*MRS Communications (2018)*, *8*, 1300–1310
© Materials Research Society, 2018
doi:10.1557/mrc.2018.155

MRS Communications

Research Letter

# Integrating exploratory data analytics into ReaxFF parameterization

**Efraín Hernández-Rivera,** U.S. Army Research Laboratory, Weapons and Materials Research Directorate, RDRL-WMM, APG, MD 21005, USA
**Souma Chowdhury,** Department of Mechanical and Aerospace Engineering, University at Buffalo, 246 Bell Hall, University at Buffalo, Buffalo, NY 14260, USA
**Shawn P. Coleman,** U.S. Army Research Laboratory, Weapons and Materials Research Directorate, RDRL-WMM, APG, MD 21005, USA
**Payam Ghassemi,** Department of Mechanical and Aerospace Engineering, University at Buffalo, 246 Bell Hall, University at Buffalo, Buffalo, NY 14260, USA
**Mark A. Tschopp,** U.S. Army Research Laboratory, Weapons and Materials Research Directorate, RDRL-WMM, APG, MD 21005, USA

Address all correspondence to Efraín Hernández-Rivera at **efrain.hernandez18.civ@mail.mil** and Mark A. Tschopp at **mark.a.tschopp.civ@mail.mil**

(Received 30 April 2018; accepted 26 July 2018)

## Abstract

We present a systematic approach to refine hyperdimensional interatomic potentials, which is showcased on the ReaxFF formulation. The objective of this research is to utilize the relationship between interatomic potential input variables and objective responses (e.g., cohesive energy) to identify and explore suitable parameterizations for the boron carbide (B–C) system. Through statistical data analytics, ReaxFF's parametric complexity was overcome via dimensional reduction (55 parameters) while retaining enough degrees of freedom to capture most of the variability in responses. Two potentials were identified which improved on an existing parameterization for the objective set if interest, showcasing the framework's capabilities.

## Introduction

The ability to computationally design high performing materials while managing tradeoffs in weight, cost, and deleterious properties is an ongoing challenge that impacts a wide range of industries (e.g., automotive,[1,2] aerospace,[3] and defense[4,5]). For example, developing lower density materials with similar or superior mechanical properties compared with conventional materials can be advantageous by decreasing the overall weight requirement of the system. Conversely, developing materials with comparable weights but higher mechanical strengths to conventional materials can decrease the amount of material necessary, leading to an overall lower system weight. For instance, Mg and Al alloys are often used to design lightweight systems, providing comparable performance to heavier (conventional) high-strength materials.[1,6]

This materials design potential is true for superhard ceramic materials, which have a number of defense-related applications including unmounted soldier protection. One class of armor materials for soldier protection are boron-based ceramics, e.g., boron carbide (B–C) and boron oxide (B–O) systems. These are both lighter, harder, and ballistically outperform previous silicon carbide systems.[7,8] Other important applications of boron-based ceramics include neutron absorption and shielding,[9] which are of unique importance to the nuclear energy industry. While lighter and harder, the performance of this class of materials can suffer due to its brittle nature and poorly quantified fracture mechanisms.[5] Although, there exist qualitative insights associating fracture with shear band amorphization.[10,11] the mechanism of initiation and propagation of this amorphization process is not well understood, especially under dynamic failure conditions. Improving our understanding of the fracture mechanisms requires atomic scale studies to highlight the conditions that lead to catastrophic failure. Once these underlying atomic physics are better known, manipulating material chemistry and/or processing conditions can be used to engineer the microstructure, properties, and performance of the material. In this approach, atomic simulations serve as the basis for integrated computational materials engineering,[12] aimed at computationally designing materials in an accelerated fashion while minimizing the risk associated with the development process.

As computational resources have increased in capacity and performance, the cost of large-scale (spatial or temporal) atomic simulations to predict mechanics and physical interaction of realistic systems has become more feasible.[13] There are two main types of simulations for studying atomistic mechanisms: quantum mechanics and molecular dynamics (MD) simulations. Quantum mechanics (QM) computations predict (sub)atomic scale interactions with a high degree of accuracy. While approximating electronic contributions results in an accurate approximation of atomic properties, this leads to expensive computations, limiting QM simulations to small volumes (number of atoms) and short time scales. Hence, properly analyzing design scale features, e.g., grain boundaries,


https://doi.org/10.1557/mrc.2018.155 Published online by Cambridge University Press

becomes prohibitively expensive. Classical MD enables studying of materials atomic performance, within appropriate spatio-temporal scales, to understand the failure of boron ceramics under ballistic impact. MD solves for Newtonian physics, often ignoring electronic influence, based on an interatomic potential that describes the forces exerted by atoms on each other. Ultimately, the fidelity of MD predictions is limited to the accuracy of the interatomic potential used. A variety of potential formulations exist that aim at reducing the computational costs by "coarse-graining" physical interactions taking place at the subatomic level (e.g., electron interactions). Each potential relies on specific assumptions about the system; thus, care needs to be practiced when parameterizing potentials to be physically predictable. Nonetheless, insights from MD simulations enable the design and improvement of materials.

Complex reactive potentials have been developed[14–16] to capture the high fidelity QM results while being able to model large-scale atomic systems. One such example is known as the Reactive Force Field (ReaxFF),[14] which was developed as a way to better describe the complex structures and thermodynamics occurring during chemical reactions. The ReaxFF formulation

$$\begin{aligned} U &= U_{\text{bond}} + U_{\text{over}} + U_{\text{under}} + U_{\text{val}} + U_{\text{pen}} + U_{\text{tors}} \\ &\quad + U_{\text{conj}} + U_{\text{vdW}} + U_{\text{Coul}} \\ &= \text{F}[f(r_{ij}), g(r_{ij}, \theta_{ijk}), h(r_{ij}, q_i, q_j), \ldots] = \text{F}(\mathbf{X}) \end{aligned} \quad (1)$$

incorporates a comprehensive set of many-body interatomic interactions, bond-order, and charge. As the ReaxFF formulation aims to model many physical interactions, accurate parameterization requires optimizing a hyperdimensional parametric design space. This becomes exceedingly expensive when considering multi-component and multi-phase systems, as with many-body potentials. One such parameterization for B–C was developed by An and Goddard,[17] where shear-band amorphization was qualitatively demonstrated. Parameterizing hyperdimensional potentials like ReaxFF can take several forms, which span the straightforward one-factor-at-a-time to more complex multi-objective driven algorithms. Examples of multi-objective parameterization approaches, such as the genetic algorithm-based reactive force field optimizer (GARFfield),[18] the multiple objective evolution strategies (MOES),[19] and the potential optimization software for materials (POSMat).[20] Another example is a statistical sampling-based generalized methodology developed to understand how parametric changes lead to output uncertainties of the lower dimensional MEAM potential.[21]

In this work, a data analytics approach for exploring and optimizing complex hyperdimensional interatomic potentials is outlined. Multiple methodologies are integrated to systematically quantify the intricate relationship between input variables (factors) and model outputs (responses). It should be noted that the terms factor, input variable, and parameter are used interchangeably through the paper. This includes adaptive sampling and sensitivity analysis (SA) approaches to enable focused parametric exploration. Lastly, a demonstration of how the proposed framework can be used to improve upon existing interatomic potentials was performed on the B–C system based on An's parameterization.

## Boron carbide complexity

Modeling B–C with classical MD is challenging due to the complex crystal structures and varying stoichiometry of its many known polymorphs. The nominal crystal structure, shown in Fig. 1(a), is described as rhombohedral consisting of primarily boron-rich icosahedra located at the unit cell vertex with a three-atom backbone chain along the [1 1 1] direction. Atomic location within these icosahedra can be subdivided into equatorial and polar sites which are shown in Fig. 1(a)'s inset as gold and gray atoms, respectively. Carbon atoms are primarily located in the backbone chain but can be substituted into distinct polar and equatorial sites of the icosahedra. Boron carbide is classified in the $R\bar{3}$ m space group. However, relaxation of the atom positions and lattice parameters occur during QM and molecular statics (MS) energy minimization that disrupts this symmetry. For example, Fig. 1(b) shows a typical energy minimization routine used in this work that cycles between stages of lattice relaxation to better reduce internal pressure and relax atomic coordinate. It should be noted that common practice consists of simpler energy minimization schemes, which would follow the blue dotted-curve. Clearly, it is possible that not allowing both relaxation procedures could lead to unoptimized atomic configurations.

Boron carbide has been experimentally observed in ranges of carbon solubility from 8 to 20 at.% without phase separation or interstitial carbon occupancy.[9,22] Its most abundant configurations have transition energies in the order of a few meV per unit cell (u.c.).[7,23,24] Figure 1(c) shows three 20 at.% C ($B_4C$) and one 13 at.% C polymorphs of boron carbide, from left to right. The two left $B_4C$ structures in Fig. 1(c) have a boron atom inserted in the middle of the carbon backbone chain and a carbon atom substituted in the icosahedra at polar ($B_{11}C^{p}CBC$) and equatorial ($B_{11}C^{e}CBC$) sites, respectively. The right two boron carbide structures, $B_{12}CCC$ and $B_{12}CBC$, have full boron icosohedra with CCC and CBC chains, respectively. Due to the atomic size difference between B and C atoms, these seemingly small variations in the polymorphic configuration lead to non-negligible structural differences.

The complexities of the boron carbide crystal structures and the thermodynamic landscape of its many polymorphs have led to difficulties in developing and parameterizing accurate interatomic potentials. Without accurate boron carbide potentials, there are no sufficient MD studies to understand the root cause of fracture nor is there clear understanding how its performance is altered due to defects.[5] While there has been success in observing amorphization in bulk boron carbide under shear,[17] this parameterization has difficulties capturing polymorphic ordering, and predicting boron carbide defect and grain boundary structures.



https://doi.org/10.1557/mrc.2018.155 Published online by Cambridge University Press

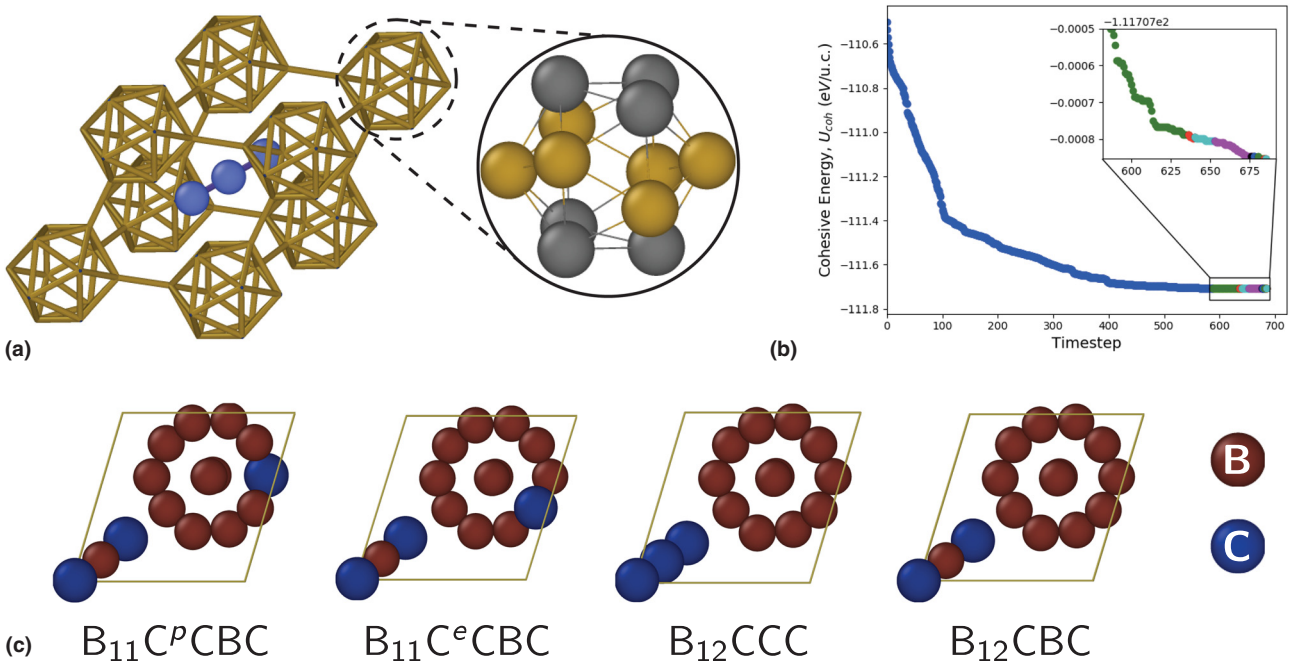



**Figure 1.** Boron carbide is a structurally complex polymorphic material. Its lightweight and toughness make it a great candidate for a wide variety of applications. (a) visualization of a B–C rhombohedral unit cell with 12-atom icosahedra at the vertices and 3-atom chain at the body, (b) energy minimization through different molecular statics stages, and (c) atomic configuration of four B–C polymorphs known to have the lowest energy (note $B_{12}CBC$ is boron-rich compared with the others).

## Framework for parameterizing interatomic potentials

Different parameterization approaches are routinely employed in the development of interatomic potentials. Nonetheless, systematic and transferable methodologies for potential parameterization are limited. When existing potentials need to be improved upon, parameter is modified based on assumptions of potential behavior. Due to its simplicity, a common practice consists of varying parameters in a one-factor-at-a-time (OFAT) fashion. The main limitation to OFAT is that there could be hidden effects (e.g., factor interactions) which it fails to capture. Factor interactions are of interest in hyperdimensional ($\mathbb{H}^D$) force fields where a number of factors could influence the property of interest. Herein, a systematic statistics-based approach is proposed to parameterize and optimize these potentials.

Figure 2 outlines the proposed systematic approach for parameterizing and optimizing interatomic potentials. The approach begins by adopting an initial-guess parameterization ($\mathbf{X}_0$). For instance, this could be a previous parameterization which needs to be improved upon to enhance its fidelity. Herein, the potential developed by An and Goddard[17] was chosen. Optimization of a large number of factors within the ReaxFF formulation can become cumbersome and expensive due to its hyperdimensionality. Therefore, the first step is to perform a SA to determine which parameters best capture response uncertainty. It is important to note that the goal is to identify a subset of factors which are exhibit *importance* (sensitivity) with respect to the responses of interest. Hence, parameter-space transformation approaches to dimensionality reduction, such as principle component analysis (which seek to project the data onto a reduced-order linearly uncorrelated factor space) are not directly applicable here. Identification of the most *sensible* parameters enables a reduction in the domain $\mathrm{H}^D \rightarrow \mathrm{H}^{d_r}$, where $D \gg d_r$. The dimensional reduction is achieved by comparing sensitivity metrics. This step should be repeated until a desirable level of variability is captured. Then, a traditional *design of experiments* (e.g., Latin Hypercube sampling) is performed on $\mathrm{H}^{d_r}$ yielding statistically representative data describing how this reduced domain can predict responses (desired material properties). The sampled reduced domain is then used to determine an optimal parameterization. For instance, the sampled set could be tested against several properties of interest to determine their accuracy. Alternatively, surrogate models can be fit to the sampled data set and optimized to determine an optimal parameterization from this reduced-order models. The latter could be performed at orders of magnitude lower computational expense required of computing MS for sampled candidate usually needed of methods typically used to optimize potentials.


https://doi.org/10.1557/mrc.2018.155 Published online by Cambridge University Press

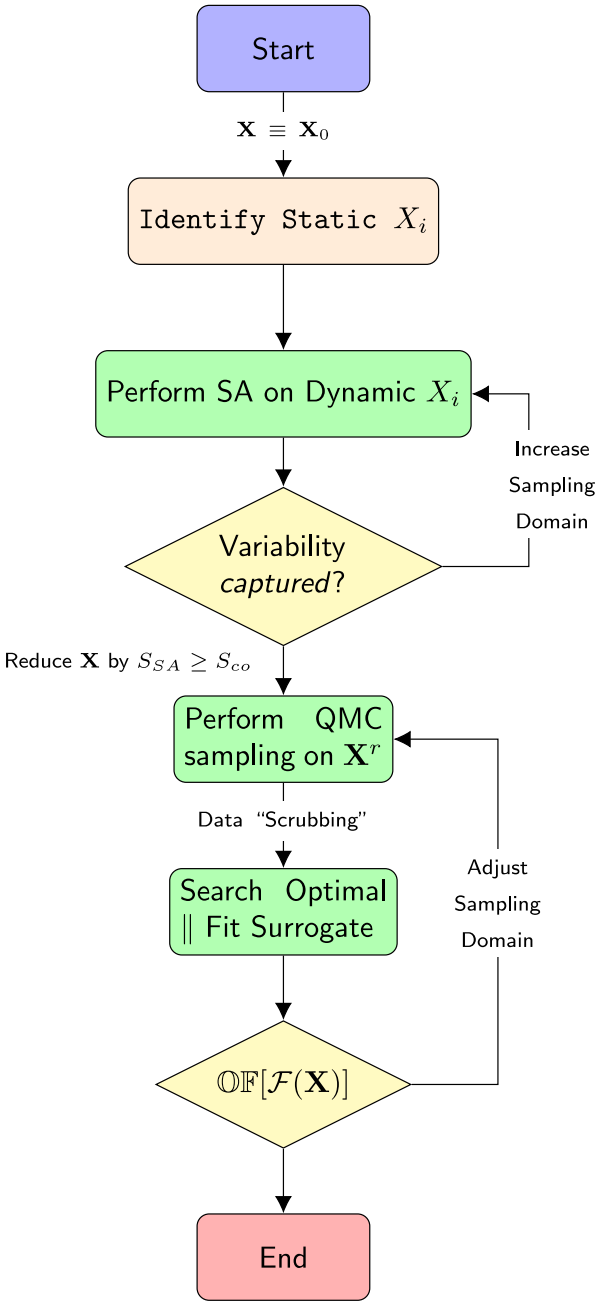



**Figure 2.** Framework workflow for parameterizing multidimensional interatomic potentials. The framework employs several statistical techniques, resulting in optimized potential(s) that predict properties calculated from molecular statics.

## Global SA enabled parameter screening

SA is a practice that helps understand how changes in independent input variables influence a model response. Understanding these relationships is paramount to being able to efficiently parameterize high fidelity interatomic potentials, especially for complex atomic systems like B–C. The ReaxFF potential permits parameterization of over 200 factors to accurately model B–C potentials, which lends itself to SA studies. However, not all these factors can be tunable as many correspond to material-specific properties (e.g. atomic mass).

**Table I.** Subset of parameters held constant from a comparative study of three prior B–C ReaxFF parameterizations from the literature[17,25]

| Potential | $N_{\text{total}}$ | Constant |
|---|---|---|
| General | 39 | 01, 13, 15, 19, 25, 29 |
| Atom | 64 | 02, 03, 08, 11, 13, 23, 27, 28, 29, 34, 35, 40, 43, 45, 49, 59, 60, 61 |
| Bond | 48 | 12, 15, 22, 28, 31, 37, 38, 39, 42, 44, 47 |
| Off-diagonal | 6 | 06 |
| Angle | 35 | 33 |
| Torsion | 42 | – |
| Total | 234 | 37 |

ReaxFF clearly presents dimensional challenges, and a complex computational and mathematical implementation, as evident from Eq. 1. Three prior B–C ReaxFF potentials[17,25] were used to identify ReaxFF factors should be held constant in the current study, deeming these "material specific" or insensitive factors. Table I summarizes the findings from this comparative analysis. The integers correspond to a parameter from the given functional subset and follow the potential template nomenclature in the Supplementary Information. For example, subset atom factor 03→ ATO03 corresponds to the C atomic mass. This approach suggests that 37 of the 234 fitting factors should remain constant, resulting in a $\mathbb{H}^{197}$ hyperdimensional domain. In other words, we need to determine a set of 197 parameters that yield an optimal ReaxFF parameterization to describe atomic interactions between B and C atoms.

Optimizing the resulting $\mathbb{H}^{197}$ still remains computationally expensive and challenging for tractable materials design. Therefore, SA is essential for determining which of these parameters should be targeted in parameterizing the potential. Multiple techniques have been developed to perform SA on different types of mathematical systems. These include OFAT, local methods, variance-based methods, etc.[26] However, the simplicity of some techniques like OFAT can miss interactions between factors.

This work focuses on the use of the derivative-based global sensitivity measure (DGSM), which systematically changes factors yielding a measure of the *global* sensitivity.[26,27] The DGSM is a derivative-based global method which consists of averaging the squared partial-derivative of each independent variable using quasi-random sampling techniques,

$$S^{i,m}_{\text{DGSM}} = \int_{\text{H}^{197}} \left( \frac{\partial f_m(\mathbf{X})}{\partial x_i} \right)^2 d\mathbf{X} \qquad (2)$$



https://doi.org/10.1557/mrc.2018.155 Published online by Cambridge University Press

where $\mathbf{X} = \{x_1, \ldots, x_i, \ldots, x_{197}\}$ is the distribution of randomly sample factors, $x_i$ is the independent input variable (factor) and $f_m(\mathbf{X})$ is the MS response of interest (e.g., for $B_{12}CCC$ 's cohesive energy $m = U_{\text{coh}}^{B_{12}CCC}$) obtained from MS minimization at each sampled potential. This approach has several commonalities to other better-known SA approaches like Morris screening and the Sobol sensitivity index but is able to overcome some of their limitations.[28] In fact, it is been shown that for uniformly and normally distributed input variables, the DGSM and Sobol's total sensitivity index ($S_{\text{tot}}$) are linked.[27] Further discussion on SA generalities and how it was implemented here can be found in the Supplementary Information.

The DGSMs were calculated by using the Sensitivity Analysis Library in Python (SALib) implementation.[29] SALib's finite difference sampler was used to probe $\mathbb{H}^{197}$, which performs a finite difference with a predefined delta ($\delta$) around a *base* Sobol sequence design. This sequence is determined by assigning upper and lower sampling bounds, which defines the hyperdimensional sampling volume. These bounds were chosen by defining a bounding sampling delta ($\Delta$) around the initial-guess potential

$$\{(1 - \Delta) \cdot \mathbf{X}_0 \leq \mathbf{X}_S \leq (1 + \Delta) \cdot \mathbf{X}_0\}$$

where $\mathbf{X}_S$ is the Sobol sequence and $\mathbf{X}_0$ is the origin potential developed by *Qi*.[21] The conventional finite difference sampling requires that multiples of $D + 1$ samples be evaluated. Therefore, a design of $N_T = n \cdot (D + 1) = 1980$ samples was generated with a $\delta = 0.015$ and a variable $\Delta$. The former delta influences the sampling distance from each Sobol sample, while the latter controls the domain volume to be sampled. It should be noted that derivative-based measures have been found to accurately identify influential input variables from small ($n \leq 10$) sample sizes.[30] Through MS energy minimizations on the sampled potentials, SA is performed on the objective response space ($O_{DGSM}$). The objective response space is composed of a set of "inexpensive" material properties

$$O_{\text{DGSM}}^{p} = \{U_{\text{coh}}^{p}, V^{p}, a^{p}, \alpha^{p}\}$$

where $p$ is a specific polymorph, $U_{\text{coh}}^{p}$ is the cohesive energy per unit cell, $V^{p}$ is the unit cell volume, $a^{p}$ is a lattice parameter and $\alpha^{p}$ is a lattice angle.

Limited to the 50 factors with the highest DGSM, the SA results are shown in Fig. 3, where the factors have been rank-ordered by their normalized DGSM. An example of $B_{11}C^{p}CBC$'s cohesive energy ($U_{\text{coh}}^{B_{11}C^{p}CBC}$) is shown in Fig. 3(a). SALib calculates a DGSM (blue bars) and a Sobol index (magenta bars) for each parameter as it relates to a specific response. The cumulative sum of the normalized DGSM (green) is representative of the variance captured. It is shown that these metrics do not agree in identifying which factor presents the highest impact on the energy response ($\max(S_{\text{Sobol}}^{f}) = S_{\text{DGSM}}^{7} \neq \max(S_{\text{DGSM}}^{f})$). Although the asymptotic saturation of the cumulative index supports our reasoning behind pursuing this approach (for dimensional reduction), interestingly it does not provide a clear directive on exactly

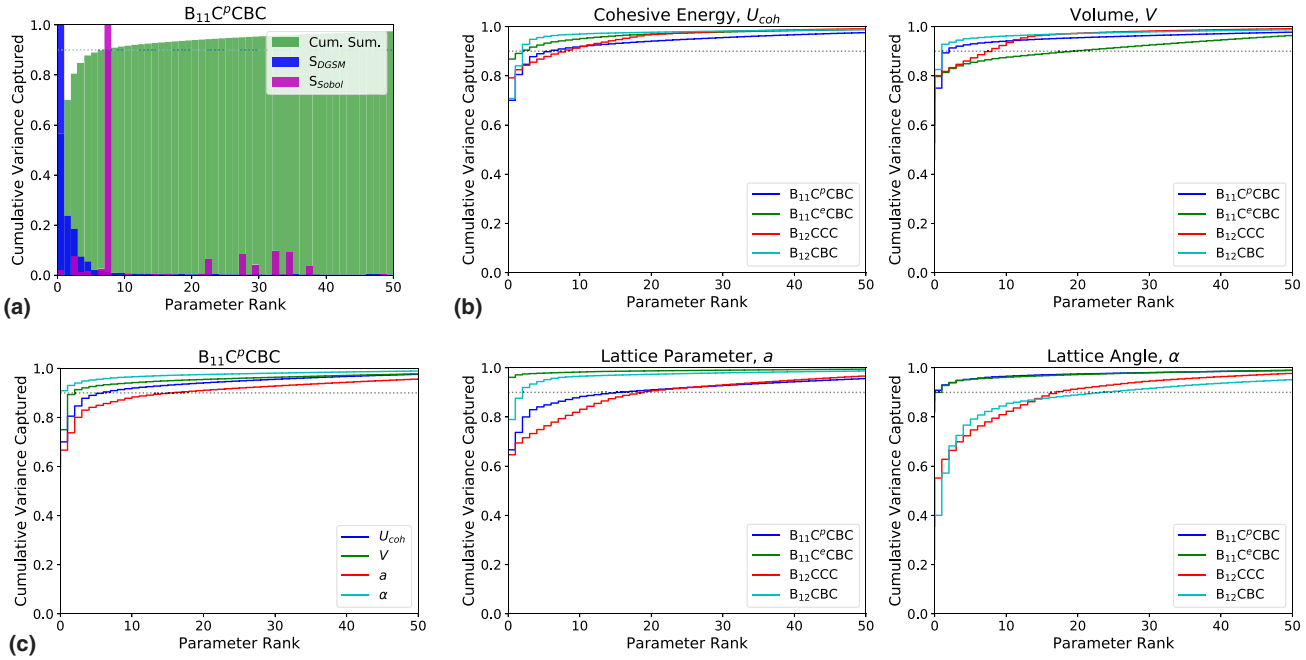



**Figure 3.** Rank ordered DGSM values show how a limited number of factors explain most of the variance. (a) Rank-ordered top 60 parameters based on the normalized DGSM ($S_{\text{DGSM}}$), its cumulative sum and normalized Sobol index ($S_{\text{Sobol}}$), (b) cumulative variability-measure for responses and polymorphs of interest, and (c) cumulative variability-measure for the different responses of $B_{11}C^{p}CBC$.


https://doi.org/10.1557/mrc.2018.155 Published online by Cambridge University Press

how many parameters should be considered. This manifest itself with a relatively long tail to the rank-ordered DGSM. Therefore, a set of rules must be employed in determining which factors are of relevance. After rank ordering, the cumulative sum of the DGSM was plotted. Then, a threshold of 90% of the cumulative sum was chosen as the cutoff criteria. This cutoff should contain enough information to explain 90% of the variability in the sampled data. This approach was performed across $O^{p}_{DGSM}$ and polymorphs, as shown in Fig. 3(b). From the $U_{coh}$ analysis, it is observed that about seven parameters are needed to explain 90% of the observed variability in $U^{B_{11}C^{p}CBC}_{coh}$, while about three parameters are needed for $U^{B_{12}CBC}_{coh}$. The number and type of factor that best explain the variability also changes as a function of response within a single polymorph, as shown in Fig. 3(c). Hence, to be able to properly describe variability for such a complicated system like B–C, a set of factors is taken from the union of all these subsets. The set of factors identified as of importance are listed in Table II, which are then defined as the reduced domain ($\mathbb{H}^{d_r} \rightarrow d_r = 55$). It should be noted that torsion factors are excluded from the table since none were classified as important.

## Objective and reduced domain sampling selection

The DGSM provides a measure of how much a factor is able to explain a response's variability, which allows the reduction of very large domains. Each parameter $S^{i,m}_{DGSM}$ was then used to enable screening a factor's ability to explain variability in objective response $Y_m$. Further, a good parameterization requires a potential to predict multiple responses, meaning parameterizing these potentials is a multi-objective optimization exercise. Here, the focus is on inexpensive properties as defined by the response set $O_{DGSM}$. Furthermore, the analysis focuses on the four lowest energy polymorphs, considering that these are the most abundant.[28]

However, it is important to understand the relationship between factors and responses, and how correlation between responses could be exploited in other ways. The Pearson correlation heatmap was plotted from the SA dataset, Fig. 4(a), which shows how responses correlate across the multiple polymorphs. It is shown that cohesive energies are clustered into a high correlation region, as are the volume and lattice parameter $a$. For instance, the cohesive energies for the $B_{11}C^{p}CBC$ and $B_{11}C^{e}CBC$ polymorphs, $|\rho(U(^{11}C^{p}),\ U(^{11}C^{e}))| = 0.96$, are highly linearly correlated. However, the lattice parameter for $B_{12}CCC$ and lattice angle for $B_{12}CBC$, $|\rho(a(^{12}C_l),\ \alpha(^{12}B_l))| = 0.26$, are not correlated. Therefore, within this sampling domain, the number of responses that must be fitted can be reduced, assuming that these correlations can be used to predictably estimate changes on $Y_j$ based on the observed changes on $Y_i$. For example, instead of using an objective function that aims to predict the volume of all the polymorphs, the volume of $B_{11}C^{p}CBC$ could be exclusively considered. Nonetheless, as these are fairly inexpensive properties, the set of properties analyzed ($U_{coh}$, $V$, $a$, and $\alpha$ for each polymorph) remained unchanged.

Interestingly, the lattice angle for all polymorphs are mostly uncorrelated. Nonetheless, it can be seen that for $B_{12}CCC$ its lattice angle and parameter are moderately correlated, $|\rho(a(^{12}C_l),\ \alpha(^{12}C_l)| = 0.87$, unlike other polymorphs. This could be due to the "symmetry disruption" that arises by the presence of C in other polymorphs' icosahedral or backbone sites. Further, $\alpha(^{11}C^{e})$ is the least correlated response as analyzed from the SA results, again likely due to the symmetry challenges and structural complexity created by adding a larger C atom into an equatorial site.

A proper exploration of the $\mathbb{H}^{197}$ domain requires that the $\Delta$ parameter is chosen accordingly. This $\Delta$ value must sample a domain that yields a response domain where all the responses are contained within the dataset. This means that the sampled potentials yield a response domain that contains the response target (i.e., DFT) values within it, e.g., $\min(U^{MS}_{coh,}) \leq U^{DFT}_{coh} \leq \max(U^{MS}_{coh})$. However, for very large $\Delta$, more Sobol samples are needed to properly sample the larger hyperdimensional volume. To determine whether the sampled volume is accurately capturing the responses, boxplots of the normalized responses are shown in Fig. 4(b) for two $\Delta$ settings. This normalization approach is similar to the "standard score", however, instead of *regressing* to the mean the data was normalized to its corresponding target

$$Y'_i = \frac{\boldsymbol{Y}^i - \boldsymbol{Y}^i}{\sigma_{\boldsymbol{Y}^i}} = \frac{\boldsymbol{Y}^i_{MS} - Y_{DFT}}{\sigma_{\boldsymbol{Y}^i_{Ms}}} \tag{3}$$

where $\boldsymbol{Y}^i_{MS}$ are the responses obtained from MS minimization for the $i$-th property, $Y_{DFT}$ is the target (DFT) value,[24] and $\sigma_{\boldsymbol{Y}^i_{MS}}$ is the response's standard deviation. Since the response distribution is "regressed to target", an ideal normalized response distribution encompassing the target values would be centered at zero. Therefore, the sampled potentials should yield a response distribution where the boxplot medians (red lines at notches) are approximately zero. From the $\Delta = 5\%$ case, it can be seen that the sampled potentials yield responses that are unlikely to accurately predict target values. Furthermore, the response distribution does not encompass all the objectives since $\max(a^{B_{11}C^{p}CBC}_{MS}) < a^{B_{11}C^{p}CBC}_{DFT}$. This is illustrated in Fig. 4(b), where the upper whisker of the $a^{B_{11}C^{p}CBC}_{MS}$ distribution (first red boxplot) falls below zero. This suggests that the sampling parameters ($\Delta$ and $\delta$) are not optimal for the objective set. While maintaining $\delta$ constant, the $\Delta$ parameter is increased to sample a larger domain on $\mathbb{H}^{197}$. The $\Delta = 7.25\%$ case is shown in Fig. 4(b) where the boxplots are shown to have shifted towards the zero line. While not perfectly aligned at zero, it is an improvement on $\Delta = 5\%$. Most importantly, the response distribution fully encompasses the objective set target. Therefore, the latter could be considered a better sampling of the hyperdimensional domain as applied to B–C, which contains more robust information.



https://doi.org/10.1557/mrc.2018.155 Published online by Cambridge University Press

https://doi.org/10.1557/mrc.2018.155 Published online by Cambridge University Press



**Table II.** Parameters identified as most sensitive by the DGSM analysis

| Response | | General | Atom | Bond | Off-diagonal | Angle |
|---|---|---|---|---|---|---|
| $B_{11}C^pCBC$ | $U_{coh}$ | 07 | 17, 20 | 26, 33 | 02 | 19, 30 |
| | $V$ | | 17, 20 | 02 | | 30 |
| | $a$ | 07, 31 | 01, 17, 20, 33, 62 | 02, 17, 18 | 02, 04, 05 | 16, 17, 19, 23, 28 |
| | $\alpha$ | | 20 | | | 30 |
| $B_{11}C^eCBC$ | $U_{coh}$ | | 20 | 17 | 02 | 16 |
| | $V$ | 10, 39 | 01, 12, 20, 22, 30, 32, 33, 41, 47 | 09, 17, 25, 26, 29, 46 | 02 | 17, 30, 35 |
| | $a$ | | 20 | | | |
| | $\alpha$ | | 20, 22 | 17 | | |
| $B_{12}CCC$ | $U_{coh}$ | 07, 31 | 20, 33, 64 | 26, 33, 46 | 02, 03 | |
| | $V$ | 07, 31 | 06, 14, 20 | 18, 45, 46 | 03 | 15 |
| | $a$ | 07, 31 | 06, 14, 20, 33, 54, 64 | 18, 26, 29, 30, 33, 45, 46 | 02, 03, 05 | 15, 16, 30 |
| | $\alpha$ | 07, 31 | 06, 14, 20, 25, 33, 64 | 17, 18, 30, 45, 46 | 03, 04, 05 | 15, 16, 30 |
| $B_{12}CBC$ | $U_{coh}$ | 07, 31 | 25, 46 | | | |
| | $V$ | 07 | 25, 46 | | | |
| | $a$ | 07, 31 | 25, 46 | | | |
| | $\alpha$ | 07, 18, 21, 31, 32, 33 | 07, 09, 12, 15, 17, 20, 21, 22, 25, 33, 46, 57 | 01, 02, 17, 46 | 03 | 08, 15, 16 |

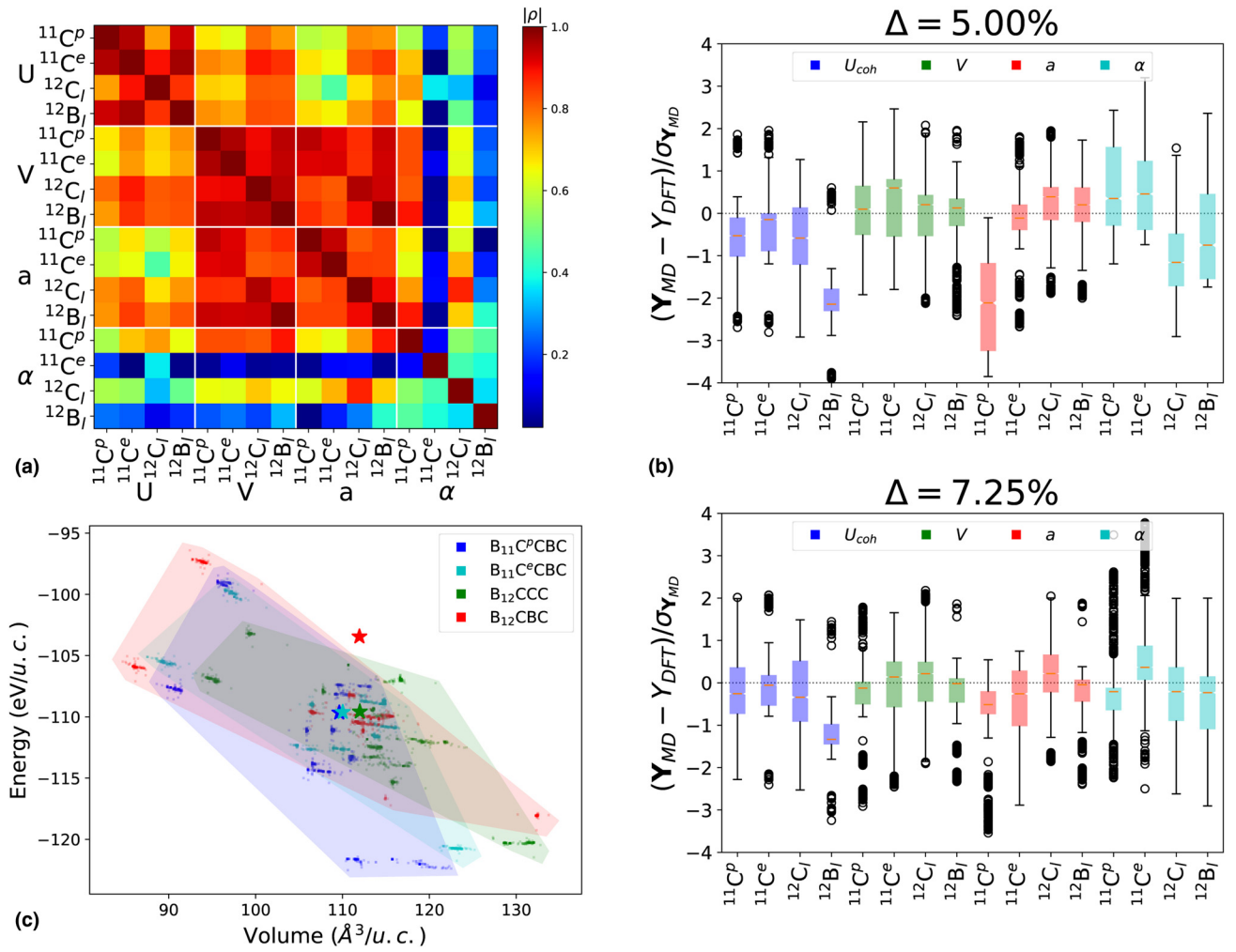



**Figure 4.** A global sensitivity analysis provides insights into the responses of interest. (a) Pearson correlation coefficient heatmap for energy, volume lattice constant, and angle for the polymorphs of interest, (b) response box-plots for two sampling bounds, and (c) Ashby-like plots of the resulting molecular statics energy minimization for the Δ = 7.25% case. (The following nomenclature was adopted: $^{11}C_p = B_{11}C^pCBC$, $^{11}C_e = B_{11}C^eCBC$, $^{12}C_l = B_{12}CCC$, and $^{12}B_l = B_{12}CBC$).

Understanding how factor modification results in property trade-offs are essential in optimizing multi-objective potentials. Ashby-like diagrams (convex-hulls) were generated for the volume and cohesive energy of the different polymorphs, Fig. 4(c). The target DFT values are shown as stars, with colors corresponding to the respective polymorph's Ashby diagram. This analysis clearly shows how the three lowest energy polymorphs fall within the center of the diagrams intersection (ca. 110 $Å^3$/u.c.,− 110 eV/u.c.). Hence, the resulting dataset contains enough information to properly describe the relationship between factors and these responses. However, the target values for $B_{12}CBC$ falls just outside of this information-rich region. While some of the sampled potentials within $\mathbf{X}_{DGSM}$ were able to predict the target values $U_{coh}^{B_{12}CBC}$ and $V^{B_{12}CBC}$ independently (see Fig. 4(b)), these potentials did not yield a response space that contained both simultaneously, i.e., $U_{coh}(V)|_{B_{12}CBC} \notin O_{DGSM}$. Nonetheless, it is expected that the selected sampling domain will contain enough information for a proper analysis of the ReaxFF potential. Last, it should be noted that the finite difference nature of the sampling can be appreciated from Fig. 4(c), where the scatter plots reflect the discreteness of the sampling technique.

## Reduced domain parametric exploration

With a significantly reduced domain ($\mathbb{H}^{55}$) to query, a focused search can be performed on the smaller hyperdimensional domain whose bounding volume was determined from the SA. This reduced domain was sampled using 1100 Latin Hypercube samples (LHS), which is a statistical technique that generates quasi-random samples for multidimensional independent variables. The LHS technique is often used to statistically characterize high-dimensional spaces since it can yield accurate results with a reduced number of samples, with respect



https://doi.org/10.1557/mrc.2018.155 Published online by Cambridge University Press

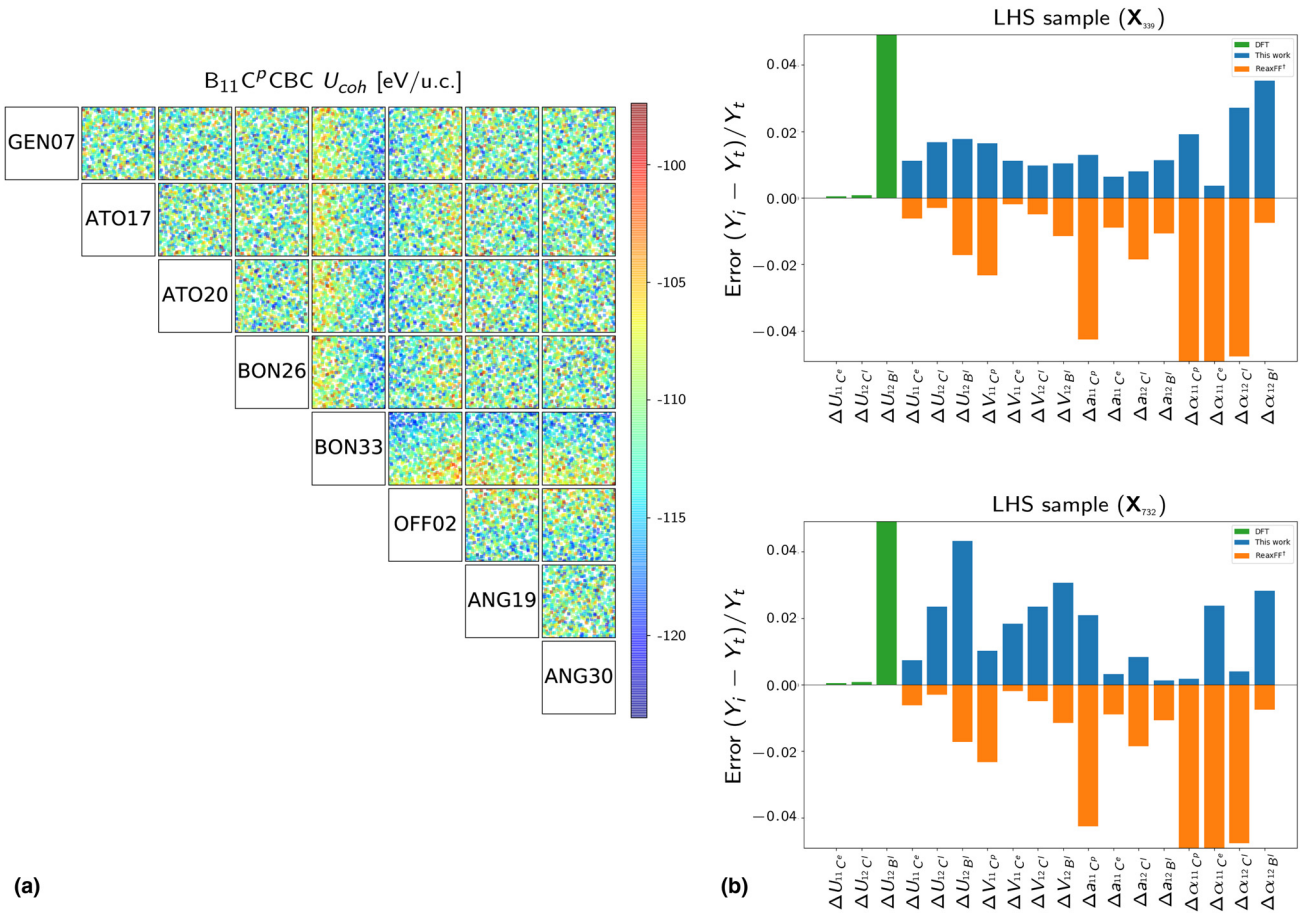



**Figure 5.** Exploratory and surrogate modeling-based techniques can be used to determine optimal parameterizations. (a) scatter plot matrix for $B_{11}C^pCBC$'s cohesive energy across an 8-dimensional space, and (b) performance of two candidate potentials (which are compared to the $\mathbf{X}_0$[17]).

to traditional Monte Carlo sampling. Since LHS produces a normalized design for each dimension with values between [0, 1], the designed samples were linearly transformed to distributions bounded by $\mathbf{X}_0 \pm \Delta$

$$\begin{aligned}\mathbf{X}_s &= (1 + \Delta \cdot (1 - 2\mathbf{X}_{\mathrm{LHS}})) \cdot \mathbf{X}_0 \\ &= \begin{cases} (1+\Delta)\cdot \mathbf{X}_0, & \mathbf{X}_{\mathrm{LHS}} = 0 \\ (1-\Delta)\cdot \mathbf{X}_0, & \mathbf{X}_{\mathrm{LHS}} = 1 \end{cases} \end{aligned} \quad (4)$$

where $\Delta$ defines the sampling distribution and $\mathbf{X}_{\mathrm{LHS}}$ is a LHS design composed of individual experiments ($(\mathbf{X}^i_{\mathrm{LHS}})$ being uniformly distributed about 0 and 1 $\{\mathbf{X}_{\mathrm{LHS}} \in \mathrm{R}^{55} : 0 \leq \mathbf{X}^i_{\mathrm{LHS}} \leq 1\}$. As for the SA, this sampling scheme is bounded by $\Delta = 7.25\%$ as this volume was found to be information-rich. While the idea to sample a larger volume is appealing, it must be noted that the input variables defining $\mathrm{H}^{d_r}$ correspond to this particular volume. Hence, exploring a larger volume might result in a different set of reduced parameters.

While the domain considered during sampling was relatively constrained, eight sampled potentials failed to converge to a stable atomic configuration during the MS energy minimization scheme. These were unsuccessful due to instabilities in the potential parameterization. Further, nine potentials converged to unphysical regimes (e.g., cohesive energies order of magnitude smaller than expected[24]). This unphysical regime was defined as sampled potentials resulting in values ± 10 "standard deviations" from the target (DFT) values, defined as $\sigma_{Y_i} \equiv \sqrt{Y_i} = (Y_i^{\mathrm{DFT}})^{1/2}$. Hence, a sampled parameterization $\mathbf{X}_i$ resulting in a cohesive energy of $|U(\mathbf{X}_i)| > \left|U_{\mathrm{DFT}} \pm 10\sqrt{U_{\mathrm{DFT}}}\right|$ would be filtered out from the 1100 sample dataset. These data were filtered to exclude potentials deemed to have resulted in *poor* convergence, yielding a dataset of 1083 successful potentials. The resulting "successful" dataset for $B_{11}C^pCBC$ is graphically shown and further discussed in the Supplementary Information.

A particular challenge to hyperdimensional models is data visualization, which can enable a further understanding of the model's behavior. To address this challenge, a matrix of scatter plots for the cohesive energy of the $B_{11}C^pCBC$ polymorph was chosen, shown in Fig. 5(a). Here, cohesive energy is represented as a function of the 8-dimensions identified in the SA as important factors (see Table II). Each box represents a 2-dimensional plane with the scattered markers representing the sampled factor


https://doi.org/10.1557/mrc.2018.155 Published online by Cambridge University Press

distribution, and the coloring representing the cohesive energy value for the corresponding potential. It can be noticed that while these factors were identified as important, that does not necessarily make them ideal for predicting the objective response at hand. For instance, no clear trend is observed from the ATO17-ATO20 plane. However, the ATO17-BON33 plane shows a clear trend where lower values of BON33 result in higher energies. Moreover, the BON33-OFF02 plane illustrates that energies become larger as OFF02 increases. Further analysis of this plane shows that there are no main interactions between these two factors (see Supplementary Information).

The approach followed in exploring the reduced hyperdimensional space yields an information-rich refined potential samples. An exploratory approach is employed as a first step to determine suitable parameterizations and provide rigorous analysis. Two tentative suitable parameterizations ($\mathbf{X}_{339}$ and $\mathbf{X}_{732}$) were identified from screening the successful samples. These were compared with the initial-guess potential $\mathbf{X}_0$, Fig. 5(b). The bar plots correspond to a property's error with respect to the target value, e.g., the volume error for the $i-$th sampled parameterization is

$$\Delta V = \left|\frac{V_{\mathbf{X}_i} - V}{V}\right| = \left|\frac{V_{\mathbf{X}_i} - V_{\mathrm{DFT}}}{V_{\mathrm{DFT}}}\right| \tag{6}$$

where $V \equiv V_{\mathrm{DFT}}$ is the target value. Note that these measures the absolute error, hence, the bar direction (sign) is not indicative of over/underestimation. Since cohesive energies depend on the DFT implementation, the corresponding cohesive energy errors were calculated with respect to its ground state $U_{\mathbf{X}_i}^{B_{11}C^p\mathrm{CBC}}$

$$\Delta U^{\mathrm{poly}} = \left|\frac{U_{\mathbf{X}_i}^{\mathrm{poly}} - U_{\mathbf{X}_i}^{B_{11}C^p\mathrm{CBC}}}{U_{\mathbf{X}_i}^{B_{11}C^p\mathrm{CBC}}}\right| \tag{7}$$

where $U_{\mathbf{X}_i}^{\mathrm{poly}}$ is the polymorph's cohesive energy for the $i-$th sampled potential. This was done since capturing the correct polymorphic energetic order/trend is more important than matching a specific cohesive energy value. Several characteristics can be observed when comparing with the $\mathbf{X}_0$ parameterization. First, DFT values yield a trend in polymorphic energy order such that $U_{\mathrm{DFT}}^{B_{11}C^e\mathrm{CBC}} < U_{\mathrm{DFT}}^{B_{12}\mathrm{CCC}} < U_{\mathrm{DFT}}^{B_{12}\mathrm{CBC}}$. This order is *not* captured by $\mathbf{X}_0$, which predicts $U_{\mathbf{X}_0}^{B_{12}\mathrm{CBC}} < U_{\mathbf{X}_0}^{B_{11}C^e\mathrm{CBC}}$. Therefore, this suggests that these sampled potentials ($\mathbf{X}_{339}$ and $\mathbf{X}_{732}$) are capturing some physics not observed on $\mathbf{X}_0$. While the refined sampling in this work yielded multiple parameterizations that capture this trend, none were as pronounced as the DFT case. It is worth noting that while parameterization $\mathbf{X}_{339}$ captured an average lower error across all the objectives, the polymorphic energy trend is very subtle. Parameterization $\mathbf{X}_{732}$ does have a more pronounced energy difference, but it results in higher errors for many of the responses.

## Conclusions

A systematic approach to optimizing hyperdimensional complex interatomic potentials is presented. The framework herein employs multiple statistical techniques, including SA and statistical sampling techniques. The ReaxFF formulation was used to showcase the proposed framework due to its parametric complexity, providing a challenging test case. Further, the framework was applied to the B–C system, which introduces structural and thermodynamic complexities external to the force field formulation. It is shown that a derivative-based SA enables dimensionality reduction (from about 200 to 55 factors), leading to a more manageable domain space. A matrix scatter plot was presented as a method for visualizing multidimensional data. It was shown that factors identified as important through the SA do not imply that these factors would make great predictors. Two trial parameterizations were compared with the initial-guess parameterization $\mathbf{X}_0$. The trial potentials obtained through the proposed framework capture the correct energetic order, as opposed to $\mathbf{X}_0$. An optimal parameterization likely requires multi-objective optimization with a response set that includes more expensive properties, e.g., elastic tensor.

## Supplementary material

The supplementary material for this article can be found at https://doi.org/10.1557/mrc.2018.155.

## References


1. J. Benedyk: *Aluminum Alloys for Lightweight Automotive Structures in Materials, Design and Manufacturing for Lightweight Vehicles* (Woodhead Publishing Series, Cambridge, 2010).
2. W. Joost: Reducing vehicle weight and improving U.S. energy efficiency using integrated computational materials engineering. *JOM* **64**, 1032 (2012).
3. H.M. Flower: *High Performance Materials in Aerospace* (Springer Science & Business Media, London, 2012).
4. J. McCauley, A. Crowson, W. Gooch, A. Rajendran, S. Bless, K. Logan, M. Normandia, and S. Wax: *Ceramic Armor Materials by Design* (John Wiley & Sons, Maui, 2012).
5. S.P. Coleman, E. Hernandez-Rivera, K.D. Behler, J. Synowczynski-Dunn, and M.A Tschopp: Challenges of engineering grain boundaries in boron-based armor ceramics. *JOM* **68**, 1605 (2016).
6. T.M. Pollock: Weight loss with magnesium alloys. *Science* **328**, 986 (2010).
7. V. Domnich, S. Reynaud, R.A. Haber, and M. Chhowalla: Boron carbide: structure, properties, and stability under stress. *JACerS* **94**, 3605 (2011).
8. C. Kaufmann, D. Cronin, M. Worswick, G. Pageau, and A. Beth: Influence of material properties on the ballistic performance of ceramics for personal body armour. *Shock Vib.* **10**, 51 (2003).
9. F. Thevenot: Boron carbide – a comprehensive review. *J. Euro. Cer. Soc.* **6**, 205 (1990).
10. M. Chen, J.W. McCauley, and K.J. Hemker: Shock-induced localized amorphization in boron carbide. *Science* **299**, 1563 (2003).
11. K.M. Reddy, P. Liu, A. Hirata, T. Fujita, and M.W. Chen: Atomic structure of amorphous shear bands in boron carbide. *Nature Comms.* **4**, 2483 (2013).
12. M.F. Horstemeyer: *Integrated Computational Materials Engineering (ICME) for Metals: Using Multiscale Modeling to Invigorate Engineering Design with Science* (John Wiley & Sons, Hoboken, 2012).
13. S.J. Plimpton and A.P. Thompson: Computational aspects of many-body potentials. *MRS Bull.* **37**, 513 (2012).




https://doi.org/10.1557/mrc.2018.155 Published online by Cambridge University Press

14. A.C.T. van Duin, S. Dasgupta, F. Lorant, and W.A. Goddard: ReaxFF: a reactive force field for hydrocarbons. *J. Phys. Chem. A* **105**, 9396 (2001).
15. D.W. Brenner, O.A. Shenderova, J.A. Harrison, S.J. Stuart, B. Ni, and S. B. Sinnott: A second-generation reactive empirical bond order (REBO) potential energy expression for hydrocarbons. *J. Phys.: Cond. Matt.* **14**, 783 (2002).
16. J. Yu, S.B. Sinnott, and S.R. Phillpot: Charge optimized many-body potential for the Si/$SiO_2$ system. *Phys. Rev. B* **75**, 085311 (2007).
17. Q. An, and W.A. Goddard III: Atomistic Origin of Brittle Failure of Boron Carbide From Large-Scale Reactive Dynamics Simulations: Suggestions Toward Improved Ductility. *Phys. Rev. Letts.* **115**, 105501 (2015).
18. A. Jaramillo-Botero, S. Naserifar, and W.A. Goddard III: General multiobjective force field optimization framework, with application to reactive force fields for silicon carbide. *J. Chem. Theo. Comp.* **10**, 1426 (2014).
19. B.M. Rice, J.P. Larentzos, E.F. Byrd, and N.S. Weingarten: Parameterizing complex reactive force fields using Multiple Objective Evolutionary Strategies (MOES): Part 2: transferability of ReaxFF models to C-H-N-O energetic materials: *J. Chem. Theo. Comp.* **11**, 392 (2015).
20. J.A. Martinez, A. Chernatynskiy, D.E. Yilmaz, T. Liang, S.B. Sinnott, and S.R. Phillpot: Potential optimization software for materials (POSMat). *Comp. Phys. Comms.* **203**, 201 (2016).
21. M.A. Tschopp, K. Solanki, M.I. Baskes, F. Gao, X. Sun, and M. F. Horstemeyer: Generalized framework for interatomic potential design: application to Fe-He system. *J. Nuc. Mats.* **425**, 22 (2012).
22. D. Gosset and M. Colin: Boron carbides of various compositions: an improved method for X-rays characterisation. *J. Nuc. Mats.* **183**, 161 (1991).
23. G. Fanchini, J.W. McCauley, and M. Chhowalla: Behavior of disordered boron carbide under stress. *Phys. Rev. Letts.* **97**, 035502 (2006).
24. D.E. Taylor, J.W. McCauley, and T.W. Wright: The effects of stoichiometry on the mechanical properties of icosahedral boron carbide under loading. *J. Phys.: Cond. Matter* **24**, 505402 (2012).
25. R. Paupitz, C.E. Junkermeier, A.C.T. van Duin, and P.S. Branicio: Fullerenes generated from porous structures. *Phys. Chem. Chem. Phys.* **16**, 25515 (2014).
26. A. Saltelli, M. Ratto, T. Andres, F. Campolongo, J. Cariboni, D. Gatelli, M. Saisana, and S. Tarantola: *Global Sensitivity Analysis: The Primer* (John Wiley & Sons, New York, 2008).
27. I.M. Sobol and S. Kucherenko: Derivative based global sensitivity measures and their links with global sensitivity indices. *Math. Comps. Sim.* **79**, 3009 (2009).
28. S. Kucherenko and B. Iooss: *Derivative-Based Global Sensitivity Measures in Handbook of Uncertainty Quantification* (Springer International Publishing, Berlin, 2016).
29. J. Herman and W. Usher: SALib: an open-source Python library for sensitivity analysis. *J. Open Source Software* **2**, 97 (2017).
30. W.E. Becker, S. Tarantola, and G. Deman: Sensitivity analysis approaches to high- dimensional screening problems at low sample size. *J. Stat. Comp. Sim.* **88**, 2089 (2018).


https://doi.org/10.1557/mrc.2018.155 Published online by Cambridge University Press