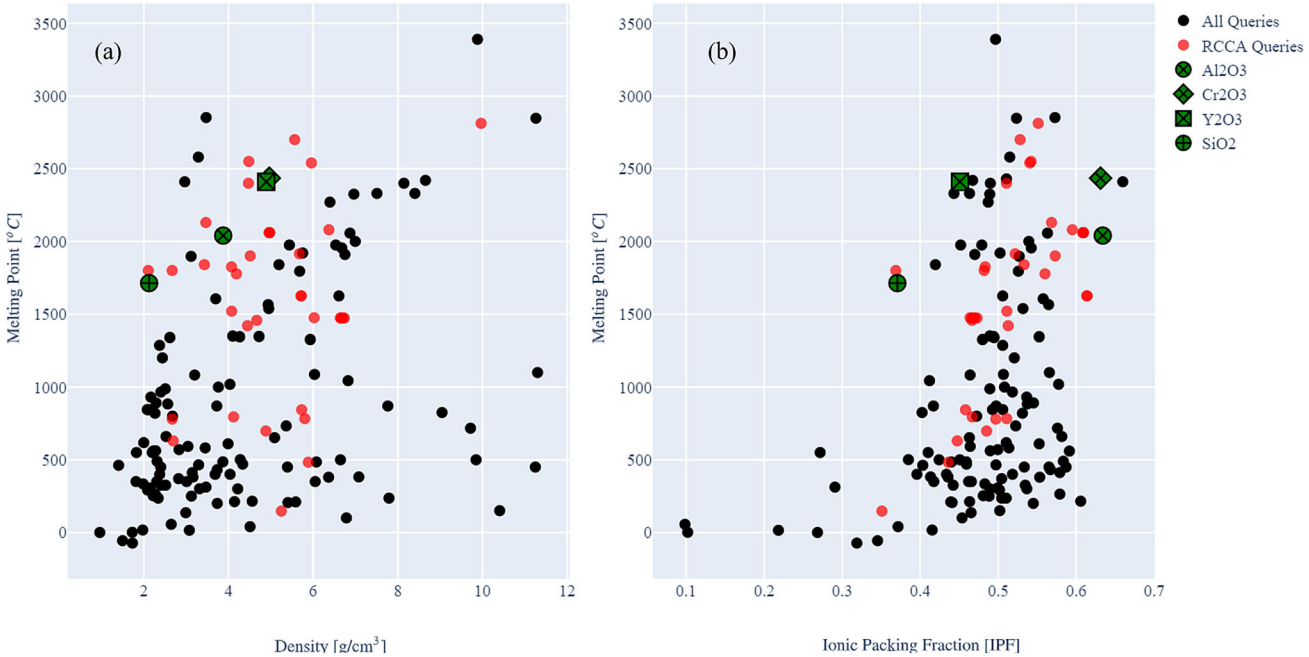

Fig. 2. Queried melting points from WolframAlpha with bulk modulus (a) and IPF (b) properties.

Since filling in the gaps in data discussed in "Currently Available Data" section via first principles calculations or experiments would be prohibitive in terms of time and cost, we will explore using data-driven machine learning tools like neural networks[42,43] and random forests.[44] One could attempt to train models that relate composition to the final QoI (e.g., melting temperatures) from the existing data. However, standard ML approaches are not applicable directly because of the scarcity of the data. This is a common challenge in the field of materials. Feature engineering, which involves feeding additional data to the model that can be easily obtained from the raw input data, can be used to address this challenge. For example, we could include electronegativity and ionic radii of the elements as inputs to the model; thus, we include information about bonding and packing. In addition to such *periodic table* data, one can further increase the information fed into the model by adding physics-based modeling results or material properties that are easy to obtain and that are expected to correlate with it. It has been shown that even with limited training data physics-based descriptors have had a significantly higher impact than models that only rely on raw volumes of data.[29,45] Here we will build on two surrogate pieces of information. First since melting temperature and stiffness are both governed by similar physics, stiffness (available from first principles calculations for 1000+ oxides) is added as an input parameter. Second, we can use Lindemann's melting temperature model to estimate values of the QoI from basic properties and add that estimate as an input. Lindemann's melting law is based on the approximation that melting occurs when atomic oscillations reach a critical value relative to the materials lattice parameter. The amplitude of atomic oscillations can be easily obtained using statistical mechanics, and the resulting expression for the melting temperature is:

$$T_m' = \alpha(4\pi^2 k/9h^2 N^{5/3}) f^2 \overline{M} V^{2/3} \Theta_D^2 \tag{1}$$

where T_m' is the Lindemann melting temperature; α is a structural factor generally taken as 1; h, k, and N are the Boltzmann, Planck and Avogadro constants, respectively; $\overline{M}$ is the mean atomic mass, V the molar volume, and Θ_D the Debye temperature.

The Debye temperature is taken from the approximation proposed by Blackman[46] using the expression:

$$\Theta_D = (h/k)(3N/4\pi)^{1/3} V^{-1/3} v_m \tag{2}$$

with v_m as the average acoustic velocity given by:

$$3/v_m^3 = 1/v_p^3 + 2/v_s^3 \tag{3}$$

with v_p and v_s the P and S wave velocities calculated from the root of the bulk and shear modulii with their densities (ρ):

$$v_p = \sqrt{\frac{K + 4/3G}{\rho}} \tag{4}$$

$$v_s = \sqrt{\frac{G}{\rho}} \tag{5}$$

Descriptors

As mentioned above, a common way of building physics into ML models is to use periodic table data of the elements involved as inputs. We primarily use the *composition featurizer* from Matminer[47] to generate a variety of properties with composition as the only input. As shown in previous work by Ward et al.,[48] statistical descriptors based on the chemical formula are useful for machine learning features.

The descriptors we use are described as follows:

1. A stoichiometric calculation of fractions of elements without considering the actual composition. This calculation includes the number of elements in the compound and normalizations of the respective fractions.
2. Periodic table-type descriptors including mean, mean absolute deviation, range, minimum, maximum, and mode of elemental properties. These values include maximum row on the periodic table, average atomic number, and the difference in atomic radii in all elements present.
3. As previously shown by Meredig et al.,[49] electronic structure attributes with averages of s, p, d, and f valence shell electron concentrations are useful as descriptor inputs.
4. Assuming that the ionic species in the oxide can form a single oxidation state, an adaption of the fractional ionic character of a compound can be used based on an electronegativity-based measure.[50]
5. The fraction of the transition metal elements.
6. The cohesive energy per atom using elemental cohesive energies.
7. An estimation of the band gap center based on electronegativity.
8. Number of available oxidation states in the compound.
9. For mechanical properties models, we also extend descriptors to include properties queried from the MP database such as density, space group number, and calculated ionic packing fractions.
10. Importantly, we added two descriptors for the meting temperature models: predicted stiffness properties and estimate of the melting temperature according to Lindemann's law discussed above.

These descriptors are able to characterize the output properties for VFE sufficiently, and we do not see evidence of over parameterization of the models. For stiffness we add additional descriptors queried from MP, and for the melting point we use the full knowledge of composition descriptors, queried MP properties, predicted stiffness, and Lindemann law melting predictions.

Predictive Models for Melting Temperature Using Random Forests

Random forests (RFs) approach regression methods through a series of decision trees[51,52] whose outputs are averaged. This averaging is done to overcome the limitation of individual tree predictions, which may have difficulty assessing noise or non-linearities in the data. Importantly, progress has been made in the quantification of uncertainties in RFs by Efron[53] and Wager et al.,[54] and more recently by Ling et al.[44] with the addition of an explicit bias term to the uncertainty. Neural networks, often outperforming random forest predictions, were considered for this study, but quantification of uncertainty in their outputs is still an active field of research.[55] Due to the accessibility of uncertainty quantification, we choose to implement random forest models with the state of the art uncertainty calibration proposed by Ling et al.[44] It involves sample-wise variance defined as the average of the jackknife-after-bootstrap and infinitesimal jackknife variance estimates with a Monte Carlo sampling correction. The RF models implemented in this study are available in the Lolo scala library.[56]

We set the maximum number of trees to match the number of samples collected in each model for our RFs while allowing for unrestricted maximum depth. While saturation of averaged prediction can occur beyond 200 or more trees,[57] the uncertainties in Lolo will not be well calibrated. The maximum depth parameter cutoff is defined by the nodes increasing until the leaves become pure or until the all leaves contain less than two samples. This is the default parameter for Lolopy.

As is common practice, each descriptor is normalized by standard normalization, and data were split into 80% training and 20% testing to evaluate performance. Assessment of the model was performed for each material property by reshuffling the dataset ten different times and taking an aggregated MAE.

When assessing uncertainty estimates for an individual output x, the residuals, $r(x)$, of the prediction when normalized by the uncertainty $\sigma(x)$ ($N = \frac{r(x)}{\sigma(x)}$), should have a Gaussian distribution with zero mean and unit standard deviation. This metric can help quantify whether the random forest uncertainty predictions are well calibrated with respect to the inherent error predicted.

Using the set of descriptors and architecture detailed in "Descriptors", we implement random forest models to predict the set of desired properties using databases from MP, WolframAlpha, and Citrination. All reported MAE values are taken as an aggregate mean after shuffling the training and testing sets ten times.

Random forest performance for VFE

Using the curated Citrination dataset, we developed a RF model for VFE. Composition-based

descriptors obtained via Matminer were used for model predictions. For ten shuffling samples we report an aggregated MAE of 0.17 eV/atom (Fig. 3).

Random forest performance for stiffness

In addition to the Matiminer featurizers described above, we added additional descriptors such as IPF and space group number since these were easily queried. Figure 4 shows a parity plots and normalized residuals for bulk and shear moduli. An aggregated testing MAE score of 18 and 10 GPa for bulk and shear modulus, respectively, was obtained after ten shufflings of samples.

Random forest performance for melting temperature

Our dataset of 162 melting points with corresponding elasticity data was used to create a predictive model for varied oxides. Figure 5 shows the performance of both the training and testing data before and after adding stiffness Lindemann properties into the model. As we can see adding stiffness and Lindemann's law as descriptors improves the accuracy of the model to a significant degree with a reduction of the MAE from 368 to 303°C. While uncertainties are not negligible, these models are promising for an initial sweep of potential oxides. A noticeable reduction in uncertainty can be seen between Fig. 5a and b, and after adding stiffness and Lindemann's law, fewer points lie outside the linear fit in the parity plot. After training the model, we use identical descriptors for the remaining compounds that we were unable to easily obtain melting points for and extend our predictions using the information gained from stiffness and melting point models. In "Uncertainty Propagation on the Melting Temperature Calculation" section, we will assess some the sensitivity of this prediction with varied UQ methods. In the outlook section of this paper, we will discuss the implications and results of extrapolating our predictions to other oxide melting points outside of our initial query with WolframAlpha.

MATERIALS SELECTION FOR PROTECTIVE OXIDE SCALES

Using the models above, we begin to extend our search space of potential oxides from our initial query of 162 melting points and 855 points with elasticity data and move into the space of the remaining 11,000 stable oxides from MP. First, we predict the elasticity data of the remaining 11,000 oxides that did not have this data to begin with. Then we use those descriptors to expand our melting point database from 158 queried data points to nearly 11,000 data points: a two order of magnitude increase.

Figure 6a shows the 11,000 oxides and their respective properties. We show the melting temperature and the oxygen VFE; bulk modulus is shown as the color of the symbol, and the IPF is represented by size. Figure 6b filters radioactive elements and lanthanides out and also removes bulk and shear modulus values $<$ 125 and 25 GPa, respectively. The plot highlights common and effective protective oxides. As expected, Cr_2O_3, Al_2O_3, and SiO_2 are among the top performers. However, our study reveals other oxides predicted to perform equally well or outperform them. Figure 6c shows the final filtering of outlier properties such as low VFE, low melting point, and IPF values $<$ 0.4, and

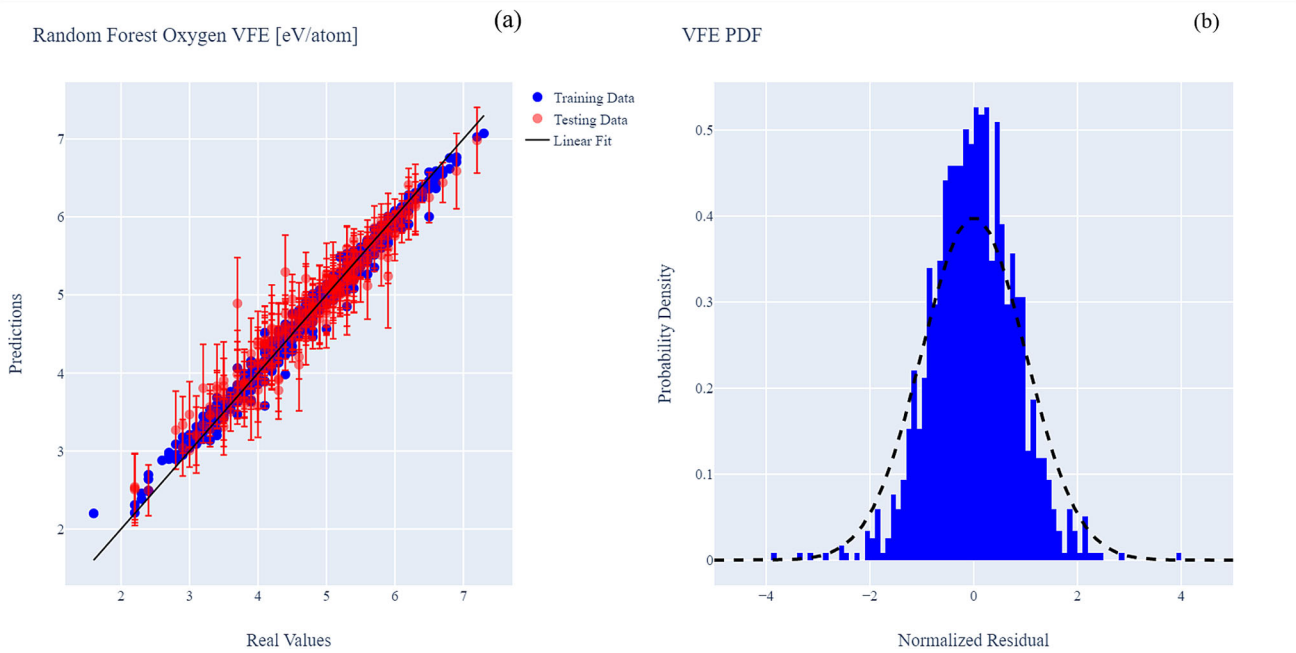

Fig. 3. (a) Parity plot diagram for predicted and real values of oxide VFE. Values directly on the line are a perfect match. (b) Normalized residuals for VFE with Gaussian-like distribution.

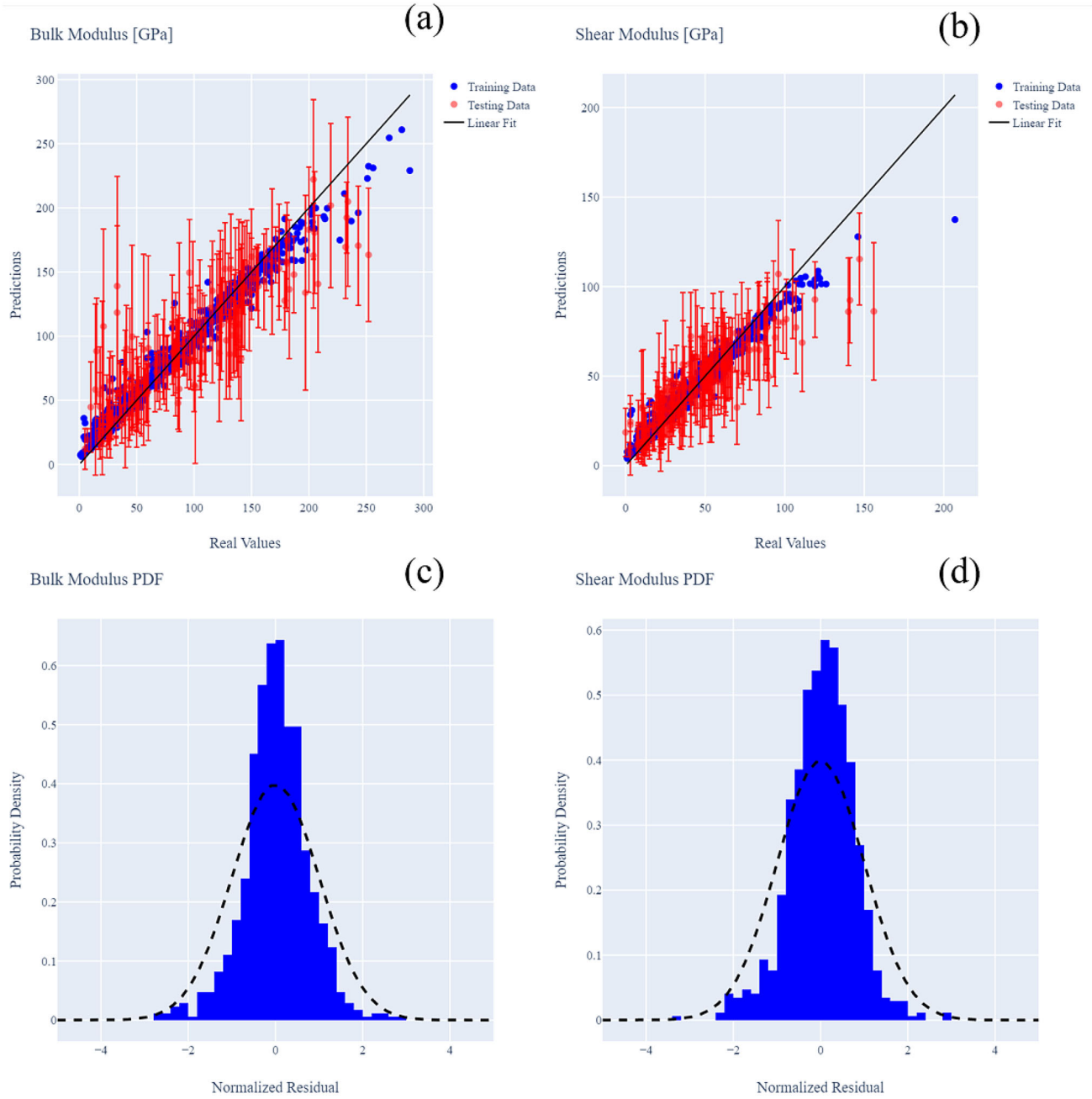

Fig. 4. (a) Parity plot diagram for predicted and real values of oxide bulk modulus; (b) shear modulus. Included are normalized residual calculations for (c) bulk modulus and (d) shear modulus.

Fig. 6d shows these final points including the uncertainties in the RF model. Data points with a cross represent materials with existing melting temperatures from WolframAlpha, and empty symbols are predictions. Values without error bars in either direction indicate database collected values.

The top oxides identified and their properties are summarized on Table I. In the case of the design of refractory CCAs, HfO_2 (T_{melt} = 2812°C, VFE = 5.9 eV/atom) is an attractive candidate since Hf is a common element. Our results indicate that the addition of Y as a dopant to RCCAs could result in the formation of $Y_2Hf_2O_7$, $YTaO_4$, $Y_3Al_5O_{12}$, or Y_6WO_{12}. While many of these have lower predicted melting points (in the 1900–2000° range), they may stabilize as complex oxides between the outer scale and substrate. Each of these oxides coupled with the RCCA substrate could be engineered to form a stabilized complex oxide of one or more of these structures. Quite interestingly, even though the Y-containing compounds in Table I were not present in our initial database of melting points, they have been investigated as promising candidates for thermal barrier coatings[58] and scales in high-

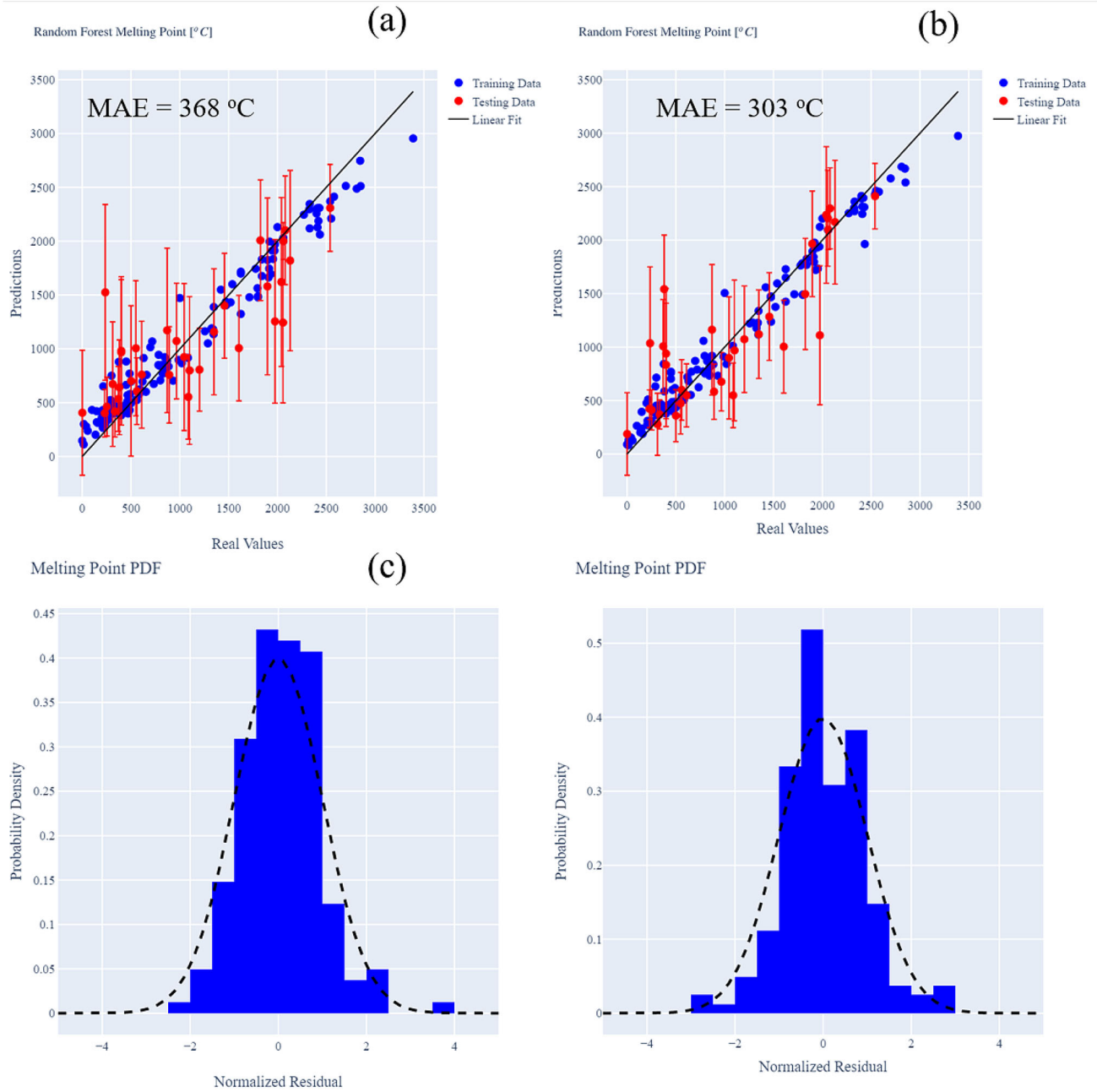

Fig. 5. (a) Parity plot diagram for predicted and real values of oxide melting temperature. Values directly on the line are a perfect match. (b) Adding stiffness properties to the model causes a decrease of 65°C with respect to MAE and a noticeable decrease in uncertainty. (c) and (d) Normalized residuals for models trained without and with additional descriptors.

temperature applications. Synthesis routes have been discussed in the literature. Reported melting temperatures include 2300°C for $YTaO_4$, [59] the well-studied yttrium aluminum garnet (YAG) compound melting at roughly 1940°C, [60] stability of the single phase under ablation temperatures > 2000° for $Y_2Hf_2O_7$, [61] and finally Y_6WO_{12} melting at 2360°C. [62] Our predicted values of the melting point through random forests with predicted uncertainties fall very close to experimental results, which is encouraging for potential extrapolation of other materials for RCCA protective scales.

Two of the oxides predicted to be of interest for high-temperature applications lack experimental melting temperatures. Also containing Y in their structure, $Y_3Al_3Cr_2O_{12}$ and Y_3ReO_8 at this time do not have reported melting points. Each of them has predicted melting points > 1900°C, and each has elements that could be used as base components in RCCA applications. Already proven to be an excellent candidate for scale formation and physical properties is Cr, but more interesting is the presence of Re in Y_3ReO_8. In the second generation of Ni-based superalloys, Re was proven to be an

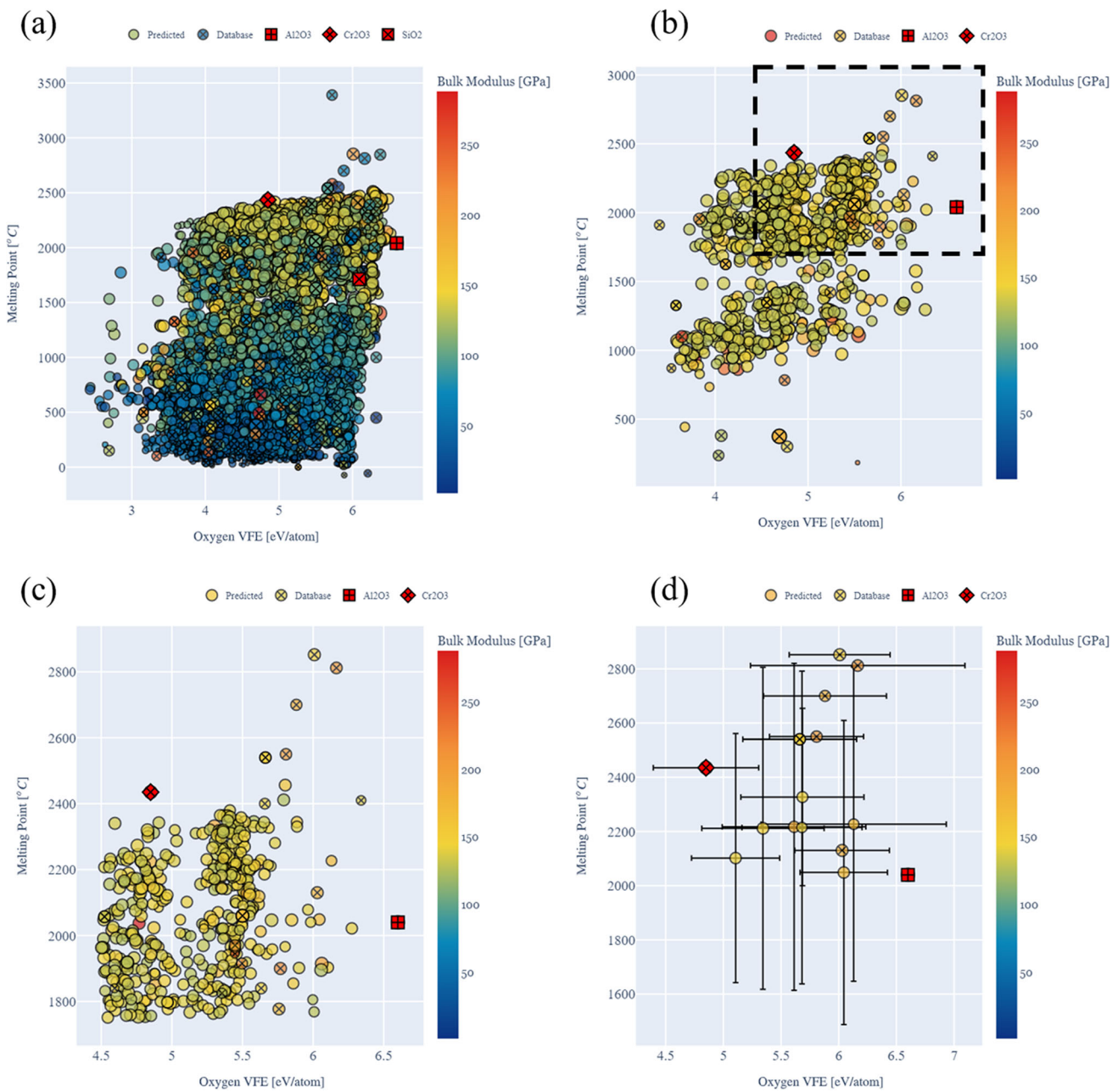

Fig. 6. Comparison of melting point and vacancy formation energy of oxide compounds. Coloring corresponds to stiffness of the material, and marker size indicates IPF where larger markers are a higher IPF. Points with an 'x' are melting points collected from queryable sources where open circles are predicted values. (a) Predicted results for original 11,000 query. (b) Results filtered to remove radioactive and lanthanide compounds and bulk and shear modulus values $<$ 125 and 25 GPa, respectively. (c) Additional filtering of properties with remaining values including IPF $>$ 0.4, T_{melt} 1750°C, and VFE $>$ 4.5 eV/atom. (d) Selected compounds for final application consideration. These compounds are listed in Table I as potential complex or native scale formers. Values that have database values do not show error in respective direction. Note the slightly different scales in the filtered figures (Color figure online).

excellent dopant for extending the creep lifetime in alloys.[63] While expensive, the addition of such an element to RCCA-type materials could have potential for oxide scale formation as well as modifying the overall mechanical properties of the material.

Other notable oxides that we found with excellent properties include: MgO (T_{melt} = 2852°C, VFE = 6.0 eV/atom), $MgAl_2O_4$ (T_{melt} = 2130°C, VFE = 6.05 eV/atom), ZrO_2 (T_{melt} = 2700°C, VFE = 5.79 eV/atom), $BaZrO_3$ (T_{melt} = 2450°C, VFE = 5.63 eV/atom), $ZrSiO_4$ (T_{melt} = 2550°C, VFE = 5.70 eV/atom), and $SrZrO_3$ (T_{melt} = 2305°C, VFE = 5.6 eV/atom). Of these other promising candidates, we note the experimental melting point for $SrZrO_3$ to be recorded as 2610°C,[64] well within the predicted value with random forest uncertainties. We would

Table I. Final compounds with uncertainties

	RF T_m (°C)	Exp. T_m (°C)	Oxygen VFE [eV/atom]
Al_2O_3	–	2040†	6.6 ± 0
Cr_2O_3	–	2435†	4.91 ± 0.48
HfO_2	–	2812†	5.99 ± 0.89
MgO	–	2852†	6.0 ± 0.43
$MgAl_2O_4$	–	2130†	6.05 ± 0.38
ZrO_2	–	2700†	5.79 ± 0.52
$ZrSiO_4$	–	2550†	5.71 ± 0.39
$BaZrO_3$	–	2540†	5.63 ± 0.48
$SrZrO_3$	2326 ± 327	2610[64]	5.62 ± 0.49
$Y_2Hf_2O_7$	2226 ± 579	2000*[61]	5.66 ± 0.77
Y_6WO_{12}	2214 ± 576	2360[62]	5.66 ± 0.45
$YTaO_4$	2217 ± 603	2300[59]	5.59 ± 0.55
$Y_3Al_5O_{12}$	2048 ± 561	1940[60]	6.04 ± 0.39
$Y_3Al_3Cr_2O_{12}$	2101 ± 459	?	4.39 ± 0.42
Y_3ReO_8	2211 ± 594	?	6.89 ± 0.44

If zero uncertainty is reported, the value was obtained from database results. Experimental validation of results for predicted values is shown in the middle column. Entries with '?' indicate an unknown melting point at this time in the literature† Experimental value from queried dataset * Ablation study.

like to stress that additional variables need to considered in the design of oxide scales, such as processability and kinetics; these are not considered in this first study.

UNCERTAINTY PROPAGATION ON THE MELTING TEMPERATURE CALCULATION

As in any decision-making exercise, uncertainties are critical in materials selection and optimization. Several sources of uncertainties must be accounted for in workflows such as the one used here. These include uncertainties in the ML models and in the input and output data fed to them. An additional challenge in our approach is the combination of experimental (e.g., melting temperatures) and first principles data (e.g., elastic constants). One could expect for systematic errors in surrogate data, like our DFT stiffness values, to be of relatively low importance as they are only used to *help* the ML models. For example, if gradient corrected exchange and correlation functionals used in DFT tend to underestimate binding, the ML models should be able to easily compensate for such discrepancies. The effects of non-systematic errors in input data such as Lindemann's melting law predictions are harder to estimate. Here, we focus on a specific kind of uncertainty that originates in transfer learning, i.e., how uncertainties are propagated across models in the transport process.

When using RF-predicted values for bulk and shear modulus as input descriptors to the melting point model, it is critical to assess how the uncertainties in elastic constants affect the predicted T_m. We note that the majority of the compounds in the final list of oxides selected in "Materials Selection for Protective Oxide Scales" section had first principles elastic constant data. One exception is $BaTi_2O_5$ so we use this material to study uncertainty propagation. The predicted mean melting point for this specific compound was 2144 ± 435°C; this was obtained with mean bulk and shear moduli. Since the elasticity models yield mean and the associated deviations, we can assess how sensitive the predicted melting point is to uncertainties in the moduli parameters. Our trained random forest models predict mean values of 142 ± 27 GPa and 75 ± 16 GPa for the bulk and shear modulus, respectively.

To propagate uncertainties in elastic constants through the melting temperature model, we use a brute force random sampling of the Gaussian distribution for each stiffness property. The resulting distribution from 10,000 samples is shown in black in Fig. 7a. The predicted distribution shows a sharp peak at 2150°C, very close to the mean prediction, and extends towards lower values with a second peak at 1950°C, and a third smaller distribution centered at 1700°C. The predicted RF distribution of melting temperatures with mean stiffness values is shown in red. Importantly, the uncertainties originating from the propagation of uncertainties in the stiffness are small compared with the intrinsic uncertainties in the prediction of melting temperature. This is, perhaps, not surprising since the melting temperature model has larger uncertainties that that for stiffness. The multi-peak nature of the distributions indicates large non-linearities in the T_m model. To assess this, we plot the melting temperature as a function of shear and bulk modulus in Fig. 7b with all other paramters fixed to

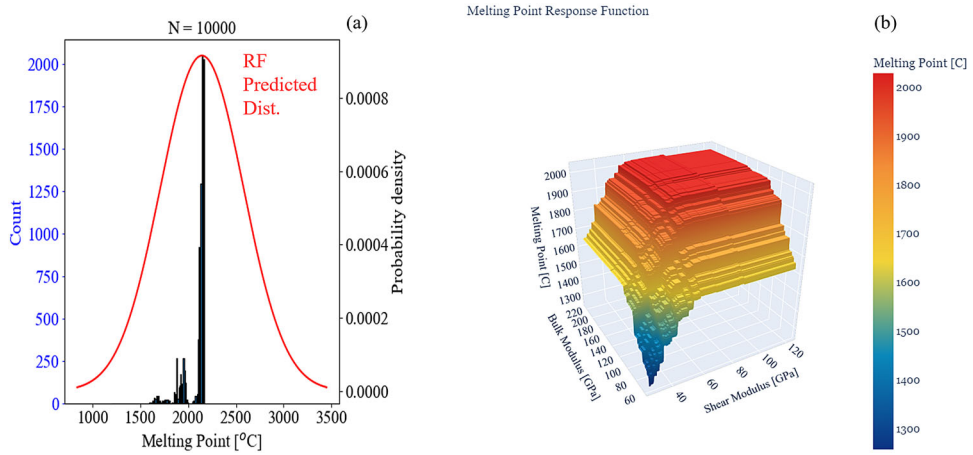

Fig. 7. (a) Histogram results for Monte Carlo (MC) sampling of bulk and shear modulus compared to the original random forest (RF) predicted distribution. (b) Response surface for shear (x-axis) and bulk modulus (y-axis) with respect to melting temperature (z-axis).

those of $BaTi_2O_5$. We find that the melting temperature drops quite significantly for low values of shear and bulk moduli. This is not surprising given the positive correlation between stiffness and melting temperature, but such extrapolations using machine learning models should be done with care.

SUMMARY AND OUTLOOK

We showed that by leveraging queryable open repositories and the use of machine learning tools with infused physics, one can greatly expand the information available for materials design or selection. Our specific goal was to find oxides for high-temperature applications with high melting temperature, high oxygen vacancy formation energy (to minimize O transport) with the coefficient of thermal expansion and stiffness as secondary design variables. Machine learning models with physics insight built in via feature engineering and surrogate properties enables us to take sparse existing data and fill in gaps in knowledge. Specifically, we found that by adding elastic constants (known for a relatively large number of oxides) as an input descriptor and easily calculated Lindemann melting laws, we could develop accurate models for melting temperature, which are harder to obtain and exist for relatively few materials.

Through transfer learning we were able to expand an initial query of 162 melting points to $>$ 10,000 compounds. The effort resulted in several candidate oxides with properties comparable to those of the protective scales of high-temperature metals such as Al_2O_3 and Cr_2O_3 with respect to melting point, VFE, IPF, and stiffness. Candidate materials include: HfO_2, $Y_2Hf_2O_7$, $YTaO_4$, $Y_3Al_5O_{12}$, Y_6WO_{12}, $Y_3Al_3Cr_2O_{12}$, Y_3ReO_8, MgO, $MgAl_2O_4$, ZrO_2, $BaZrO_3$, $ZrSiO_4$, and $SrZrO_3$.

Quantifying uncertainties in such efforts is critical in materials selection and optimization efforts. In this article we focus on the uncertainties propagated through predicted stiffness parameters and the uncertainties in the random forest. Additional work on uncertainties originating from combining information from different sources (e.g., DFT and experiments) would be very valuable.

Gathering the initial information from materials informatics platforms is a key step in our workflow and many similar efforts. This is enabled by recent progress on materials cyberinfrastructure, and community contributions to these repositories remains key. In the case of oxides, additional elastic constant calculations and melting temperatures would be beneficial. The models developed in this article can be accessed online through the US National Science Foundation's nanoHUB.[20] The tool High Temperature Oxide Property Explorer[65] includes live Jupyter notebooks with all models and data. The final curated data and models can be downloaded but they can also be modified and executed online.

ACKNOWLEDGEMENTS

Insightful discussions with Prof. M. Titus and K. Sandhage of Purdue University are gratefully acknowledged. This effort was supported by the US National Science Foundation, DMREF program, under Contract Number 1922316-DMR. We acknowledge the computational resources from nanoHUB and Purdue University.

CONFLICT OF INTEREST

The authors declare that they have no conflict of interest.

REFERENCES

1. T.M. Pollock, and S. Tin, *J. Propuls. Power* 22(2), 361–374 (2006).
2. General Electric. General Electric Annual Report. Technical report, (2018).
3. F. Falempin, E. Daniau, N. Getin, F. Bykovskii, and S. Zhdan, in *14th AIAA/AHI space planes and hypersonic systems and technologies conference*, p. 7956, (2006).
4. J. Doychak, and M.G. Hebsur, *Oxid. Metals* 36(1–2), 113–141 (1991).

5. J.F. Justin, and A. Jankowiak, *J. Aerosp. Lab* 3, 1–11 (2011).
6. J. L. Smialek, and N. S. Jacobson, Oxidation of high-temperature aerospace materials. *High temperature materials and mechanisms*, pp. 95–162, (2014).
7. B. Cantor, I.T.H. Chang, P. Knight, and A.J.B. Vincent, *Mater. Sci. Eng. A* 375, 213–218 (2004).
8. J.-W. Yeh, S.-K. Chen, S.-J. Lin, J.-Y. Gan, T.-S. Chin, T.-T. Shun, C.-H. Tsau, and S.-Y. *Adv. Eng. Mater.* 6(5), 299–303 (2004).
9. S. Gorsse, J.-P. Couzinié, and D.B. Miracle, *Comptes Rendus Physique* 19(8), 721–736 (2018).
10. O.N. Senkov, G.B. Wilks, D.B. Miracle, C.P. Chuang, and P.K. Liaw, *Intermetallics* 18(9), 1758–1765 (2010).
11. C.-H. Chang, M.S. Titus, and J.-W. Yeh, *Adv. Eng. Mater.* 20(6), 1700948 (2018).
12. G. D. Smith, and J. J. Fischer, in *ASME 1990 International Gas Turbine and Aeroengine Congress and Exposition. Brussels, Bel-gium*, pp. 1–7. Citeseer, (1990).
13. T.M. Butler, K.J. Chaput, J.R. Dietrich, and O.N. Senkov, *J. Alloys Compd.* 729, 1004–1019 (2017).
14. R.E. Bedworth, and N.B. Pilling, *J. Inst. Meter.* 29(3), 529–582 (1923).
15. A. Jain, S.P. Ong, G. Hautier, W. Chen, W.D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, and K.A. Persson, *APL Materials*, 1(1):011002, 2013. ISSN 2166532X. https://doi.org/10.1063/1.4812323. URL http://link.aip.org/link/AMPADS/v1/i1/p011002/s1&Agg=doi.
16. National Science and Technology Council (US). *Materials genome initiative for global competitiveness*. Executive Office of the President, National Science and Technology Council, (2011).
17. J.E. Saal, S. Kirklin, M. Aykol, B. Meredig, and C. Wolverton, *Jom* 65(11), 1501–1509 (2013).
18. Citrine Informatics. Citrination Database, citrination.com, (2020). URL https://citrination.com/.
19. OpenKIM. Open Knowledgebase of Interatomic Models https://openkim.org/, 2018. URL https://openkim.org/.
20. A. Strachan, G. Klimeck, and M. Lundstrom, *Comput. Sci. Eng.* 12(2), 12–17 (2010).
21. Google Colaboratory. Google colaboratory. *Google*, 2020. URL https://research.google.com/colaboratory/faq.html.
22. S.T. Reeve, D.M. Guzman, L. Alzate-Vargas, B. Haley, P. Liao, and A. Strachan, *MRS Adv.*, 1–16, (2019).
23. M.D. Wilkinson, M. Dumontier, I.J.J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L.B. da Silva Santos, P.E. Bourne, et al, *Scientific data*, 3, (2016).
24. M.F. Ashby, and D. Cebon, Materials selection in mechanical design. *Le Journal de Physique IV*, 3(C7):C7–1, (1993).
25. M.F. Ashby, *Acta Materialia* 48(1), 359–369 (2000).
26. M.F. Ashby, *Acta Metallurgica et Materialia* 41(5), 1313–1335 (1993).
27. V. Cutello, G. Narzisi, and G. Nicosia, in *Workshops on Applications of Evolutionary Computation*, pages 54–63. Springer, (2005).
28. E. van der Giessen, P.A. Schultz, N. Bertin, V.V. Bulatov, W. Cai, G. Csányi, S.M. Foiles, M.G.D. Geers, C. González, M. Hütter et al., *Modell. Simul. Mater. Sci. Eng.* 28(4), 043001 (2020).
29. E.D. Cubuk, A.D. Sendek, E.J. Reed, *J. Chem. Phys.* 150(21), 214701 (2019).
30. A. Seko, A. Togo, H. Hayashi, K. Tsuda, L. Chaput, and I. Tanaka, *Phys. Rev. Lett.* 115(20), 205901 (2015).
31. L.M. Ghiringhelli, J. Vybiral, S.V. Levchenko, C. Draxl, and M. Scheffler, *Phys. Rev. Lett.* 114(10), 105503 (2015).
32. W.M. Brown, S. Martin, M.D. Rintoul, and J.-L. Faulon, *J. Chem. Inf. Model.* 46(2), 826–835 (2006).
33. C.J. Churchwell, M.D. Rintoul, S. Martin, D.P. Visco Jr., A. KotuKotu, R.S. Larson, L.O. Sillerud, D.C. Brown, and J.-L. Faulon, *J. Molecul. Graph. Model.* 22(4), 263–273 (2004).
34. F.A. Lindemann, *Z. Phys.* 11, 609–612 (1910).
35. J.P. Poirier, *Phys. Earth Planetary Interiors* 54(3–4), 364–369 (1989).
36. G.H. Wolf, and R. Jeanloz, *J. Geophys. Res. Solid Earth* 89(BB9), 7821–7835 (1984).
37. Wolfram Research Inc. Mathematica, version 12.0. URL https://www.wolfram.com/wolfram-alpha-notebook-edition. Champaign, IL, (2019).
38. S.P. Ong, S. Cholia, A. Jain, M. Brafman, D. Gunter, G. Ceder, and K.A. Persson, *Comput. Mater. Sci.* 97, 209–215 (2015). https://doi.org/10.1016/j.commatsci.2014.10.037.
39. A. Jain, G. Hautier, C.J. Moore, S.P. Ong, C.C. Fischer, T. Mueller, K.A. Persson, and G. Ceder, *Comput. Mater. Sci.* 50(8), 2295–2310 (2011).
40. A.M. Deml, A.M. Holder, R.P. O'Hayre, C.B. Musgrave, and V. Stevanović, *J. Phys. Chem. Lett.* 6(10), 1948–1953 (2015).
41. J.T. Schick, A.M. Gopakumar, and A.M, Rappe, Descriptors for thermal expansion in solids. arXiv:1701.03966, (2017).
42. C.W. Coley, W. Jin, L. Rogers, T.F. Jamison, T.S. Jaakkola, W.H. Green, R. Barzilay, and K.F. Jensen, *Chem. Sci.* 10(2), 370–377 (2019).
43. D.C. Elton, Z. Boukouvalas, M.S. Butrico, M.D. Fuge, and P.W. Chung, *Sci. Rep.* 8(1), 1–12 (2018).
44. J. Ling, M. Hutchinson, E. Antono, S. Paradiso, and B. Meredig, *Integr. Mater. Manufact. Innov.* 6(3), 207–217 (2017).
45. M. Raissi, P. Perdikaris, and G.E. Karniadakis, *J. Comput. Phys.* 378, 686–707 (2019).
46. M. Blackman, *HDP* 3, 325–382 (1955).
47. L. Ward, A. Dunn, A. Faghaninia, N.E.R. Zimmermann, S. Bajaj, Q. Wang, J. Montoya, J. Chen, K. Bystrom, M. Dylla et al., *Comput. Mater. Sci.* 152, 60–69 (2018).
48. L. Ward, A. Agrawal, A. Choudhary, and C. Wolverton, *npj Comput. Mater.* 2, 16028 (2016).
49. B. Meredig, A. Agrawal, S. Kirklin, J.E. Saal, J.W. Doak, A. Thompson, K. Zhang, A. Choudhary, and C. Wolverton, *Phys. Rev. B* 89(9), 094104 (2014).
50. W.D. Callister, and D.G. Rethwisch, *Materials science and engineering: an introduction* (Wiley, New York, 2018).
51. Tin Kam Ho, *IEEE Trans. Pattern Anal. Mach. Intell.* 20(8), 832–844 (1998).
52. L. Breiman, *Mach. Learn.* 45(1), 5–32 (2001).
53. B Efron. Model selection estimation and bootstrap smoothing. division of biostatistics, (2012).
54. S. Wager, T. Hastie, and B. Efron, *J. Mach. Learn. Res.* 15(1), 1625–1651 (2014).
55. R.K. Tripathy, and I. Bilionis, *J. Comput. Phys.* 375, 565–588 (2018),.
56. M Hutchinson. Citrine informatics lolo, (2016).
57. T.M. Oshiro, P.S. Perez, and J.A. in *International workshop on machine learning and data mining in pattern recognition*, pages 154–168. Springer, (2012).
58. Y. Jun Wang, X.Y.C. Zhou, R. Zhou, and J. Feng, *Ceramics Int.* 42(12), 13876–13881 (2016).
59. A.K. Bhattacharya, V. Shklover, W. Steurer, G. Witz, H.-P. Bossmann, and O. Fabrichnaya, *J. Eur. Ceram. Soc.* 31(3), 249–257 (2011).
60. J.L. Caslavsky, and D.J. Viechnicki, *J. Mater. Sci.* 15(7), 1709–1718 (1980).
61. G. Shengyue, S. Zhang, F. Liu, and W. Li, *J. Eur. Ceram. Soc.* 38(15), 5082–5091 (2018).
62. K. Kuribayashi, M. Yoshimura, T. Ohta, and T. Sata, *J. Am. Ceram. Soc.* 63(11–12), 644–647 (1980).
63. R.C. Reed, *The superalloys: fundamentals and applications* (Cambridge University Press, Cambridge, 2008).
64. S. Yamanaka, K. Kurosaki, T. Oyama, H. Muta, M. Uno, T. Matsuda, S.-I. Kobayashi, *J. Am. Ceram. Soc.* 88(6), 1496–1499 (2005).
65. Z.D. McClure, and A. Strachan, High temperature oxide property explorer, (Jun 2020). https://nanohub.org/resources/htoxideprop.