Statistical approach to obtaining vacancy formation energies in high-entropy crystals from first principles calculations: Application to a high-entropy diboride

S. E. Daigle and D. W. Brenner
Department of Materials Science and Engineering, North Carolina State University, Raleigh, North Carolina 27695, USA

(Received 17 June 2020; accepted 25 September 2020; published 7 December 2020)

A reduced pair approximation model for vacancy formation energy in multicomponent materials is proposed as an alternative to the commonly used cluster expansion method. By imposing physical constraints to the interaction coefficients, lower rank models are obtained with improved accuracy as measured by Bayesian information criterion and cross validation. Additionally, reduced models can outperform the full parametrization in high-entropy compounds with as much as 50% less training data. The results are presented for cation vacancies in the high-entropy diboride $Hf_{0.2}Zr_{0.2}Ti_{0.2}Ta_{0.2}Nb_{0.2}B_2$ calculated by density functional theory simulations of large cell special quasirandom structures. Further, the calculation of vacancy concentrations from a distribution of energies is considered, wherein the chemical disorder on lattice sites gives rise to non-Arrhenius temperature dependence. Preferential clustering and the possibility of short-range order in the high-entropy lattice are explored through pair affinities derived from model coefficients.

DOI: 10.1103/PhysRevMaterials.4.123602

I. INTRODUCTION

High-entropy ceramics, a relatively recent variation on high-entropy alloys, have emerged as a new class of material with unique and potentially important functional and structural properties. Possible applications for these materials include catalysis, barriers for thermal and environmental protection, substrates for water splitting, and as energy storage materials [1]. The high-entropy ceramics reported to date have two sublattices: an ordered sublattice with one or two species, and a second sublattice with multiple, randomly arranged species. High-entropy rocksalt oxides [2] and carbides [3], for example, consist of an fcc sublattice occupied by oxygen or carbon, respectively, and a chemically disordered interpenetrating fcc sublattice containing multiple metallic species. Other examples of high-entropy ceramics include spinel and perovskite oxides, metal diborides, nitrides, and silicides [2–22].

While the presence of multiple species is important for thermochemical stability as well as structural and functional properties, their presence also creates special challenges to predicting phase stability and defect properties. In general, as the number of species in a given phase increases, the availability of reliable thermodynamic phase data tends to decrease. This has led to the development of quantities such as the entropy forming ability [23] introduced by Curtarolo and coworkers that can be used to efficiently screen for potentially stable high-entropy compounds from first principles data. Similarly, the calculation of vacancy formation energies becomes complex due to the range of compositions and configurations available to a vacancy site on the chemically disordered lattice. Unlike in chemically ordered crystals, vacancies in crystals with random mixing cannot be represented by a single defect energy. Each vacancy must instead be treated individually with respect to not only the species removed, but also the corresponding occupations of the local environment [24,25]. The number of possible occupations can be exceedingly large for these systems, particularly when multiple neighbor shells need to be considered to accurately represent the vacancy formation energies. Furthermore, the way in which diverse vacancies influence the bulk vacancy concentration must be carefully considered where a single energy would have otherwise been used in the relevant thermodynamic equations.

Much of the literature on vacancies in disordered systems has centered on the cluster expansion method and its application to the simplest case of binary systems [24–27]. This method takes the linear sum of interactions for every atom cluster in a material to a given distance and accuracy (e.g., pair, triplet, or n-tuple) to calculate the relative energy of ordering in the local atomic structure. When applied to vacancies, the vacancy can be treated as an $N + 1$ species in the cluster expansion. While the cluster expansion method can predict accurate energies across the configurational and compositional space, the number of coefficients generally scales as roughly N^x with N as the number of components and x as the number of atoms in the highest-order cluster [28]. For a high-entropy compound with five or more principal components, this rapidly becomes intractable, requiring a tremendous number of simulations to properly fit the interaction coefficients.

This work proposes a simplified neighbor interaction model derived from generalizations of the physical interactions in the local environment and guided by parameter space statistics to enable efficient prediction of cation vacancy formation energies, vacancy concentrations, and cluster affinities in many-component materials. The model is applied to the high-entropy diboride $Hf_{0.2}Zr_{0.2}Ti_{0.2}Ta_{0.2}Nb_{0.2}B_2$, which has a crystalline planar AlB2 structure with one plane containing

boron and the other plane containing the randomly arranged metal cations. Parameter sets within this model were determined using the Bayesian information criterion (BIC) [29] through a goodness of fit to first principles calculations combined with a penalty imposed on the size of the parameter space. The predictive performance of the model was tested via regression across a range of training sample sizes through repeated k-fold and leave-one-out cross validation (LOOCV). Using the model, a dense sampling of the discrete distribution of cation vacancy formation energies was generated and used to estimate vacancy concentrations as a function of temperature. The method, which includes uncertainty estimates for the interaction parameters, is general and therefore can be applied to other high-entropy compounds.

II. METHODS

DFT energies of bulk and vacancy structures were calculated using plane-wave projector-augmented wave pseudopotential methods [30] implemented in the Vienna *Ab initio* Simulation Package [31–33]. The generalized gradient approximation as parameterized by Perdew *et al.* [34] was used for the exchange-correlation potential. The cutoff energy for the plane-wave basis was set to 415 eV with a $3 \times 3 \times 3$ Γ-centered k-point mesh. Five distinct configurations of a 135 atom (45 cation) special quasirandom structure (SQS) [35] were constructed on a skewed hexagonal lattice such that vacancy-vacancy distances were greater than 10 Å and artifacts from periodic lattice strains on tiled vectors were minimized. Vacancy simulations were relaxed with respect to their structures while fixing the supercell volume to that of the bulk in order to capture the relevant strains in the local lattice.

Vacancy formation energies E^{vf} were calculated for each cation site in the five SQS configurations and for each constituent binary diboride using the expression

$$E^{vf} = E^{v} - E^{\text{bulk}} + \mu. \tag{1}$$

Here, E^{v} is the energy of the vacancy-containing supercell, E^{bulk} is the energy of the bulk supercell, and μ is the chemical potential of the vacating species taken as the DFT energy per atom of the elemental solid phase. Interaction parameters were fit to 225 vacancies simulated at every cation site in the SQS supercells (45 vacancies for each cation type).

A. Pair approximation model

Unlike the cluster expansion method, which references the summation to the total energy, solving explicitly for the vacancy formation energy allows one to consider instead the contribution of each neighbor interaction to the vacancy formation directly. Following the pair approximation model in Ref. [36], the vacancy formation energy can be represented as

$$E^{vf} = nE^{b} + E^{r} - \mu, \tag{2}$$

where E^{b} is the energy required to break each of n bonds between the atom and its neighbors, E^{r} is the electronic and structural relaxation of the local environment surrounding the vacancy, and μ is the chemical potential of the vacating atom (alternatively in this framework, the atom can be placed on the surface of an infinite bulk with half of its bonds restored). In the case of a multicomponent system, a linear contribution from each neighbor is assumed and the equation is rewritten as

$$E^{vf} = \sum_{j=1}^{N} n_j \left(E^{b}_{ij} + E^{r}_{j}\right) + E^{r_0}_{i} - \mu_i \tag{3}$$

or, for multiple neighbor shells,

$$E^{vf} = \sum_{j=1}^{N} \sum_{k=1}^{K} n_{jk} \left(E^{b}_{ijk} + E^{r}_{jk}\right) + E^{r_0}_{i} - \mu_i. \tag{4}$$

Here, i represents the atomic species of the removed atom, j is the species of a given neighboring atom, and k is the index of the neighbor shell. Each bond E^{b}_{ijk} is represented by an ij pair in shell k, whereas the relaxation component E^{r}_{jk} is independent of the vacating atom i, and $E^{r_0}_{i}$ is independent of any bonds.

In the case of the diboride (or for any other ionic and covalent materials), the terms in the equation also have embedded in them the neighbor-dependent contributions to the i-type bonding with the anion sublattice, as this cannot be isolated in the DFT energies. Because the boron sites are compositionally invariant, they do not need to be considered explicitly. Rather, the bonding and relaxation terms can be considered as the effective interactions of the variants in the composition.

With Eq. (4) as a starting point, physical constraints can be applied to reduce the parameter space. First, the bond energy between two atoms in the compound must be nondirectional, i.e.,

$$E^{b}_{ijk} = E^{b}_{jik}. \tag{5}$$

Additionally, while bonded interactions are unlikely to extend beyond the first few neighbor shells, structural and electronic relaxation contributions can be longer ranged. From this, the vacancy interactions of distant neighbors can be reduced to only their relaxation components by imposing

$$E^{b}_{ijk} = 0 \tag{6}$$

for any shell k that is determined to have a noninfluential bonded interaction.

A final reduction of the coefficients is made with the introduction of semiarbitrary parameters for each interaction term. Specifically, it can be shown that for any vector U:

$$U + U^{T} = M, \tag{7}$$

where U^{T} is the transpose of U and M is a symmetric matrix $M_{ij} = M_{ji}$. Thus,

$$A + U + U^{T} = B$$

and

$$A + U = B - U^{T} \tag{8}$$

for some arbitrary symmetric matrices A and B. Applied to the bond matrices for a given shell, this yields

$$E^{b} + E^{r} = E^{b*} - [E^{r}]^{T}, \tag{9}$$

where $E^{b*} - [E^{r}]^{T}$ is an equivalent solution to the pair approximation model with the relaxation terms indexed over vacating atom i. Defining the set of neighbor shells B such

that bonding parameters E^b_{ijk} are symmetric and nonzero for all $k \in B$, the relaxation terms can now be summed to a single i-wise vector with elements E^{r*}_i such that

$$E^{r*}_i = E^{r_0}_i - \mu_i - \sum_j^N \sum_{k\in B} n_{jk} E^r_{ik}. \tag{10}$$

Because $\sum_j n_{jk}$ is constant for each shell k in a given crystal structure, E^{r*}_i is independent of the neighbor shell compositions.

Applying these assumptions to the pair approximation in Eq. (4), the total parametrization for the five-cation case ($N = 5$) is reduced to

$$E^{vf} = \sum_{j=1}^{5} \left(\sum_{k\in B} n_{jk} E^{b*}_{ijk} + \sum_{k\in R} n_{jk} E^r_{jk} \right) + E^{r*}_i \tag{11}$$

for neighbor shells B that involve energy contributions from broken bonds, and for shells R that only contribute to the relaxation energy. The determination of shells included in B and R is made through statistical evaluation of the model with different parameter sets. In the form presented in Eq. (11), there remains an additional degree of freedom for each contributing neighbor shell. These can be set as constant for a fixed reference point in the results, e.g., by setting each E^{b*}_{11k} and E^r_{1k} equal to zero.

This reduced pair approximation (RPA) model considers only the composition of neighbor shells, representing a simplification of other common methods that introduce additional coefficients to account for the configurations of neighbor pairs, triplets, and n-tuples. In the case of the five-cation high-entropy diboride, this model can represent the 3.45×10^{16} possible compositions of the first seven neighbor shells of a vacancy (cutoff radius of 7 Å) with as few as 53 linear degrees of freedom. As demonstrated in the following section, imposing the generalized physical constraints allows for an accurate representation of vacancy formation energies in high-entropy systems without the need to consider the many possible decorations of the local environment.

III. RESULTS AND DISCUSSION

For the selection of bonded and relaxation sets B and R in Eq. (11), the BIC is compared for each disjoint combination of $k \in B$, $k \in R$, and $k \in$ reject including all shell indices $k \leqslant 8$. The BIC for a given set of parameters is calculated from the goodness of fit of the DFT results with a penalty imposed on the size of the parameter space, specifically

$$BIC = k_M \ln n - 2 \ln \hat{L}, \tag{12}$$

for sample size n, model dimension k_M, and maximum likelihood $\hat{L}$. For BIC calculation in this work, errors are assumed to be independent and identically distributed. For the Gaussian case, the log likelihood is calculated from the variance estimate $\hat{\sigma}^2$ of multiple linear regression as

$$\hat{L} = -\tfrac{1}{2} n [\ln 2\pi + \ln \hat{\sigma}^2 + 1]. \tag{13}$$

Further, the formulation in Eq. (11) is expanded to include $k \in U$ for unconstrained parametrization of neighbor shells wherein the assumptions in Sec. II are ignored in favor of the formulation in Eq. (4). This allows for rejection of the proposed framework for any shell k during model selection.

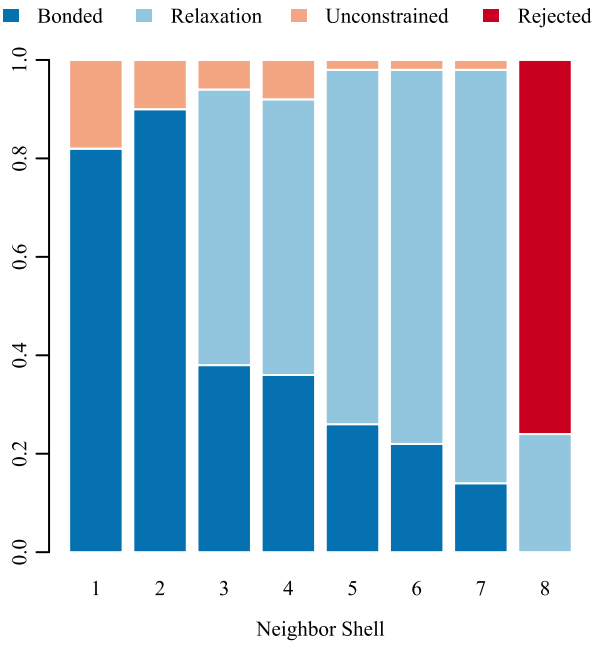

FIG. 1. Selection frequency of each parameter set in the top 50 models by BIC.

Within the 50 best models (as ranked by BIC) in the 65536 parameter set combinations up to eight neighbor shells, the selection frequency of each shell classification is calculated and plotted in Fig. 1. In a given shell, a high rate of unconstrained parameters might suggest that the proposed model assumptions are flawed, while a low but nonzero rate points to the information contained in that shell being valuable enough to outweigh an inefficient parametrization. The selection frequencies show that the reduced parametrizations dominate in the models with the best expected performance, as well as a trend toward parameter sets without a bond energy term (and, eventually, rejected parameters) with increasing distance from the vacancy site.

To quantify the predictive performance of the selected models for $Hf_{0.2}Zr_{0.2}Ti_{0.2}Ta_{0.2}Nb_{0.2}B_2$, cross-validation scores were calculated on models fit to the DFT results using multiple regression across a range of training sample sizes through repeated k-fold cross validation and LOOCV. Plotted in Fig. 2 is the performance and convergence behavior of an optimized parametrization ($B = \{1, 2, 3, 4\}$, $R = \{5, 6, 7\}$, $U = \emptyset$, and reject $= \{8\}$) compared against the rest of the top 50 models and the fully unconstrained model for $K = 7$. Additionally, cross-validation error is plotted against the largest shell K for the optimized and unconstrained models. This shows improved predictive performance with the inclusion of each of the first seven neighbor shells, in agreement with the parameter selection frequency in Fig. 1.

For all sample sizes and cutoff radii, the RPA outperforms the unconstrained model. In particular, the RPA model is shown to be more resilient to overfitting when including

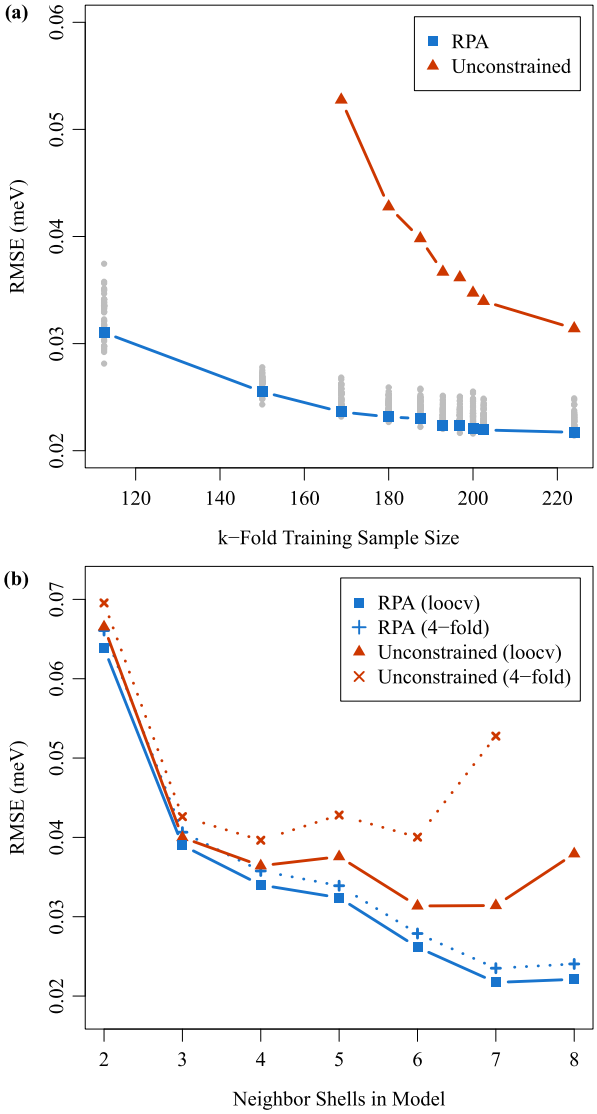

FIG. 2. (a) Convergence of the reduced pair approximation model ($B = \{1, 2, 3, 4\}$, $R = \{5, 6, 7\}$, and $U = \emptyset$) and the unconstrained model ($U = \{1, \ldots, 7\}$) as a function of training sample size for $5\times$ k-fold cross validation up to and including LOOCV; and (b) cross-validation error as a function of largest shell K in the model for fourfold and LOOCV (training sample size of 169 and 224, respectively). Cross-validation errors for the 50 top models by BIC shown in gray.

distant pairs, even with smaller sample sizes. The resilience of the model is, in part, gained from the symmetry of the parameters linking the coefficients from different vacancy types. Whereas the unconstrained parametrization acts as separate, independent models for each i-type vacancy, the RPA coefficients are symmetric across ij terms, effectively increasing the size of the training set available to each parameter while removing unnecessary degrees of freedom. Of note for models with $K = 2$ and relatively few parameters, the unconstrained model is only able to approach the performance of the optimized RPA, not surpass it, suggesting that the RPA assumptions do not remove information from the model.

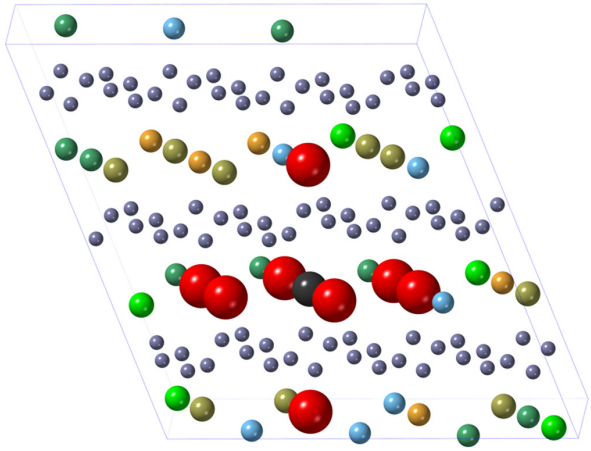

FIG. 3. A sample construction of the SRO structure in which a near-neighbor shell (red) is replaced with a single element.

The optimizations of the RPA model converge quickly to root mean square prediction errors on the order of 20 meV across a 2 eV range of calculated formation energies. RPA models yield 30% lower LOOCV errors than those of the unconstrained model for training size $n = 224$, and 50% lower fourfold cross-validation errors ($n = 169$). Comparing the fitting errors and LOOCV prediction errors within each model, the RPA values are overfit by less than 50%, compared to 185% overfitting in the unconstrained case. Model parameters and fitness metrics are presented in Table I.

The RPA model is shown to predict the DFT results for the stoichiometric and maximally disordered SQS; however, that result says little about the applicability of the model to non-ideally mixed local arrangements. By its formulation, every site in the SQS should closely approximate ideal mixing in a random compound. To investigate if the model can be extrapolated to atomic clusters that do not follow the assumptions imposed by the SQS formulation, a set of pseudo-ordered structures is constructed in which the nearest-neighbor shells are selectively replaced, in turn, with each constituent metal species to approximate certain short-range order (SRO) clusters as illustrated in Fig. 3.

Due to supercell size constraints, each shell is calculated separately, and the surrounding bulk is left unchanged from the SQS, that is, we do not impose a deficient region surrounding the SRO shell to maintain the equimolar stoichiometry of the high-entropy compound. As a result, the supercells are no longer fully consistent with the assumptions of our model, most importantly that the chemical potential reference assumes the compound to be compositionally invariant. It might also be expected that any ignored cluster interactions would play a larger role when present in such large concentrations, in addition to any volume or strain effects of the composition changes. These measurements are, as such, only useful as a qualitative guide to assess the applicability of the model as configurations deviate from ideal mixing.

TABLE I. Root mean squared error (in meV) from least squares fitting as well as cross-validation errors from LOOCV and 5× k-fold CV for unconstrained and RPA models. SRO is the prediction error for the set of pseudo-ordered simulations. Sets B and R denote the indices of shells modeled as bonded and relaxation interactions, respectively.

Assumption	Parameters	BIC	RMSE	LOOCV	fourfold CV	SRO
$U = \{1, ..., 7\}$	145	−601.7	11.0	31.4	52.8	101.4
$B = \{1, 2, 4\}$; $R = \{3, 5, 6, 7\}$	63	−878.6	15.9	22.3	24.4	77.0
$B = \{1, 2, 5\}$; $R = \{3, 4, 6, 7\}$	63	−876.6	16.0	22.4	24.2	74.3
$B = \{1, 2, 3, 6\}$; $R = \{4, 5, 7\}$	73	−869.3	14.4	21.4	23.2	90.6
$B = \{1, 2\}$; $R = \{3, ..., 7\}$	53	−864.9	18.5	24.3	25.8	79.4
$B = \{1, 2, 3, 4\}$; $R = \{5, 6, 7\}$	73	−863.1	14.6	21.7	23.6	81.2

Despite the deficiencies of both the model and the simulations for representing these SRO cases, the predictions are largely in agreement with the simulated results. The modeled vacancy formation energies and the SRO predictions are plotted against the DFT energies in Fig. 4. RMSE around 75–80 meV for the SRO structures across 3 eV of predictions suggests reasonable extensibility of the RPA model beyond the explicitly formulated case.

Based on this result, the remaining analysis is conducted across the entire range of local compositions. Any errors that might emerge as the local order deviates from the ideal random SQS structures are omitted for the purposes of this work.

A. Vacancy concentration

Because the vacancy formation energy in disordered multicomponent compounds does not take on a single value, the traditional formalisms for calculating the vacancy concentration in a material are no longer straightforward. In the case of the proposed neighbor model for a five-component diboride, there are over 3.45×10^{16} possible compositions of the first seven neighbor shells. Each of these neighbor environments should be considered in the derivation of the vacancy concentration.

In Fig. 5 the histogram of vacancy formation energies from the 225 DFT calculations, representing a set of ideal random arrangements, is compared to the predicted distribution of formation energies across all combinatorial compositions of the local environment. The binary diboride cation vacancy formation energies of each constituent metal are included for reference. Using only the explicitly calculated SQS structures results in a clear bias of the formation energy distribution toward the peaks. The peaks of the distribution represent the most abundant configurations in an ideally mixed compound; however, these are not the most energetically favorable configurations and therefore may not be the largest contributors to the total vacancy concentration. By applying the model across the entire composition space, a complete picture of the vacancy formation energy distribution can be used to calculate the precise contributions to the global vacancy concentration.

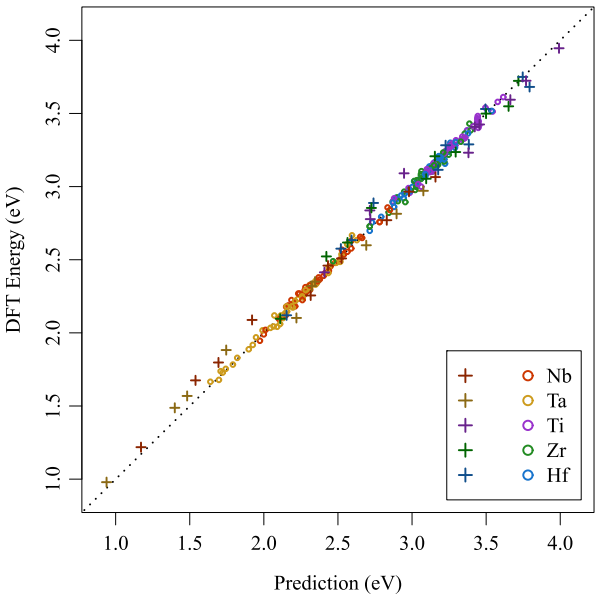

FIG. 4. DFT simulated vacancy formation energies (○) fit to the RPA model with $B = \{1, 2, 3, 4\}$ and $R = \{5, 6, 7\}$. SRO predictions (+) included to illustrate an extrapolation of the model.

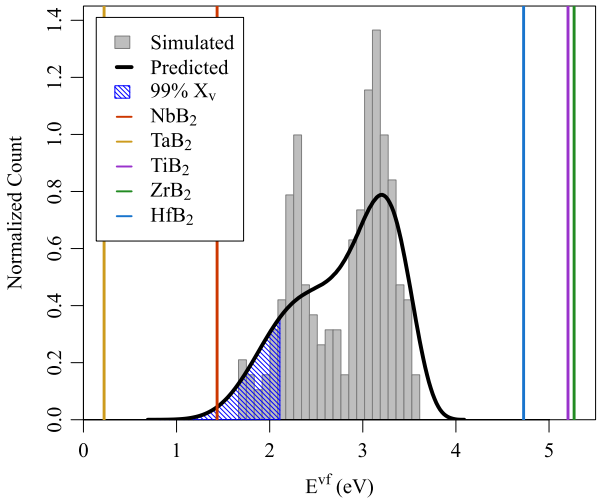

FIG. 5. Vacancy formation energies of all simulations and the predicted total distribution using the RPA model. Here, the parameter set of $B = \{1, 2, 3\}$ is used for illustrative purposes. The shaded region represents the vacancy configurations, which account for 99% of the total concentration at 2000 K. Vertical lines indicate the binary diboride vacancy formation energies of each constituent element.

An important issue when dealing with a multicomponent system is how to reconcile the identity of a vacancy. When approaching the issue from a bulk reference, as in the case of the cluster expansion method, it can be stated that vacancies do not have an identity. While this is ostensibly true in all cases, when using a chemical potential reference it becomes critical to consider what previously occupied the vacancy site.

Further, Zhang and Sluiter have suggested that the composition can be constrained such that the vacancy concentration can be solved with the use of Lagrange multipliers [25]. While it might be mathematically convenient to impose such constraints (and in the case of the cluster expansion method, perhaps necessary), it is not necessarily an accurate representation of a system that interacts with its environment.

Alternatively, a framework can be constructed in which each local environment (as defined by the RPA model) is considered to be a distinct lattice site, and the occupation of each lattice site is calculated independently. Applying traditional Boltzmann statistics to each site configuration yields an equation for the total vacancy concentration:

$$X_v = \sum_{\sigma} p(\sigma) e^{\frac{-E(\sigma)}{kT}}, \tag{14}$$

summed over all configurations σ, where $p(\sigma)$ and $E(\sigma)$ represent the probability and vacancy formation energy, respectively, of a given site. Here, $p(\sigma)$ is calculated using the purely combinatorial approach, however, it would be trivial to add another layer of Boltzmann statistics to this term to reflect the relative probabilities of certain clusters in the case where SRO is a significant factor. This approach is similar to the method employed by Ruban [24], differing only in the use of a discrete distribution over a continuous approximation.

The vacancy concentrations derived from the predicted distribution were converged by random sampling of the available configuration space for 2×10^{11} arrangements. The temperature dependence of X_v is plotted in Fig. 6. Due to the unequal contributions of the varied energies, the total vacancy concentration deviates from Arrhenius behavior. The logarithm of X_v now takes on a nonlinear dependence on inverse temperature wherein the lower-energy vacancies dominate at low temperature and the concentration converges to that of the mean formation energy at ultrahigh temperatures. From this behavior, the calculation of the entire energy distribution becomes critical, as the energies in the left tail of the distribution have the largest contribution to the vacancy characteristics of the bulk. To further illustrate this uneven contribution of the different local environments, Fig. 5 plots the vacancy configurations that contribute 99% of the total vacancies in the bulk at 2000 K, all of which occur below the first peak of the distribution (primarily Nb and Ta vacancies).

Generally, whether by this isolated site method or by an alternative approach explored in the literature [24,25,37,38], the vacancy concentration in a high-entropy system is expected to be within the range of the constituent compounds, but above the average. Early experimental results from a recent positron annihilation study of a high-entropy alloy support this finding [39].

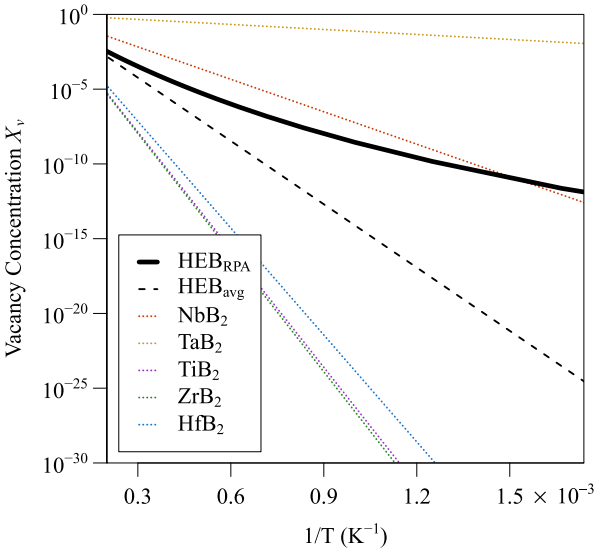

FIG. 6. Cation vacancy concentration of the high-entropy diboride structure as a function of inverse temperature, calculated by individual site contributions as compared to the vacancy concentration of the uniform mean formation energy and the concentration in each of the binary transition-metal diborides.

B. Configurational entropy

An additional factor in calculating the concentration of vacancies in a disordered material is the effect of a disordered multicomponent lattice on the entropic contribution to vacancy formation. This contribution was incorrectly derived in Ref. [40] as scaling with a factor of $e^{(N-1)}/N$ for an N-component system. Rather, the entropic contribution scales with $1/N$ for small concentrations as derived in the Appendix.

This solution is intuitive if the vacancy is considered to be an $N+1$ species, as suggested in Refs. [25,26], due to the reduced configurational entropy contribution of the nth species as illustrated for high-entropy oxides in the supplemental materials of Ref. [2]. Specifically, for each additional component added to a compound, there is a smaller gain in configurational entropy for the equivalent increase in concentration, always reaching a maximum entropy at the equiatomic composition. The decrease in the entropic contribution to the vacancy formation energy, in turn, results in a lower vacancy concentration, calculated as

$$X_v = \frac{1}{N} \sum_{\sigma} p(\sigma) e^{\frac{-E(\sigma)}{kT}}. \tag{15}$$

This decrease is significant, though not as large as the differences in concentration between frameworks for addressing the enthalpy distribution.

C. Pair affinities

Because the pair approximation method implemented in this work is compositionally invariant and fixed to a chemical potential reference, it does not provide the same insight as the cluster expansion method into the bulk energies across the composition space of high-entropy materials. It does, how-

TABLE II. Atom pair affinities (meV/bond) and standard deviations ($\pm 2\sigma$) for cation near neighbors in the first four shells of the RPA model with $B = \{1, 2, 3, 4\}$ and $R = \{5, 6, 7\}$.

(i, j,1)	Ti	Zr	Hf	Nb	Ta
Ti	4.6 ± 5.3	20.1 ± 3.9	8.2 ± 4.4	-13.3 ± 3.8	-19.6 ± 3.7
Zr		-5.2 ± 5.8	0.5 ± 4.1	-5.2 ± 4.0	-10.1 ± 4.5
Hf			7.5 ± 4.5	-7.5 ± 3.9	-8.8 ± 3.5
Nb				12.2 ± 6.6	13.9 ± 3.9
Ta					24.5 ± 5.0
(i, j,2)					
Ti	5.7 ± 7.6	14.5 ± 6.6	9.9 ± 6.2	-11.1 ± 5.7	-19.1 ± 5.5
Zr		0.9 ± 9.4	0.0 ± 6.9	-2.1 ± 6.0	-13.2 ± 5.8
Hf			11.6 ± 5.8	-13.3 ± 6.1	-8.2 ± 5.9
Nb				14.7 ± 9.1	11.9 ± 5.9
Ta					28.6 ± 8.3
(i, j,3)					
Ti	4.5 ± 5.2	-0.2 ± 3.3	-2.8 ± 2.9	-1.8 ± 3.0	0.3 ± 2.9
Zr		4.1 ± 3.7	2.8 ± 2.7	-1.5 ± 3.0	-5.3 ± 2.9
Hf			0.8 ± 2.9	2.0 ± 2.9	-2.8 ± 2.9
Nb				-0.2 ± 4.7	1.5 ± 3.4
Ta					6.3 ± 4.0
(i, j,4)					
Ti	-4.2 ± 7.1	9.0 ± 4.4	2.0 ± 4.4	-2.6 ± 3.9	-4.2 ± 4.1
Zr		-4.8 ± 7.0	-6.3 ± 4.1	1.2 ± 4.1	0.9 ± 4.7
Hf			-6.1 ± 4.7	5.3 ± 4.2	5.1 ± 4.0
Nb				1.5 ± 5.0	-5.4 ± 3.7
Ta					3.6 ± 5.4

ever, retain the ability to approximate pair affinities from the interaction coefficients within the constraints of the formulation. The pseudo-ordered structure calculations in this work demonstrate that while the prediction accuracy may suffer to some extent with varying composition and cluster order, the constraints on disorder and composition are not overly restrictive for application of the model to configurations that deviate from the ideally mixed SQS structure.

To find any pair affinities that might lead to short-range order in the compound, the relative energy of two configurations is taken as the bonds gained minus the bonds lost. To form an arbitrary additional ij bond, an iX and a jX bond must be broken, restoring an XX bond in the process (where X matches the composition of the bulk). Recalling from Eq. (9) that for each fitted bonding parameter

$$E^{b*}_{ijk} = E^{b}_{ijk} + E^{r}_{ik} + E^{r}_{jk}\,, \tag{16}$$

it is apparent that the relaxation terms will cancel out in any redistribution of bonds, so long as composition is conserved. As such, the arbitrary parameters E^{b*}_{ijk} can be used directly for this analysis.

Furthermore, variances of the pair affinities can be computed using approximated local sensitivities. The covariance matrix of the parameters is found using

$$V = \sigma_0^2 [S^T S]^{-1} \tag{17}$$

for model variance σ_0^2 from the sample size n, number of parameters p, and residuals matrix R as

$$\sigma_0^2 = \frac{1}{n-p} R R^T \tag{18}$$

and local sensitivity matrix S, calculated as the partial derivatives of the n samples with respect to each parameter. The terms of the local sensitivity matrix are defined as

$$S_{np} = \frac{dE^{vf}_n}{d\theta_p} \tag{19}$$

for each simulated vacancy formation energy E^{vf}_n and each interaction parameter θ_p. Applied to the linear RPA model, the terms reduce to the parameter coefficients for each configuration, i.e., the number of each neighbor interaction n_{ijk} and n_{jk} in Eq. (11). The variances are extended to pair affinities using the expression

$$\sigma^2_{ij} = c^T V c, \tag{20}$$

where c is a vector representing the linear combination of the total bonds added and removed. The derived pair affinities and uncertainties are presented in Table II.

The largest positive affinities are noted for Ti-Zr, Ta-Nb, and Ta-Ta pairs in both the first and second shells, as well as negative affinities for the first and second Ti-Nb, Ti-Ta, and Zr-Nb pairs. While the magnitude of pair affinities generally decreases after the first and second neighbors, there is not a simple distance dependence of the interactions. Given the similar bond radii of the in-plane and out-of-plane neighbors that comprise the first two neighbor shells (3.1 Å and 3.4 Å, respectively), orientational and boron sublattice bonding effects likely play a role.

A separate analysis would be necessary to determine the degree of short-range order that could result from these relative energies. Above a certain magnitude, one might expect to see clustering among high affinity pairs and segregation of

negative affinity pairs; however, the tradeoff between entropic stabilization and short-range order is beyond the scope of this work.

IV. CONCLUSIONS

An efficient method for prediction of diverse vacancy formation energies across the vast composition space of high-entropy materials is applied to the high-entropy diboride $Hf_{0.2}Zr_{0.2}Ti_{0.2}Ta_{0.2}Nb_{0.2}B_2$. A reduced set of bonding and relaxation coefficients with 73 linear degrees of freedom is shown to outperform the unconstrained parametrization, achieving an RMSE for k-fold cross validation of less than 25 meV on training sets as small as 150 vacancy simulations. This model framework enables rapid evaluation of vacancies in compounds across the high-entropy composition space.

Accurate evaluation of the vacancy formation energy distribution, particularly with respect to the left tail, is presented as critical for the calculation of vacancy concentrations in disordered materials. The bulk vacancy concentration is estimated as the combined total of the isolated contributions across a dense sampling of the available vacancy configurations.

Pair affinities derived from pair interaction coefficients in the RPA model are presented as a means to investigate possible short-range order in high-entropy materials and likely cluster pairs are highlighted. Further derivation of the interplay between entropy and clustering interactions is left for future analyses.

ACKNOWLEDGMENTS

This work was supported by a Multi-Disciplinary University Research Initiative through Award No. N00014-15-1-2863 from the U.S. Office of Naval Research and by the National Science Foundation under Grant No. DGE-1633587 (Science and Engineering of the Atomic Structure at North Carolina State University).

APPENDIX: CONFIGURATIONAL ENTROPY DERIVATION

The derivation for vacancy concentration in a multicomponent material mirrors the classical thermodynamic derivation, starting from the configurational entropy S_c for an equiatomic compound as

$$S_c = k\left[\ln N! - m\ln\left(\frac{N-n}{m}\right)! - lnn!\right] \tag{A1}$$

for number of atoms N, number of species m, and number of vacancies n where composition is maintained. Applying Sterling's approximation and taking the derivative with respect to n:

$$S_c = k\left[N\ln N - (N-n)\ln\left(\frac{N-n}{m}\right) - n\ln n\right] \tag{A2}$$

$$\Delta S_c = k\left[\ln\left(\frac{N-n}{m}\right) - \ln n\right], \tag{A3}$$

and for small n:

$$\Delta S_c = -k\ln\left(\frac{mn}{N-n}\right) \approx -k\ln\left(\frac{mn}{N}\right). \tag{A4}$$

Setting the free energy equal to zero and rearranging to solve for vacancy concentration X_v,

$$\Delta H + kT\ln\left(\frac{mn}{N}\right) = 0 \tag{A5}$$

$$X_v = \frac{1}{m}\exp\frac{-\Delta H}{kT}. \tag{A6}$$

[1] C. Oses, C. Toher, and S. Curtarolo, Nature Rev. Mater. **5**, 295 (2020).
[2] C. M. Rost, E. Sachet, T. Borman, A. Moballegh, E. C. Dickey, D. Hou, J. L. Jones, S. Curtarolo, and J. P. Maria, Nature Commun. **6**, 8485 (2015).
[3] T. J. Harrington, J. Gild, P. Sarker, C. Toher, C. M. Rost, O. F. Dippo, C. McElfresh, K. Kaufmann, E. Marin, L. Borowski *et al.*, Acta Mater. **166**, 271 (2019).
[4] S. Jiang, T. Hu, J. Gild, N. Zhou, J. Nie, M. Qin, T. Harrington, K. Vecchio, and J. Luo, Scr. Mater. **142**, 116 (2018).
[5] A. Sarkar, Q. Wang, A. Schiele, M. R. Chellali, S. S. Bhattacharya, D. Wang, T. Brezesinski, H. Hahn, L. Velasco, and B. Breitung, Adv. Mater. **31**, 1970189 (2019).
[6] C. M. Rost, Z. Rak, D. W. Brenner, and J. P. Maria, J. Am. Ceram. Soc. **100**, 2732 (2017).
[7] D. Bérardan, S. Franger, A. K. Meena, and N. Dragoe, J. Mater. Chem. A **4**, 9536 (2016).
[8] A. Sarkar, R. Djenadic, N. J. Usharani, K. P. Sanghvi, V. S. K. Chakravadhanula, A. S. Gandhi, H. Hahn, and S. S. Bhattacharya, J. Eur. Ceram. Soc. **37**, 747 (2017).
[9] A. Sarkar, C. Loho, L. Velasco, T. Thomas, S. S. Bhattacharya, H. Hahn, and R. Djenadic, Dalton Trans. **46**, 12167 (2017).
[10] A. Sarkar, L. Velasco, D. Wang, Q. Wang, G. Talasila, L. de Biasi, C. Kübel, T. Brezesinski, S. S. Bhattacharya, H. Hahn *et al.*, Nature Commun. **9**, 1 (2018).
[11] Q. Wang, A. Sarkar, Z. Li, Y. Lu, L. Velasco, S. S. Bhattacharya, T. Brezesinski, H. Hahn, and B. Breitung, Electrochem. Commun. **100**, 121 (2019).
[12] R. Witte, A. Sarkar, R. Kruk, B. Eggert, R. A. Brand, H. Wende, and H. Hahn, Phys. Rev. Materials **3**, 034406 (2019).
[13] J. Gild, Y. Zhang, T. Harrington, S. Jiang, T. Hu, M. C. Quinn, W. M. Mellor, N. Zhou, K. Vecchio, and J. Luo, Sci. Rep. **6**, 37946 (2016).
[14] G. Tallarita, R. Licheri, S. Garroni, R. Orru, and G. Cao, Scr. Mater. **158**, 100 (2019).
[15] D. Liu, T. Wen, B. Ye, and Y. Chu, Scr. Mater. **167**, 110 (2019).
[16] P. H. Mayrhofer, A. Kirnbauer, P. Ertelthaler, and C. M. Koller, Scr. Mater. **149**, 93 (2018).
[17] E. Castle, T. Csanádi, S. Grasso, J. Dusza, and M. Reece, Sci. Rep. **8**, 8609 (2018).
[18] X. Yan, L. Constantin, Y. Lu, J. F. Silvain, M. Nastasi, and B. Cui, J. Am. Ceram. Soc. **101**, 4486 (2018).
[19] J. Zhou, J. Zhang, F. Zhang, B. Niu, L. Lei, and W. Wang, Ceram. Int. **44**, 22014 (2018).
[20] Y. Yang, W. Wang, G. Y. Gan, X. F. Shi, and B. Y. Tang, Phys. B **550**, 163 (2018).

[21] J. Dusza, P. Švec, V. Girman, R. Sedlák, E. G. Castle, T. Csanádi, A. Kovalčíková, and M. J. Reece, J. Eur. Ceram. Soc. **38**, 4303 (2018).
[22] J. Gild, J. Braun, K. Kaufmann, E. Marin, T. Harrington, P. Hopkins, K. Vecchio, and J. Luo, J. Materiomics **5**, 337 (2019).
[23] P. Sarker, T. Harrington, C. Toher, C. Oses, M. Samiee, J. P. Maria, D. W. Brenner, K. S. Vecchio, and S. Curtarolo, Nature Commun. **9**, 4980 (2018).
[24] A. V. Ruban, Phys. Rev. B **93**, 134115 (2016).
[25] X. Zhang and M. H. F. Sluiter, Phys. Rev. B **91**, 174107 (2015).
[26] A. Van der Ven and G. Ceder, Phys. Rev. B **71**, 054102 (2005).
[27] E. K. Delczeg-Czirjak, L. Delczeg, L. Vitos, and O. Eriksson, Phys. Rev. B **92**, 224107 (2015).
[28] N. Artrith, A. Urban, and G. Ceder, Phys. Rev. B **96**, 014112 (2017).
[29] G. Schwarz *et al.*, Ann. Statist. **6**, 461 (1978).
[30] P. E. Blöchl, Phys. Rev. B **50**, 17953 (1994).
[31] G. Kresse and J. Hafner, Phys. Rev. B **47**, 558 (1993).
[32] G. Kresse and J. Furthmüller, Phys. Rev. B **54**, 11169 (1996).
[33] G. Kresse and J. Furthmüller, Comput. Mater. Sci. **6**, 15 (1996).
[34] J. P. Perdew, K. Burke, and M. Ernzerhof, Phys. Rev. Lett. **77**, 3865 (1996).
[35] A. Van de Walle, P. Tiwary, M. De Jong, D. L. Olmsted, M. Asta, A. Dick, D. Shin, Y. Wang, L. Q. Chen, and Z. K. Liu, Calphad **42**, 13 (2013).
[36] B. T. M. Loh and C. T. Liu, Scr. Metall. **6**, 565 (1972).
[37] W. Chen, X. Ding, Y. Feng, X. Liu, K. Liu, Z. P. Lu, D. Li, Y. Li, C. T. Liu, and X. Q. Chen, J. Mater. Sci. Technol. **34**, 355 (2018).
[38] A. Esfandiarpour and M. N. Nasrabadi, Calphad **66**, 101634 (2019).
[39] K. Sugita, N. Matsuoka, M. Mizuno, and H. Araki, Scr. Mater. **176**, 32 (2020).
[40] Z. Wang, C. T. Liu, and P. Dou, Phys. Rev. Materials **1**, 043601 (2017).