Supplementary Material: Integrating Exploratory Data analytics into ReaxFF Parameterization

Efrań Hernández–Rivera,[*,†] Souma Chowdhury,[‡] Shawn P. Coleman,[†] Payam Ghassemi,[‡] and Mark A. Tschopp[†]

[†] *U.S. Army Research Laboratory, Weapons and Materials Research Directorate, APG, MD 21005*

[‡] *University at Buffalo, Department of Mechanical and Aerospace Engineering, Buffalo, NY 14260*

E-mail: efrain.hernandez18.civ@mail.mil

Sensitivity Analysis: A Brief Description

It is often important to understand how uncertainty in input variables (factors) can lead to uncertainty in model responses (output). Sensitivity analysis (SA) is a statistical technique for quantifying this relationship. Multiple SA methods have been developed, to include the Morris method (a one-factor-at-a-time global sensitivity method) and the Sobol method (a variance-based global sensitivity measure). The Morris method[1] employs the use of the so called elementary effect (`EE`) which for each i$-th$ factor is given as the finite differentiation

$$\texttt{EE}_i = \frac{f(\mathbf{X}_1, \cdots, \mathbf{X}_{i-1}, \mathbf{X}_i + \Delta, \cdots, \mathbf{X}_k) - f(\mathbf{X}_0)}{\Delta} \tag{1}$$

where Δ is a predetermined multiple of $1/(p-1)$. It should be noted that for a $k-$dimensional, the sampling range for each factor is divided into $p-$levels essential forming a gridded sampling. The Sobol index[2] is a variance-based measure of the first-order sensitivity of a given factor

$$S_i = \frac{\mathtt{Var}_i\left(\mathbb{E}_{\mathbf{X}_{j\neq i}}(\mathbf{Y}|\mathbf{X}_i)\right)}{\mathtt{Var}(\mathbf{Y})} \tag{2}$$

where the index is given by the variance of the expected values divided by the total variance. These indices aim to decompose the variance into attributable to each model variable, i.e., how much (percentage wise) does variable $\mathbf{X}_i$ contribute to the variance in $\mathbf{Y}$.

An alternative variance-based method is the derivative based global sensitivity measure (DGSM),[3] which is becoming popular among practitioners for sensitivity analysis of high-dimensional models. The DGSM has links to the Morris factor and Sobol index, but can outperform these for high-dimensional problem[4] and is defined as[3–5]

$$S^i_{DGSM} = \int_{\mathbb{H}^k} \left(\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}_i}\right)^2 d\mu(\mathbf{x}) = \mathbb{E}\left[\left(\frac{\partial f(\mathbf{X})}{\partial \mathbf{X}_i}\right)^2\right] \tag{3}$$

where $\mu(\mathbf{x})$ is the distribution of the $\mathbf{X}$ independent variables. Each variable sensitivity is then calculated through numerical integral by quadratures as $\partial f/\partial \mathbf{X}_i$ is square-integrable. To implement this approach, python's Sensitivity Analysis Library in Python (SALib)[6] was adopted. The sensitivities were then calculated through the use of the `SALib.analyze.dgsm` module. It should be noted that these are a few of the many SA methods available. Nonetheless, the DGSM has been shown to be robust in converging for high-dimensional low sample size analysis.[7]

Sampling for the DGSM was carried out by using the `SALib.sample.finite_diff` module. This is similar to sampling techniques used for the Morris method, but instead of changing factors one-at-a-time all factors are changed simultaneously. Then, the domain around the new sample position is probed. Figure S1 shows a schematic representation of this sampling mechanics for a 3-dimensional model with five samples. An original parameterization ($\mathbf{X}_0 = (0.5, 0.5, 0.5)$, blue in Figure S1) and a predetermined $\Delta = 0.5$ defines the bounding domain. The first step then is to sample the space around that point by another predetermined parameter δ, which is represented by the "axes" about $\mathbf{X}_0$. Once sampled, a new sample is randomly generated within the sampling domain and the same exploration about δ is performed. This process is repeated $N_T = n \cdot (D+1)$ in order to properly calculate the sensitivity measures.

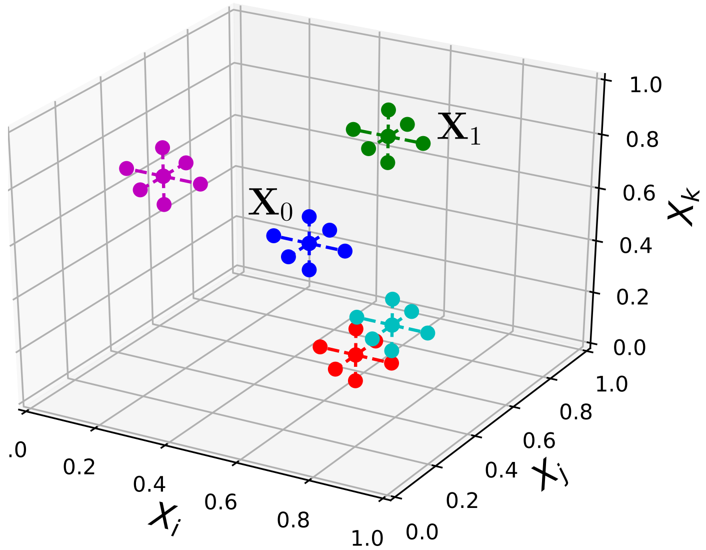

Figure S1: Schematic of the sampling scheme used to perform the sensitivity analysis.

Data Scrubbing

As multiple parameterizations resulted in "bad" potentials, e.g., failed convergence, the resulting dataset required curation. This process needs to be carried our carefully as to not introduce artificial biases. The resulting dataset, after scrubbing out bad parameterizations for the $\mathrm{B}_{11}\mathrm{C}^p\mathrm{CBC}$ objectives, is shown in Figure S2 as scatter plots. These show that the resulting data set is able to capture the expected values (plotted as orange stars) and these are reasonably normally distributed. It is also clear, as in Figure 4(a), that V and a linearly correlate, while others are largely uncorrelated.

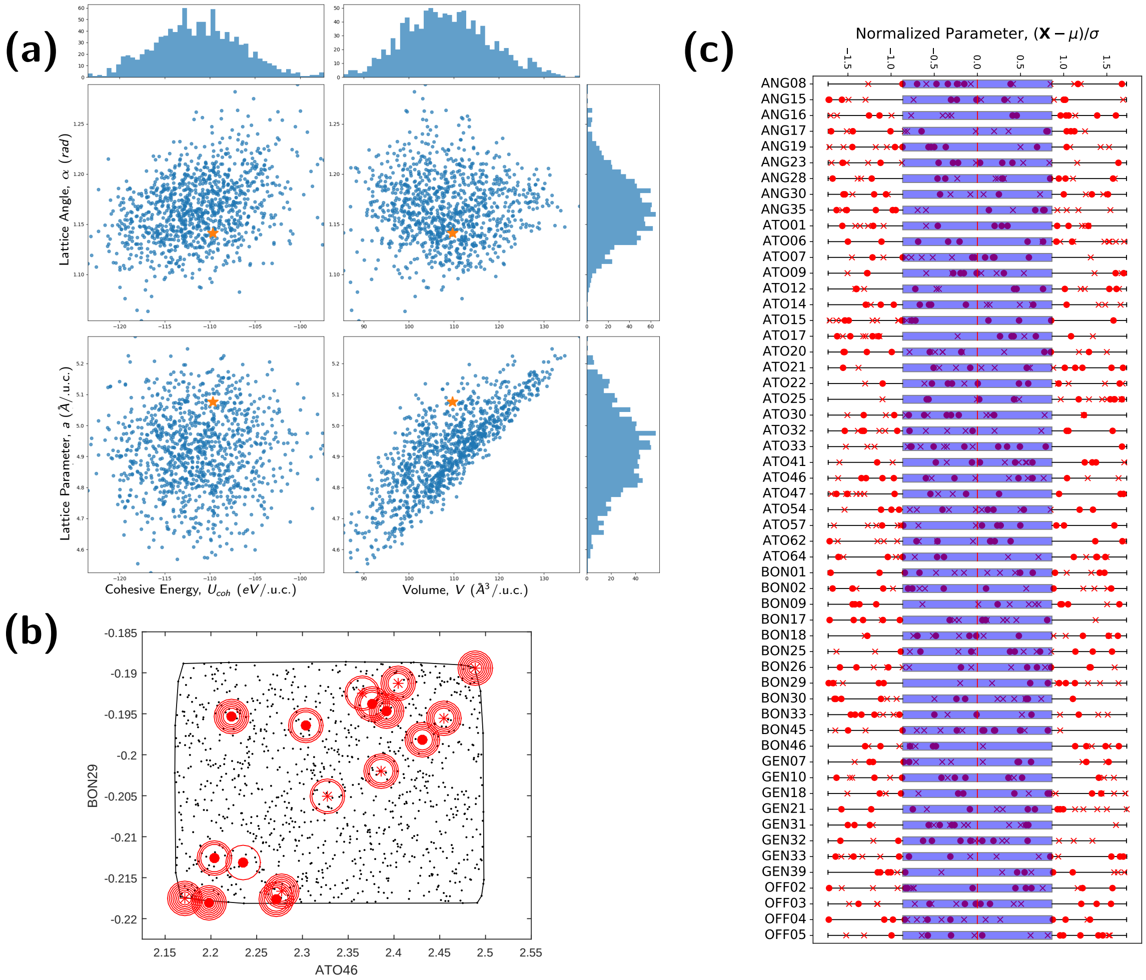

Figure S2: Filtering LHS data to remove outliers and failed potentials. (a) distribution of successful sampled potentials for $B_{11}C^pCBC$, (b) convex-hull based screening of failed samples, and (c) box plots of normalized factors with failed samples overlaid (marked as red: cross for non-convergence and circles for minimization outliers).

Identifying where $\mathbb{H}^{55}$ fails to converge to a minimized structure could enable a focused sampling scheme as not all parameters must be sampled under identical conditions, e.g., $\pm\Delta$. A visual representation of the highlighted failed potentials is shown in Figure S2(b-c). Due to the dimensional complexity, a hyperdimensional convex hull analysis was performed to identify the smallest set of factors that can predict the failed samples, Figure S2(b). In this manner, the hyperdimensional convex hull could be used to screen samples in undesirable parts of the design space. This concept could be used for resampling the domain space or as non-orthogonal bounds for the optimization space as well.

The analysis is as follows. First, the convex hulls were created from the successful samples for the n-dimensional space. Then, the number of failed samples outside the convex hull is assessed (each red dot or red cross corresponds to a failed sample). The goal is to compose a hyperdimensional convex hull that can predict failed samples with as few dimensions as possible. Each circle around the failed samples represents that it lied outside of the convex hull (e.g., 5 circles means that it lied outside in 4th, 5th, 6th, 7th, and 8th dimensional convex hulls). For two dimensions, the `ATO46`–`BON29` plane explains the most failed samples, two at the bottom left and one at the top right. In other words, within this 2D slice through our 55-dimensional space, the convex hull created from the successful samples is able to exclude the most failed samples (in this case, 3 failed samples). For three dimensions, the `ATO46`–`BON29` plane was retained and all other potential parameters were iterated over to find the best one. Subsequent dimensions (input variables) were added (retaining the prior dimensions parameters) until all failed samples were outside of the multi-dimensional convex hull. A total of 8-dimensions was needed to create a convex hull volume that fully excludes all failed parameterizations. It should be noted that these are not unique sets and other parameter combinations could explain failure within the sampled data. A 1-dimensional equivalent analysis would be to visualize box plots for each input parameter, Figure S2(c). Clearly, no single parameter can effectively predict whether a sampled parameterization will fail. In fact, these failed samples are fairly distributed along the sampled volume. Focusing on the variables identified by the 2-dimensional convex hull, it is clear that the 1-dimensional approach is not able to refine search spaces.

Surrogate Modeling of Cohesive Energy

As suggested by the proposed framework, surrogate models could be used to predict molecular statics optimized properties. Here, a two-dimensional surrogate linear model predicting $B_{11}C^{p}$CBC's cohesive energy was fit to the successful dataset to showcase this approach. An ordinary least squares regression was used to fit a simple linear model to the dataset X_S

$$U_{coh}^{11C^p} = \beta_0 + \sum_{j=1}^{2} \beta_j \mathbf{X}_j$$

where β_j is the regression coefficient of parameter$-j$ (with distribution $\mathbf{X}_j$) and β_0 is the intercept. Figure S3(a) shows the resulting response surface overlaid by the scatter data from the molecular static minimizations. A good fit results in identical coloring between the scatter data and the background response surface. While the correct trend is captured, higher BON33 and BON17 values lead to lower energies, a significant amount of noise is present in the data. This noise emerges since the molecular statics results were obtained from changing 55 parameters, and not just these two. Hence, a bootstrapping approach to fit the linear function was implemented, as shown in Figure S3(b). This approach yields understanding of the expected error on the fitted linear coefficients.

The coefficient (β) distributions largely depend on the randomly sampled data subset size (n_{LHS}) during the bootstrap analysis. In fact, for $n_{LHS} < 300$, bimodal distributions were observed for all three $\beta-$coefficients. This suggests that a sample size $n \geq 300$ should be analyzed in fitting a linear model to this property. By examining the overlaid scatter data on the response surface, Figure S3(a), it can be appreciated that small sample sets lead to widely divergent fits. For instance, focusing on the lower end of the surface plot (yellow region), a significant amount of light green scatter dots can be observed (i.e., lower energy than predicted). The bimodal nature of the scatter data with its non-negligible overestimated cohesive energy leads to different linear models. Therefore, a small n_{LHS} increases the likelihood of biasing the regression towards outlying data.

A potential extension of this approach will be to use nonlinear surrogate models such as Radial basis functions, Neural Networks or Kriging, trained through an automated modeling framework,[8] in order to provide more reliable representation of the functional relationships between the outputs of interest (e.g., cohesive energy and unit cell volume) and the interatomic potentials.

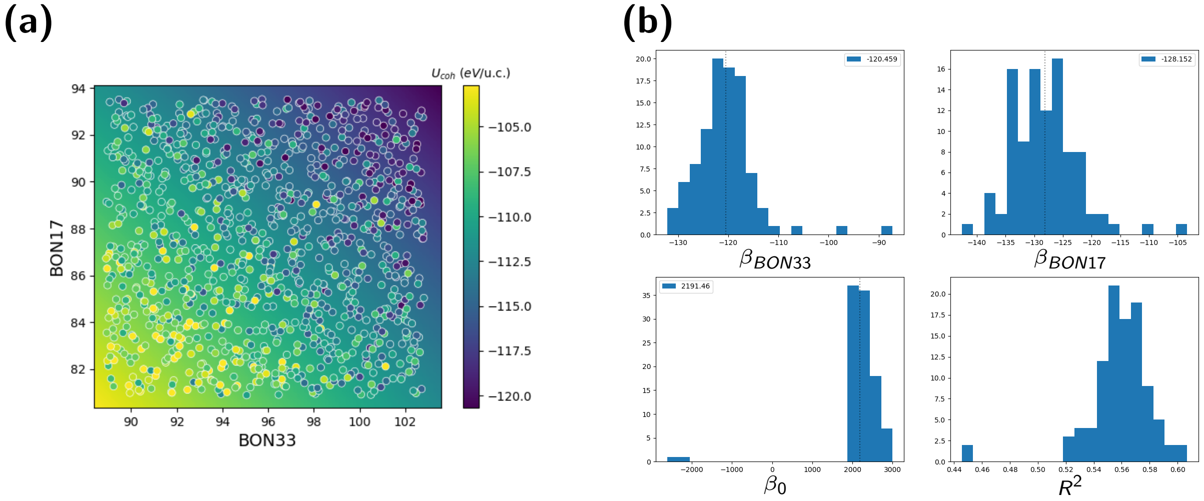

Figure S3: Surrogate modeling based techniques can be used to determine optimal parameterizations. (a) two-dimensional response surface for the B11CpCBC cohesive energy, and (b) linear model coefficients determined from bootstrapped fitting.

An alternative use of these surrogate models can be their use to determine whether important factors interact. As shown in Figure 5(a), the `BON33-OFF02` plane illustrates a clear trend between these factors and the objective of interest. Therefore, fitting the data to different linear formulations can enable further understanding of the interaction between factors in predicting objectives. This is shown in Figure S4.

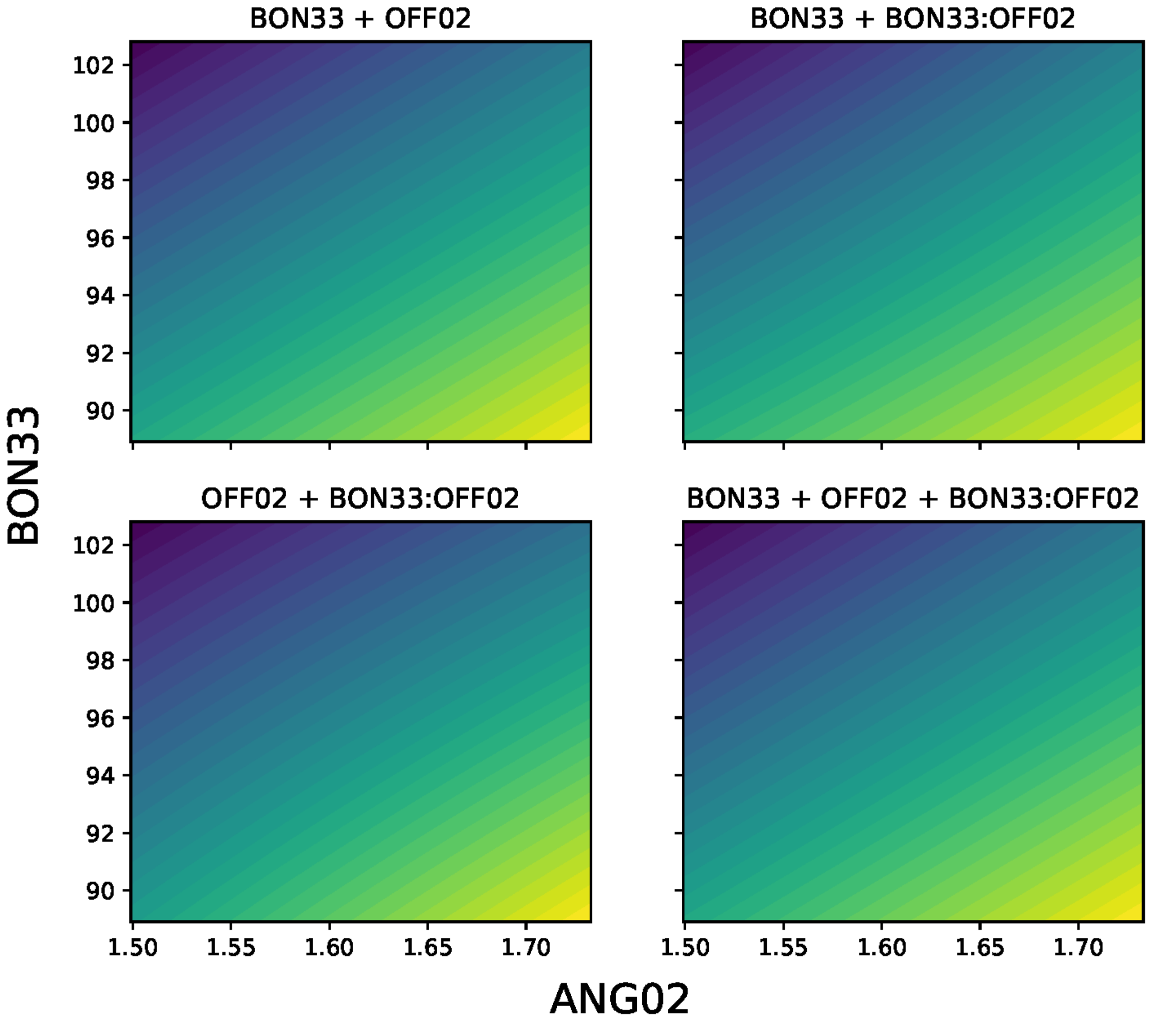

Figure S4: Different surrogate linear formulations showing the factors in the plane at hand do no interact. Each subplot title describes the surrogate model used, e.g., the one in the bottom right is $U = \beta_0 + \sum_{i=0}^{B33,A02} \left(\beta_i X_i + \sum_{j \neq i}^{B33,A02} \beta_{ij} X_i X_j \right)$.

Data Archive Directory Tree

A zipped file containing the required data to reproduce the results presented in the manuscript are provided. Figure S5 shows the files contained, their location, and brief descriptions. The dataset include MD energy minimization for each polymorph and sampled potential. The polymorphs (ID = `poly`) are numbered as: 1) $B_{11}C^{p}CBC$, 2) $B_{11}C^{e}CBC$, 3) $B_{12}CCC$, and 4) $B_{12}CBC$.

- supp: Supplemental information archive
  - input: LAMMPS input script
    - B4C.in: template input file for molecular energy minimization
  - DGSM
    - POTS: ReaxxFF potentials sampled for the sensitivity analysis
      - 5_perc: sampled potentials for $\Delta = 5\%$
        - reax.dgsm-0001.txt
        - reax.dgsm-0002.txt
        - reax.dgsm-0003.txt
        - ⋮
        - reax.dgsm-1980.txt
      - 7-25_perc: sampled potentials for $\Delta = 7.25\%$
        - reax.dgsm-0001.txt
        - reax.dgsm-0002.txt
        - reax.dgsm-0003.txt
        - ⋮
        - reax.dgsm-1980.txt
    - SA_DGSM.txt: compiled dataset with independent input variables and responses for the $\Delta = 7.25\%$ case
  - OptSearch
    - POTS: ReaxxFF potentials sampled for focused optimal search
      - reax.lhs-0001.txt
      - reax.lhs-0002.txt
      - reax.lhs-0003.txt
      - ⋮
      - reax.lhs-1110.txt
    - DoE_LHS.txt: compiled dataset with independent input variables and responses
  - misc
    - reaxff.template.txt: ReaxFF template with variable values replaced with corresponding labels

Figure S5: Schematic of the directory tree within the supp.zip archive. The archive includes all sampled potentials (for SA and reduced domain optimization search), resulting databases, and input file for LAMMPS minimization scheme.

References

(1) Morris, M. *Technometrics* **1991**, *33*, 161–174.

(2) Sobol, I. **2001**, *55*, 271–280.

(3) Sobol, I.; Kucherenko, S. *Mathematics and Computers in Simulation* **2009**, *79*, 3009–3017.

(4) Kucherenko, S.; Song, S. *Monte Carlo and Quasi-Monte Carlo Methods*; Springer, 2016; pp 455–469.

(5) Lamboni, M.; Iooss, B.; Popelin, A.; Gamboa, F. *Mathematics and Computers in Simulation* **2013**, *87*, 45–54.

(6) Herman, J.; Usher, W. *Journal of Open Source Software* **2017**, *2*.

(7) Becker, W.; Tarantola, S.; Deman, G. *Journal of Statistical Computation and Simulation* **2018**, 2089–2110.

(8) Mehmani, A.; Chowdhury, S.; Meinrenken, C.; Messac, A. *Structural and Multidisciplinary Optimization* **2017**, 1–22.