Computational Materials Science 230 (2023) 112431

Contents lists available at ScienceDirect

Computational Materials Science

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

Full Length Article

Voxelized atomic structure framework for materials design and discovery

Matthew C. Barry [a], Jacob R. Gissinger [b], Michael Chandross [c], Kristopher E. Wise [b], Surya R. Kalidindi [a,*], Satish Kumar [a,*]

[a] G.W. Woodruff School of Mechanical Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA
[b] Advanced Materials and Processing Branch, NASA Langley Research Center, Hampton, VA 23681, USA
[c] Material, Physical and Chemical Sciences Center, Sandia National Laboratories, Albuquerque, NM 87123, USA

ARTICLE INFO

Keywords:
Structure-property relationship
Density functional theory
Charge density field
Atomistic modeling
Materials Informatics
Multicomponent

ABSTRACT

We present a computational framework for developing physics-based, high-fidelity structure–property relationships with atomic systems. In this framework, atomic structure is quantified by directionally resolved two-point spatial correlations of the charge density field, projected to a salient low-dimensional feature space via principal component analysis (PCA), and correlated to physical properties by Gaussian process regression (GPR). The charge density field provides a complete, purely physics-based definition of the atomic structure that is independent of chemical species information and does not require additional feature engineering or idealizations beyond those of first-principles computations. The two-point spatial correlations capture the salient spatial features underlying the atomic structure that dictate the physics underlying the material response. Since the feature engineering approach explored in this work is universally applicable to all atomic structures independent of the chemical species present in the structure, it offers new avenues for efficiently exploring the space of atomic structures for desired property combinations. A further contribution of this work comes from utilizing the uncertainty quantification inherently provided by GPR to deploy a Bayesian experiment design strategy to minimize the number of computationally expensive physics simulations required to achieve the desired accuracy. In this work, we demonstrate the proposed framework to elucidate the relationship between the chemical composition and bulk modulus in AlNbTiZr high entropy alloys. It is shown that a highly accurate structure–property relationship with less than 2% average error can be established using a small training dataset of less than 30 samples.

1. Introduction

The relationship between material structure and property is governed by complex physical processes over a hierarchy of length scales, making it computationally expensive to model the underlying linkages with physics-driven simulation methods alone. Atomic systems at the first-principles level are typically considered the smallest relevant length scale in materials modeling, and thus form the foundation of the multiscale physics governing the material response [1,2]. However, it is well known that first-principles simulation methods are limited to small atomic systems and short length scales due to their high computational cost and poor scaling with system size [3–6]. Machine learning (ML) approaches offer new opportunities to efficiently learn the desired structure–property relationships in atomic systems using low-computational cost surrogate models trained on expensive first-principles-based computations [7–40]. The successful development of such an approach promises to efficiently elucidate practically useful material structure–property relationships that can enable the rapid exploration of materials space and aid in the design of multifunctional materials with tailored properties [41–44].

The development of reliable low-computational cost structure–property relationships from data collected from computationally expensive physics-based simulations encounters two main challenges: (i) a suitable definition of the material structure and subsequent quantification of its salient features (i.e., feature engineering) to serve as inputs for the ML models, and (ii) a learning strategy that maximizes the accuracy and reliability of the developed relationships while minimizing the need for the computationally expensive training data. The material structure feature engineering challenge has been addressed in prior literature from the first-principles length scale [20–32] to the mesoscale [45–48] using a number of different approaches that rely primarily on expert-guided selection of the salient material structure descriptors.

* Corresponding authors.
E-mail addresses: mbarry31@gatech.edu (M.C. Barry), surya.kalidindi@me.gatech.edu (S.R. Kalidindi), satish.kumar@me.gatech.edu (S. Kumar).

https://doi.org/10.1016/j.commatsci.2023.112431
Received 14 January 2023; Received in revised form 23 July 2023; Accepted 8 August 2023
Available online 24 August 2023

While such approaches have demonstrated notable successes, the central challenge in active selection of material structure descriptors from intuition is that the information loss resulting from the chosen low-dimensional representation is not necessarily minimized and may be difficult to quantify. For example, it is not clear to what extent the original structure can be recovered from such choices of the low-dimensional representation. Furthermore, it is difficult to design such representations to capture the material structure at more than one length scale, limiting potential applications in a generalized hierarchical multiscale materials modeling framework. Efforts to establish protocols that provide a consistent representation of the material structure across multiple material length scales and material classes have led to the development of approaches in which the material structure is quantified by its spatial correlations [49–62]. The n-point spatial correlation function in particular has emerged as a powerful solution to the material structure feature engineering challenge due to its comprehensive statistical description of the material structure that allows for the systematic and hierarchical inclusion of increasingly complex higher-order spatial correlations [56–62].

In recent years, the materials knowledge system (MKS) framework [56–58] has formally addressed the challenge of material structure engineering using a systematic approach comprising the following sequence of steps: (i) defining the material structure as a microstructure function (i.e., mapping of material local states to spatial locations), (ii) quantifying the spatial features underlying the material structure with directionally resolved two-point spatial correlations (and high-order spatial correlations as needed) [60,61], and (iii) projecting these spatial features to a salient low-dimensional feature space utilizing principal component analysis (PCA) [63]. The MKS framework has been primarily focused on structure–property relationships at the mesoscale, with only limited explorations thus far at the atomic scale [37–40]. In the context of atomic systems, the MKS framework has defined the material local state as a vector of atomic attributes (including labels for chemical species, heat of fusion, ionization energy, Pauling electronegativity, etc.); each vector corresponding to a specific chemical species present in the atomic system. However, the selection of such atomic attributes is largely *ad hoc*, and it is impossible to determine the best combination of atomic attributes that offer the most critical value for building high-fidelity ML models. Furthermore, since these representations of the atomic structure depend directly on the chemical species present in the atomic system, the resulting structure–property relationships will be incapable of reliably extrapolating to atomic systems containing chemical species that were not present in the training dataset.

In this article, we explore the feasibility of directly utilizing the charge density field as the definition of atomic structure. Unlike atomic attribute-based definitions, the charge density field provides a complete, purely physics-based definition of the atomic structure and therefore does not require additional feature engineering or idealizations beyond those of the first-principles computations. Most importantly, since the charge density field implicitly embeds the chemical species information in the atomic system, we argue that this approach offers the best avenue for reducing the material local state from a high-dimensional vector of possible atomic attributes to a single field variable – charge density. As a result, structure–property relationships utilizing the charge density field definition of the atomic structure do not depend explicitly on the chemical species present in the atomic systems, and therefore are capable of extrapolating to atomic systems containing chemical species not present in the training dataset.

A few studies have attempted to develop structure–property relationships by utilizing the charge density field directly as the essential definition of the material structure at the molecular length scale [33–35]. In one such study, Pilania et al. [33] explored the feasibility of predicting effective properties of polymer chains utilizing the Fourier coefficients of the corresponding 1-dimensional charge density fields (averaged along the plane normal to the chain axis). However, they found that structure–property relationships developed with this definition of atomic structure performed worse (on the same dataset) than those developed utilizing an atomic attribute-based definition of atomic structure. Other approaches [34,35] have directly utilized the 3-dimensional charge density field as the input to a convolutional neural network (CNN) [64,65]. However, these approaches have practical and physical limitations. From a ML perspective, CNNs require substantial training data to achieve sufficient accuracy, making them impractical for computationally expensive material properties. From the perspective of materials design/discovery, the black-box nature of CNNs make it challenging to reverse engineer novel material structures with tailored physical properties. Finally, from a physics perspective, it should be expected that the effective properties of a given atomic structure will depend mainly on certain spatial patterns that underly the charge density field. In other words, the salient regressors for predicting the effective properties should be a projection of the charge density field onto a metric space – called the feature space – which quantifies the "distance" between two atomic structures based on the differences in their spatial patterns.

Towards this goal, we build on our recent experience [8] to introduce, for the first time, a new voxelized atomic structure (VASt) framework for formulating physics-based, high-fidelity reduced-order structure–property relationships in which the salient spatial features underlying the atomic structure that dictate the material response are quantified using the directionally-resolved two-point spatial correlations of the charge density field and subsequently projected to a low-dimensional feature space using PCA. We claim that the mathematical and physical theory on which the VASt framework is built addresses all of the abovementioned deficiencies in existing approaches. Even though the protocols employed in the VASt framework (i.e., microstructure function definition of material structure, two-point spatial correlations, and PCA) are mathematically equivalent to those utilized in the MKS framework, the VASt framework is novel in applying these protocols to the first-principles length scale. As a result, the VASt framework is not only a significant standalone contribution towards the development of reduced-order material structure–property relationships with atomic systems, but also a critical step towards the development of a comprehensive hierarchal multiscale materials modeling framework.

The second contribution of this paper lies in exploring the benefits of combining active learning strategies and GPR models for building reliable structure–property linkages at the molecular scale with relatively small, but optimally selected, training data. Because GPR models the output as a probability distribution rather than a single point, they offer new avenues for the implementation of active learning strategies that minimize the need for large collections of training data, which for the present application requires the execution of computationally expensive density functional theory (DFT) computations. The proposed VASt framework is designed to take full advantage of these emergent toolsets. As a result, the structure–property models formulated by the proposed VASt framework promise high accuracy and reliability, while minimizing the computational cost of the training data needed. Although similar concepts have been explored in the MKS framework at the mesoscale with great success [66,67], their feasibility and utility has not been evaluated systematically at the atomic scale.

The feasibility and utility of the proposed VASt framework are evaluated in this paper by developing a structure–property relationship for predicting the bulk modulus of AlNbTiZr refractory high entropy alloys (RHEAs) as a function of their chemical composition. We demonstrate that two-point spatial correlations of the charge density field provide a high-fidelity statistical representation of the spatial features underlying the global atomic structure that dictate the material response. We then show that PCA projects the spatial features captured by the two-point spatial correlations of the charge density field to a salient low-dimensional feature space that minimizes information loss and allows for high-throughput exploration of materials design space for desired properties. We then evaluate the accuracy of the VASt

structure–property relationship and the reliability of its uncertainty quantification. Finally, we demonstrate the significant improvements/advantages provided by the abovementioned Bayesian experiment design strategy compared to random selection.

2. Methodology

In this section, we present the theoretical basis of the VASt framework. We first define the two-point spatial correlation function for charge density fields. We then detail a protocol for efficiently computing the discrete two-point spatial correlation function of a discrete charge density field using fast Fourier transforms. We then briefly discuss PCA and its value in reverse-engineering novel atomic structures with desired physical properties. We then review GPR and its use in the VASt framework. Finally, we present a Bayesian experiment design protocol implemented in the VASt framework to minimize the number of computationally expensive physics simulations required for training highly accurate models.

2.1. Two-point spatial correlation function

The material structure can be defined mathematically as a mapping $m : \Omega \times \boldsymbol{H} \rightarrow \mathbb{R}_{\geq 0}$ of material local states, $\boldsymbol{h} \in \boldsymbol{H}$, over a continuous spatial domain, $\Omega \subset \mathbb{R}^3$ [57,60]. Physically, the charge density field is a continuous function $\rho : \Omega \rightarrow \mathbb{R}^{+}_{\geq 0}$, where $\Omega \subset \mathbb{R}^3$ is the spatial domain of the atomic system. As previously mentioned, the charge density field provides a complete, purely physics-based definition of the atomic structure that implicitly embeds the chemical species information present in the atomic system. As a result, the material local state is defined completely by a single field variable – charge density. Therefore, in the VASt framework, we assume that the atomic structure (i.e., the material internal structure) is adequately represented by the charge density field (i.e., $m(\boldsymbol{x}, \boldsymbol{h}) = \rho(\boldsymbol{x})$). It should be noted that in DFT computations it is typical to employ the frozen core approximation and only consider the valence electrons, since the core electrons often have a negligible effect on the physics. Therefore, we have only considered the charge density field resulting from the valence electrons in this work. Although it is possible for multiple chemical species to have the same number of valence electrons, such chemical species are not permutationally invariant (i.e., are not interchangeable) since their contributions to the valence charge density field of a given system are unique due to their complex and distinct interactions with the other valence electrons in the system. We also note that although the VASt framework depends on the DFT-computed charge density field of new systems to predict their effective properties, this requirement does not increase the computational cost compared to descriptors based only on atomic positions and chemical species. This is because for any new system whose properties one wants to predict, in nearly all practical applications, the ground state atomic structure (i.e., atomic positions) would be unknown. Therefore, evaluating a new system with a descriptor based on atomic positions and chemical species still requires a structure minimization performed with DFT (of which the charge density field is a direct output) to obtain the ground state atomic positions needed to construct the descriptors/features required to predict the desired properties.

In establishing practically useful structure–property relationships, we are not necessarily interested in the material structure directly, but rather the spatial features underlying the material structure. Our interest lies in predicting the effective response (i.e., property) associated with a given atomic structure. Since effective physical properties of a material are determined by the material structure as a whole, suitable global measures of the material structure are needed to correlate a given material to its effective properties. Such global measures can be constructed by quantifying the relative spatial positioning of two or more material local states (i.e., spatial correlations). Since this type of spatial feature can be present at multiple locations throughout the atomic structure, the intensity (i.e., signal value) of such a spatial feature should quantify the frequency with which it occurs in the material structure. In the VASt framework, a novel global measure of the atomic structure is defined using the two-point spatial correlation function of the charge density field, defined as

$$f(\boldsymbol{r}) = \frac{1}{|\Omega_{\boldsymbol{r}}|} \int_{\boldsymbol{x} \in \Omega_{\boldsymbol{r}}} \rho(\boldsymbol{x})^{1/2} \rho(\boldsymbol{x} + \boldsymbol{r})^{1/2} d\boldsymbol{x} \tag{1}$$

where $\Omega_{\boldsymbol{r}} = \{\boldsymbol{x} \in \Omega \,|\, \boldsymbol{x} + \boldsymbol{r} \in \Omega\}$ and $|\Omega_{\boldsymbol{r}}|$ is the measure of $\Omega_{\boldsymbol{r}}$. For periodic atomic systems, $\Omega_{\boldsymbol{r}} = \Omega$ and $|\Omega_{\boldsymbol{r}}|$ is equal to the volume of the periodic unit cell. We chose to define Eq. (1) using the square root of the charge density field so that $f(\boldsymbol{0})$ is the average charge density of the atomic system. Mathematically, Eq. (1) is the autocorrelation of the charge density field (or more specifically, its square root) normalized by the volume of the atomic system. Physically, $f(\boldsymbol{r})$ is a measure of the charge density found throughout the global atomic system separated by the vector $\boldsymbol{r} \in \mathbb{R}^3$. Hence, the two-point spatial correlation function offers a meaningful statistical quantification of the global atomic structure and, as a result, the distance between the two-point spatial correlation functions of two charge density fields is not sensitive to local fluctuations in their respective charge density fields. Consequently, statistically similar charge density fields will have similar two-point spatial correlation function representations, and therefore should have similar predicted properties. We note that although the two-point spatial correlation function is inherently translationally and permutationally invariant, it is not inherently rotationally invariant. However, rotational invariance can be enforced on the two-point spatial correlations by (1) imposing a standardized reference frame defined procedurally in a way that is independent of the initial origin and orientation of the given charge density field, such as with a suitable modification of the procedure established in [19] or (2) removing the complex argument (i.e., phase angle information) from the Fourier transform of the two-point spatial correlations [68].

The two-point spatial correlation function is often a sufficient low-order approximation of the complete statistical description of the material structure. However, a key feature of the VASt framework is that higher-order spatial correlations of the charge density can be included in a systematic and hierarchical manner as needed. Furthermore, although the charge density field is an excellent representation of the atomic structure for many applications, it may be insufficient in more complex applications such as those involving spin-polarized charge densities. Another key feature of the VASt framework is that the (total) charge density field is just one specific instance of a microstructure function (in which the material local state is defined completely by total charge density). Therefore, it is trivial to extend the VASt framework to account for complexities such as spin-polarized charge densities by simply extending the material local state from one field variable (i.e., total charge density) to as many variables as needed to fully account for the physics imposed by such complexities on the material response. In the case of spin-polarized charge densities, for example, the material local state can be defined completely by splitting the total charge density into the spin up and spin down charge densities. The full set of two-point spatial correlations would then be given by the autocorrelations of the spin up and spin down charge density fields, respectively, along with the cross-correlation between them.

2.2. Numerical implementation of two-point spatial correlation function for charge density fields

The numerically computed charge density field is a discrete function $\rho : \mathbf{S} \rightarrow \mathbb{R}^{+}_{\geq 0}$, where $\boldsymbol{S} = \{(s_1, s_2, s_3) | s_i \in [0, 1, \cdots, S_i - 1], i = 1, 2, 3\}$ is the set of all voxel indices, $\boldsymbol{s} = (s_1, s_2, s_3)$, that discretize the three-dimensional spatial domain of the atomic structure, and $S_1, S_2, S_3 \in \mathbb{N}$ are the number of voxels along each dimension. The discrete two-point spatial correlation function of the charge density field is thus given by

$$f[\boldsymbol{t}] = \frac{1}{|\boldsymbol{S}_t|}\sum_{s\in S_t}\rho[\boldsymbol{s}]^{1/2}\rho[\boldsymbol{s}+\boldsymbol{t}]^{1/2}, \tag{2}$$

where $\mathbf{S}_t = \{\boldsymbol{s}\in\mathbf{S}\,|\boldsymbol{s}+\boldsymbol{t}\in\mathbf{S}\}$ and $|\mathbf{S}_t|$ is the cardinality of the set $\boldsymbol{S}_t$ [62,69]. For periodic atomic systems, we can take $\mathbf{S}_t = \mathbf{S}$ and in this case, Eq. (2) is the discrete autocorrelation of the discrete charge density field (or more specifically, its square root), which can be written in terms of discrete Fourier transforms as

$$f[\boldsymbol{t}] = \frac{1}{|\boldsymbol{S}|}\mathscr{F}^{-1}\left[\mathscr{F}\left(\rho[\boldsymbol{s}]^{1/2}\right)^{*}\mathscr{F}\left(\rho[\boldsymbol{s}]^{1/2}\right)\right], \tag{3}$$

where $\mathscr{F}$ and $\mathscr{F}^{-1}$ denote the discrete Fourier transform and inverse discrete Fourier transform, respectively, and * denotes the complex conjugate. By defining the discrete two-point spatial correlation function in the form of Eq. (3), we can take advantage of computationally efficient fast Fourier transform algorithms to compute the two-point spatial correlations. The components of the discrete two-point spatial correlation function defined in Eq. (3) form a high-dimensional vector space of material structure descriptors that encode the spatial features of a given atomic system.

For the discrete two-point spatial correlation function to provide meaningful representations of the atomic structure for building structure–property relationships, the discrete charge density fields must all be defined on a standardized discretization of their respective spatial domains (i.e., the spatial domain on which each charge density field is defined must be discretized using a consistent, standardized voxel size). However, it is impossible to find a (reasonable) voxel size that will perfectly discretize the spatial domain of every atomic system under consideration. Thus, the charge density fields must be approximated on a standardized spatial domain prior to computing their two-point spatial correlations using a suitable interpolation scheme. Details of the methods used to standardize the charge density field are provided in Appendix A. The standardized charge density field can also be obtained directly by reconstructing and sampling the basis function expansion of the charge density. Here, we have employed a computationally efficient scheme that does not depend on the DFT simulation method or the specific basis used.

2.3. Principal component analysis

The discrete two-point spatial correlation function representation of the charge density field could be used directly as the input to a CNN without suffering from the same feature-based limitations as utilizing the charge density field directly. However, this approach would still be limited by both the substantial quantities of training data required to achieve highly accurate CNNs and the challenges of reverse engineering novel material structures from CNNs. Thus, in practice we are typically forced to use models such as GPR, which have significantly lower demands on training data. However, the discrete two-point spatial correlation function is a high-dimensional representation of the charge density field and GPR models typically require that the input be of relatively low-dimensionality. Therefore, it is first necessary to project the high-dimensional two-point spatial correlations of the charge density field to a low-dimensional feature space that can then be used as the input to the GPR model. In the VASt framework, a salient low-dimensional feature space is obtained by applying PCA to the collection of two-point spatial correlations obtained for a representative set of charge density fields. PCA is a linear, distance-preserving transformation that projects the data onto a new orthogonal basis whose basis vectors are aligned with and ordered by the directions of maximum variance in the dataset. There are two main advantages of PCA. First, there is typically a strong decay in the amount of variance captured by each successive basis vector. As a result, the subspace of the full PC space determined by the first few basis vectors provides a salient low-dimension approximation of the charge density field that inherently minimizes information loss. Second, given any set of theoretically feasible charge density fields, any point in the convex hull of the PC space representation of their two-point spatial correlations can be interpreted to correspond to a theoretically feasible charge density field. The set of two-point spatial correlations corresponding to all theoretically feasible heterogeneous but statistically homogeneous structures that can be produced from a prescribed set of material local states (e.g., charge density fields) is a convex set [61]. Since PCA represents a simple rotational transformation, the PC space representation of this complete set of two-point spatial correlations will also be convex. Hence, any point in the convex hull of the PC space representation of the two-point spatial correlations of a given set of known charge density fields corresponds to a theoretically feasible charge density field. Therefore, the PC space representation of the charge density field provides a practical avenue for reverse engineering atomic systems with tailored properties. Here, we demonstrate the novel use of PCA as a low-dimensional representation of two-point spatial correlations of the charge density field to predict material properties.

2.4. Gaussian process regression

Gaussian process regression offers a nonparametric, kernel-based ML method for developing probabilistic models using Bayesian inference. Let $\boldsymbol{X}$ and $\boldsymbol{X}_*$ be $N\times D$ and $N_*\times D$ matrices of input features corresponding to the training and testing samples, respectively, where N is the number of training samples, N_* is the number of testing samples, and D is the number of input features used to represent a sample. Let $\boldsymbol{y}$ and $\boldsymbol{f}_*$ be the vectors of outputs (i.e., targets) corresponding to the training samples and testing samples, respectively. In GPR, the mapping from the inputs to the outputs is modeled as a joint multivariate Gaussian distribution defined as

$$\begin{bmatrix}\boldsymbol{y}\\ \boldsymbol{f}_*\end{bmatrix} \sim \mathscr{N}\left(\begin{bmatrix}\boldsymbol{\mu}(\boldsymbol{X})\\ \boldsymbol{\mu}(\boldsymbol{X}_*)\end{bmatrix}, \begin{bmatrix}\boldsymbol{K}(\boldsymbol{X},\boldsymbol{X}) & \boldsymbol{K}(\boldsymbol{X},\boldsymbol{X}_*)\\ \boldsymbol{K}(\boldsymbol{X}_*,\boldsymbol{X}) & \boldsymbol{K}(\boldsymbol{X}_*,\boldsymbol{X}_*)\end{bmatrix}\right), \tag{4}$$

where $\mathscr{N}$ denotes a multivariate normal distribution completely specified by its mean function, $\boldsymbol{\mu}(\boldsymbol{X})$, and its covariance function, $\boldsymbol{K}(\boldsymbol{X},\boldsymbol{X}')$. Typically, we can assume $\boldsymbol{\mu}\equiv 0$. In this case, the predictive distributions for the test samples are given by

$$P(\boldsymbol{f}_*|\boldsymbol{X},\boldsymbol{y},\boldsymbol{X}_*) \sim \mathscr{N}(\bar{\boldsymbol{f}}_*, \mathrm{cov}(\boldsymbol{f}_*)), \tag{5}$$

where

$$\bar{\boldsymbol{f}}_* \triangleq \mathbb{E}[\boldsymbol{f}_*|\boldsymbol{X},\boldsymbol{y},\boldsymbol{X}_*] = \boldsymbol{K}(\boldsymbol{X}_*,\boldsymbol{X})\boldsymbol{K}(\boldsymbol{X},\boldsymbol{X})^{-1}\boldsymbol{y} \tag{6}$$

$$cov(\boldsymbol{f}_*) = \boldsymbol{K}(\boldsymbol{X}_*,\boldsymbol{X}_*) - \boldsymbol{K}(\boldsymbol{X}_*,\boldsymbol{X})\boldsymbol{K}(\boldsymbol{X},\boldsymbol{X})^{-1}\boldsymbol{K}(\boldsymbol{X},\boldsymbol{X}_*). \tag{7}$$

In the VASt framework, the covariance matrix is computed using the automatic relevance determination squared exponential (ARDSE) kernel [70] defined as

$$k(\boldsymbol{x},\boldsymbol{x}') = \sigma_s^2 exp\left[-\frac{1}{2}\sum_{d=1}^{D}\left(\frac{(x_d - x_d')^2}{l_d^2}\right)\right] + \sigma_n^2\delta_{\boldsymbol{x}\boldsymbol{x}'}, \tag{8}$$

where $\boldsymbol{x}$ and $\boldsymbol{x}'$ represent the feature vectors of two samples, σ_s is a scaling factor for the output variance, l_d is the length-scale parameter corresponding to the d^{th} input feature, σ_n is a noise hyperparameter that captures the variance in the training dataset, and $\delta_{\boldsymbol{x}\boldsymbol{x}'}$ is the Kronecker delta function.

2.5. Bayesian experiment design

In materials design/discovery applications, the computational cost of obtaining the atomic structure of a sample (i.e., charge density field) is typically orders of magnitude lower than the computational cost of computing the ground-truth (i.e., physical) value of a desired property. Therefore, while it is possible to access an arbitrary number of candidate structures, we can only compute the ground-truth properties for a small

subset of them. Thus, a successful ML framework for exploring a large compound space must not only be capable of achieving high accuracy, but also minimize the number of computationally expensive physics simulations required to achieve that accuracy. In the VASt framework, this is accomplished by utilizing the uncertainty quantification provided by GPR to implement an active learning strategy based on Bayesian experiment design [71,72], as shown in Fig. 1. Let $X = \{\boldsymbol{x}_n\}_{n=1}^{N}$ be the set of all available atomic structures (or more specifically, their feature space representations). As previously mentioned, we assume that the number of available atomic structures, N, is sufficiently large for training. We first compute the ground-truth properties for a small subset, $X_{K,0} \subset X$, to obtain the initial training dataset $T_{K,0} = \{(\boldsymbol{x}, \boldsymbol{y}) | \boldsymbol{x} \in X_{K,0}\}$, where $\boldsymbol{y}$ is the ground-truth property (or properties) of the atomic structure represented by $\boldsymbol{x}$, and the subscript K is used to denote samples whose ground-truth properties are known. A GPR model is then trained with $T_{K,0}$ and used to estimate the predictive distributions on the remaining candidate structures, $X_{U,0} = X \backslash X_{K,0}$, where the subscript U denotes samples whose ground-truth properties are unknown. The expected information gain, $I(\boldsymbol{x})$, for the GPR model from each candidate atomic structure, $\boldsymbol{x} \in X_{U,0}$, is then evaluated from its predictive distribution. The ground-truth properties of the k atomic structures with the largest information gain, $X_{C,0} = \underset{\boldsymbol{x} \in X_{U,0}}{\max_k} I(\boldsymbol{x})$, are then computed and the chosen samples are added to the training dataset. This process is repeated until the accuracy of the GPR model converges. In this study, we define the expected information gain for a candidate sample as $I(\boldsymbol{x}) = |\sigma(\boldsymbol{x})/\mu(\boldsymbol{x})|$, where $\boldsymbol{x}$ is the feature vector representation of the candidate sample and $\mu(\boldsymbol{x})$ and $\sigma(\boldsymbol{x})$ are the mean and standard deviation of the corresponding GPR predicted distribution for the output, respectively. Under this definition, the candidate sample with the largest expected information gain is the one with the largest predicted uncertainty (i.e., variance) relative to the magnitude of its predicted property (i.e., mean). A Bayesian experiment design procedure with this definition of information gain will therefore aim to minimize the (normalized) uncertainty in the model predictions.

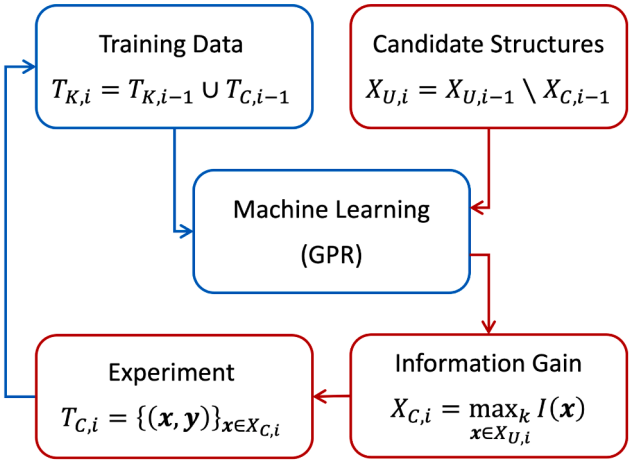

Fig. 1. The Bayesian experiment design active learning procedure used to develop VASt structure–property relationships. For every iteration, a new GPR model is trained using the current training dataset, $T_{K,i}$. This GPR model is then used to estimate the predictive distributions on the remaining candidate atomic structures, $X_{U,i}$. The set of k candidate atomic structures that maximize the information gain to the current GPR model, $X_{C,i}$, are chosen for the next set of computationally expensive experiments (i.e., first-principles computations). Once the experiments are completed, the set of chosen atomic structures with corresponding ground-truth properties, $T_{C,i}$, is added to the training dataset and the process is repeated until the model accuracy converges.

3. Case Study: Bulk modulus of refractory high entropy alloys

The VASt framework is demonstrated by developing a structure–property relationship for predicting the bulk modulus of AlNbTiZr RHEAs as a function of their chemical composition. RHEAs have gained significant interest for their excellent mechanical properties at high temperatures [73,74]. The composition of RHEAs, typically consisting of 4+ elements present in varying compositions, can have a significant effect on their properties [75,76]. As a result, the composition space of possible RHEAs is extremely large, making it impractical to optimize their chemical composition for tailored properties using computationally expensive physics simulations alone. Furthermore, because RHEAs are typically studied at extreme temperatures, classical simulations with currently available empirical interatomic potentials may not produce accurate results and new empirical potentials may be challenging to develop. Previous studies have attempted to overcome these limitations by developing ML-based interatomic potentials to study RHEAs using molecular dynamics (MD) simulations [76,77]. Here, we show that the VASt framework is capable of not only producing a high-fidelity reduced-order homogenization model for the effective properties of RHEAs, but also that it can do so with a relatively small training dataset.

3.1. Computational methods

To sufficiently cover the composition space of AlNbTiZr RHEAs, we consider the set

$$\left\{\mathrm{Al}_\alpha \mathrm{Nb}_\beta \mathrm{Ti}_\gamma \mathrm{Zr}_\delta | \alpha, \beta, \gamma, \delta \in [0, 4, \cdots, 128] \text{ with } \alpha + \beta + \gamma + \delta = 128\right\} \tag{9}$$

of 128-atom RHEAs. This set has a total of 6545 compositions. For each composition, we first obtain the special quasirandom structure (SQS) [78–81] corresponding to a BCC atomic structure. It should be noted that the use of an SQS assumes that the RHEAs are a perfect random solid solution. Therefore, the properties modeled by the VASt framework in this study correspond to those of a perfectly random solid solution alloy. Although this assumption is common and reasonable when modeling RHEAs [74–77], it should be noted that high entropy alloys can also often be multiphase [82] and that Al and Zr have shown some tendency for ordering in high entropy alloys [83]. We then obtain the ground state atomic structure (atomic positions and lattice constant) and corresponding charge density field using DFT [3,5]. The DFT calculations are performed using the Vienna Ab initio Simulation Package (VASP) [84,85]. A plane-wave basis set and the projector augmented wave method [86,87] are used with the Perdew-Burke-Ernzerhof generalized gradient approximation exchange–correlation functional [88]. The VASP-recommended pseudopotentials are used for each element. Specifically, for Nb, Ti, and Zr, we use the "_sv" pseudopotentials, which treat the semi-core p- and s-states as valence. For Al, we use the standard pseudopotential, in which no semi-core states are treated as valence. We note that this is the only available pseudopotential for Al. The calculations are performed on a $2 \times 2 \times 2$ grid of k-points using a plane-wave basis cutoff energy of 400 eV. The convergence criteria for the electronic self-consistency and ionic relaxation steps are 10^{-4} eV and 10^{-2} eV, respectively. The chosen tolerances were found to be sufficiently converged with respect to bulk modulus (stricter tolerances resulted in a change of less than 1 GPa — or less than 0.2% — in both the elemental Nb and equiatomic AlNbTiZr systems). The equilibrium lattice constant is obtained by fitting a Birch-Murnaghan Equation of State (EOS) [89,90] to 9 energy-volume points taken about the equilibrium volume. The change in volume is applied hydrostatically and the atomic positions are relaxed at each volume. The ground state atomic structure and corresponding charge density field are then obtained by performing a final relaxation of the atomic positions using the equilibrium lattice constant. The corresponding ground-truth (i.e., DFT-computed) bulk modulus is obtained from the Birch-Murnaghan EOS. We found that the bulk moduli of the elemental

systems computed using these parameters were in good agreement with available experimental [91–93] and computational [94,95] values as shown in Table 1. For example, the DFT-computed bulk modulus of BCC Nb in this study, 165.3 GPa, is within less than 3% of the experimentally measured value at 273 K, 170.0 GPa [91].

3.2. Two-point spatial correlations

The discrete two-point spatial correlations of the charge density fields are computed using a standardized voxel size, λ, of 0.2 Å and a cutoff distance, R_c, of 6.0 Å. Examples of the charge density field and corresponding standardized two-point spatial correlations for BCC Nb and the equiatomic TiZr BCC SQS are shown in Fig. 2. As expected, the charge density field of Nb is that of a perfect, single-species BCC atomic system, whereas the charge density field of the equiatomic TiZr SQS resembles that of a BCC crystal subjected to some quantity of disorder resulting from the varying valency of the different chemical species, the inherent disorder in an SQS, and the lattice distortions present in the ground-state atomic structure. These spatial features are well-captured by the corresponding two-point spatial correlations. For the Nb charge density field, the periodic pattern of the corresponding two-point spatial correlations reflects the expected symmetry in a geometrically perfect, single-species BCC atomic structure. For the equiatomic TiZr SQS, the underlying BCC structure is still evident in the two-point spatial correlations, but there is a steady decay in the peak values with an increasing magnitude of $\boldsymbol{r}$ (up to the periodicity of the atomic structure simulated). This is due to the previously mentioned sources of disorder in the equiatomic TiZr SQS. Hence, the two-point spatial correlations can provide a measure of the distortion of a given atomic system from that of a perfect one. Furthermore, because the two-point spatial correlation function is a statistical quantification of the global atomic structure, atomic systems with statistically similar distortions will have similar two-point spatial correlations even though their charge density fields may not appear similar when compared directly. Additionally, as previously mentioned, the zero-vector of the two-point spatial correlation function (located in the center of the two-point spatial correlation maps shown in Fig. 2) corresponds to the average charge density of a voxel in the standardized discrete spatial domain of the atomic system. Comparing the Nb and equiatomic TiZr two-point spatial correlations, it is seen that the average charge density in the Nb atomic system is greater than that of the equiatomic TiZr system. This is reasonable since Nb has a larger valency than Ti and Zr. We note that since all the RHEA systems considered here are BCC, it was deemed adequate to establish a standardized reference axis by simply aligning the lattice vectors of each system with the DFT simulation reference axis.

3.3. Principal component analysis

The two-point spatial correlations of the charge density fields are projected onto a low-dimensional feature space using PCA. The individual and cumulative variance in the dataset captured by the first 4 PCs is shown in Fig. 3. We find that over 98% of the variance in the dataset is captured by the first PC and 99.8% is captured by the first 3 PCs. Insight into salient spatial features captured by each PC can be obtained from its corresponding basis vector. The ensemble-averaged two-point spatial correlation function, $\bar{f}$, and the basis vectors of the first 3 PCs, φ_i, are shown in Fig. 4. As expected, the spatial features in the ensemble-averaged two-point spatial correlation, $\bar{f}$, are representative of a BCC crystal subjected to some quantity of disorder resulting from the varying valency of the different chemical species, the inherent disorder in an SQS, and the lattice distortions present in the ground-state atomic structure. The first PC basis, φ_1, appears to suggest that the first PC score is capturing mainly the variations in the average charge density, $f(\mathbf{0})$, and certain aspects of the disorder in the different atomic structures considered. The second and higher PC basis appear to systematically capture other salient aspects of the variations in the disorder in the different atomic structures considered. A complete quantitative interpretation of the PC basis is currently not possible.

As previously mentioned, one advantage of PCA is that any point belonging to the convex hull of the PC space representation of the dataset corresponds to a valid atomic structure, thereby providing an opportunity for exploring new atomic structures exhibiting desired or tailored properties. This can be seen in the projections of the dataset onto its first 2 PCs shown in Fig. 5(a) and 5(b). In Fig. 5(a), the dataset is colored by the unique chemical species present in the atomic systems. We find that the projection of the dataset onto its first 3 PCs closely approximates a tetrahedron with vertices, edges, faces, and interior corresponding to the elemental, binary, ternary, and quaternary atomic systems, respectively. In Fig. 5(b), the dataset is colored by the DFT-computed bulk modulus. Together, Fig. 5(a) and 5(b) provide a visualization of the relationship between chemical composition and bulk modulus. By extending this from a single property to a set of multiple properties (which may be computed using either physics or a structure–property relationship), one can identify a subset of the convex hull of the dataset in PC space which optimizes the desired set of properties and reverse engineer the corresponding material systems. A key feature of this approach for materials design is that rather than inversely modeling an exact charge density field or atomic positions (which material designers cannot fully control), points in PC space that are predicted to exhibit one or more desired properties are correlated to practically useful and controllable design parameters such as chemical composition, crystal structure, etc. Although a point in the convex hull of the PC space representation of the dataset corresponds to a theoretically feasible charge density field, there is no guarantee that the corresponding hypothetical material structure is stable. This can be addressed in the established VASt framework by treating stability as another physical property of interest to be modeled/predicted.

Table 1
The DFT-computed bulk moduli of the BCC Nb, Ti, and Zr elemental systems compared with previously reported DFT-computed and experimentally measured values. All values are in GPa.

	DFT – This Study	DFT – Previous Studies	Experimental
Nb	165.3	171.4 [94] 173.0 [95]	170.0 (273 K) [91]
Ti	107.7	107.9 [94] 108.0 [95]	118.0 (1293 K) [92]
Zr	85.4	91.0 [94] 91.0 [95]	66.0 (973 K) [93]

3.4. Results

For this study, we use the first 3 PCs to represent the atomic structure as the input to the GPR model. The GPR model is trained using the Bayesian experiment design active learning process described in section 2.5 Here, we take the initial training dataset to be the set of 4 elemental atomic systems and, as previously mentioned in section 2.5, define the expected information gain for a candidate sample as $I(\boldsymbol{x}) = |\sigma(\boldsymbol{x})/\mu(\boldsymbol{x})|$, where $\boldsymbol{x}$ is the feature vector representation of the candidate sample and $\mu(\boldsymbol{x})$ and $\sigma(\boldsymbol{x})$ are the mean and standard deviation of the corresponding GPR predicted distribution for the output, respectively. The convergence of the GPR model accuracy with respect to the total number of training samples using active selection (i.e., Bayesian experiment design) is shown in Fig. 6(a). For each iteration, the validation dataset is taken to be the set of all samples not currently in the training dataset, and the model accuracy is evaluated by computing the mean absolute percentage error (MAPE) of the VASt bulk modulus predictions for the validation samples. We find that with active selection, the VASt structure–property relationship converges to a MAPE of $< 3\%$ and $< 2\%$ with just 10 and 26 total training samples (including the 4 elemental

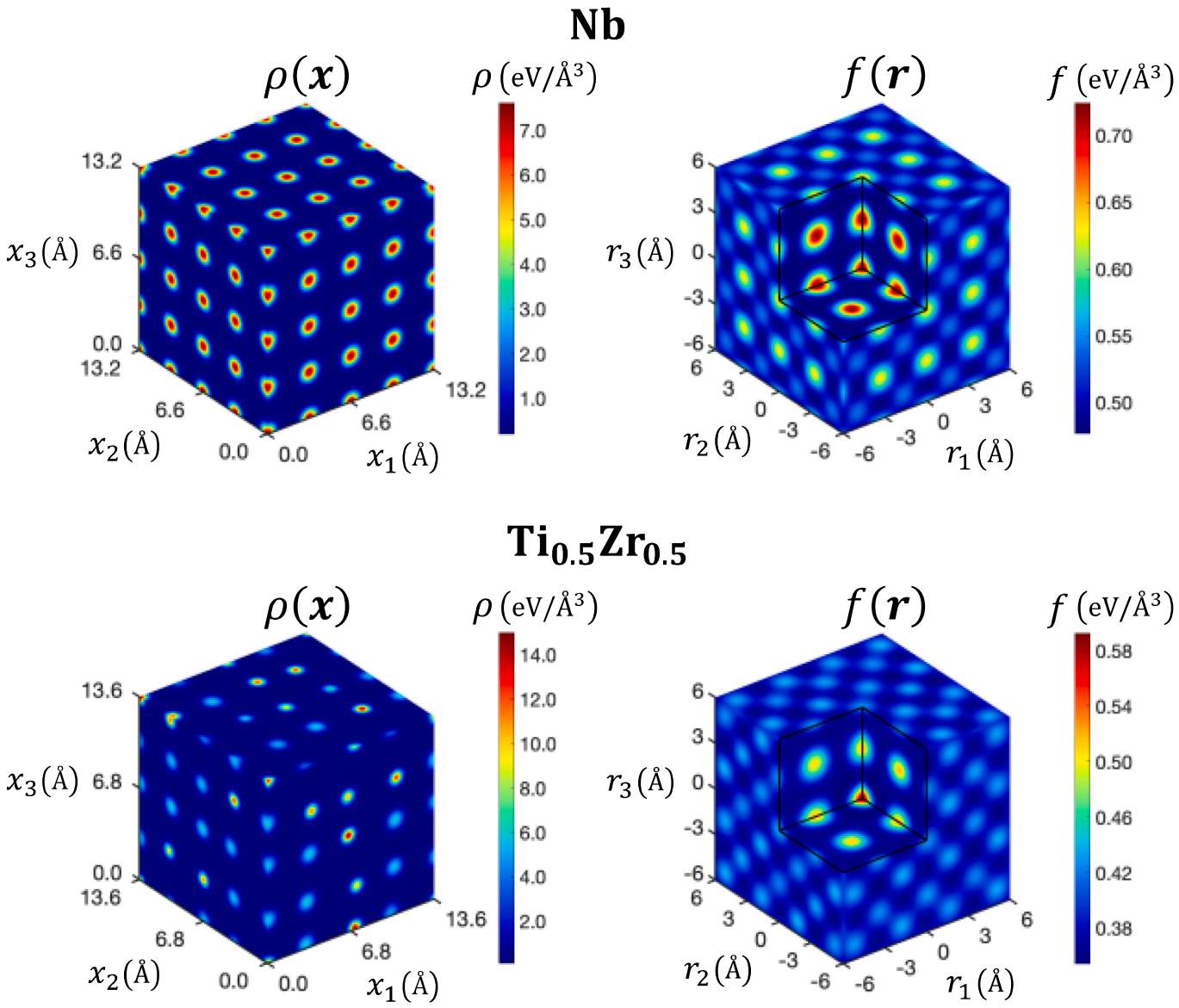

Fig. 2. The first-principles computed charge density field and corresponding standardized two-point spatial correlations of (Top) BCC Nb and (Bottom) the equiatomic TiZr BCC SQS. Note that the scale of the color bars is different for each system.

atomic systems in the initial training dataset), respectively. For comparison, we also show the convergence of the GPR model accuracy with respect to the total number of training samples using random selection. We find that with random selection, the VASt structure–property relationship converges to a MAPE of $< 3\%$ and $< 2\%$ with 20 and 31 total training samples, respectively. Although the VASt structure–property relationships trained with active or random selection converge to a MAPE of $< 2\%$ with relatively small training datasets, it is apparent that active selection with Bayesian experiment design still yields a significant improvement over random selection. In particular, we note that a MAPE of $< 2\%$ is first reached with just 13 total training samples using active selection compared to 31 with random selection. Although the MAPE with active selection does not fully converge to $< 2\%$ at 13 samples, it remains below 3% until finally converging to $< 2\%$ with 26 training samples. Furthermore, we note that the noise in MAPE seen with active selection is equally present in random selection but has been smoothed in Fig. 6(a) by averaging over multiple trials. Therefore, we conclude that the noise is likely due to the overall small number of training samples. This suggests that active selection could present an even more significant improvement over random selection in cases where more training samples are needed and the effects of noise become insignificant.

In a similar study, Tran et al. used multi-fidelity (MF) ML to predict the bulk modulus of 3-component AlNbTi alloys as a function of chemical composition [76]. They computed the low-fidelity bulk moduli using a separately trained Spectral Neighborhood Analysis Potential (SNAP) ML-based interatomic potential and predicted the corresponding high-fidelity (i.e., DFT level of accuracy) bulk moduli using a MF Gaussian-process (MFGP). The performance of the VASt structure–property relationships trained using both active and random selection are compared to that of the MFGP in Fig. 6(b). We find that the R^2 of the converged VASt structure–property relationship for 4-component RHEAs ($R^2 = 0.9811$) significantly outperforms that of the fully-trained SNAP ML-based interatomic potential for 3-component alloys ($R^2 = 0.7122$) and is nearly the same as that of the MFGP for 3-component alloys ($R^2 = 0.9954$). Furthermore, we find that the R^2 of the VASt structure–property relationship for 4-component RHEAs converges with fewer training samples than the MFGP for 3-component alloys. For example, with 19 training samples, the R^2 of the VASt structure–property relationships trained using active and random selection are

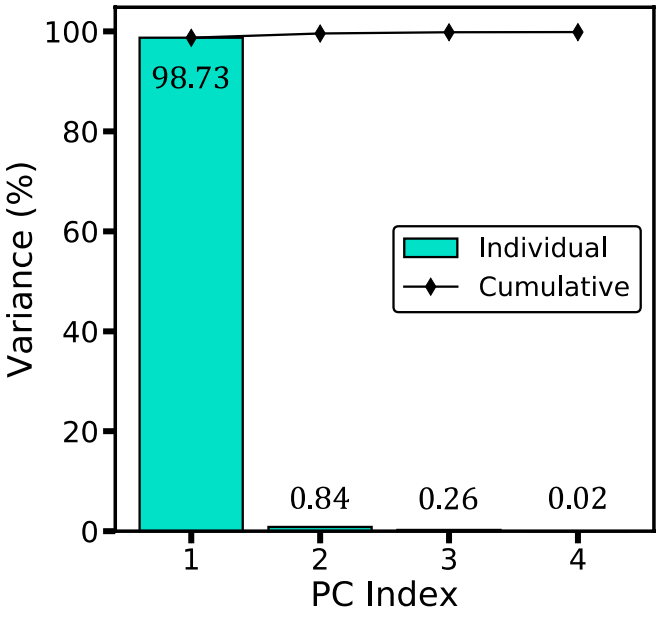

Fig. 3. The individual and cumulative variance in the dataset captured by the first 4 PCs.

$R^2 = 0.9690$ and $R^2 = 0.8507$, respectively, compared to $R^2 = 0.7235$ for the MFGP.

The bulk moduli of the test samples predicted using the VASt structure–property relationship trained on 26 actively selected samples (i.e., converged to a MAPE $< 2\%$) are plotted against their DFT-computed bulk moduli in Fig. 7(a). The mean absolute error and root mean square error of the predictions are 2.0 GPa and 2.7 GPa, respectively. To quantify the reliability of the VASt structure–property relationship, we elucidate the accuracy of its uncertainty quantification capabilities in Fig. 7(b). We first plot the standard deviations of the VASt predictions as a cumulative histogram. We find that over 98% of the predicted standard deviations are less than 3.2 GPa, with the largest being 3.78 GPa. Furthermore, the predicted standard deviation is on average less than 2.6% of the predicted bulk modulus value, suggesting that the model has high confidence in its predictions. We also show in Fig. 7(b) the percentage of predictive distributions that contain the corresponding DFT-computed bulk modulus within 1 and 2 of the predicted standard deviations (which correspond to the 68% and 95% confidence intervals (CIs), respectively). We find that 77.3% and 95.7% of the predicted distributions contain the corresponding DFT-computed bulk modulus within 1 and 2 standard deviations, respectively, which verifies that the VASt structure–property relationship is not only accurate, but also reliable.

Although bulk modulus is a simpler property, the results of this study suggest that it may be possible to develop accurate VASt structure–property relationships for complex physical properties that are computationally prohibitive to compute from first-principles for more than a few samples. The VASt framework can be readily utilized for any such property or properties of interest without modification. Even for bulk modulus, the VASt framework offers a significant computational savings over DFT when the material design space of interest is large since each DFT-computed bulk modulus will require computing the ground-state energy for a minimum of 5 volumes (and often more than 5 are used) [96,97]. This computational savings is particularly significant when the atomic systems of interest are large and disordered, as is the case with RHEAs. We also note that although the atomic systems of interest in this study are all crystalline, the underling physics and mathematics on which the VASt framework is developed (i.e., charge density fields, two-point spatial correlations, PCA, etc.) are general and readily applicable to atomic systems of any type (i.e., amorphous systems, surfaces, individual molecules, molecular crystals, etc.).

4. Conclusions

We have presented a comprehensive framework for developing reduced-order homogenization models for effective properties of atomic

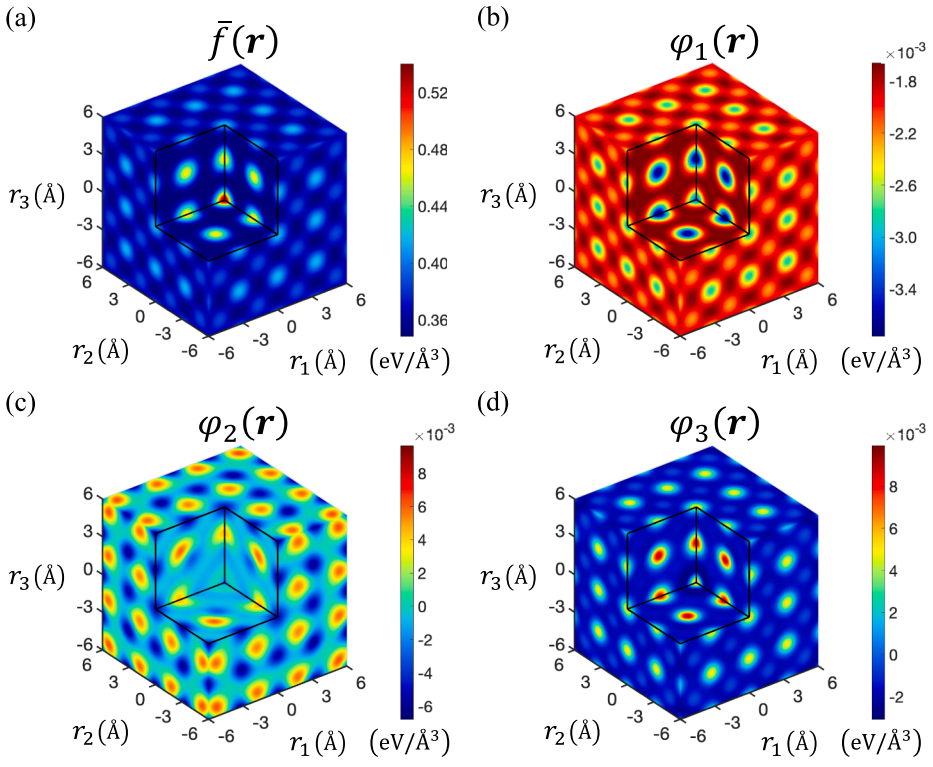

Fig. 4. (a) The ensemble averaged two-point spatial correlation function. (b-d) The basis vectors of the PC 1, PC 2, and PC 3, respectively.

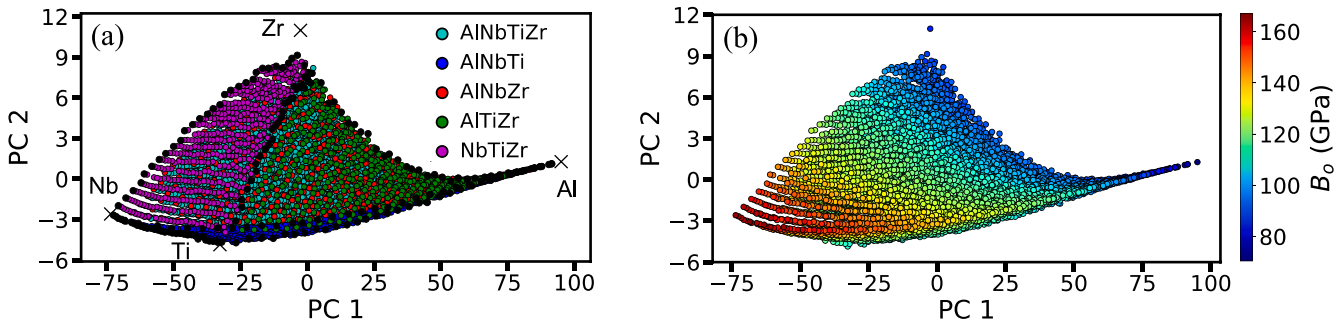

Fig. 5. The salient low-dimensional representation of the AlNbTiZr RHEA dataset visualized as the projection of the two-point spatial correlations of the charge density fields onto the first 2 PCs, colored by (a) the unique chemical species present in the atomic systems, and (b) the DFT-computed bulk modulus, B_o. In (a) the Xs correspond to the pure elemental compositions as labeled.

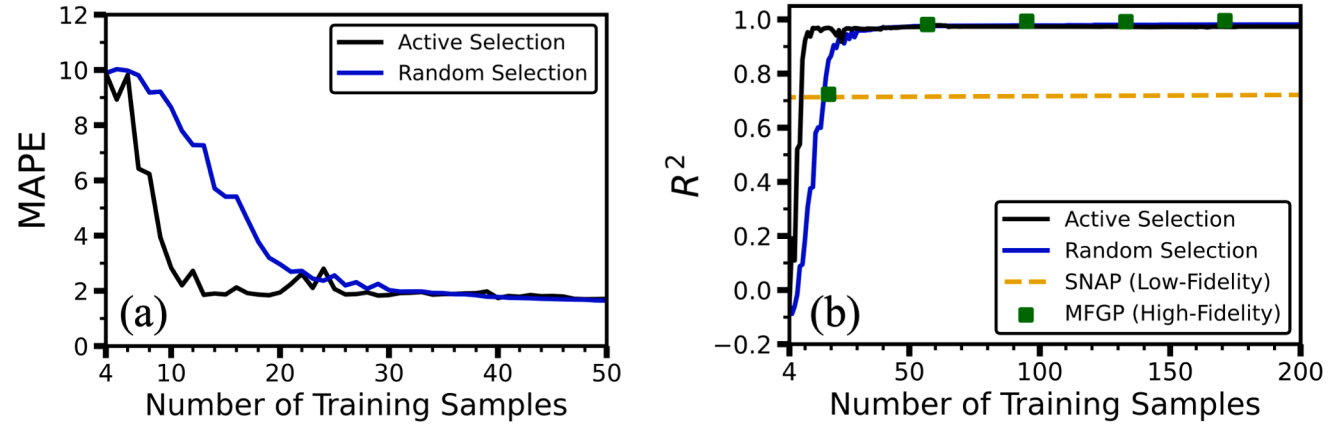

Fig. 6. (a) The convergence of the VASt structure–property relationship MAPE with respect to the number of training samples using active (i.e., Bayesian experiment design) and random selection. The initial training dataset consists of only the 4 elemental systems. (b) The convergence of R^2 with respect to the number of training samples for the VASt structure–property relationship developed for 4-component RHEAs compared to that of the MFGP developed for 3-component alloys [76]. The yellow horizontal dashed line denotes the R^2 of the low-fidelity bulk moduli predicted using a fully trained SNAP ML-based interatomic potential for 3-component alloys [76]. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

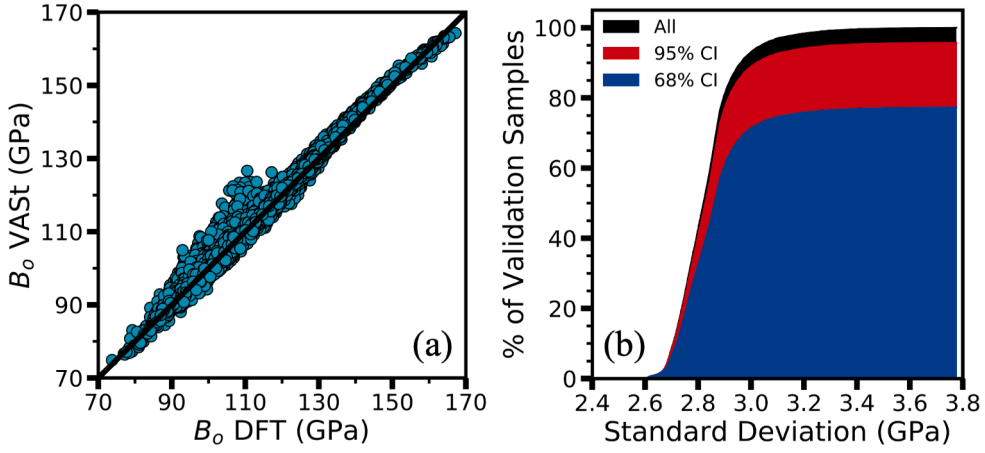

Fig. 7. (a) A parity plot of the DFT-computed and VASt-predicted bulk moduli, B_o, of the validation samples using the first converged GPR model from the Bayesian experiment design active learning procedure. (b) The reliability and accuracy of the VASt structure–property relationship uncertainty quantification. A cumulative histogram of the predicted standard deviations of the validation samples is shown in black and the percentage of predictive distributions that contain the corresponding DFT-computed bulk modulus within 1 or 2 of the predicted standard deviations – corresponding to the 68% and 95% CIs, respectively – are shown in blue and red, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

systems. In this framework, the atomic structure is defined by the charge density field, eliminating the need for additional *ad hoc* feature engineering of atomic attributes. The spatial features underlying the charge density field are then statistically quantified in the form of two-point spatial correlations and projected to a low-dimensional feature space using PCA. The low-dimensional PC representation then serves as the input to a GPR model. The uncertainty quantification inherent in GPR is utilized to deploy a Bayesian experiment design procedure for active learning to minimize the number of training samples required for achieving sufficiently accurate ML models. Finally, we demonstrated the VASt ML framework by developing a structure–property relationship to predict the bulk modulus of AlNbTiZr RHEAs for all possible chemical

compositions. We found that the VASt structure–property relationship reaches an error of $< 2\%$ with fewer than 30 total training samples and that the uncertainty of these predictions is well-captured.

CRediT authorship contribution statement

Matthew C. Barry: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Jacob R. Gissinger:** Conceptualization, Investigation, Writing – review & editing. **Michael Chandross:** Conceptualization, Investigation, Writing – review & editing, Supervision, Funding acquisition. **Kristopher E. Wise:** Conceptualization, Investigation, Writing – review & editing, Supervision, Funding acquisition. **Surya R. Kalidindi:** Conceptualization, Methodology, Validation, Resources, Writing – review & editing, Supervision, Funding acquisition. **Satish Kumar:** Conceptualization, Methodology, Validation, Resources, Writing – review & editing, Supervision, Project administration, Funding acquisition.

Declaration of Competing Interest

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

Data availability

Data will be made available on request.

Acknowledgements:

This work was supported by a NASA Space Technology Research Fellowship. This work used computer resources supported by the National Science Foundation under Grant No. 1828187. Work at Sandia was funded by the Laboratory Directed Research and Development program at Sandia National Laboratories, a multi-mission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC., a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the views of the U.S. Department of Energy or the United States Government.

Appendix A

Given an atomic system with side length L (for simplicity we will assume that the atomic systems are cubes) whose first-principles computed charge density field is defined on a discrete spatial domain with voxel indices, $\boldsymbol{S}$, and voxel length, l, we define the standardized discrete spatial domain, $\boldsymbol{\Sigma}$, as the closest approximation whose side length, Λ, is an integer multiple of the standardized voxel length, λ (see Fig. A.1(a)). The distance $|L-\Lambda|$ is minimized by defining $\Lambda = \lambda \cdot \lfloor \frac{L}{\lambda} + 0.5 \rfloor$, where $\lfloor \bullet \rfloor$ denotes the floor function. However, this definition allows for $\Lambda < L$, and thus has the potential to "destroy" information near the boundary of the original, nonstandardized discrete spatial domain. Therefore, we instead define $\Lambda = \lambda \cdot \lceil \frac{L}{\lambda} \rceil$, where $\lceil \bullet \rceil$ is the ceiling function. Although this definition of Λ does not minimize the distance $|L-\Lambda|$ for $\Lambda > L + 0.5\lambda$, it guarantees that $\Lambda > L$ and therefore will never destroy information.

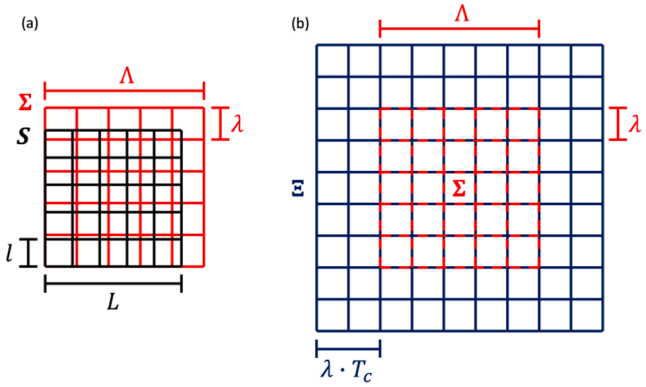

Fig. A1. (a) A nonstandardized discrete spatial domain (black) with voxel indices, $\boldsymbol{S}$, side length, L, and voxel length, l, and the corresponding standardized discrete spatial domain (red) with voxel indices, $\boldsymbol{\Sigma}$, side length, Λ, and standardized voxel length, λ. (b) The standardized discrete spatial domain (red) and corresponding extended standardized discrete spatial domain (blue) with voxel indices, $\boldsymbol{\Xi}$, and $T_c = \lceil R_c/\lambda \rceil$. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Given Λ, we define the standardized voxel indices as the set $\boldsymbol{\Sigma} = \{(\sigma_1, \sigma_2, \sigma_3) | \sigma_i \in [0, 1, \cdots, \Sigma - 1], i = 1, 2, 3\}$, where $\Sigma = \Lambda/\lambda$ is the number of standard voxels that discretize each dimension of the standardized discrete spatial domain. The charge density field can then be approximated on the standardized discrete spatial domain using an interpolation function. Since the atomic systems considered in this work are periodic, a natural choice is trigonometric interpolation [98,99]. In this approach, a continuous, periodic (with period equal to the length, L, of the original, nonstandardized spatial domain) approximation of the first-principles computed charge density field is given by the trigonometric polynomial

$$\widetilde{\rho}(\boldsymbol{x}) = \frac{1}{|\boldsymbol{S}|} \sum_{k_1=-\frac{S}{2}+1}^{\frac{S}{2}} \sum_{k_2=-\frac{S}{2}+1}^{\frac{S}{2}} \sum_{k_3=-\frac{S}{2}+1}^{\frac{S}{2}} \widehat{\rho}_{k_1 k_2 k_3} e^{\frac{2\pi i k_1 x_1}{L}} e^{\frac{2\pi i k_2 x_2}{L}} e^{\frac{2\pi i k_3 x_3}{L}}, \tag{A1}$$

where $\widehat{\rho}_{k_1 k_2 k_3}$ are the discrete Fourier coefficients given by the discrete Fourier transform of the first-principles computed discrete charge density field, which is defined as

$$\widehat{\rho}_{k_1 k_2 k_3} = \sum_{s_1=0}^{S-1} \sum_{s_2=0}^{S-1} \sum_{s_3=0}^{S-1} \rho_{s_1 s_2 s_3} e^{-\frac{2\pi i s_1 k_1}{S}} e^{-\frac{2\pi i s_2 k_2}{S}} e^{-\frac{2\pi i s_3 k_3}{S}}. \tag{A2}$$

Since the approximated charge density field given by Eq. (A.1) is a continuous, periodic function, the charge density at any standardized voxel,

$\boldsymbol{\sigma} \in \boldsymbol{\Sigma}$, can be approximated by evaluating $\widetilde{\rho}(\boldsymbol{x})$ for $\boldsymbol{x} = \boldsymbol{\sigma}\cdot\lambda$. We note that the summations in Eq. (A.1) can be computationally expensive to evaluate for large S. However, the magnitude of the discrete Fourier coefficients, $\widehat{\rho}_{k_1k_2k_3}$, decays rapidly and therefore, a sufficiently accurate approximation of Eq. (A.1) can be obtained by evaluating the summation for a small subset of the discrete Fourier coefficients with the largest magnitudes. We found that a sufficiently accurate approximation of Eq. (A.1) is given by summing over the N largest magnitude discrete Fourier coefficients such that the ratio between the first (i.e., largest) and N^{th} coefficients is $\geq 10^3$. Finally, we note that when $\boldsymbol{\Sigma} = \boldsymbol{S}$, Eq. (A.1) is simply the inverse discrete Fourier transform of $\widehat{\rho}$ and $\widetilde{\rho}[\boldsymbol{\sigma}] = \rho[\boldsymbol{s}]$.

Although the approximated charge density field will no longer be periodic on the standardized discrete spatial domain, one can still compute the two-point spatial correlation function by suitably extending the standardized discrete spatial domain on which the charge density field is approximated. Typically, we are only interested in the spatial correlations up to some cutoff distance, $R_c > 0$. In this case, we define the extended standardized discrete spatial domain by the set $\boldsymbol{\Xi} = \{(\xi_1, \xi_2, \xi_3) | \xi_i \in [-T_c, \cdots, 0, \cdots, \Sigma - 1 + T_c], i = 1, 2, 3\}$, where $T_c = \lceil R_c/\lambda \rceil$ is the minimum number of voxels such that $T_c\lambda \geq R_c$. Thus, $\boldsymbol{\Sigma} \subseteq \boldsymbol{\Xi}$ and the two-point spatial correlation function is given by

$$f[\boldsymbol{t}] = \frac{1}{|\boldsymbol{\Sigma}|}\sum_{\boldsymbol{\xi}\in\boldsymbol{\Xi}} \widetilde{\rho}[\boldsymbol{\xi}]^{1/2}\chi_{\boldsymbol{\Sigma}}(\boldsymbol{\xi})\widetilde{\rho}[\boldsymbol{\xi}+\boldsymbol{t}]^{1/2}, \tag{A3}$$

where $\chi_{\boldsymbol{\Sigma}}(\boldsymbol{\xi})$ is the indicator function. For the standardized discrete spatial domain shown in Fig. A.1(a), the corresponding extended standardized discrete spatial domain is shown in Fig. A.1(b). The product $\widetilde{\rho}[\boldsymbol{\xi}]\chi_{\boldsymbol{\Sigma}}(\boldsymbol{\xi})$ is equivalent to zero-padding a T_c-voxel thick boarder around the charge density field approximated on $\boldsymbol{\Sigma}$. The discrete Fourier transform based formulation of the two-point spatial correlation function is then given by

$$f[\boldsymbol{t}] = \frac{1}{|\boldsymbol{\Sigma}|}\mathscr{F}^{-1}\left[\mathscr{F}\left(\widetilde{\rho}[\boldsymbol{\xi}]^{1/2}\chi_{\boldsymbol{\Sigma}}(\boldsymbol{\xi})\right)^{*}\mathscr{F}\left(\widetilde{\rho}[\boldsymbol{\xi}]^{1/2}\right)\right]. \tag{A4}$$

Hence, one can still take advantage of computationally efficient discrete Fourier transforms to compute the two-point spatial correlations.

We note that although the protocol outlined above assumes a cubic and periodic atomic system, the underlying physics and mathematics on which the VASt framework is developed (i.e., charge density fields and two-point spatial correlations) are general and readily applicable to atomic systems of any type (i.e., non-orthogonal crystal structures, amorphous systems, surfaces, individual molecules, molecular crystals, etc.). For example, the standardized charge density field of a non-orthogonal periodic crystal structure can be obtained by suitably modifying the trigonometric polynomial given by Eq. (A.1). The standardized charge density field of non-periodic atomic systems such as molecules and surfaces can also be obtained using suitably defined interpolation functions.

References

[1] S.R. Kalidindi, Hierarchical materials informatics: novel analytics for materials data, Elsevier, 2015.
[2] S. Ramakrishna, T.-Y. Zhang, W.-C. Lu, Q. Qian, J.S.C. Low, J.H.R. Yune, D.Z. L. Tan, S. Bressan, S. Sanvito, S.R. Kalidindi, Materials informatics, J. Intell. Manuf. 30 (6) (2019) 2307–2326.
[3] M.C. Payne, M.P. Teter, D.C. Allan, T. Arias, A.J. Joannopoulos, Iterative minimization techniques for ab initio total-energy calculations: molecular dynamics and conjugate gradients, Rev. Mod. Phys. 64 (4) (1992) 1045.
[4] R. Car, M. Parrinello, Unified Approach for Molecular Dynamics and Density-Functional Theory, Phys. Rev. Lett. 55 (22) (1985) 2471–2474.
[5] W. Kohn, L.J. Sham, Self-Consistent Equations Including Exchange and Correlation Effects, Phys. Rev. 140 (4A) (1965) A1133–A1138.
[6] C. Fonseca Guerra, J.G. Snijders, G. te Velde, E.J. Baerends, Towards an order-N DFT method, Theor. Chem. Acc. 99 (6) (1998) 391–403.
[7] M.C. Barry, N. Kumar, S. Kumar, Boltzmann transport equation for thermal transport in electronic materials and devices, Annual Review of Heat Transfer 24 (2022).
[8] M.C. Barry, K.E. Wise, S.R. Kalidindi, S. Kumar, Voxelized atomic structure potentials: predicting atomic forces with the accuracy of quantum mechanics using convolutional neural networks, The Journal of Physical Chemistry Letters 11 (21) (2020) 9093–9099.
[9] A.P. Bartók, R. Kondor, G. Csányi, On representing chemical environments, Phys. Rev. B 87 (18) (2013), 184115.
[10] A.P. Bartók, M.C. Payne, R. Kondor, G. Csányi, Gaussian Approximation Potentials: The Accuracy of Quantum Mechanics, without the Electrons, Phys. Rev. Lett. 104 (13) (2010), 136403.
[11] J. Behler, M. Parrinello, Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces, Phys. Rev. Lett. 98 (14) (2007), 146401.
[12] V. Botu, R. Ramprasad, Learning scheme to predict atomic forces and accelerate materials simulations, Phys. Rev. B 92 (9) (2015), 094306.
[13] Z. Li, J.R. Kermode, A. De Vita, Molecular Dynamics with On-the-Fly Machine Learning of Quantum-Mechanical Forces, Phys. Rev. Lett. 114 (9) (2015), 096405.
[14] N. Lubbers, J.S. Smith, K. Barros, Hierarchical modeling of molecular energies using a deep neural network, J. Chem. Phys. 148 (24) (2018), 241715.
[15] K.T. Schütt, H.E. Sauceda, P.-J. Kindermans, A. Tkatchenko, K.-R. Müller, SchNet – A deep learning architecture for molecules and materials, The Journal of Chemical Physics 148 (24) (2018), 241722.
[16] A.V. Shapeev, Moment tensor potentials: A class of systematically improvable interatomic potentials, Multiscale Model. Simul. 14 (3) (2016) 1153–1173.
[17] J.S. Smith, O. Isayev, A.E. Roitberg, ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost, Chem. Sci. 8 (4) (2017) 3192–3203.
[18] A.P. Thompson, L.P. Swiler, C.R. Trott, S.M. Foiles, G.J. Tucker, Spectral neighbor analysis method for automated generation of quantum-accurate interatomic potentials, J. Comput. Phys. 285 (2015) 316–330.
[19] L. Zhang, J. Han, H. Wang, R. Car, W. e,, Deep Potential Molecular Dynamics: A Scalable Model with the Accuracy of Quantum Mechanics, Phys. Rev. Lett. 120 (14) (2018), 143001.
[20] D.K. Duvenaud, D. Maclaurin, J. Iparraguirre, R. Bombarell, T. Hirzel, A. Aspuru-Guzik, R.P. Adams, Convolutional networks on graphs for learning molecular fingerprints, Adv. Neural Inf. Proces. Syst. 28 (2015).
[21] J. Gilmer, S.S. Schoenholz, P.F. Riley, O. Vinyals, G.E. Dahl, Neural message passing for quantum chemistry, International Conference on Machine Learning, PMLR (2017) 1263–1272.
[22] K. Hansen, F. Biegler, R. Ramakrishnan, W. Pronobis, O.A. Von Lilienfeld, K.-R. Müller, A. Tkatchenko, Machine learning predictions of molecular properties: Accurate many-body potentials and nonlocality in chemical space, The Journal of Physical Chemistry Letters 6 (12) (2015) 2326–2331.
[23] J. Lee, A. Seko, K. Shitara, K. Nakayama, I. Tanaka, Prediction model of band gap for inorganic compounds by combination of density functional theory calculations and machine learning techniques, Phys. Rev. B 93 (11) (2016), 115104.
[24] M. Rupp, A. Tkatchenko, K.-R. Müller, O.A. von Lilienfeld, Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning, Phys. Rev. Lett. 108 (5) (2012), 058301.
[25] L. Ward, A. Agrawal, A. Choudhary, C. Wolverton, A general-purpose machine learning framework for predicting properties of inorganic materials, npj Comput. Mater. 2 (1) (2016) 16028.
[26] C. Ben Mahmoud, A. Anelli, G. Csányi, M. Ceriotti, Learning the electronic density of states in condensed matter, Phys. Rev. B 102 (23) (2020), 235130.
[27] K. Choudhary, B. DeCost, Atomistic Line Graph Neural Network for improved materials property predictions, npj Comput. Mater. 7 (1) (2021) 185.
[28] S. Kong, F. Ricci, D. Guevarra, J.B. Neaton, C.P. Gomes, J.M. Gregoire, Density of states prediction for materials discovery via contrastive learning from probabilistic embeddings, Nat. Commun. 13 (1) (2022) 949.
[29] P.-P. De Breuck, G. Hautier, G.-M. Rignanese, Materials property prediction for limited datasets enabled by feature selection and joint learning with MODNet, npj Comput. Mater. 7 (1) (2021) 83.
[30] R. Ouyang, S. Curtarolo, E. Ahmetcik, M. Scheffler, L.M. Ghiringhelli, SISSO: A compressed-sensing method for identifying the best low-dimensional descriptor in an immensity of offered candidates, Physical Review Materials 2 (8) (2018), 083802.
[31] Y. Shao, L. Knijff, F.M. Dietrich, K. Hermansson, C. Zhang, Modelling Bulk Electrolytes and Electrolyte Interfaces with Atomistic Machine Learning, Batteries & Supercaps 4 (4) (2021) 585–595.
[32] M.J. Willatt, F. Musil, M. Ceriotti, Feature optimization for atomistic machine learning yields a data-driven construction of the periodic table of the elements, PCCP 20 (47) (2018) 29661–29668.
[33] G. Pilania, C. Wang, X. Jiang, S. Rajasekaran, R. Ramprasad, Accelerating materials property predictions using machine learning, Sci. Rep. 3 (1) (2013) 2810.

[34] A.D. Casey, S.F. Son, I. Bilionis, B.C. Barnes, Prediction of energetic material properties from electronic structure using 3D convolutional neural networks, J. Chem. Inf. Model. 60 (10) (2020) 4457–4473.
[35] Y. Zhao, K. Yuan, Y. Liu, S.-Y. Louis, M. Hu, J. Hu, Predicting elastic properties of materials from electronic charge density using 3D deep convolutional neural networks, J. Phys. Chem. C 124 (31) (2020) 17262–17273.
[36] X. Lei, A.J. Medford, A Universal Framework for Featurization of Atomistic Systems, The Journal of Physical Chemistry Letters 13 (34) (2022) 7911–7919.
[37] J.A. Gomberg, A.J. Medford, S.R. Kalidindi, Extracting knowledge from molecular mechanics simulations of grain boundaries using machine learning, Acta Mater. 133 (2017) 100–108.
[38] S.R. Kalidindi, J.A. Gomberg, Z.T. Trautt, C.A. Becker, Application of data science tools to quantify and distinguish between structures and models in molecular dynamics datasets, Nanotechnology 26 (34) (2015), 344006.
[39] P.R. Kaundinya, K. Choudhary, S.R. Kalidindi, Machine learning approaches for feature engineering of the crystal structure: Application to the prediction of the formation energy of cubic compounds, Physical Review Materials 5 (6) (2021), 063802.
[40] P.R. Kaundinya, K. Choudhary, S.R. Kalidindi, Prediction of the electron density of states for crystalline compounds with Atomistic Line Graph Neural Networks (ALIGNN), arXiv preprint arXiv:2201.08348 (2022).
[41] K. Choudhary, K.F. Garrity, A.C. Reid, B. DeCost, A.J. Biacchi, A.R. Hight Walker, Z. Trautt, J. Hattrick-Simpers, A.G. Kusne, A. Centrone, The joint automated repository for various integrated simulations (JARVIS) for data-driven materials design, npj Comput. Mater. 6 (1) (2020) 1–13.
[42] B. Kolb, L.C. Lentz, A.M. Kolpak, Discovering charge density functionals and structure-property relationships with PROPhet: A general framework for coupling machine learning and first-principles methods, Sci. Rep. 7 (1) (2017) 1–9.
[43] C. Kim, A. Chandrasekaran, T.D. Huan, D. Das, R. Ramprasad, Polymer Genome: A Data-Powered Polymer Informatics Platform for Property Predictions, J. Phys. Chem. C 122 (31) (2018) 17575–17585.
[44] J.E. Saal, S. Kirklin, M. Aykol, B. Meredig, C. Wolverton, Materials Design and Discovery with High-Throughput Density Functional Theory: The Open Quantum Materials Database (OQMD), JOM 65 (11) (2013) 1501–1509.
[45] W.M. Garrison, A.L. Wojcieszynski, A discussion of the effect of inclusion volume fraction on the toughness of steel, Mater. Sci. Eng. A 464 (1) (2007) 321–329.
[46] L. Holzer, B. Münch, B. Iwanschitz, M. Cantoni, T. Hocker, T. Graule, Quantitative relationships between composition, particle size, triple phase boundary length and surface area in nickel-cermet anodes for Solid Oxide Fuel Cells, J. Power Sources 196 (17) (2011) 7076–7089.
[47] J. van de Lagemaat, K.D. Benkstein, A.J. Frank, Relation between Particle Coordination Number and Porosity in Nanoparticle Films: Implications to Dye-Sensitized Solar Cells, J. Phys. Chem. B 105 (50) (2001) 12433–12436.
[48] D.M. Dimiduk, P.M. Hazzledine, T.A. Parthasarathy, M.G. Mendiratta, S. Seshagiri, The role of grain size and selected microstructural parameters in strengthening fully lamellar TiAl alloys, Metall. Mater. Trans. A 29 (1) (1998) 37–47.
[49] P. Debye, H.R.A. Jr, H. Brumberger, Scattering by an Inhomogeneous Solid. II. The Correlation Function and Its Application, J. Appl. Phys. 28 (6) (1957) 679–683.
[50] E. Kröner, Statistical modelling, in: Modelling small deformations of polycrystals, Springer, 1986, pp. 229–291.
[51] G. Li, Y. Liang, Z. Zhu, C. Liu, Microstructural analysis of the radial distribution function for liquid and amorphous Al, J. Phys. Condens. Matter 15 (14) (2003) 2259.
[52] J. Schröder, D. Balzani, D. Brands, Approximation of random microstructures by periodic statistically similar representative volume elements based on lineal-path functions, Arch. Appl. Mech. 81 (7) (2011) 975–997.
[53] H. Singh, A.M. Gokhale, S.I. Lieberman, S. Tamirisakandala, Image based computations of lineal path probability distributions for microstructure representation, Mater. Sci. Eng. A 474 (1) (2008) 104–111.
[54] S. Torquato, H.W. Haslach Jr, Random Heterogeneous Materials: Microstructure and Macroscopic Properties, Appl. Mech. Rev. 55 (4) (2002) B62–B63.
[55] C.L.Y. Yeong, S. Torquato, Reconstructing random media, Phys. Rev. E 57 (1) (1998) 495–506.
[56] S.R. Kalidindi, A Bayesian framework for materials knowledge systems, MRS Commun. 9 (2) (2019) 518–531.
[57] S.R. Kalidindi, Feature engineering of material structure for AI-based materials knowledge systems, J. Appl. Phys. 128 (4) (2020), 041103.
[58] S.R. Kalidindi, A. Khosravani, B. Yucel, A. Shanker, A.L. Blekh, Data infrastructure elements in support of accelerated materials innovation: ELA, PyMKS, and MATIN, Integrating Materials and Manufacturing Innovation 8 (4) (2019) 441–454.
[59] S.R. Kalidindi, S.R. Niezgoda, A.A. Salem, Microstructure informatics using higher-order statistics and efficient data-mining protocols, JOM 63 (4) (2011) 34–41.
[60] B.L. Adams, S. Kalidindi, D.T. Fullwood, D. Fullwood, Microstructure sensitive design for performance optimization, Butterworth-Heinemann2012.
[61] S. Niezgoda, D. Fullwood, S. Kalidindi, Delineation of the space of 2-point correlations in a composite material system, Acta Mater. 56 (18) (2008) 5285–5292.
[62] D. Fullwood, S. Kalidindi, S. Niezgoda, A. Fast, N. Hampson, Gradient-based microstructure reconstructions from distributions using fast Fourier transforms, Mater. Sci. Eng. A 494 (1–2) (2008) 68–72.
[63] K. Pearson, Liii., On lines and planes of closest fit to systems of points in space, The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science 2 (11) (1901) 559–572.
[64] I. Goodfellow, Y. Bengio, A. Courville, Deep learning, MIT press2016.
[65] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proceedings of the IEEE 86(11) (1998) 2278-2324.
[66] A.R. Castillo, V.R. Joseph, S.R. Kalidindi, Bayesian Sequential Design of Experiments for Extraction of Single-Crystal Material Properties from Spherical Indentation Measurements on Polycrystalline Samples, JOM 71 (8) (2019) 2671–2679.
[67] A.R. Castillo, A. Venkatraman, S.R. Kalidindi, Mechanical Responses of Primary-α Ti Grains in Polycrystalline Samples: Part II—Bayesian Estimation of Crystal-Level Elastic-Plastic Mechanical Properties from Spherical Indentation Measurements, Integrating Materials and Manufacturing Innovation 10 (1) (2021) 99–114.
[68] A. Cecen, Y.C. Yabansu, S.R. Kalidindi, A new framework for rotationally invariant two-point spatial correlations in microstructure datasets, Acta Mater. 158 (2018) 53–64.
[69] A. Cecen, T. Fast, S.R. Kalidindi, Versatile algorithms for the computation of 2-point spatial correlations in quantifying material structure, Integrating Materials and Manufacturing Innovation 5 (1) (2016) 1–15.
[70] C.E. Rasmussen, Gaussian processes in machine learning, in: Summer School on Machine Learning, Springer, 2003, pp. 63–71.
[71] X. Huan, Y.M. Marzouk, Simulation-based optimal Bayesian experimental design for nonlinear systems, J. Comput. Phys. 232 (1) (2013) 288–317.
[72] D.V. Lindley, On a measure of the information provided by an experiment, Ann. Math. Stat. 27 (4) (1956) 986–1005.
[73] M.A. Melia, S.R. Whetten, R. Puckett, M. Jones, M.J. Heiden, N. Argibay, A. B. Kustas, High-throughput additive manufacturing and characterization of refractory high entropy alloys, Appl. Mater. Today 19 (2020), 100560.
[74] Y. Ikeda, B. Grabowski, F. Körmann, Ab initio phase stabilities and mechanical properties of multicomponent alloys: A comprehensive review for high entropy alloys and compositionally complex alloys, Mater Charact 147 (2019) 464–511.
[75] D.B. Miracle, O.N. Senkov, A critical review of high entropy alloys and related concepts, Acta Materialia 122 (2017) 448–511.
[76] A. Tran, J. Tranchida, T. Wildey, A.P. Thompson, Multi-fidelity machine-learning with uncertainty quantification and Bayesian optimization for materials design: Application to ternary random alloys, J. Chem. Phys. 153 (7) (2020), 074705.
[77] X.-G. Li, C. Chen, H. Zheng, Y. Zuo, S.P. Ong, Complex strengthening mechanisms in the NbMoTaW multi-principal element alloy, npj Comput. Mater. 6 (1) (2020) 1–10.
[78] A. Van De Walle, Multicomponent multisublattice alloys, nonconfigurational entropy and other additions to the Alloy Theoretic Automated Toolkit, Calphad 33 (2) (2009) 266–278.
[79] A. Van De Walle, M. Asta, G. Ceder, The alloy theoretic automated toolkit: A user guide, Calphad 26 (4) (2002) 539–553.
[80] A. Van de Walle, P. Tiwary, M. De Jong, D. Olmsted, M. Asta, A. Dick, D. Shin, Y. Wang, L.-Q. Chen, Z.-K. Liu, Efficient stochastic generation of special quasirandom structures, Calphad 42 (2013) 13–18.
[81] A. Zunger, S.H. Wei, L.G. Ferreira, J.E. Bernard, Special quasirandom structures, Phys. Rev. Lett. 65 (3) (1990) 353.
[82] O.N. Senkov, D.B. Miracle, K.J. Chaput, J.-P. Couzinie, Development and exploration of refractory high entropy alloys—A review, J. Mater. Res. 33 (19) (2018) 3092–3128.
[83] Y. Wu, D.L. Irving, Prediction of chemical ordering in refractory high-entropy superalloys, Appl. Phys. Lett. 119 (11) (2021).
[84] G. Kresse, J. Furthmüller, Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set, Phys. Rev. B 54 (16) (1996) 11169–11186.
[85] G. Kresse, J. Furthmüller, Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set, Comput. Mater. Sci 6 (1) (1996) 15–50.
[86] P.E. Blöchl, Projector augmented-wave method, Phys. Rev. B 50 (24) (1994) 17953–17979.
[87] G. Kresse, D. Joubert, From ultrasoft pseudopotentials to the projector augmented-wave method, Phys. Rev. B 59 (3) (1999) 1758–1775.
[88] J.P. Perdew, K. Burke, M. Ernzerhof, Generalized Gradient Approximation Made Simple, Phys. Rev. Lett. 77 (18) (1996) 3865–3868.
[89] F. Birch, Finite elastic strain of cubic crystals, Phys. Rev. 71 (11) (1947) 809.
[90] F.D. Murnaghan, Finite deformations of an elastic solid, American Journal of Mathematics 59 (2) (1937) 235–260.
[91] G. Simmons, Single crystal elastic constants and calculated aggregate properties, Southern Methodist Univ Dallas Tex, 1965.
[92] W. Petry, A. Heiming, J. Trampenau, M. Alba, C. Herzig, H.R. Schober, G. Vogl, Phonon dispersion of the bcc phase of group-IV metals. I. bcc titanium, Physical Review B 43 (13) (1991) 10933.
[93] Y. Zhao, J. Zhang, C. Pantea, J. Qian, L.L. Daemen, P.A. Rigg, R.S. Hixson, G. T. Gray III, Y. Yang, L. Wang, Thermal equations of state of the α, β, and ω phases of zirconium, Phys. Rev. B 71 (18) (2005), 184119.
[94] L.-Y. Tian, G. Wang, J.S. Harris, D.L. Irving, J. Zhao, L. Vitos, Alloying effect on the elastic properties of refractory high-entropy alloys, Mater. Des. 114 (2017) 243–252.
[95] B. Feng, M. Widom, Elastic stability and lattice distortion of refractory high entropy alloys, Mater. Chem. Phys. 210 (2018) 309–314.
[96] A. Togo, L. Chaput, I. Tanaka, G. Hug, First-principles phonon calculations of thermal expansion in Ti_3SiC_2, Ti_3AlC_2, and Ti_3GeC_2, Phys. Rev. B 81 (17) (2010), 174301.
[97] C. Toher, J.J. Plata, O. Levy, M. de Jong, M. Asta, M.B. Nardelli, S. Curtarolo, High-throughput computational screening of thermal conductivity, Debye temperature,

and Gr\"uneisen parameter using a quasiharmonic Debye model, Physical Review B 90 (17) (2014), 174107.
[98] A. Gupta, S.R. Kalidindi, Addressing biases in spectral databases for increasing accuracy and computational efficiency of crystal plasticity computations, Int. J. Plast 138 (2021), 102945.
[99] M. Zecevic, R.J. McCabe, M. Knezevic, Spectral database solutions to elasto-viscoplasticity within finite elements: Application to a cobalt-based FCC superalloy, Int. J. Plast 70 (2015) 151–165.