Machine learning approaches for feature engineering of the crystal structure: Application to the prediction of the formation energy of cubic compounds

Prathik R. Kaundinya,[1] Kamal Choudhary,[2,3] and Surya R. Kalidindi[1,4,*]

[1]*School of Computational Science and Engineering, Georgia Institute of Technology, Atlanta, Georgia 30332, USA*
[2]*Materials Science and Engineering Division, National Institute of Standards and Technology, Gaithersburg, Maryland 20899, USA*
[3]*Theiss Research, La Jolla, California 92037, USA*
[4]*G.W. Woodruff School of Mechanical Engineering, Georgia Institute of Technology, Atlanta, Georgia 30332, USA*

(Received 3 January 2021; revised 13 May 2021; accepted 20 May 2021; published 2 June 2021)

In this study, we present an approach (along with the needed computational strategies) for efficient and scalable feature engineering of the crystal structure in compounds of different chemical compositions. This approach utilizes a versatile and extensible framework for the quantification of a three-dimensional voxelized crystal structure in the form of 2-point spatial correlations of multiple atomic attributes and performs principal component analysis to extract the low-dimensional features that could be used to build surrogate models for material properties of interest. An application of the proposed feature engineering framework is demonstrated on a case study involving the prediction of the formation energies of crystalline compounds using two vastly different surrogate model building strategies; local Gaussian process regression and neural networks. Specifically, it is shown that the top 25 features (i.e., principal component scores) identified by the proposed framework serve as good regressors for the formation energy of the crystalline substance for both model building strategies.

DOI: 10.1103/PhysRevMaterials.5.063802

I. INTRODUCTION

Although physics-based modeling approaches such as density-functional theory (DFT) [1,2] offer the preferred avenue for estimating the physical and chemical properties of crystal structures, they are not ideally suited for materials discovery and innovation efforts. Such efforts demand inverse solutions such as finding the molecular chemistry and structure that meets a targeted combination of physical and chemical properties [3]. One of the most practical strategies for addressing the inverse solutions of materials design is to first produce high-fidelity low-computational cost surrogate models trained to the available data (e.g., collections of DFT computations), and then use the surrogate model for addressing the inverse problems of interest. Such a strategy can prove to be highly beneficial in rapid screening of vast design spaces [3–5]. In this context, the emerging toolsets and paradigms in machine learning (ML) offer new opportunities for building the surrogate models needed to accelerate the discovery of new crystal structures. In particular, it is now possible to train these models using collections of publicly accessible DFT data in repositories such as the Materials Project (MP), Open Quantum Materials Database (OQMD), Joint Automated Repository for Various Integrated Simulations (JARVIS), Computational 2D Materials Database, and Quantum Machine [6–16].

ML models have been used in prior work to predict a broad range of material properties based on DFT datasets. The properties modeled have included optical and electronic bandgaps [17,18], formation energy [6,8,19–24], atomization energies [25–29] and polarizability of crystalline compounds [30]. ML models have also been used in prior work for the prediction and classification of crystal compositions and structures [31–35] and in the development of many-body interatomic potentials for atomistic simulations [36–38]. One of the foundational elements of ML common to most model building approaches is feature engineering, which aims to identify a small list of transformed input variables (called features) from the original large list of input variables that potentially could influence the predictions of the output variables (called targets). As one would expect, feature engineering governs the accuracy and utility of the surrogate model. In some of the ML approaches [e.g., convolutional neural networks (CNNs)], feature engineering occurs implicitly in the model training phase. In general, the more implicit one makes the feature engineering task, the more training data would be needed. This is because the model needs to learn the salient features first before learning their quantitative relationships with the targets. In much of the prior work [12,21,25,33], large feature lists have been manually cultivated by researchers based largely on their intuition of the atomic physics. Consequently, these feature lists have included various simplified attributes of the chemical compositions (e.g., atomic fractions) as well as the atomic structure (e.g., bond lengths and bond angles). These approaches face significant challenges. First, the creation of these lists has been pursued largely in an *ad hoc* manner, which inevitably reflects the bias of the individual researchers. As such, these approaches may fail to account for certain potentially salient features controlling the target property. Second, the specification of chemical composition using distinct labels for the different atomic species hinders learning across crystal structures with different elemental compositions. One strategy to address this challenge has been to

[*] Corresponding author: surya.kalidindi@me.gatech.edu

supplement the feature lists with certain atomically weighted physical properties [8,39,40]. Additionally, there have also been various efforts at enhancing the specification of the atomic structure in the feature lists by retaining much of the actual three-dimensional (3D) atomic structure information. For instance, Faber *et al.* [20] utilized an extension of the Coulomb matrix representation to account for the periodicity of the crystal structure. In a different approach, Ward *et al.* [23] represented the crystal structure with a Voronoi decomposition and employed local atomic species-level descriptors between neighboring species in addition to global structure descriptors such as maximum packing efficiency. In similar efforts, Schutt *et al.* and Honrao *et al.* used a feature set containing values of the partial radial distribution function (RDF) between pairs of constituent species in the crystal structure [41,42]. The partial RDF is a measure of the frequency of a pair of species separated by a certain distance within the material volume. However, the partial RDF does not contain directional information about the distribution of the atoms. One approach to overcome this limitation has been to supplement the partial RDF with an angular distribution function that contains information on the distribution of interatomic angles between species [43]. Nevertheless, this approach still utilizes distinct labels for constituent species, and is thus typically useful only for description of crystal structures with few species (such as Al-Ni and Cd-Te systems [43]). Another interesting approach proposed recently relied on the development of graph embeddings that encode species-specific features (such as the periodic table group number, period number, electronegativity) and a limited set of structural features (such as the interatomic distance) as nodes and edges of a multigraph, respectively [24,44,45]. This representation is amenable to usage in graph convolutional neural networks [46,47] that sequentially build up localized representations for each node by iteratively including information from neighboring nodes.

Using the feature engineering schemes described above, current efforts have employed various regression approaches such as kernel ridge regression (KRR) [20,42], support vector machines [42,43], deep neural networks [21], gradient boosted decision trees [8], and graph convolutional neural networks [24,44,45] to build the desired reduced-order models. While the regression-based learning methods are computationally very efficient, they are often prone to over-fitting because they invariably employ a large number of implicit model parameters that need to be trained on a large collection of ground-truth data. The overfitting may be mitigated to a certain extent with the adoption of well-known techniques such as early stopping [48], loss function regularization, and dropout [49,50]. With the adoption of such techniques, the high computational efficiency of neural networks (NNs) makes them an attractive option for building reduced-order models. However, a significant challenge arises from the availability of relatively small training datasets due to the very high cost of DFT computations. The role of the feature engineering procedure is especially critical in such smaller sized datasets because of the need to restrict the number of model fit parameters that can be used in these cases.

ML approaches based on Bayesian inference offer an alternate option that is likely to prove beneficial for building surrogate models from DFT computations. More specifically, Gaussian process regression (GPR) [51] offers a powerful nonparametric Bayesian approach to building surrogate models from small training datasets. There are many potential benefits to the utilization of Gaussian processes in the context of regression. First, they allow a formal treatment of uncertainty in the model predictions. In other words, they do not just estimate the expected values for the model outputs, but also their distributions. Second, the formal treatment of model uncertainty allows the design and implementation of strategies that could potentially reduce the effort spent in the generation of the training data. More specifically, it is possible to formulate and maximize the expected information gain to the surrogate model with the addition of each specific new training data point. This is particularly important to building surrogate models with limited data from the computationally expensive DFT datasets. Furthermore, GPR models are typically more interpretable than NNs.

In this paper, we present a systematic and comprehensive approach to feature engineering of crystal structure that directly addresses the challenges described earlier. In this approach, we systematically and efficiently assemble an extremely large number of spatial correlations (specifically, 2-point correlations) [5,52] that implicitly account for atomic attributes (e.g., Pauling electronegativity, ionization energy, heat of fusion). After processing the large feature list for the complete ensemble of crystal structures of interest, we employ principal component analysis (PCA) to obtain a suitable low-dimensional representation of the feature list. This strategy offers many advantages compared to the approaches used in current literature. First, the approach presented in this work has the potential to systematically generate a very large set of physics-inspired features for the rigorous and comprehensive quantification of the crystal structure. Second, the proposed protocols utilize digital (i.e., voxelized) representations of the 3D crystal structure [53,54] along with compact Fourier representations and associated computational algorithms [e.g., fast Fourier transform (FFT)] for highly efficient computation of the features (i.e., spatial correlations). Third, the encoding of individual chemical species through their individual physical properties enables effective learning across crystal structures of different elemental compositions. Fourth, the use of PCA provides an objective (i.e., data-driven) path for establishing low-dimensional representations of the compound crystal structure that can then be used with a broad variety of model-building approaches. Fifth, the features identified by the proposed framework are independent of the target property predicted by the model.

The primary goal of this paper is to develop and demonstrate a versatile feature engineering methodology for materials problems involving different chemical compositions of compounds and their crystal structures. This methodology is aimed at extracting reliable models from small data sets using physics-inspired feature engineering approaches. We demonstrate the utility of this feature engineering scheme by building predictive models for the crystal formation energy using two drastically different model-building approaches: (i) a localized variant of GPR [55], and (ii) a feed-forward NN. A second goal of this work is to critically compare the relative merits of both model building techniques for addressing materials problems.

II. BACKGROUND

The central goal of the feature engineering step is to establish a compact set of salient inputs (i.e., features) that serve as suitable predictors for the selected targets. In the context of our problem, this would entail establishing a set of salient crystal structure descriptors for capturing high-fidelity reduced-order structure-property (SP) relationships [5,56]. In this effort, the formalism of n-point spatial correlations (also called n-point statistics) [5,57–61] offers a systematic approach to statistical quantification of the underlying morphological patterns in the heterogeneous material internal structure. Several existing statistical atomic physics models such as the Ising model (for predicting ferromagnetism) [62] and the Potts model (a general model of interacting atomic spins) [63] predict the bulk properties of materials as a sum of contributions arising from local structure features that can be easily interpreted as components of the n-point spatial correlations. Spatial correlations are indeed the features dictated by the governing physics in studies of heterogeneous material structure at all length scales, spanning from the atomistic to the mesoscale. However, the comprehensive set of n-point spatial correlations is typically too large and unwieldy to serve directly as inputs for the reduced-order models. As such, there is a critical need for a compact representation of the material structure features that can serve effectively as inputs to produce the desired high-fidelity reduced-order models. In this section, we present a brief overview of the concepts of spatial correlations, salient feature extraction (using PCA), and the strategies for building reduced-order models (i.e., GPR and NNs) used in this work.

A. Spatial correlations

Typically, efficient computations of spatial correlations for the quantification of the heterogeneity of the material structure utilize digital representations (i.e., voxelized representations with suitable assignments of material states to each voxel) and FFT algorithms for computing the convolution operations involved in these computations [5]. Mathematically, these voxelized representations can be expressed as a high-dimensional array m_s^p, whose elements capture the value of a local feature (reflecting the local material state) indexed by p in a voxel indexed by $\boldsymbol{s}$ (defined as vector of integers $\{s_1, s_2, s_3\}$ for 3D material volumes). This array of size $(P \times n(S))$, with P denoting the number of distinct material states assigned to each voxel in the material volume and $n(S)$ denoting the cardinality of the set of spatial voxels S, defines one instantiation of a material structure.

Our interest lies in extracting microstructure statistics from each instantiation (i.e., m_s^p array) that can serve as features in the formulation of surrogate models connecting material structure to the properties/performance characteristics of interest. The most systematic set of such microstructure statistics are provided by the formalism of n-point spatial correlations, which capture the relevant statistics related to the morphological details (i.e., size and shape distributions) of the distinct local states in the material structure. The most basic set of these statistics are the 2-point correlations denoted by $f_{\boldsymbol{r}}^{pq}$, which track how often distinct local states p and q are separated by a discretized vector indexed by $\boldsymbol{r}$. Mathematically, 2-point correlations may be computed as [5,61,64]

$$f_{\boldsymbol{r}}^{pq} = \frac{\sum_{s \in S} m_s^p m_{s+r}^q}{n(S)} . \quad (1)$$

One way to interpret the above definition is to recognize that the numerator denotes the number of successes in locating the local states p and q separated by the vector index $\boldsymbol{r}$, and that the denominator represents the total number of valid trials (for periodic microstructures, this is equal to the number of voxels). It should be noted that the RDF (used commonly in the quantification of the molecular structures) [59,65] is a particular variant of the 2-point correlations described in Eq. (1) in which one does not consider the orientation of the vector, but only its magnitude. Noting that the central operation in Eq. (1) is essentially a discrete cross correlation over the spatial domain of voxels S, the 2-point correlations are efficiently computed by taking advantage of FFT algorithms [5,52]. The computations naturally exploit the periodicity of the crystal structures. There also exist several redundancies that can be leveraged for obtaining a more compact set of spatial correlations. Specifically, it can be seen that Eq. (1) produces P^2 sets of spatial correlations. It has been shown that only P of these are adequate to compute the rest of the spatial correlations [61,66]. As a result, it is often adequate to compute the autocorrelations [which correspond to $n = p$ in Eq. (1)] for the dominant local state in the collected ensemble of material structures and its cross correlations with the rest of the local states.

B. Extraction of salient features

In prior work, our research group has established a generalized framework referred to as materials knowledge systems (MKS) [67–69] that has demonstrated the versatility and utility of assembling a large set of n-point spatial correlations and then applying PCA to establish suitable data-driven low-dimensional representations of the material internal structure. PCA offers a dimensionality reduction technique that performs a rotational transformation of the features into an orthogonal space organized to maximize the capture of variance in the given dataset. The transformed axes (i.e., new basis) are known as PCs, while the transformed coordinates are known as PC scores. However, since PCA maximizes the capture of variance, it tends to emphasize the features that exhibit the largest variance in the dataset. Therefore, one could apply suitable scaling factors to various subgroups of the features before performing the PCA, in order to adjust their roles (i.e., increase or decrease their importance in the PCs). In this work, the features are scaled such that the total variance in the spatial correlations corresponding to each selected pair of local states (i.e., each combination of p and q) is the same. This type of scaling equalizes the importance of each subset of spatial correlations related to a specific combination of local states in the PCA. This protocol produces an ordered list of transformed features such that the selection of each additional transformed feature maximizes the capture of the explained variance in the dataset. A truncated list of these

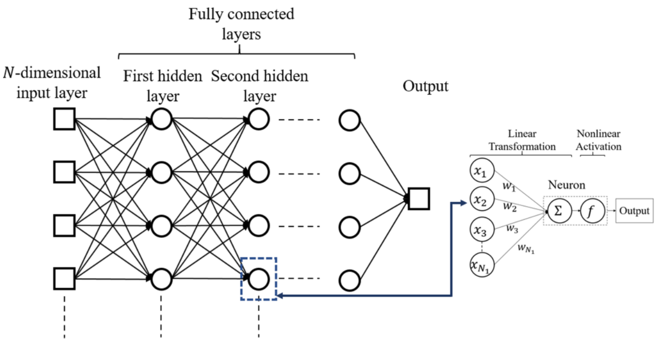

FIG. 1. Schematic description of a fully connected neural network for an N-dimensional input. The callout describes the details of the computations performed at each neuron.

ordered transformed features (i.e., PC scores) serve as suitable low dimensional representations of the material structure.

The benefits of the MKS pipeline have been expounded in prior work [5,56,69–73]. First, this approach yields an objective low-dimensional representation of the material structure that is not influenced by either the materials manufacturing process or property information, thereby providing a consistent representation of the material structure in process-structure-property (PSP) linkages [5]. Second, since PCA is essentially a Fourier representation, it allows for approximate but optimal reconstructions of the material structure statistics (controlled by the truncation levels applied in retaining the PC scores), which in turn can then be used with sophisticated algorithms [74,75] for the statistical reconstruction of the material structure. Third, the concepts based on spatial correlations and PCA are broadly applicable to virtually all different classes of material internal structures found at multiple hierarchical length scales, spanning from the atomic to the macroscale. Fourth, and perhaps the most importantly, it has been seen that only a handful of PC scores (≈5 to 10) are often adequate in producing high fidelity PSP linkages needed to drive materials innovation efforts [5,70,76]. It should be noted that the use of the PC representations of the spatial correlations as the features of the material structure has thus far been explored mostly at the mesolength scales [5,70,73,77]. This concept is only now beginning to be explored at the atomic structure length scales.

C. Reduced-order model building strategies

In the final step of the MKS framework, the reduced-order features of interest are correlated with the target property using a variety of model-building strategies. NNs provide a nonlinear modeling approach to accomplish this task and are known to be sufficient to capture any arbitrary mapping between the inputs and the output. The most basic feedforward NNs consist of multiple fully connected layers of multiple neurons, with each neuron capturing a linear transformation followed by a nonlinear activation (e.g., ReLu, sigmoid function) [78]. The architecture of a typical feed-forward NN is shown in Fig. 1. The weights and the bias (i.e., model fitting parameters) associated with each neuron are calibrated with backpropagation. This is accomplished through minimization of a user-specified loss function between the predicted output and the corresponding observed values in the training data. NNs derive their scalability from the computationally efficient algorithms used for the calibration of the model parameters. These are readily accessible in many software packages (e.g., PyTorch [79], TensorFlow [80]), most of which are amenable to computation on graphics processing units. NNs also typically iterate through multiple epochs (i.e., passes over the entire training data set) in order to optimize the model parameters. The main limitation of the NNs is that they need a substantially large training dataset, without which they are most likely to produce a model overfit (i.e., significantly larger errors for test data points compared to the training data points). Overfits to training data often occur due to an overparametrization of the NN (i.e., more trainable weights and biases to be calibrated than the size of the dataset). In this work, we utilize feedforward NNs with two hidden layers to predict the formation energies of the compounds in our dataset. For an NN with two hidden layers consisting of h_1 and h_2 neurons, respectively, the number of trainable model parameters to be calibrated (denoted by C) for a single output from an N-dimensional input is given by

$$C = (Nh_1 + h_1h_2 + h_2) + (h_1 + h_2 + 1). \tag{2}$$

An alternate modeling technique that is better suited for small datasets is GPR [51,81,82], which offers a nonparametric Bayesian approach to building surrogate models. GPR (and its variants) have successfully been applied to model structure-property linkages in the mesoscale from relatively small datasets [70,83,84]. GPR models the target as a Gaussian process (GP) that is fully defined by the specification of a mean and a covariance. The GP is then tuned using conditional distributions defined on available training data. Let $\boldsymbol{X}$, $\boldsymbol{X^*}$, and $\boldsymbol{y}$ denote the $N \times D$ matrix of training data points, $N^* \times D$ matrix of test data points and $N \times\ 1$ vector of target values in the training set, respectively. Additionally, let $\boldsymbol{K}(\boldsymbol{X}, \boldsymbol{X'})$, $\boldsymbol{K}(\boldsymbol{X^*}, \boldsymbol{X^{*'}})$ and $\boldsymbol{K}(\boldsymbol{X}, \boldsymbol{X^{*'}})$ denote the $N \times N$, $N^* \times N^*$, and $N \times N^*$ covariance matrices assembled (using a kernel function described later) using the inputs in the training dataset, test dataset, and between the training and test datasets, respectively. In GPR, the predictive mean and variance for test points is expressed as

$$\boldsymbol{\mu^*} = \boldsymbol{K}(\boldsymbol{X}, \boldsymbol{X^{*'}})^{\boldsymbol{T}} \boldsymbol{K}(\boldsymbol{X}, \boldsymbol{X'})^{-1} y, \tag{3}$$

$$\boldsymbol{\Sigma^*} = K(\boldsymbol{X^*}, \boldsymbol{X^{*'}}) - \boldsymbol{K}(\boldsymbol{X}, \boldsymbol{X^{*'}})^{\boldsymbol{T}} \boldsymbol{K}(\boldsymbol{X}, \boldsymbol{X'})^{-1} K(\boldsymbol{X}, \boldsymbol{X^{*'}}), \tag{4}$$

where T denotes the transpose of a matrix. As stated earlier, one typically uses a suitable kernel function to compute the various covariance matrices in Eqs. (2) and (3). One of the most commonly used kernel functions is the automatic relevance determination – squared exponential (ARD-SE) [51,85] function expressed as

$$k(\boldsymbol{x}, \boldsymbol{x'}) = \sigma_{\boldsymbol{s}}^2 \sum_{\boldsymbol{d}=1}^{\boldsymbol{D}} \exp\left(\frac{-(x_d - x_d^{'})^2}{2\theta_d{}^2} \right) + \sigma_n^2 \delta_{xx'}, \tag{5}$$

where $\boldsymbol{x}$ and $\boldsymbol{x'}$ denote any two input vectors selected from the data matrices for which the covariance matrix is being computed, subscript d corresponds to the d^{th} feature in the input vector, σ_s is a scaling parameter that controls the scaling of the output variance, θ_d is the correlation-length hyperparameter corresponding to the dth feature in the input vector, σ_n denotes the noise in the target, and $\delta_{xx'}$ denotes the Kronecker delta. The treatment of the noise using a single parameter added to the diagonal of the covariance indicates independence of noise with the input data (i.e., homoscedasticity) [51]. The central benefit of the ARD-SE kernel is that it allows us to tune independently the correlation-length hyperparameter for each of the input features. This allows for better interpretability of the model, because the salient features with relatively smaller length scale hyperparameters exhibit a higher sensitivity to the predicted target. One should also note that extremely small values of the hyperparameters tend to make the predictions very noisy. Therefore, it is important to tune the values of the hyperparameters in order to produce robust and reliable predictions. This is typically accomplished with maximum likelihood estimation, which lacks a closed-form solution; consequently, one must resort to iterative schemes such as quasi-Newton algorithms to optimize the hyperparameters [86–88].

One of the central challenges in the implementation of GPR comes from the computational cost, especially with large training datasets. The main computational bottleneck arises from the computation of the inverse of $\boldsymbol{K}(\boldsymbol{X}, \boldsymbol{X}')$ in Eq. (3), which typically scales as $O(N^3)$ for dense matrices. This makes traditional GPR intractable even for moderately sized datasets ($N > 1000$). This difficulty is compounded by the need to optimize the large number of hyperparameters ($D + 2$) in the ARD-SE kernel. Several methods have been suggested in literature to address these challenges. These include the utilization of low-rank sparse approximations for the covariance matrix [51,89,90], tree-based partitioning to develop smaller datasets and fit individual GP predictors [91], and localized GPR (L-GPR) proposed by Gramacy and Apley (discussed next and used in this study) [92].

L-GPR involves identifying a local subset of training data points to construct a separate GP for each predictive point. In addition to allowing the application of GPR to larger datasets, a key advantage of L-GPR is the ability to accommodate nonstationarity in the model by allowing optimization of the interpolation hyperparameters in the kernel functions depending on the location of the test point in the high-dimensional input space. Different criteria may be considered while selecting the local training points; an intuitive but suboptimal method would be using a certain number of nearest neighbors as the local neighborhood [93]. However, a more informative design criterion would be to sequentially build the local neighborhood by evaluating a tradeoff between expected information gain by adding a training point and the increase in the predictive variance of the GPR. This method of local training point selection promotes an efficient exploration of the sample space and provides a natural guidance to avoid overfitting of the models. This approach is referred to as active learning Cohn (ALC) [55,94], and is utilized in this work.

III. NOVEL FRAMEWORK FOR QUANTIFYING CRYSTAL STRUCTURES

In this section, we describe a framework for effective low-dimensional representation (i.e., fingerprints) of the crystal structure of the compounds. This is accomplished by first establishing computationally efficient protocols to arrive at voxelized representations of the atomic structure of compounds, and then suitably extending and applying the current MKS framework (described in Secs. II A and II B) on such voxelized representations. In prior work [74,95], very simple descriptors were utilized for the microstructure array, m_s^n, where it was assigned a value of one for voxels within the atomic volume and a value of zero otherwise. A key challenge encountered in this simple representation is that it does not allow an efficient interpolation between material structures (i.e., compounds) of varying chemical compositions. In this work, we expand the previous framework to allow for an enhanced representation of multiple atomic attributes of interest in each voxel. This is accomplished by representing the local material state as a vector set of attributes that take continuous values. Such continuous local states have been utilized previously in the treatment of mesoscale polycrystalline microstructures and mesoscale fields of chemical compositions [5,96]. In this work, we will extend these representations to the quantification of crystal structures of compounds.

Choudhary *et al.* [8] have studied the order of importance of various atomic attributes and their role in the formation energy of a compound. Specifically, they have identified ionization energy, Pauling electronegativity, and heat of fusion as the most important attributes. Therefore, in this work, we define the local material state in the voxel indexed by $\boldsymbol{s}$ using a vector of attributes $\langle a_{0s}, a_{1s}, a_{2s}, v_s\rangle$, where a_{0s}, a_{1s}, and a_{2s} denote the heat of fusion, ionization energy, and Pauling electronegativity, respectively, and v is a binary attribute that takes a value of zero for voxels within any of the atomic volumes (defined by a sphere with radius equal to the atomic radius of the species) and a value one otherwise. Our model building efforts (described later) found it beneficial to further enhance the attribute list using various monomials of the atomic attributes. In other words, the full list of local states included in the atomic state vector $\boldsymbol{m}_s$ may be defined as

$$\boldsymbol{m}_s = \big\langle a_{0s}, a_{1s}, a_{2s}, a_{0s}a_{1s}, a_{1s}a_{2s}, a_{0s}a_{2s}, a_{0s}a_{1s}a_{2s}, \\ a_{0s}^2, a_{0s}^2a_{1s}, \ \ldots\ a_{0s}^{l_0}a_{1s}^{l_1}a_{2s}^{l_2}, \ v_s\big\rangle, \tag{6}$$

where l_0, l_1, and l_2 represent the maximum degree considered for each of the three atomic attributes. A total of $(l_0 + 1) \times (l_1 + 1) \times (l_2 + 1)$ local states were used in the definition of the atomic state vector. Through multiple trials, it was found that the combination of l_0, l_1, and l_2 set to 5, 5, and 2, respectively, yielded the optimal performance in terms of the tradeoff between computational cost (which scales with the number of local states), and the accuracy of our models. For these selections, the total number of local states in $\boldsymbol{m}_s$ is 108. Elements of $\boldsymbol{m}_s$ are then used in the computation of the 2-point correlations using Eq. (1). As already noted, the convolution operation involved is best accomplished using the FFT algorithm. Mathematically, this is expressed as

$$F_{\boldsymbol{k}}^{pq} = \frac{1}{n(S)} M_{\boldsymbol{k}}^{p*} M_{\boldsymbol{k}}^{q}, \tag{7}$$

where $M_{\boldsymbol{k}}^{p}$ and $F_{\boldsymbol{k}}^{pq}$ denote the discrete Fourier transforms of m_s^p and f_r^{pq}, respectively, and the superscript $*$ denotes the complex conjugate. It is important to note that Eqs. (7)

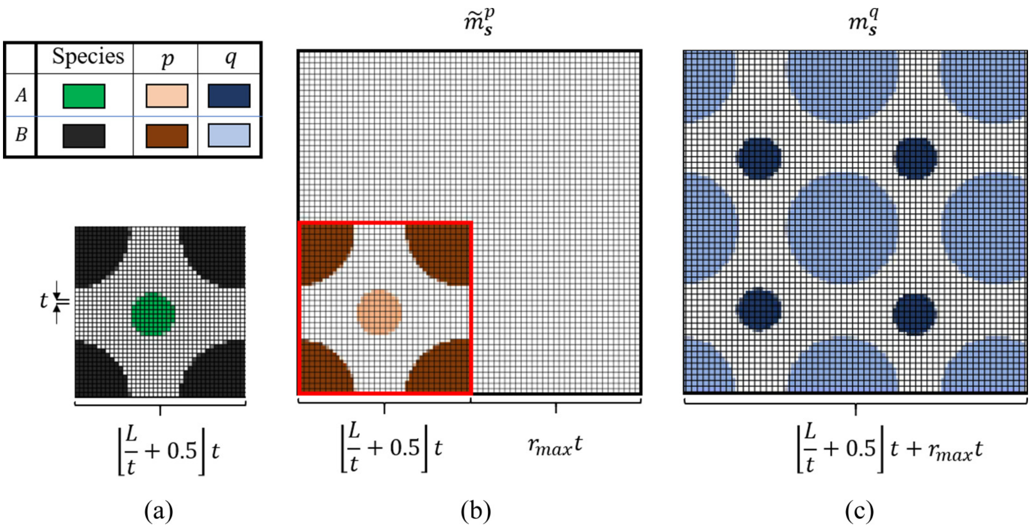
Species
p
q
A
B
t
$\tilde{m}_s^p$
m_s^q
$\left\lfloor \frac{L}{t} + 0.5 \right\rfloor t$
$\left\lfloor \frac{L}{t} + 0.5 \right\rfloor t$
$r_{max} t$
$\left\lfloor \frac{L}{t} + 0.5 \right\rfloor t + r_{max} t$
(a)
(b)
(c)

FIG. 2. (a) A typical 2D cross section of a unit cell of the compound structure, with green and black regions representing two different chemical species. (b) and (c) Extended compound structures depicting the spatial distribution of the local states p and q, respectively. These extended structures are designed such that the spatial correlations to the desired vector length (i.e., $r_{\max}$ in each of the vector components) can be computed efficiently using the FFT algorithms that implicitly treat the functions as being periodic. The red border in (b) identifies the outline of original unit cell from (a).

implicitly assumes the fields involved are spatially periodic. Although all the compounds considered in this work exhibit periodicity (i.e., all spatial fields defined based on the atomic structure of the compound are inherently spatially periodic), the sizes of their periodic unit cells vary substantially. For example, the lattice parameter (reflecting the length scale of the periodicity) for the compounds included in this study varies over the range of 1.8–12.4 Å. Therefore, if one were computing the spatial correlations using voxelization schemes based on the respective unit cell sizes, Eqs. (7) can be applied directly without any problems. However, a constant voxel size is essential for obtaining the desired low-dimensional representations using PCA in the later steps. Since the unit cell size L of the different compounds considered in this study is unlikely to be an integer multiplier of a common voxel size, t, a suitable strategy is needed to efficiently compute the spatial correlations using a standardized voxel size. Given the size of the dataset, these computations are only practical if we can continue to exploit the cost savings provided by the FFT algorithm in evaluating the spatial correlations. A suitable computational protocol has been devised to address this challenge, which is described next. This new protocol builds on prior efforts involving mesoscale material microstructures [5].

The computational scheme presented in this work is designed to use a single voxel size and a common set of corresponding vectors indexed by $\{\boldsymbol{r}\}$, for the computation of sets of 2-point correlations (i.e., $f_{\boldsymbol{r}}^{pq}$) using Eqs. (1) and (7) for all the compounds considered in this study. For 3D volumes, it is convenient to represent $\boldsymbol{r}$ itself as a vector of integers $\langle r_1, r_2, r_3 \rangle$. The 2-point correlations used in this study are computed only for the set of positive vectors up to a selected maximum length for each of the three components, i.e., $\{\boldsymbol{r} | 0 \leqslant r_1 \leqslant r_{\max}, 0 \leqslant r_2 \leqslant r_{\max}, 0 \leqslant r_3 \leqslant r_{\max}\}$. Therefore, the total number of discrete vectors for which we seek to compute $f_{\boldsymbol{r}}^{pq}$ is $r_{\max}^3$. Our approach to efficiently compute these correlations considers the numerator and denominator in Eq. (1) separately. As mentioned earlier, for each selected $\boldsymbol{r}$, the numerator represents the number of successes in locating the local states p and q separated by the selected vector. Our strategy to accurately (and efficiently) compute this quantity is to express the numerator as a cross correlation between two different material structure spatial fields, $\tilde{m}_s^p$ and m_s^q, both of which are derived from the original material structure. The construction of these structure fields is illustrated in Fig. 2 for a simple crystalline compound comprising two chemical species A and B. Note that all structure fields shown in Fig. 2 are 2D sections of the 3D material volume of this compound. Figure 2(a) shows the unit cell of the original compound structure within a volume that is closely approximated by an integer number of voxels of the selected size, t. Therefore, the voxelated structure shown in Fig. 2(a) is no longer exactly periodic. In fact, the size of this structure is $\lfloor \frac{L}{t} + 0.5 \rfloor t$, where $\lfloor \cdot \rfloor$ indicates the greatest integer function (also referred to as the floor function). Clearly, the size of the voxelized unit cell shown in Fig. 2(a) is not equal to L. This discrepancy introduces a small error in the computation of the spatial correlations that scales with the value of t. In other words, one would have to select the value of t carefully, as a compromise between minimizing the voxelization error and keeping the computational cost reasonable.

The structure fields $\tilde{m}_s^p$ and m_s^q are defined over larger spatial domains as shown in Figs. 2(b) and 2(c). The size of these fields is specified as $\lfloor \frac{L}{t} + 0.5 \rfloor t + r_{\max} t$, while the corresponding set of voxels is labeled as S^*. Note that the fields are extended so that when any of the vectors of interest (identified earlier as $\{\boldsymbol{r}\}$) are placed with their tails in the unit cell structure shown in Fig. 2(a), their heads are guaranteed to lie within the extended spatial domains shown in Figs. 2(b) and 2(c). Furthermore, the two fields in these figures are associated with local states p and q identified in the definition of the subset of spatial correlations $f_{\boldsymbol{r}}^{pq}$. The structure fields shown in Figs. 2(b) and 2(c) have been specifically designed

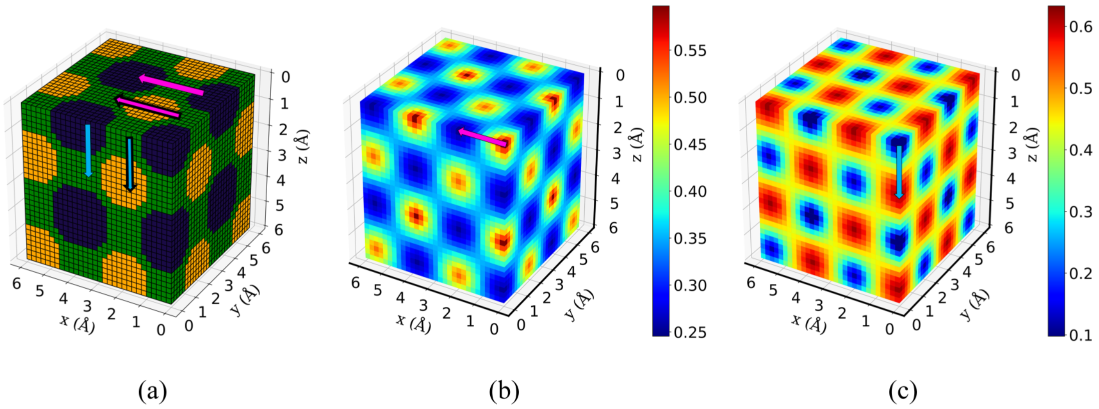

FIG. 3. (a) Voxelized 3D crystal structure of $AlNi_3$ with the green, orange, and dark blue colors representing the voxels occupied by the void, Al, and Ni species. (b) The autocorrelation map for the void state (v_s). (c) The v_s(void) $- a_{2s}$ (Pauling electronegativity) cross-correlation map.

so that the cross correlation defined as $\sum_{s\in S^*} \tilde{m}_s^p m_{s+\boldsymbol{r}}^q$ provides a sufficiently accurate estimate of the numerator to within the inherent voxelization error discussed earlier. This is because $\tilde{m}_s^p$ returns the value of the local state p only when the tail of the vector indexed by $\boldsymbol{r}$ lies within the original unit cell shown in Fig. 2(a). This region is indicated by the red outline in Fig. 2(b) More importantly, this structure field returns a zero value otherwise, which nullifies any possible contributions arising from the wrap-around vectors implicit in the Fourier transform operations performed in Eq. (7). Note that this nullification only works for vectors within the selected set $\{\boldsymbol{r}\}$. In other words, the use of Eq. (7) on the fields shown in Figs. 2(b) and 2(c) produces estimates of the numerator in Eq. (1) for many more vectors not included in $\{\boldsymbol{r}\}$. In the strategy described here, one retains only the results for the vectors in $\{\boldsymbol{r}\}$ and discards the rest of the results. Although one computes more values than needed in this protocol, the overall computational cost is much cheaper than the direct computation of the correlations exclusively for the vector set of interest. The denominator in Eq. (1) is simply the number of voxels in the original unit cell shown in Fig. 2(a), and is given by $(\lfloor \frac{L}{t} + 0.5\rfloor)^3$. The protocol described here was found to provide an excellent approximation for the fast computation of the desired spatial correlations. In our work, we implemented the aforementioned protocol with the value of t and $r_{\max}$ selected as 0.2 Å and 30, respectively, which corresponds to the computation (and retention) of the 2-point correlations for vectors with components up to 6 Å. This selection was motivated by the fact that 92% of the crystalline compounds present in our dataset had a lattice parameter less than 6 Å.

Figures 3(a)–3(c) show an example crystal structure consisting of two atomic species ($AlNi_3$), its autocorrelation map for v_s (the void state), and its $v_s - a_{2s}$ cross-correlation map, respectively. Note that the auto- and cross-correlation maps shown correspond only to a single octant in the vector space (with positive values of r_1, r_2, and r_3). It can be observed that the periodicity of the crystal structure is indeed captured in these maps. The value of the autocorrelation corresponding to the zero vector [0.58 in Fig. 3(b)] reflects the void volume fraction in the crystal structure. Since the void state in the present application is set to either zero or one in each voxel, the autocorrelation map in Fig. 3(a) actually provides statistical information for a large set of vectors, whose heads can be selected anywhere in the depicted vector space, but tails fixed at $\langle 0, 0, 0\rangle$. As a specific example, the pink-colored vector in Fig. 3(b) corresponds to $\langle 3.8,\ 0,\ 0\ \rangle$Å, with a void autocorrelation of 0.58. This autocorrelation value simply reflects the probability of finding two void (green) voxels in the crystal structure shown in Fig. 3(a) separated by the selected vector. Figure 3(a) shows two example placements of the selected vector – a successful placement (pink vector with a black outline) connecting two void (green) voxels and an unsuccessful placement (pink vector without a black outline). Due to the periodicity of the crystal structure, it is seen that the autocorrelation value for the pink-colored vector is the same as autocorrelation for the zero vector (i.e., the void volume fraction). The cross-correlation map shown in Fig. 3(c) similarly captures the spatial correlations between states v_s and a_{2s} present in the structure. Note that since we do not allow the void state and any material state to coexist in a single voxel, the cross correlation for the zero vector exhibits a zero value. A cross-correlation peak is observed for $\langle 0, 0, 1.8\rangle$ Å in Fig. 3(c), shown using a blue-colored vector. As before, two example placements of this vector are shown in Fig. 3(a), with one reflecting a successful sampling of the desired spatial correlation.

IV. DATASET

A dataset of compound crystal structures and formation energies was obtained from the publicly accessible JARVIS-DFT repository [7,9–12]. The extracted data included the crystal structure information (i.e., Bravais lattice type and the coordinates of the atoms present in the unit cell) and their DFT-computed chemical properties. Specifically, the formation energies of ∼1740 cubic crystalline compounds that were computed with the OPTB88VDW functional [12] were extracted from this repository, and were used to train our models and evaluate the utility of our feature engineering scheme. The compounds included in this dataset comprised 70 different chemical species, thereby offering opportunities for the extraction of surrogate models applicable to a broad range of chemical compositions. The most frequently occurring nonmetallic species in the dataset were oxygen and nitrogen,

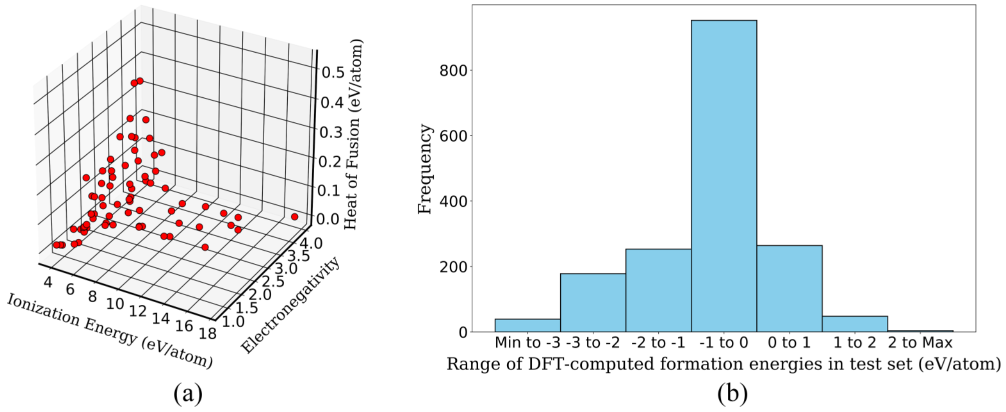

FIG. 4. (a) Distribution of the attribute values for the different atomic species present in the dataset. (b) Histogram of the formation energies of the compounds in the dataset.

while the most frequently appearing metallic species were aluminum, palladium, and rubidium. The least frequently appearing metal atoms are cesium, rhenium, and technetium. Among the crystalline compounds in the dataset, there are 1160 intermetallic compounds, 408 metal oxides, 95 metal halides, and 37 metal sulfides. Moreover, the dataset consisted of 21, 950, 743, and 27 unary, binary, ternary, and quaternary crystalline compounds.

Figure 4(a) shows the distribution of the values of the three local state attributes, i.e., heat of fusion (a_0) ionization energy (a_1), Pauling electronegativity (a_2) for all of the different chemical species present in the dataset. The electronegativity of the different chemical species in the dataset appears to be fairly evenly distributed over the range 0.79 to 3.98. In contrast, the distributions of the heat of fusion and ionization energy appear to be nonuniformly distributed, with more of the chemical species having attribute values in the lower end of their respective ranges of ($2.3e{-}3$ to 0.52) eV/atom and (3.89 to 17.42) eV/atom. Figure 4(b) represents the distribution of the DFT-computed formation energies of compounds (the target or output for the models) present in the dataset. Note that the samples are unevenly distributed in the values of the formation energy, with most of the samples having a formation energy between -3 and 1 eV/atom. From this distribution of target properties, one would expect a higher accuracy in the prediction of the formation energy within this region, and a lower accuracy outside it.

The feature engineering framework developed in Sec. III is applied to the selected dataset. As described in Sec. II A, the complete set of P^2 (or 108^2) spatial statistics contains several redundancies. The following sets of spatial correlations were used in this work: (i) autocorrelations of v_s, a_{0s}, and a_{1s} (this were the most dominant signals in the dataset), and (ii) the cross correlations of these three local states with the rest of the local states, while eliminating the trivially related sets (the $f_{\boldsymbol{r}}^{pq}$ are trivially related to $f_{\boldsymbol{r}}^{qp}$). This resulted in a total of 321 sets of spatial correlations for each crystal compound. The dimensionality reduction procedure described in Sec. II B was then applied on the complete set of spatial statistics to obtain 25 PC scores for each compound. Since each PC basis in our application carries (weighted) information from a total of ~3 million 2-point statistics, its precise interpretation is currently impractical.

V. RESULTS AND DISCUSSION

As mentioned earlier, two different modeling strategies have been explored in this study to evaluate critically the efficacy our feature engineering framework. These modeling strategies were specifically chosen to represent distinctly different approaches to building surrogate models. Therefore, our main goal here is to evaluate the performance of the proposed feature engineering framework (involving PCA on a large feature vector of spatial correlations that account rigorously for the atomic details in the crystal structure) on distinctly different model building strategies. The selected model building strategies for this study include L-GPR (implemented using the laGP package in the R programming language [55]) and NN (implemented using PyTorch [79]). The inputs and the output to both models were the set of the top 25 PC scores (scaled to unit variance) generated using the feature engineering framework presented in this work and the DFT-computed formation energy, respectively. Both modeling strategies benefited significantly from the scaling of inputs (to unit variance), due to their usage of gradient-based optimization in the regression procedure (i.e., hyperparameter tuning in L-GPR, and loss function minimization in NNs).

In this study, mean absolute error (MAE) and median absolute error (MedAE) have been chosen as error metrics for quantifying the predictive accuracy of the surrogate models. For a set of J predictive test samples, these error measures are expressed as

$$\mathrm{MAE} = \frac{1}{J}\sum_{j\in J}\left|y_{\mathrm{pred}}^{(j)} - y_{\mathrm{actual}}^{(j)}\right|, \tag{8}$$

$$\mathrm{MedAE} = \mathrm{Median}\big(\left|y_{\mathrm{pred}}^{(1)} - y_{\mathrm{actual}}^{(1)}\right|, \left|y_{\mathrm{pred}}^{(2)} - y_{\mathrm{actual}}^{(2)}\right|, \ldots, \left|y_{\mathrm{pred}}^{(J)} - y_{\mathrm{actual}}^{(J)}\right|\big) \tag{9}$$

in which $y_{\mathrm{actual}}^{(j)}$ and $\mathrm{y}_{\mathrm{pred}}^{(\mathrm{j})}$ indicate the DFT-computed and the surrogate model predicted values of the formation energy for the sample j, respectively. While the same error metrics are used to evaluate performance of both modeling strategies, the training and test protocols are substantially different because of the very different underlying philosophies

TABLE I. The predictive accuracy of the five L-GPR models built in this work to estimate the DFT-computed formation energies. The input to all models is the set of the 25 PC scores of the spatial correlations of the crystal structure.

Model No.	χ	Percentage of entire dataset (%)	$\text{MAE}_{\text{L-GPR}}$ (eV/atom)	$\text{MedAE}_{\text{L-GPR}}$ (eV/atom)
LGPR1	50	2.84	0.393	0.256
LGPR2	100	5.68	0.378	0.244
LGPR3	150	8.52	0.352	0.219
LGPR4	200	11.36	0.330	0.192
LGPR5	250	14.20	0.329	0.187

involved in these model building strategies. These are described next.

A. L-GPR modeling approach

In this study, five different models were built using the L-GPR strategy. Recall that in this strategy, a separate GP is formulated for each point in the dataset using training points selected from the local neighborhood based on the ALC criterion described in Sec. II C. Since predictions were made at each point independently, the MAE and MedAE were computed directly with Eqs. (8) and (9) over all samples (indexed by j) in the dataset (consisting of J samples). These values are denoted as $\text{MAE}_{\text{L-GPR}}$ and $\text{MedAE}_{\text{L-GPR}}$, respectively. The main difference in the five L-GPR models built for this study was in the number of points selected from the local neighborhood (denoted by χ). Table I summarizes the accuracies of these models, as evaluated by the error metrics described previously. In this table, the value of χ is also shown as a percentage of the total number of points in the dataset. The main purpose of this analysis was to understand the effectiveness of the ALC criterion in the L-GPR scheme for selecting optimal training points for the desired model. Since GPR is essentially an interpolation strategy, L-GPR offers significant computational savings compared to the conventional GPR strategy (which utilizes all points in the dataset). This is because the computational cost of the GPR scales as $O(N^3)$, as mentioned previously. Also, the use of ALC in the L-GPR to select the training points offers an organic strategy for regularization and avoiding overfit. More specifically, since the ALC criterion selects neighborhood points based on the goal of maximizing information gain, one would expect that after the training dataset has included most of the “important” points, there would only be an incremental improvement in performance with a further increase in χ. This tendency was indeed observed in our results. Table I shows that a selection of about 200 points in the neighborhood of each test point in the 25-dimensional input space (i.e., with $\chi = 200$) provides a robust prediction of the formation energy. Figure 5(a) shows the histogram of errors of the formation energies predicted by this model. As seen in this figure, around 80% of the formation energies are predicted to within an error of 0.4 eV/atom. Only a small improvement in overall predictive performance was observed when χ was further increased to 250 points.

Figure 5(b) shows the parity plot of the L-GPR predicted vs DFT-computed formation energies corresponding to Model 4 from Table I (i.e., $\chi = 200$). Figure 5(c) presents a bar plot of the distribution of the average error for the different values of the formation energies. As seen from the bar plot, the average error in the predicted values is lowest for samples with DFT-computed formation energies in the range of (−2 to 0) eV/atom. This range is also the region with the highest concentration of training data points [see Fig. 4(b)]. As expected, the accuracy of the L-GPR models built is quite sensitive to the amount of available training data close to test data points. For example, the average error for the samples with DFT-computed formation energies greater than 2 eV/atom is quite high, as there are only 20 data points with the corresponding formation energies. The predictive uncertainty of the L-GPR model also followed a similar distribution over these ranges of formation energies, with the highest uncertainty associated with the crystals with DFT-computed formation energies greater than 2 eV/atom. This indicates that prediction confidence in regions with more data is significantly higher than in the regions with sparse data.

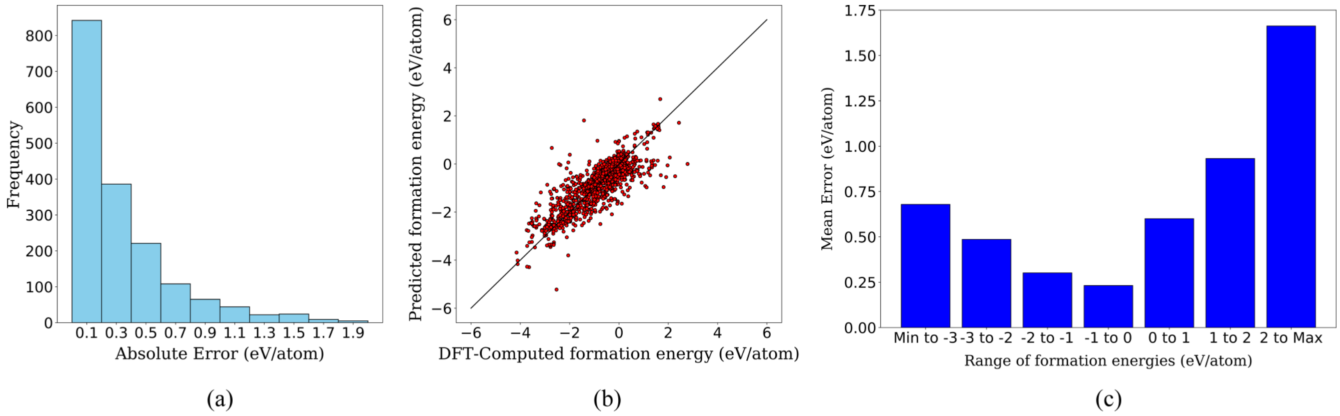

FIG. 5. (a) Histogram of errors obtained in prediction of the DFT-computed formation energy for L-GPR Model 4 in Table I with the local training dataset size χ of 200 points. (b) Parity plot of the L-GPR predicted vs DFT-computed formation energies for the same L-GPR model for all data points used in this study. (c) Bar plot of the mean prediction error (of the same model) in the formation energy for different ranges of DFT-computed formation energies.

TABLE II. The predictive accuracy of the five NN models built in this work to estimate the DFT-computed formation energies. The input to all models is the set of the 25 PC scores of the spatial correlations of the crystal structure. Note that the error metrics ($\mathrm{MAE}_{\mathrm{NN}}$, $\mathrm{MedAE}_{\mathrm{NN}}$) were computed only for the ∼345 Data points in the test set (represented by the green points in Fig. 6).

Model No.	Number of hidden neurons	Number of parameters	$\mathrm{MAE}_{\mathrm{NN}}$ (eV/atom)	$\mathrm{MedAE}_{\mathrm{NN}}$ (eV/atom)
NN1	(20,10)	741	0.361	0.232
NN2	(20,14)	829	0.352	0.225
NN3	(20,20)	961	0.341	0.198
NN4	(25,20)	1191	0.340	0.201
NN5	(25,25)	1326	0.350	0.195

B. NN modeling approach

Five different feedforward NN models (all consisting of two hidden layers) were evaluated in this study. For each model, the loss function (to be minimized) was chosen as L1 loss [97], which reflects the mean absolute error in the prediction of the target for the training dataset. This minimization was performed using the Adam gradient-based optimizer [98] with a learning rate set to $5e{-}5$. The available dataset was partitioned into separate training and test datasets in a 80% to 20% split. This partitioning was done while maintaining similar distributions of the output values in the train and test datasets [similar to Fig. 4(b)]. A small fraction of the training dataset (about 5%) was designated as a validation set, and utilized to implement an early stopping criterion [48], designed to mitigate overfitting of the NN models produced. The performance of the model on the validation dataset is used to make decisions on when to stop the minimization iterations (epochs) on the loss function. This is especially important in the present work because of the relatively small training datasets, and relatively large number of trainable model parameters inherent to the NN models. The MAE and MedAE error metrics (denoted by $\mathrm{MAE}_{\mathrm{NN}}$ and $\mathrm{MedAE}_{\mathrm{NN}}$, respectively) were computed using Eqs. (8) and (9) on the test data points for all NN models built in this study to evaluate their predictive accuracy.

Table II summarizes the accuracies of the five different NN models produced in this study. The numbers of neurons in the hidden layers were varied across all the NN models that were built in this study, in order to explore their influence on model performance. The number of the trained model parameters in each NN model [calculated using Eqs. (2)] are also shown in Table II. It is seen that the predictive accuracy of the model does not increase significantly beyond the NN3 model, in spite of large increases in the number of trainable model parameters. Indeed, with the number of trainable model parameters approaching the number of training data points, there is clear evidence of model overfit, especially with models NN4 and NN5. Figure 6 shows a parity plot of the train and test predictions from the NN3 model. It can be seen in this plot that the prediction quality on the train and test sets are similar (with MAEs of 0.28 eV/atom and 0.34 eV/atom on each of these sets, respectively), indicating that the early stopping criterion provides an effective regularization method to help mitigate an overfit to the training set. Out of the set of compounds exhibiting a high prediction error (greater than 1 eV/atom), 28% are metal oxides, 11% are metal halides, and 35% are intermetallic compounds. Further analysis of the crystals having the ten highest prediction errors indicates that the prediction quality is affected by the proportion of occurrence of the atoms in the dataset. In each case, at least one of the constituent atoms occurs in less than 3% (i.e., fewer than 52 samples) from the dataset.

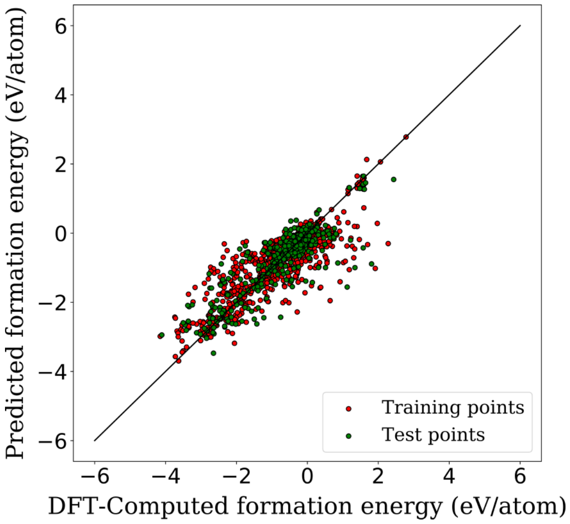

FIG. 6. Parity plot of the NN-predicted vs DFT-computed formation energies for NN3 Model (see Table II).

C. Comparison of the L-GPR and NN surrogate model building strategies

While the accuracy of both modeling strategies explored in this study are reported using the same metrics (MAE and MedAE), the performance of these models are not easily compared directly because of the different underlying philosophies involved in the two strategies. However, they clearly point to the efficacy of the feature engineering paradigm presented in this work. It is indeed remarkable that the 25 features identified by our protocols (from an extremely large initial set of approximately three million 2-point correlations) provided reasonably accurate models when used as inputs to two completely different modeling strategies (i.e., the L-GPR strategy utilizing Bayesian local interpolations and the NN providing a global regression). This observation strongly supports our claim that the protocols involving spatial correlations and PCA are capable of identifying salient features that exhibit high utility in the formulation of surrogate structure-property models, independent of the modeling strategy employed.

The results from the models produced in this work are quite reasonable when compared with prior modeling efforts in literature. Faber *et al.* [20] utilized three different structure representation schemes (Ewald sum matrix, extended Coulomb matrix, and Sine matrix) together with a KRR modeling scheme and reported test set MAEs of 0.49 eV/atom, 0.64 eV/atom and 0.37 eV/atom, respectively, for the predicted values of formation energies. However, the authors also reported significantly lower training set MAEs of the order of

0.005 eV/atom for each of these models, indicating that the models obtained in this study are likely overfit to the training data. These models were trained on 3000 data points taken from the MP database [16]. In a different study, Choudhary *et al.* [8] curated a large set of ~1550 chemical and structural descriptors as input features for establishing a model using a much larger dataset consisting of ~24 500 materials. This study reported a test set MAE of 0.12 eV/atom with the use of an ensemble-based gradient-boosting decision tree regression model consisting of ~1150 estimators with up to 270 leaves per estimator and an unconstrained tree depth (which was decided based on an early stopping criterion). This correlated with the use of ~500 nodes for almost all estimators built, indicating that a large number of parameters were employed in the final model. While the test MAE reported in this study is lower, the feature engineering procedure used in this study is significantly more computationally intensive and difficult to scale to larger datasets.

It is emphasized that in comparison with the prior studies, the strategy developed in this work relies on unsupervised (i.e., independent of the selected output variable) feature engineering to identify a small number of salient features (PC scores). These features are then used as inputs to models with relatively fewer fit parameters. This confirms the value and utility of the feature engineering protocols presented in this work. Both models produced in this study are likely to find uses in potentially different applications. The main strength of the L-GPR strategy is that it can provide objective guidance on what new training data points should be generated to improve the model fidelity. This is because the GPR models not only predict the expected values of the output for test data points, but also their variance. This attribute can be used in a suitable design of experiments strategy [51,99] to identify which specific inputs exhibit the highest potential for model improvement. This is particularly useful in the initial stages of generating the training data sets, especially in situations where the cost of data generation is high (e.g., data generated using DFT models). In contrast, NNs could become the preferred approach after a substantial amount of data has been collected. This is mainly because of their ability to cover a much richer space of model functions, and their superior computational efficiency in handling large data sets. The data set used here is small enough that the L-GPR strategy is likely to be the better option for its size.

The overall predictive accuracy obtained by the models in this work is not sufficient for the accurate computation of the crystal formation energy. We believe that the model accuracy can be improved by (i) implementing the feature engineering scheme on a larger dataset (by lifting the constraint of cubic crystal symmetry), and (ii) using the obtained spatial correlations directly as inputs to other model architectures like CNNs. Larger training datasets would allow for models with a larger number of model fit parameters, and thereby improve the model accuracy.

VI. CONCLUSIONS

In this study, we have introduced a systematic and computationally efficient feature engineering framework based on 2-point spatial correlations for the quantification of cubic crystal structures with varied chemistries. The approach presented in this work employed a voxelized representation of the crystal structure and computed the spatial correlations on suitably selected atomic attributes (here taken as heat of fusion, ionization energy, and Pauling electronegativity). The proposed approach offers an avenue for feature engineering that will allow interpolations across different chemistries of the compounds. This work demonstrates that the proposed feature engineering approach combined with PCA for dimensionality reduction is capable of generating a compact set of salient features (i.e., PC scores) representing the many details of the crystal structure. These PC scores can be used effectively in building surrogate models needed to screen for materials exhibiting potential for improved properties. In this work, this was demonstrated by using the same features as inputs to two very different model-building strategies (i.e., L-GPR and NN) for the predictions of crystal formation energy. The results of this work indicate that the utility of the generated features is independent of the model-building strategy. Moreover, the distribution of errors in predictions (and their associated uncertainties in the case of Bayesian approaches) offer objective guidance on where additional training data should be targeted to further improve the performance of the surrogate models. This work also found that the L-GPR modeling strategy produces more robust predictions when dealing with relatively smaller datasets, as the one utilized in this study.

The dataset used in this study is available to download from Ref. [100].

ACKNOWLEDGMENTS

P.R.K. and S.R.K. gratefully acknowledge support from ONR N00014-18-1-2879. The Hive cluster at Georgia Institute of Technology (supported by NSF 1828187) was used for this work.

The authors declare that no known competing financial interests have influenced the work reported in this paper.

[1] W. Kohn and L. J. Sham, Self-consistent equations including wxchange and correlation effects, Phys. Rev. **140**, A1133 (1965).

[2] P. Hohenberg and W. Kohn, Inhomogeneous electron gas, Phys. Rev. **136**, B864 (1964).

[3] A. Zunger, Inverse design in search of materials with target functionalities, Nat. Rev. Chem. **2**, 0121 (2018).

[4] A. Agrawal and A. Choudhary, Perspective: Materials informatics and big data: Realization of the "fourth paradigm" of science in materials science, APL Mater. **4**, 053208 (2016).

[5] S. Kalidindi, *Hierarchical Materials Informatics: Novel Analytics for Materials Data* (Elsevier, Amsterdam, 2015), p. 219.

[6] S. Kirklin *et al.*, The open quantum materials database (OQMD): Assessing the accuracy of DFT formation energies, npj Comput. Mater. **1**, 15010 (2015).

[7] K. Choudhary: Jarvis-DFT (2014), https://www.nist.gov/system/files/documents/2018/02/02/jarvis-dft_1_31_2017.pdf (accessed May 10, 2021).

[8] K. Choudhary, B. DeCost, and F. Tavazza, Machine learning with force-field inspired descriptors for materials: Fast screening and mapping energy landscape, Phys. Rev. Mater. **2**, 083801 (2018).

[9] K. Choudhary, K. F. Garrity, and F. Tavazza, Data-driven discovery of 3D and 2D thermoelectric materials, J. Phys.: Condens. Matter **32**, 475501 (2020).

[10] K. Choudhary *et al.*, High-throughput Identification and Characterization of Two-dimensional Materials using Density functional theory, Sci. Rep. **7**, 5179 (2017).

[11] K. Choudhary and F. Tavazza, Convergence and machine learning predictions of Monkhorst-Pack k-points and plane-wave cut-off in high-throughput DFT calculations, Comput. Mater. Sci. **161**, 300 (2019).

[12] K. Choudhary *et al.*, Computational screening of high-performance optoelectronic materials using OptB88vdW and TB-mBJ formalisms, Sci. Data **5**, 180082 (2018).

[13] S. Haastrup *et al.*, The Computational 2D Materials Database: High-throughput modeling and discovery of atomically thin crystals, 2D Mater. **5**, 042002 (2018).

[14] L. Ruddigkeit *et al.*, Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17, J. Chem. Inf. Model. **52**, 2864 (2012).

[15] R. Ramakrishnan *et al.*, Quantum chemistry structures and properties of 134 kilo molecules, Sci. Data **1**, 140022 (2014).

[16] A. Jain *et al.*, Commentary: The Materials Project: A materials genome approach to accelerating materials innovation, APL Mater. **1**, 011002 (2013).

[17] J. Lee *et al.*, Prediction model of band gap for inorganic compounds by combination of density functional theory calculations and machine learning techniques, Phys. Rev. B **93**, 115104 (2016).

[18] G. Pilania *et al.*, Machine learning bandgaps of double perovskites, Sci. Rep. **6**, 19375 (2016).

[19] A. M. Deml *et al.*, Predicting density functional theory total energies and enthalpies of formation of metal-nonmetal compounds by linear regression, Phys. Rev. B **93**, 085142 (2016).

[20] F. Faber *et al.*, Crystal structure representations for machine learning models of formation energies, Int. J. Quantum Chem. **115**, 1094 (2015).

[21] W. Ye *et al.*, Deep neural networks for accurate predictions of crystal stability, Nat. Commun. **9**, 3800 (2018).

[22] A. Agrawal *et al.*, A Formation Energy Predictor for Crystalline Materials Using Ensemble Data Mining, in *IEEE 16th International Conference on Data Mining Workshops (ICDMW)* (IEEE, New York, 2016).

[23] L. Ward *et al.*, Including crystal structure attributes in machine learning models of formation energies via Voronoi tessellations, Phys. Rev. B **96**, 024104 (2017).

[24] T. Xie and J. C. Grossman, Crystal Graph Convolutional Neural Networks for an Accurate and Interpretable Prediction of Material Properties, Phys. Rev. Lett. **120**, 145301 (2018).

[25] M. Rupp *et al.*, Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning, Phys. Rev. Lett. **108**, 058301 (2012).

[26] H. Huo and M. Rupp, Unified representation of molecules and crystals for machine learning, arXiv:1704.06439 (2018).

[27] G. Montavon *et al.*, Learning invariant representations of molecules for atomization energy prediction, *Advances in Neural Information Processing Systems*, Vol. 25 (Elsevier, 2012), p. 449.

[28] L. Ward *et al.*, Machine learning prediction of accurate atomization energies of organic molecules from low-fidelity quantum chemical calculations, MRS Commun. **9**, 891 (2019).

[29] R. Ramakrishnan *et al.*, Big data meets quantum chemistry approximations: The delta-machine learning approach, J. Chem. Theory Comput. **11**, 2087 (2015).

[30] D. M. Wilkins *et al.*, Accurate molecular polarizabilities with coupled cluster theory and machine learning, Proc. Natl. Acad. Sci. **116**, 3401 (2019).

[31] A. Ziletti *et al.*, Insightful classification of crystal structures using deep learning, Nat. Commun. **9**, 2775 (2018).

[32] K. Ryan, J. Lengyel, and M. Shatruk, Crystal structure prediction via deep learning, J. Am. Chem. Soc. **140**, 10158 (2018).

[33] J. Graser, S. K. Kauwe, and T. D. Sparks, Machine learning and energy minimization approaches for crystal structure predictions: A review and new horizons, Chem. Mater. **30**, 3601 (2018).

[34] B. Meredig *et al.*, Combinatorial screening for new materials in unconstrained composition space with machine learning, Phys. Rev. B **89**, 094104 (2014).

[35] O. Egorova *et al.*, Multifidelity statistical machine learning for molecular crystal structure prediction, J. Phys. Chem. A **124**, 8065 (2020).

[36] E. V. Podryabinkin *et al.*, Accelerating crystal structure prediction by machine-learning interatomic potentials with active learning, Phys. Rev. B **99**, 064114 (2019).

[37] K. Hansen *et al.*, Machine learning predictions of molecular properties: Accurate many-body potentials and nonlocality in chemical space, J. Phys. Chem. Lett. **6**, 2326 (2015).

[38] J. Behler, Perspective: Machine learning potentials for atomistic simulations, J. Chem. Phys. **145**, 170901 (2016).

[39] V. Gladkikh *et al.*, Machine learning for predicting the band gaps of ABX3 perovskites from elemental properties, J. Phys. Chem. C **124**, 8905 (2020).

[40] D. Jha *et al.*, ElemNet: Deep learning the chemistry of materials from only elemental composition, Sci. Rep. **8**, 17593 (2018).

[41] K. T. Schütt *et al.*, How to represent crystal structures for machine learning: Towards fast prediction of electronic properties, Phys. Rev. B **89**, 205118 (2014).

[42] S. Honrao *et al.*, Machine learning of *ab initio* energy landscapes for crystal structure predictions, Comput. Mater. Sci. **158**, 414 (2019).

[43] S. J. Honrao, S. R. Xie, and R. G. Hennig, Augmenting machine learning of energy landscapes with local structural information, J. Appl. Phys. **128**, 085101 (2020).

[44] M. Karamad *et al.*, Orbital graph convolutional neural network for material property prediction, Phys. Rev. Mater. **4**, 093801 (2020).

[45] C. W. Park and C. Wolverton, Developing an improved crystal graph convolutional neural network framework for accelerated materials discovery, Phys. Rev. Mater. **4**, 063801 (2020).

[46] A. Micheli, Neural network for graphs: A contextual constructive approach, IEEE Trans. Neural Netw. **20**, 498 (2009).

[47] T. N. Kipf and M. Welling, Semi-supervised classification with graph convolutional networks, arXiv:1609.02907.

[48] L. Prechelt, Early Stopping — But When? in *Neural Networks: Tricks of the Trade: Second Edition*, edited by G. Montavon, G. B. Orr, and K.-R. Müller (Springer, Berlin, 2012), p. 53

[49] N. Srivastava *et al.*, Dropout: A simple way to prevent neural networks from overfitting, J. Mach. Learn. Res. **15**, 1929 (2014).

[50] F. Girosi, M. Jones, and T. Poggio, Regularization theory and neural networks architectures, Neural Comput. **7**, 219 (1995).

[51] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning. Adaptive Computation and Machine Learning* (MIT Press, Cambridge, MA, 2006), p. 248

[52] A. Cecen, T. Fast, and S. R. Kalidindi, Versatile algorithms for the computation of 2-point spatial correlations in quantifying material structure, IMMI **5**, 1 (2016).

[53] S. Kajita *et al.*, A universal 3D voxel descriptor for solid-state material informatics with deep convolutional neural networks, Sci. Rep. **7**, 16991 (2017).

[54] Y. Zhao *et al.*, Predicting elastic properties of mater. from electronic charge density using 3d deep convolutional neural networks, J. Phys. Chem. C **124**, 17262 (2020).

[55] R. B. Gramacy, LaGP: Large-scale spatial modeling via local approximate gaussian processes in R, J. Stat. Softw. **72**, 1, (2016).

[56] S. R. Kalidindi, S. R. Niezgoda, and A. A. Salem, Microstructure informatics using higher-order statistics and efficient data-mining protocols, JOM **63**, 34 (2011).

[57] A. G. Gray and A. W. Moore, N-body' problems in statistical learning, in *Proceedings of the 13th International Conference on Neural Information Processing Systems* (MIT Press, 2000), pp. 500–06.

[58] A. W. Moore *et al.*, *Fast Algorithms and Efficient Statistics: N-Point Correlation Functions* (Springer, Berlin, 2001).

[59] S. Torquato and H. Haslach Jr., Random heterogeneous materials: Microstructure and macroscopic properties, Appl. Mech. Rev. **55**, B62 (2002).

[60] A. Cecen, Y. C. Yabansu, and S. R. Kalidindi, A new framework for rotationally invariant two-point spatial correlations in microstructure datasets, Acta Mater. **158**, 53 (2018).

[61] S. R. Niezgoda, D. T. Fullwood, and S. R. Kalidindi, Delineation of the space of 2-point correlations in a composite material system, Acta Mater. **56**, 5285 (2008).

[62] W. P. Wolf, The Ising model and real magnetic materials, Braz. J. Phys. **30**, 794 (2000).

[63] F. Y. Wu, The Potts model, Rev. Mod. Phys. **54**, 235 (1982).

[64] A. Gupta *et al.*, Structure–property linkages using a data science approach: Application to a nonmetallic inclusion/steel composite system, Acta Mater. **91**, 239 (2015).

[65] P. Debye, H. R. Anderson Jr., and H. Brumberger, Scattering by an Inhomogeneous Solid. II. The Correlation Function and Its Application, J. Appl. Phys. **28**, 679 (1957).

[66] A. M. Gokhale, A. Tewari, and H. Garmestani, Constraints on microstructural two-point correlation functions, Scr. Mater. **53**, 989 (2005).

[67] D. B. Brough, D. Wheeler, and S. R. Kalidindi, Materials knowledge systems in python - a data science framework for accelerated development of hierarchical materials, IMME **6**, 36 (2017).

[68] T. Fast and S. R. Kalidindi, Formulation and calibration of higher-order elastic localization relationships using the MKS approach, Acta Mater. **59**, 4595 (2011).

[69] S. R. Kalidindi, Computationally efficient, fully coupled multiscale modeling of materials phenomena using calibrated localization linkages, ISRN Mater. Sci. **2012**, 305692 (2012).

[70] P. Fernandez-Zelaia, Y. C. Yabansu, and S. R. Kalidindi, A comparative study of the efficacy of local/global and parametric/nonparametric machine learning methods for establishing structure–property linkages in high-contrast 3D elastic composites, IMME **8**, 67 (2019).

[71] S. R. Kalidindi, A Bayesian framework for materials knowledge systems, MRS Commun. **9**, 518 (2019).

[72] J. A. Gomberg, A. J. Medford, and S. R. Kalidindi, Extracting knowledge from molecular mechanics simulations of grain boundaries using machine learning, Acta Mater. **133**, 100 (2017).

[73] N. H. Paulson *et al.*, Reduced-order structure-property linkages for polycrystalline microstructures based on 2-point statistics, Acta Mater. **129**, 428 (2017).

[74] D. T. Fullwood, S. R. Niezgoda, and S. R. Kalidindi, Microstructure reconstructions from 2-point statistics using phase-recovery algorithms, Acta Mater. **56**, 942 (2008).

[75] P.-E. Chen *et al.*, Hierarchical n-point polytope functions for quantitative representation of complex heterogeneous materials and microstructural evolution, Acta Mater. **179**, 317 (2019).

[76] B. Yucel *et al.*, Mining the correlations between optical micrographs and mechanical properties of cold-rolled HSLA steels using machine learning approaches, Integr. Mater. Manuf. Innov. **9**, 240 (2020).

[77] S. R. Niezgoda *et al.*, Optimized structure based representative volume element sets reflecting the ensemble-averaged 2-point statistics, Acta Mater. **58**, 4432 (2010).

[78] C. Nwankpa *et al.*, Activation functions: Comparison of trends in practice and research for deep learning, arXiv:1811.03378 (2018).

[79] A. Paszke *et al.*, PyTorch: An imperative style, high-performance deep learning library, *Advances in Neural Information Processing Systems*, Vol. 32 (2019).

[80] M. Abadi *et al.*, TensorFlow: A system for large-scale machine learning, in *Proceedings of the 12th USENIX conference on Operating Systems Design and Implementation* (USENIX Association, Savannah, GA, 2016), p. 265–283

[81] D. G. Krige, A statistical approach to some basic mine valuation problems on the Witwatersrand, J. South. Afr. Inst. Min. Metall. **52**, 119 (1951).

[82] E. Schulz, M. Speekenbrink, and A. Krause, A tutorial on Gaussian process regression: Modelling, exploring, and exploiting functions, J. Math. Psych. **85**, 1 (2018).

[83] Y. C. Yabansu *et al.*, Application of Gaussian process regression models for capturing the evolution of microstructure statistics in aging of nickel-based superalloys, Acta Mater. **178**, 45 (2019).

[84] Y. C. Yabansu *et al.*, Application of Gaussian process autoregressive models for capturing the time evolution of microstructure statistics from phase-field simulations for sintering of polycrystalline ceramics, Model. Simul. Mater. Sci. Eng. **27**, 084006 (2019).

[85] D. Duvenaud, Automatic model construction with Gaussian processes, Doctoral thesis, University of Cambridge, 2014.

[86] W. E. Leithead and Y. Zhang, O(N 2)-operation approximation of covariance matrix inverse in gaussian process regression based on quasi-newton BFGS method, Commun. Stat. – Simul. Comput. **36**, 367 (2007).

[87] J. Nocedal, Updating quasi-newton matrices with limited storage, Math. Comput. **35**, 773 (1980).

[88] J. A. Nelder and R. Mead, A simplex method for function minimization, Comput. J. **7**, 308 (1965).

[89] M. McIntire, D. Ratner, and S. Ermon, Sparse Gaussian processes for Bayesian optimization, in *Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence* (AUAI Press, Jersey, NJ, 2016), p. 517–526

[90] J. Quiñonero-Candela and C. E. Rasmussen, A unifying view of sparse approximate gaussian process regression, J. Mach. Learn. Res. **6**, 1939 (2005).

[91] R. B. Gramacy and H. K. H. Lee, Bayesian treed gaussian process models with an application to computer modeling, J. Am. Statist. Assoc. **103**, 1119 (2008).

[92] R. B. Gramacy and D. W. Apley, Local gaussian process approximation for large computer experiments, J. Comput. Graph. Statist. **24**, 561 (2015).

[93] A. V. Vecchia, Estimation and model identification for continuous spatial processes, J. R. Stat. Soc. B **50**, 297 (1988).

[94] D. A. Cohn, Neural network exploration using optimal experiment design, Neural Netw. **9**, 1071 (1996).

[95] S. R. Kalidindi *et al.*, Application of data science tools to quantify and distinguish between structures and models in molecular dynamics datasets, Nanotechnology **26**, 344006 (2015).

[96] Y. C. Yabansu and S. R. Kalidindi, Representation and calibration of elastic localization kernels for a broad class of cubic polycrystals, Acta Mater. **94**, 26 (2015).

[97] F. Nie, H. Zhanxuan, and X. Li, An investigation for loss functions widely used in machine learning, Commun. Inf. Sys. **18**, 37 (2018).

[98] D. P. Kingma and J. Ba, Adam: A Method for Stochastic Optimization, arXiv:1412.6980.

[99] B. Weaver *et al.*, Computational enhancements to bayesian design of experiments using gaussian processes, Bayesian Anal. **11**, 191 (2016).

[100] P. K. Kaundinya, K. Choudhary, and S. R. Kalidindi, https://materialsdata.nist.gov/handle/11256/994.