





RESEARCH ARTICLE | APRIL 10 2024

# Multivariate Gaussian process surrogates for predicting basic structural parameters of refractory non-dilute random alloys

Cesar Ruiz   ; Anshu Raj  ; Shuozhi Xu 

 Check for updates

*APL Mach. Learn.* 2, 026107 (2024)

<https://doi.org/10.1063/5.0186045>



**APL Energy**  
**Latest Articles Online!**  
**Read Now**



# Multivariate Gaussian process surrogates for predicting basic structural parameters of refractory non-dilute random alloys

Cite as: APL Mach. Learn. 2, 026107 (2024); doi: 10.1063/5.0186045

Submitted: 3 November 2023 • Accepted: 21 March 2024 •

Published Online: 10 April 2024



View Online



Export Citation



CrossMark

Cesar Ruiz,<sup>1,a)</sup>  Anshu Raj,<sup>2</sup>  and Shuozhi Xu<sup>2</sup> 

## AFFILIATIONS

<sup>1</sup>School of Industrial and Systems Engineering, University of Oklahoma, Norman, Oklahoma 73019, USA

<sup>2</sup>School of Aerospace and Mechanical Engineering, University of Oklahoma, Norman, Oklahoma 73019, USA

<sup>a)</sup> Author to whom correspondence should be addressed: [caruiz@ou.edu](mailto:caruiz@ou.edu)

## ABSTRACT

Refractory non-dilute random alloys consist of two or more principal refractory metals with complex interactions that modify their basic structural properties such as lattice parameters and elastic constants. Atomistic simulations (ASs) are an effective method to compute such basic structural parameters. However, accurate predictions from ASs are computationally expensive due to the size and number of atomistic structures required. To reduce the computational burden, multivariate Gaussian process regression (MVGPR) is proposed as a surrogate model that only requires computing a small number of configurations for training. The elemental atom percentage in the hyper-spherical coordinates is demonstrated to be an effective feature for surrogate modeling. An additive approximation of the full MVGPR model is also proposed to further reduce computations. To improve surrogate accuracy, active learning is used to select a small number of alloys to simulate. Numerical studies based on AS data show the accuracy of the surrogate methodology and the additive approximation, as well as the effectiveness and robustness of the active learning for selecting new alloy designs to simulate.

© 2024 Author(s). All article content, except where otherwise noted, is licensed under a Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>). <https://doi.org/10.1063/5.0186045>

## I. INTRODUCTION

In a dilute random alloy, one or more elements (the solutes) are randomly present in much smaller proportions (usually less than 10%) compared to the primary metal or base (i.e., the solvent).<sup>1</sup> Properties of the dilute alloy are often dominated by the base metal, with the solute making only minor modifications to those properties.<sup>2</sup> However, non-dilute random alloys (NDRAs) consist of two or more principal elements with equal- or near-equal molar ratios that are randomly located at simple lattice sites.<sup>3</sup> Because of the complex interactions among elements, properties of NDRAs are not mere interpolations of those of their constituent elements.<sup>4</sup>

Among the alloy properties, basic structural properties, including lattice parameters and elastic constants, play a key role in controlling alloy structures and properties.<sup>5</sup> The lattice parameters define the unit cell of an alloy; by knowing them, one can identify the crystal system and the specific crystal structure. The elastic constants give insights into an alloy's stiffness, toughness, strength,

and phase stability under different loading conditions.<sup>6</sup> In addition, all structural properties can change due to the presence of lattice defects and/or thermal expansion.<sup>7</sup> Thus, monitoring changes in them can provide information on internal defect structures and/or thermal properties such as specific heat capacity and melting point. Understanding basic structural parameters aids in predicting and tailoring an alloy's properties, ensuring optimal performance in target applications.<sup>8</sup>

Atomistic simulations (ASs) are an effective method to compute the basic structural parameters. ASs create a representation of atomic configuration within an alloy and then use interatomic potentials to evaluate the interactions among atoms.<sup>9</sup> However, direct, accurate computation of basic structural parameters of even a single random alloy necessitates carrying out ASs of either a small set of large structures or a large set of small structures,<sup>10</sup> both of which can be computationally intensive.

To reduce the computational burden when studying the properties of alloys, machine learning (ML) is an attractive alternative for

surrogate modeling due to its flexibility, low computational cost, and relatively low complexity of use. In the literature, several regression models have been proposed as surrogates for the prediction of key alloy properties.<sup>11</sup> For example, Mohanty *et al.*<sup>12</sup> proposed boosted regression trees to design alloys with good plastic properties for high-temperature applications; Wang *et al.*<sup>13</sup> use neural networks to predict tensile strength and electrical conductivity of copper alloys; Linton and Aidhy<sup>14</sup> use linear, gradient-boosted, and random forest regression to predict elastic properties of multi-principal element alloys (MPEAs). For the prediction of general material properties, deep learning methods have successfully predicted the material properties based on composition and structural features.<sup>15,16</sup>

However, most of these studies rely on a large number of samples, i.e., alloys, to construct and train the relatively complex surrogate models. Since our objective is to limit the number of simulations as much as possible, multivariate Gaussian process regression (MVGPR) is proposed to construct the surrogate model for basic structural properties. GPR is a popular ML methodology for surrogate modeling of smooth functions,<sup>17</sup> making them ideal for the prediction of materials properties.<sup>18</sup> In addition, Gaussian processes are the underlying statistical methodology used in Bayesian optimization, which has been successfully utilized for alloy design optimization of mechanical or structural properties by sequentially exploring the design space.<sup>19–22</sup> Different from such methodologies for alloy design, active learning<sup>17</sup> will be employed to reduce the overall prediction uncertainty of the surrogate by sequentially choosing new alloy configurations to simulate.

The rest of this paper is organized as follows: Sec. II introduces the MVGPR methodology for surrogate modeling of average structural properties and active learning for the sequential design of experiments to improve the surrogate accuracy. Section III demonstrates the prediction accuracy and robustness of the MVGPR with active learning for selecting the next alloy design to simulate. Section IV summarizes this work and discusses future research directions.

## II. MVGPR FOR SURROGATE MODELING

### A. Univariate surrogate model

GPR is a popular non-parametric ML technique that assumes the value of an output function  $y$  follows a Gaussian distribution with correlation given by the distance of the predictors that belong to a possibly multi-dimensional space  $\mathcal{X}$ .<sup>23</sup> Therefore, for a given set  $X$  of  $N$  predictor points, the distribution of the output  $y$  is

$$y \sim \mathcal{N}(\mu, \Sigma), \tag{1}$$

where  $\mu$  is the mean function evaluated at  $X$  and  $\Sigma$  is an  $N \times N$  variance-covariance matrix. In general, the mean function is modeled as  $\mu(x) = b^T(x)\beta$  for a vector of basis functions  $b$ , and a vector of coefficients  $\beta$ . The special case with  $b(x) = 1$  indicates a constant mean function with a value of  $\beta$ . The elements  $\Sigma_{ij}$  are calculated as  $\sigma^2 \kappa(x_i, x_j | \theta)$ , where  $\sigma$  is the standard deviation parameter,  $\kappa$  is a kernel function parameterized by  $\theta$  that measures the correlation between points  $x_i, x_j$  in the input space  $\mathcal{X}$ . In practice,  $\kappa$  is often chosen to be stationary such that it is only a function of the distance between predictors. This assumption reduces the number of

parameters to estimate and improves computational efficiency at the expense of modeling flexibility.

Given an estimate of the parameters  $\beta, \sigma$ , and  $\theta$ , the prediction of  $y^*$  and  $\Sigma^*$  on a new set of locations  $X^*$  are given by the following universal Kriging equations:<sup>24</sup>

$$\begin{aligned} y^* &= \hat{\mu}^* + R^T K^{-1} (y - \hat{\mu}), \\ \Sigma^* &= \hat{\sigma}^2 \left[ \hat{K}^* - \hat{R}^T \hat{K}^{-1} \hat{R} + (B^* - \hat{R}^T \hat{K}^{-1} B)^T \right. \\ &\quad \left. \times (B^T \hat{K}^{-1} B)^{-1} (B^* - \hat{R}^T \hat{K}^{-1} B) \right] \end{aligned} \tag{2}$$

where  $\hat{\mu}^*$  and  $\hat{\mu}$  are the mean of the new and old predictors, respectively, computed using  $\hat{\beta}$ ;  $\hat{K}$  is the correlation between points in  $X$ ;  $\hat{K}^*$  is the correlation between points in  $X^*$ ; and  $\hat{R}$  is the correlation matrix between  $X$  and  $X^*$ , all computed with  $\hat{\theta}$ .

For the prediction of structural properties in refractory NDRAs, the simplest input predictor is the concentration of each atom in the alloy  $\tilde{x} \in [0, 1]^m$ , where  $m$  is the number of possible elements considered; this feature is also known as the elemental atom percentage.<sup>12</sup> However, the direct use of this representation can lead to numerical instability due to rank deficiency of  $\tilde{X}$  since the total concentration of all elements has to add up to one. To reduce the numerical instability of ML regression models, the elemental atom percentage is transformed to the  $(m - 1)$  hyper-spherical coordinates using the following equation:

$$\begin{aligned} x_i &= \tan^{-1} \left( \frac{\sqrt{\tilde{x}_{i+1} + \tilde{x}_{i+1}, \dots + \tilde{x}_m}}{\tilde{x}_i} \right), \quad i = 1, \dots, m - 2, \\ x_{m-1} &= 2 \tan^{-1} \left( \frac{\tilde{x}_m}{\tilde{x}_{m-1} + \sqrt{\tilde{x}_{m-1} + \tilde{x}_m}} \right). \end{aligned} \tag{3}$$

### B. Multivariate GPR and parameter estimation

The AS for each alloy configuration outputs an  $n$ -dimensional vector of basic structural properties  $y \in \mathbb{R}^n$ . To predict this vector, a constant mean is assumed for each basic structural property such that, for each configuration,  $\mu = I_n \beta$ , where  $I_n$  is the  $n$ -dimensional identity matrix and  $\beta \in \mathbb{R}^n$ . For a set of alloy configurations, the covariance matrix will be given by  $\Sigma = K \otimes \Omega$ , where  $\Omega$  is the variance-covariance matrix between structural properties and  $\otimes$  is the standard Kronecker product. This is equivalent to a separable correlation structure, where  $K$  represents the common correlation across alloy configurations, while  $\Omega$  represents the correlation across basic structural properties.

The training dataset has  $N$  distinct simulated alloy configurations. Let  $1_N$  be the  $N$ -dimensional vector of ones. Then,  $B = 1_N \otimes I_n$  and the parameters of the model can be estimated by minimizing the scaled profiled negative log-likelihood,

$$\begin{aligned} \min_{L_\Omega, \theta} \log(2\pi|\Sigma|) + (y - B\hat{\beta})^T \Sigma^{-1} (y - B\hat{\beta}), \\ \hat{\beta} &= [B^T \hat{\Sigma}^{-1} B]^{-1} B^T \hat{\Sigma}^{-1} y, \\ \Sigma &= K(X, \theta) \otimes L_\Omega L_\Omega^T, \end{aligned} \tag{4}$$

where  $L_\Omega$  is the Cholesky factor of  $\Omega$ . The Cholesky factor parameterization improves the numerical stability of optimization methods commonly used for parameter estimation of Gaussian processes.<sup>23,25</sup>

### C. GPR approximation as an additive model

The GPR model for a single structural property  $j$  can be rewritten as an additive model of the following form:

$$y_i = \beta_j + \eta_j(X) + \varepsilon, \quad (5)$$

where  $\eta$  is a mean zero GP with covariance kernel  $\kappa$ . If we define the matrix  $A_K$  as the decomposition  $A_K A_K^T = K$  (such as a Cholesky decomposition), the GPR problem can be transformed into a linear regression model of the following form:<sup>25</sup>

$$y_i = \beta_j + A_K(X, \theta)\alpha + \varepsilon, \quad (6)$$

where  $\alpha$  is a vector of random effects with mean zero and variance  $\sigma_j^2$ . Note that this model still has the same statistical distribution as standard one-dimensional GPR, but is expressed as a hierarchical model, see Chapter 4.4 of Wikle *et al.*<sup>26</sup> for more details.

To reduce the computational cost of estimating  $\theta$  or introducing the knowledge of the covariance of the output,<sup>26</sup> the matrix  $A_K$  can be approximated using basis functions expansion as  $A_K \approx \Psi(X)$  with each row  $i$  as a basis expansion for alloy configuration  $i$ . This regression problem can be efficiently solved using penalized least squares<sup>27</sup> of the following form:

$$\min_{\sigma_j^2, \beta_j, \alpha_j} |y_j - \beta_j - \Psi\alpha_j|^2 + \sigma_j^2 \alpha_j^T S \alpha_j, \quad (7)$$

where  $S$  is a penalty matrix constructed based on the complexity of the basis functions. To construct the basis functions  $\psi$ , the Kronecker product of univariate basis is used such that<sup>27</sup>

$$\psi(x) = \text{vec}\left(\psi^{(1)}(x_1) \otimes \cdots \otimes \psi^{(m-1)}(x_{m-1})\right), \quad (8)$$

where  $\text{vec}$  is the vector form of a matrix. Note that such construction allows for automatic modeling of non-stationary patterns<sup>26</sup> in the data that require more complex kernel functions in the standard GPR. The predictive mean and standard error can be computed as follows:<sup>28</sup>

$$\begin{aligned} y_j^* &= \hat{\beta}_j + \Psi(X^*)\hat{\alpha}_j, \\ \Sigma_{jj}^* &= \sigma_j^2 V_{jj}, \end{aligned} \quad (9)$$

where  $V_{jj}$  is the Bayesian posterior covariance of the parameter estimates of  $\beta_j$  and  $\alpha_j$ . Such estimates are readily available in the R package *mgcv*.<sup>27</sup>

There are several advantages of the GPR additive approximation (GPRAA) model: (1) it reduces the number of parameters to be estimated; (2) there are several implementations of such models reducing the adoption barriers for practitioners; (3) the computational cost increases based on the number of basis functions rather than the number of data points (i.e., inverting  $\Sigma^{-1}$ ); and (4) it is easier to encode expert knowledge in the choice of basis functions than in the construction of custom covariance kernel functions. However, the main drawback is that information is not shared between structural properties since the  $\sigma_j^2$  is equivalent to imposing a diagonal structure to  $\Omega$  in the full MVGPR model.

### D. Active learning to minimize prediction uncertainty

Active learning with surrogate models refers to the sequential design of experiments (DOE) to reduce the prediction error and/or uncertainty. The sequential design allows for the reduction of the total number of alloys to simulate.<sup>29</sup> The simplest criterion for sequential DOE is to find the design that maximizes the uncertainty of the current surrogate model, i.e., the predictive variance for the univariate case or the determinant of the predictive variance-covariance matrix  $|\Sigma|^*$  for the multivariate case.<sup>17</sup> However, this criterion is not well suited for prediction with multiple structural properties with different scales, even orders of magnitude, since the design will be dominated by uncertainty of the structural property with the largest scale. Therefore, sequential design is carried out in this paper by maximizing the scaled determinant of the predictive covariance as

$$x_{N+1} = \arg \max_{x \in \mathcal{X}} |M(x)^T \Sigma^*(x) M(x)| \quad (10)$$

where  $M$  is a diagonal matrix with entries  $1/y^*$  at  $x$ . This determinant is equivalent to the squared coefficient of variation for the univariate case. This nonlinear optimization problem can be solved using numerical search methods or by maximizing the predictive uncertainty on a set of candidate alloys. Note that  $\Sigma^*$  and  $y^*$  are computed as in Eq. (2) for the full MVGPR model and as in Eq. (9) for the GPRAA.

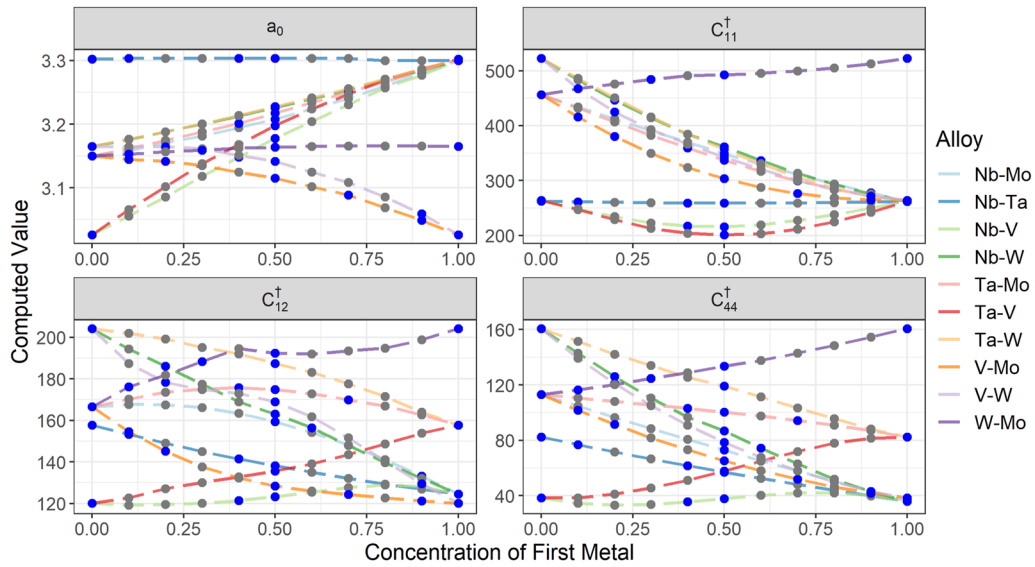
## III. NUMERICAL STUDY

There are six body-centered cubic (BCC) refractory metals in total: Cr, Mo, Nb, Ta, V, and W. However, only the last five are used to construct refractory NDRAAs because ASs do not consider Cr's anti-ferromagnetism, which is key for computing basic structural parameters of magnetic alloys.<sup>30</sup> We emphasize that the five elements considered here—Mo, Nb, Ta, V, and W—have complete mutual solubility with respect to each other,<sup>31–35</sup> meaning that any percentile of any of the five pure metals in any alloy that is based on them is possible.

### A. Generated dataset

The dataset of basic structural properties was compiled from three sources: (1) the study by Xu *et al.*<sup>10</sup> for 16 equal-molar MPEAs, i.e., an equal proportion of every principal element in a ternary, a quaternary, or a quinary; (2) the study by Mamun *et al.*<sup>36</sup> that investigated equal- or unequal-molar Mo-W, Nb-W, and Ta-W random binary alloys; (3) the remaining seven sets of random binaries are newly calculated in this paper. The new calculations largely follow Mamun *et al.*<sup>10</sup> As a result, the dataset for this study contains the computations for 5 pure metals, 90 binaries, 10 ternaries, 5 quaternaries, and 1 quinary, for a total of 111 unique metallic materials. For each alloy, ASs are used to compute the lattice parameter  $a_0$  and the effective BCC elastic constants  $C_{11}^\dagger$ ,  $C_{12}^\dagger$ , and  $C_{44}^\dagger$  based on the stiffness tensor  $C$ .<sup>37</sup> We remark that all atomistic simulations, old or new, use the same interatomic potential,<sup>10</sup> making the data self-consistent.

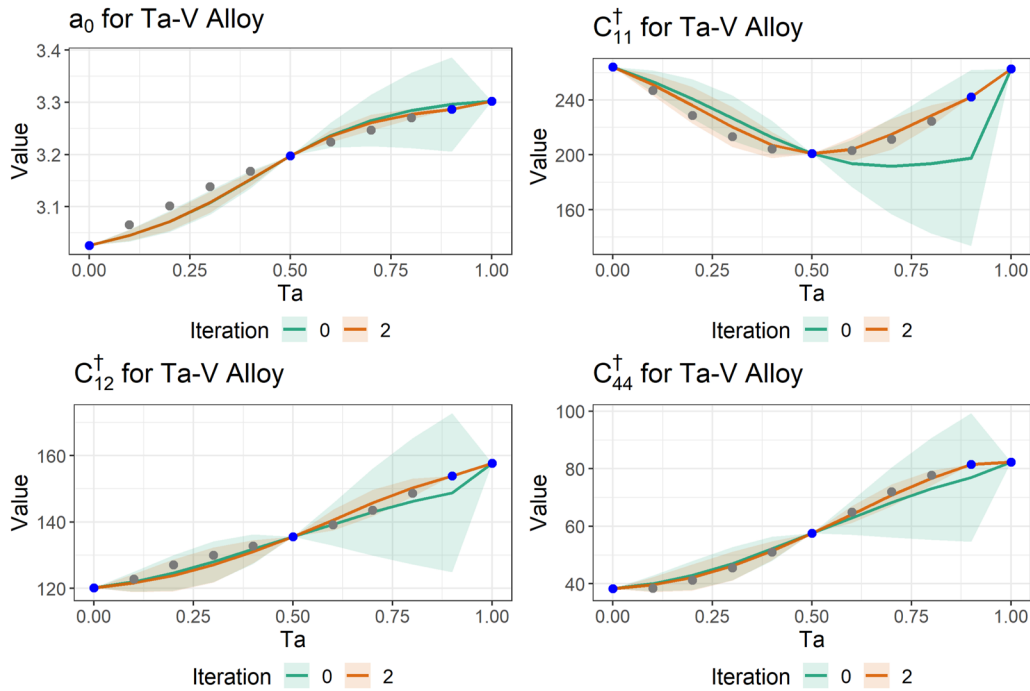
Figure 1 shows the computed structural properties for all ten types of binary alloys. The difficulty of creating surrogate models for the structural properties with a few data points increases with the complexity and heterogeneity of the property profile as a function



**FIG. 1.** Computed structural properties for binary alloys combining two of the five refractory metals: Mo, Nb, Ta, V, and W. The x axis shows the concentration of the first metal in each set of binary alloys. Points in blue are used for the initial training dataset.

of element concentration. From the figure, the model for  $a_0$  should be the simplest since most profiles seem to follow a linear relationship while the model for  $C_{12}^\dagger$  is the most complex due to the different levels of smoothness of the functions. For  $C_{11}^\dagger$ , the addition of V,

which has the smallest  $a_0$  and  $C_{12}^\dagger$  among all pure metals, to the alloy seems to cause a reduction in structural properties below to any of the two pure metals, while for other alloys, this value is always between that of the two pure constituents. Due to the heterogeneity



**FIG. 2.** Predicted value of the basic structural properties of Ta-V binaries and their 95% confidence interval using the MVGPR surrogate. The 0.9 Ta-0.1 V binary was added to the training set in the second iteration of the active learning.

10 April 2024 14:03:10

of the functional profiles, a negative transfer of knowledge between basic structural properties may occur in the surrogate models.

### B. Surrogate model training and prediction results

The data for the computed structural properties were partitioned into training and testing sets. The training set contains the basic structural properties for pure metals, equal-molar binary alloys, and the equal-molar MoTaV, NbTaW, NbVW, MoNbW, MoNbVW, and MoNbTaW alloys that provide information for the three types of W-based binaries studied by Mamun *et al.*<sup>36</sup> In addition, a random sample of 20% of the remaining alloys was added to the initial training set for a total of 37 alloys and 5 pure metals. The points included in the initial training set are shown in blue in Fig. 1. The initial training set was expanded through active learning by sequentially selecting five more alloy designs from the test dataset based on the current working model. In each iteration, the datasets were modified by removing the selected point from the test and adding it to the training set before re-training the surrogate model.

The full MVGPR model is constructed using the separable Matérn kernel function between points  $x_i$  and  $x_{i'}$  parameterized as

$$\kappa(x_i, x_{i'}) = \prod_{j=1}^{m-1} \exp\left(-\frac{d_j}{\theta_j}\right) \left(1 + \frac{d_j}{\theta_j} + \frac{5}{3} \frac{d_j^2}{\theta_j^2}\right), \quad (11)$$

where  $d_j = \sqrt{(x_{ij} - x_{i'j})^2}$  is the distance and  $\theta_j$  is the length-scale parameter in the  $j$ th dimension. Due to their capability to approximate any function, the GPRAA model is constructed using

four cubic b-spline basis functions<sup>27</sup> for each spherical coordinate. The penalized regression problem is solved using generalized cross-validation.

For illustration purposes, Fig. 2 shows the prediction and uncertainty for the structural properties of Ta-V binary alloys using the full MVGPR surrogate model. The uncertainty is measured by the point-wise confidence interval based on the posterior standard deviation of the basic structural properties. Around the 90% Ta region, the initial prediction is quite biased due to negative knowledge transfer between basic structural properties and has high uncertainty due to lack of data. Consequently, that point is selected for simulation during the first iteration of the active learning algorithm, and the local prediction accuracy improves and uncertainty is reduced by the second iteration. Since the active learning algorithm seeks to improve the global accuracy of the surrogate model, additional points to simulate are selected in regions of the design space far away from the Ta-V binary, resulting in no further improvement to the prediction accuracy of the Ta-V binaries in subsequent iterations. Figure 3 shows the equivalent results using the GPRAA surrogate. The predictions based on the initial training set have a fairly piece-wise linear pattern due to the lack of data points in this region of the design space. As more data points are added, the complexity of the prediction pattern increases. The main difference between the two surrogate models is the higher uncertainty on the full MVGPR model, which allows it to cover most of the true property values within its confidence interval while the additive approximation has severe coverage issues in the 0%–50% Ta region for all structural properties.

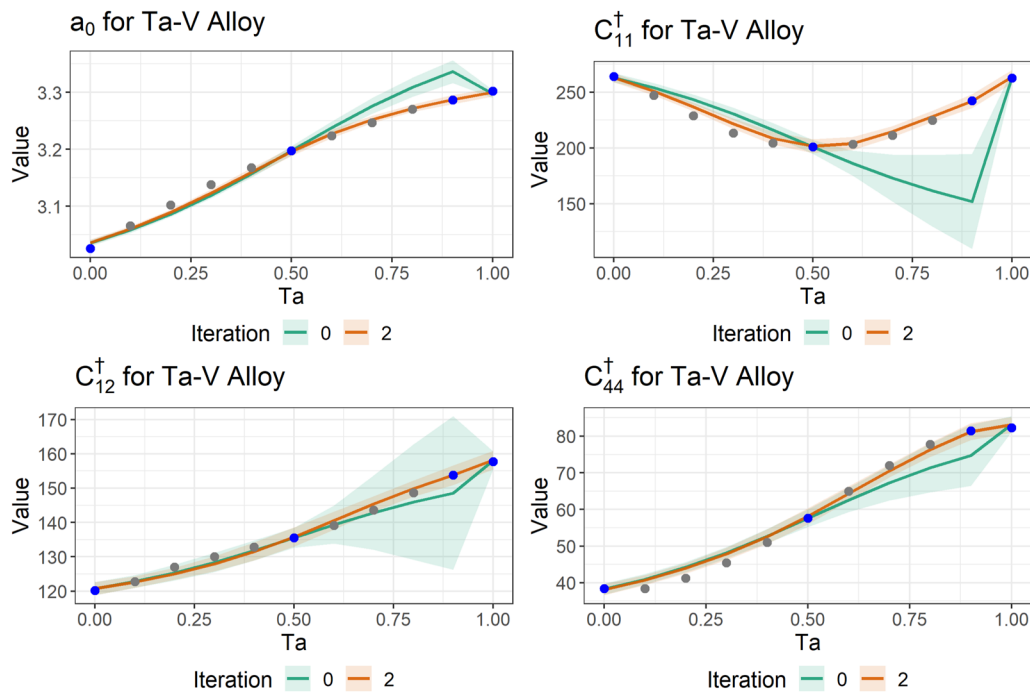
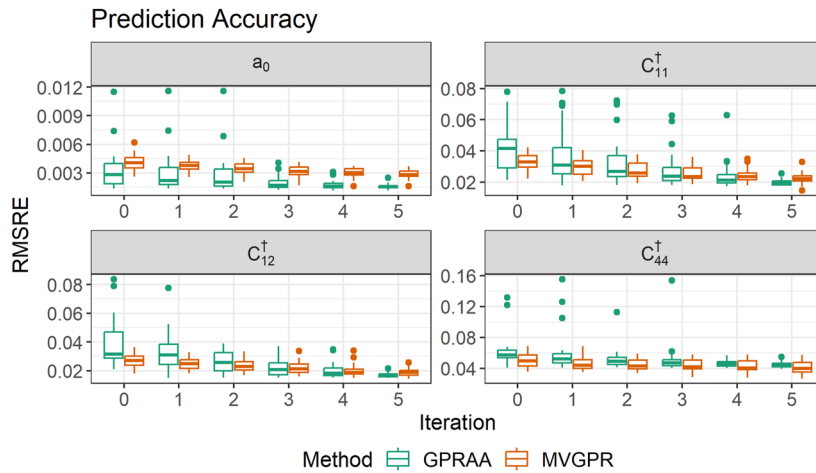


FIG. 3. Predicted value of the basic structural properties of Ta-V binaries and their 95% confidence interval using the GPRAA surrogate. The 0.9 Ta-0.1 V binary was added to the model in the second iteration of the active learning.



**FIG. 4.** Prediction accuracy measured by the RMSRE on the test set after adding new points to the training set; the experiment was repeated 30 times with different training sets.

The overall prediction accuracy for each structural property is measured by the root mean squared relative error (RMSRE),

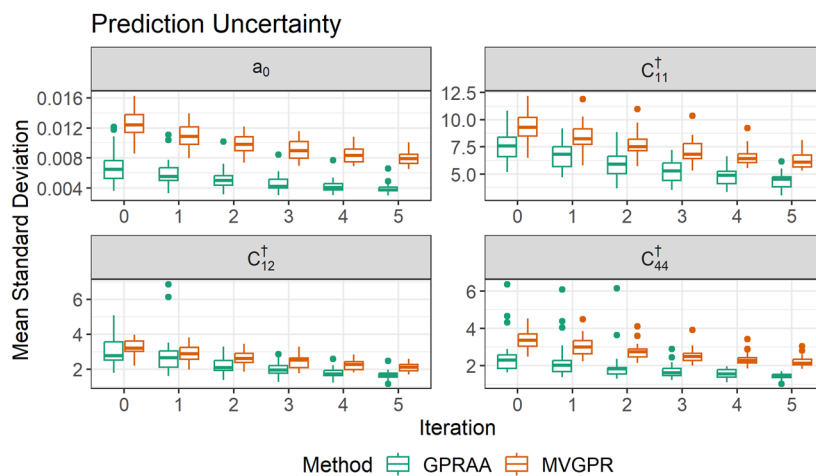
$$\text{RMSRE} = \sqrt{\frac{1}{L} \sum_{\ell=1}^L \left( \frac{\hat{y}_\ell - y_\ell}{y_\ell} \right)^2}, \quad (12)$$

where  $y$  is the true value and  $\hat{y}$  is the estimate. The prediction uncertainty is measured by the mean standard deviation for each model. Both are evaluated on the testing set during each iteration, i.e., reduced set due to moving one point to the training set during each iteration.

The active learning experiment was repeated 30 times with a different set of initial data. Each initial set was constructed by drawing a new sample of the 20% of remaining alloys described

at the beginning of this subsection. Figure 4 shows the boxplot of the RMSRE for all structural properties under the full and additive approximation surrogate models in each iteration of the active learning algorithm. The full MVGPR model performs worse for  $a_0$ , while it does better for  $C_{44}^\dagger$  than the additive approximation GPRAA. This was expected since the  $a_0$  shows a mostly linear relationship, while  $C_{44}^\dagger$  shows more complex patterns on the binary alloys (see Fig. 1), which represent the bulk of the dataset. For  $C_{11}^\dagger$  and  $C_{12}^\dagger$  both models have a similar accuracy with the additive approximation being slightly more accurate.

Figure 5 shows the model uncertainty measured by averaging the point-wise standard deviation of predictions on the test dataset. While both models reduce their uncertainty at roughly the same pace, the additive approximation consistently shows less



**FIG. 5.** Prediction uncertainty measured by the mean  $\Sigma_{jj}^*$  on the test set after adding new points to the training set; experiment repeated 30 times with different initial training sets.

uncertainty, which can lead practitioners to believe it is a better model. However, as illustrated in Figs. 2 and 3, the 95% point-wise confidence intervals often do not cover the true value with only a few simulation experiments. This is caused by significant bias introduced due to the significant difference between the monotone patterns observed in other binary alloys and the convex pattern in the Ta-V data, see Fig. 1. This behavior suggests that the uncertainty due to data heterogeneous behavior is better captured in the MVGPR model than by the additive approximation, which does not share information across structural properties. Note that this might not affect the active learning procedure as long as the uncertainty is equally underestimated for all alloys, as is the case in Figs. 2 and 3.

#### IV. CONCLUSIONS

In this paper, we constructed two ML surrogate models to predict basic structural parameters in refractory NDRAs, which have intricate interactions between elements and are crucial in modern alloy design. The MVGPR is introduced as a surrogate model, offering a more efficient alternative to traditionally computationally expensive ASs. The approach of using the elemental atom percentage in hyper-spherical coordinates as a feature for surrogate modeling offers enhanced computational stability and potentially improved prediction accuracy. MVGPR represents a single model to predict all material properties of interest. The information shared across structural properties allows for conservative uncertainty quantification on the prediction of new alloy properties. To reduce the computational cost, the additive approximation model utilizes the basis expansion to mimic the correlation structure of individual properties. However, the uncertainty quantification is overoptimistic. Active learning further refines the surrogate model, ensuring that only a few select, most informative alloy configurations need to be simulated. The combination of physics-based and data-driven models makes this paper a valuable resource for materials scientists seeking efficient, accurate tools for the prediction of alloy basic structural properties.

There are many exciting directions for further research. First, the model can be extended to incorporate information on the atomistic configuration to improve the accuracy of the surrogate and enhance the uncertainty quantification of the prediction. Second, transfer learning between different physic-based computation approaches can be explored. A common transfer learning approach is to model discrepancies between predictions of heterogeneous sources.<sup>38,39</sup> For example, by learning the discrepancy between ASs and density functional theory (DFT) predictions, it is possible to utilize the computationally cheapest method between AS or DFT to compute an initial estimate of basic properties and then adjust it by the predicted discrepancy.

#### ACKNOWLEDGMENTS

This research was supported by seed funding from the University of Oklahoma (OU) Data Institute for Societal Challenges (DISC). In addition, we are grateful for the startup funds provided by OU. This work used Bridges-2 at the Pittsburgh Supercomputing Center through Allocation No. MAT230058 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support

(ACCESS) program, which is supported by the National Science Foundation (Grant Nos. 2138259, 2138286, 2138307, 2137603, and 2138296). Some of the computing for this project was performed at the OU Supercomputing Center for Education & Research (OSKER) at OU.

#### AUTHOR DECLARATIONS

##### Conflict of Interest

The authors have no conflicts to disclose.

##### Author Contributions

**Cesar Ruiz:** Conceptualization (equal); Methodology (lead); Software (lead); Visualization (lead); Writing – original draft (lead); Writing – review & editing (lead). **Anshu Raj:** Data curation (lead); Formal analysis (equal); Investigation (equal). **Shuozhi Xu:** Conceptualization (equal); Funding acquisition (lead); Methodology (equal); Resources (lead); Validation (equal); Writing – review & editing (equal).

##### DATA AVAILABILITY

The data and ML code to generate the results presented in this paper are publicly available at [https://github.com/caruzto/mvgpr\\_as](https://github.com/caruzto/mvgpr_as). All files used in atomistic simulations to generate new data in this paper can be found at [https://github.com/shuozhixu/Binary\\_2024](https://github.com/shuozhixu/Binary_2024).

#### REFERENCES

- 1 A. Brezini, R. Bouamrane, F. Hamdache, and C. Depolier, "Theoretical model of the density of states for random dilute binary alloys," *Phys. Status Solidi B* **188**, 697–710 (1995).
- 2 M. Laurent-Brocq, L. Perrière, R. Pirès, F. Prima, P. Vermaut, and Y. Champion, "From diluted solid solutions to high entropy alloys: On the evolution of properties with composition of multi-components alloys," *Mater. Sci. Eng.: A* **696**, 228–235 (2017).
- 3 S. Nag and W. A. Curtin, "Effect of solute-solute interactions on strengthening of random alloys from dilute to high entropy alloys," *Acta Mater.* **200**, 659–673 (2020).
- 4 A. Bansil, L. Schwartz, and H. Ehrenreich, "The electronic structure of non-dilute alloys," *Phys. Condens. Matter* **19**, 391–403 (1975).
- 5 S. A. Khandy, I. Islam, D. C. Gupta, R. Khenata, A. Laref, and S. Rubab, "DFT understandings of structural properties, mechanical stability and thermodynamic properties of BaCfO<sub>3</sub> perovskite," *Mater. Res. Express* **5**, 105702 (2018).
- 6 Y. Su, M. Ardeljan, M. Knezevic, M. Jain, S. Pathak, and I. J. Beyerlein, "Elastic constants of pure body-centered cubic Mg in nanolaminates," *Comput. Mater. Sci.* **174**, 109501 (2020).
- 7 P. K. Tripathi, Y.-C. Chiu, S. Bhowmick, and Y.-C. Lo, "Temperature-dependent superplasticity and strengthening in CoNiCrFeMn high entropy alloy nanowires using atomistic simulations," *Nanomaterials* **11**, 2111 (2021).
- 8 A. J. Zaddach, C. Niu, C. C. Koch, and D. L. Irving, "Mechanical properties and stacking fault energies of NiFeCrCoMn high-entropy alloy," *JOM* **65**, 1780–1789 (2013).
- 9 E. Tadmor and R. Miller, *Modeling Materials: Continuum, Atomistic and Multiscale Techniques*, 1st ed. (Cambridge University Press, Cambridge, NY, 2012).



- <sup>10</sup>S. Xu, S. Z. Chavoshi, and Y. Su, “On calculations of basic structural parameters in multi-principal element alloys using small atomistic models,” *Comput. Mater. Sci.* **202**, 110942 (2022).
- <sup>11</sup>X. Liu, P. Xu, J. Zhao, W. Lu, M. Li, and G. Wang, “Material machine learning for alloys: Applications, challenges and perspectives,” *J. Alloys Compd.* **921**, 165984 (2022).
- <sup>12</sup>T. Mohanty, K. Chandran, and T. D. Sparks, “Machine learning guided optimal composition selection of niobium alloys for high temperature applications,” *APL Mach. Learn.* **1**, 036102 (2023).
- <sup>13</sup>C. Wang, H. Fu, L. Jiang, D. Xue, and J. Xie, “A property-oriented design strategy for high performance copper alloys via machine learning,” *npj Comput. Mater.* **5**, 87 (2019).
- <sup>14</sup>N. Linton and D. S. Aidhy, “A machine learning framework for elastic constants predictions in multi-principal element alloys,” *APL Mach. Learn.* **1**, 016109 (2023).
- <sup>15</sup>R. E. Goodall and A. A. Lee, “Predicting materials properties without crystal structure: Deep representation learning from stoichiometry,” *Nat. Commun.* **11**, 6280 (2020).
- <sup>16</sup>A. Y.-T. Wang, S. K. Kauwe, R. J. Murdock, and T. D. Sparks, “Compositionally restricted attention-based network for materials property predictions,” *npj Comput. Mater.* **7**, 77 (2021).
- <sup>17</sup>R. B. Gramacy, *Surrogates: Gaussian Process Modeling, Design, and Optimization for the Applied Sciences* (CRC Press, 2020).
- <sup>18</sup>V. L. Deringer, A. P. Bartók, N. Bernstein, D. M. Wilkins, M. Ceriotti, and G. Csányi, “Gaussian process regression for materials and molecules,” *Chem. Rev.* **121**, 10073–10141 (2021).
- <sup>19</sup>F. Tancret, I. Toda-Caraballo, E. Menou, P. E. J. Rivera Díaz-Del-Castillo *et al.*, “Designing high entropy alloys employing thermodynamics and Gaussian process statistical analysis,” *Mater. Des.* **115**, 486–497 (2017).
- <sup>20</sup>S. Karumuri, Z. D. McClure, A. Strachan, M. Titus, and I. Bilonis, “Hierarchical Bayesian approach to experimental data fusion: Application to strength prediction of high entropy alloys from hardness measurements,” *Comput. Mater. Sci.* **217**, 111851 (2023).
- <sup>21</sup>M. Hu, Q. Tan, R. Knibbe, M. Xu, B. Jiang, S. Wang, X. Li, and M.-X. Zhang, “Recent applications of machine learning in alloy design: A review,” *Mater. Sci. Eng., R* **155**, 100746 (2023).
- <sup>22</sup>N. Chiba, K. Masuda, K.-i. Uchida, and Y. Miura, “Designing composition ratio of magnetic alloy multilayer for transverse thermoelectric conversion by Bayesian optimization,” *APL Mach. Learn.* **1**, 026114 (2023).
- <sup>23</sup>C. K. Williams and C. E. Rasmussen, *Gaussian Processes for Machine Learning* (MIT Press, Cambridge, MA, 2006), Vol. 2.
- <sup>24</sup>O. Roustant, D. Ginsbourger, and Y. Deville, “DiceKriging, DiceOptim: Two R packages for the analysis of computer experiments by Kriging-based metamodeling and optimization,” *J. Stat. Software* **51**, 1–55 (2012).
- <sup>25</sup>D. Bates, M. Mächler, B. Bolker, and S. Walker, “Fitting linear mixed-effects models using lme4,” *J. Stat. Software* **67**, 1–48 (2015).
- <sup>26</sup>C. K. Wikle, A. Zammit-Mangion, and N. Cressie, *Spatio-Temporal Statistics with R* (CRC Press, 2019).
- <sup>27</sup>S. N. Wood, *Generalized Additive Models: An Introduction with R* (CRC Press, 2017).
- <sup>28</sup>G. Marra and S. N. Wood, “Coverage properties of confidence intervals for generalized additive model components,” *Scand. J. Stat.* **39**, 53–74 (2012).
- <sup>29</sup>A. Sauer, R. B. Gramacy, and D. Higdon, “Active learning for deep Gaussian process surrogates,” *Technometrics* **65**, 4–18 (2023).
- <sup>30</sup>S. Xu, A. S. Kulathuvayal, L. Xiong, and Y. Su, “Effects of ferromagnetism in ab initio calculations of basic structural parameters of Fe-A (A = Mo, Nb, Ta, V, or W) random binary alloys,” *Eur. Phys. J. B* **95**, 167 (2022).
- <sup>31</sup>J. F. Smith and O. N. Carlson, “The Nb–V (niobium–vanadium) system,” *Bull. Alloy Phase Diagrams* **4**, 46–49 (1983).
- <sup>32</sup>F. Zheng, B. B. Argent, and J. F. Smith, “Thermodynamic computation of the Mo–V binary phase diagram,” *J. Phase Equilib.* **20**, 370 (1999).
- <sup>33</sup>H. Okamoto, “Ta–V (tantalum–vanadium),” *J. Phase Equilib. Diffus.* **26**, 298–300 (2005).
- <sup>34</sup>H. Okamoto, “V–W (vanadium–tungsten),” *J. Phase Equilib. Diffus.* **31**, 324 (2010).
- <sup>35</sup>M. Widom, W. P. Huhn, S. Maiti, and W. Steurer, “Hybrid Monte Carlo/molecular dynamics simulation of a refractory metal high entropy alloy,” *Metall. Mater. Trans. A* **45**, 196–200 (2014).
- <sup>36</sup>A. A. Mamun, S. Xu, X.-G. Li, and Y. Su, “Comparing interatomic potentials in calculating basic structural parameters and Peierls stress in tungsten-based random binary alloys,” *Phys. Scr.* **98**, 105923 (2023).
- <sup>37</sup>S. Xu, E. Hwang, W.-R. Jian, Y. Su, and I. J. Beyerlein, “Atomistic calculations of the generalized stacking fault energies in two refractory multi-principal element alloys,” *Intermetallics* **124**, 106844 (2020).
- <sup>38</sup>A. Sabbaghi, Q. Huang, and T. Dasgupta, “Bayesian model building from small samples of disparate data for capturing in-plane deviation in additive manufacturing,” *Technometrics* **60**, 532–544 (2018).
- <sup>39</sup>Y. Wang, C. Ruiz, and Q. Huang, “Learning and predicting shape deviation of smooth and non-smooth 3D geometries through mathematical decomposition of additive manufacturing,” *IEEE Trans. Autom. Sci. Eng.* **20**, 1527 (2023).