Five Degree-of-Freedom Property Interpolation of Arbitrary Grain Boundaries via Voronoi Fundamental Zone Octonion Framework: Supplementary Information

Sterling G. Baird[a,*], Eric R. Homer[a], David T. Fullwood[a], Oliver K. Johnson[a]

[a]*Department of Mechanical Engineering, Brigham Young University, Provo, UT 84602, USA*

Contents

S1. Euclidean and Arc Length Distances

The close correlation between Euclidean and arc length distances in the Voronoi Fundamental Zone octonion (VFZO) sense is shown in Figure S1 using pairwise distances of 10 000 VFZOs. This justifies our use of Euclidean distance as an approximation of hyperspherical arc length (and by extension, that a scaled Euclidean distance approximates a non-symmetrized octonion distance, see Eqs. (1)–(3) of the main paper). However, comparison with the original octonion metric [1] gives overestimation for some boundaries. This is an inherent feature of the VFZO framework that can be addressed via use of the ensemble methods described in Section 2.1.3 (see also Figures 3 and 4).

*Corresponding author.
Email address: ster.g.baird@gmail.com (Sterling G. Baird)

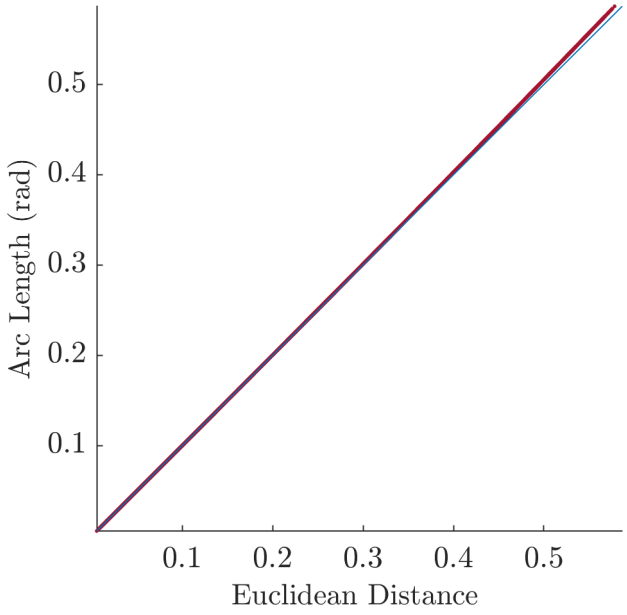

Figure S1: Parity plot of 8D Cartesian hyperspherical arc length vs. 8D Cartesian Euclidean distance for pairwise distances in a $(m\bar{3}m)$ symmetrized set of 10 000 randomly sampled VFZOs. The max arc length is approximately 0.58 rad, indicating a max octonion distance of approximately 1.16 rad or 66.5° between any two points in the Voronoi Fundamental Zone (VFZ). The close correlation between arc length and Euclidean distance supports the validity of using Euclidean distance instead of arc length in the interpolation methods. This is *separate* from the correlation between VFZO Euclidean or arc length distances with the traditional octonion distance [2].

S2. Ensemble Interpolation Results

Ensemble interpolation is a classic technique that can be used to enhance predictive performance of models. Here we describe our methods (Section S2.1), results (Section S2.2), and the potential of integrating ensemble interpolation with a Gaussian process regression mixture (GPRM) scheme (Section S2.3).

S2.1. Methods

VFZO ensemble[1] interpolation occurs by:

1. generating multiple reference octonions to define multiple VFZs
2. obtaining multiple VFZO representations for a set of grain boundaries (GBs) based on the various reference octonions
3. performing an interpolation (e.g. Gaussian process regression (GPR)) for each of the representations
4. homogenizing the ensemble of models (e.g. by taking the mean or median of the various models)

[1] Ours is a "bagging"-esque ensemble scheme because the same interpolation method (GPR) is used but with different representations for the input data.

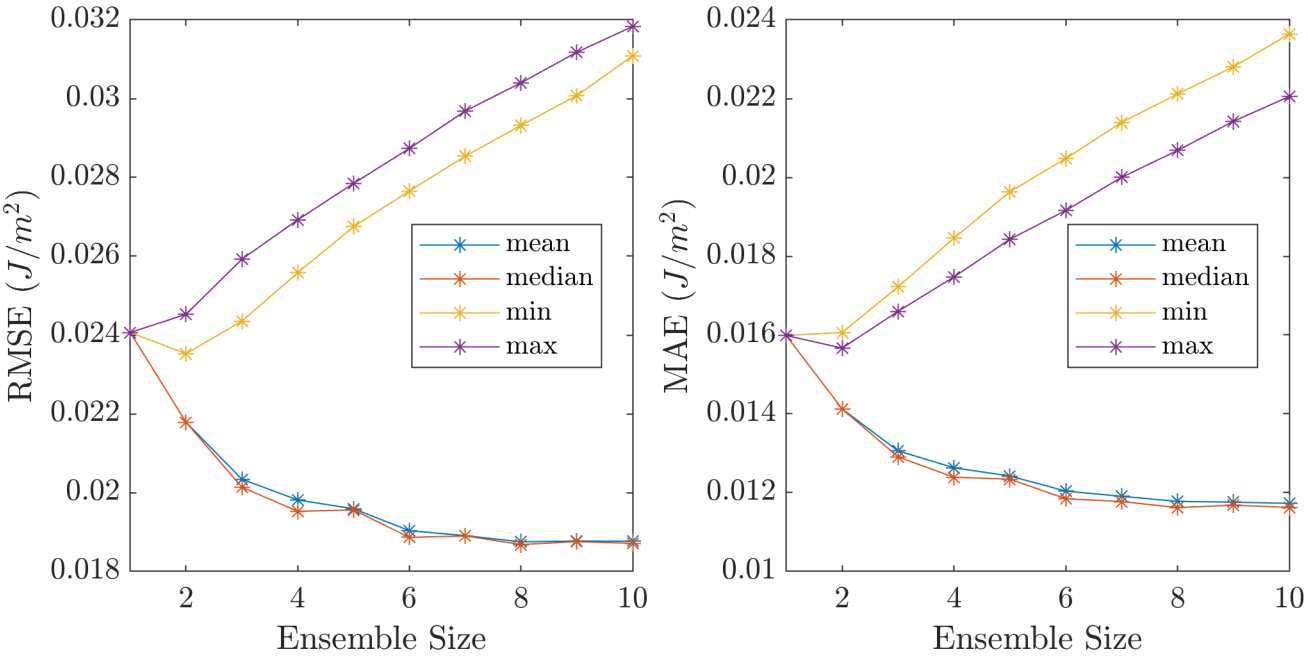

Figure S2: (a) RMSE and (b) MAE vs. ensemble size for mean, median, minimum, and maximum homogenization functions. A GPR model with 50 000 input and 10 000 prediction VFZOs was used.

S2.2. Results

Use of an ensemble interpolation scheme decreases interpolation error for a GPR model with 50 000 input and 10 000 prediction VFZOs. By using an ensemble size of 10 (i.e. 10 GPR models each with different reference octonions and therefore different VFZs), root mean square error (RMSE) and mean absolute error (MAE) decreased from 0.0241 J m^{-2} and 0.0160 J m^{-2} to 0.0187 J m^{-2} and 0.0116 J m^{-2}, respectively, using the median homogenization function (Figure S2).

Figure S3 shows the hexagonally binned parity plots for predictions made using the mean, median, minimum, and maximum predicted values over an ensemble of 10 VFZs. Qualitatively, the ensemble mean and ensemble median parity plots look similar to those from the main text (Figure 6), though the distributions of the ensemble scheme are somewhat tighter. The ensemble minimum produces better predictions of low grain boundary energy (GBE) than any of the other models, but underestimates high GBE as expected. Naturally, the ensemble maximum overestimates in general. Diminishing returns manifest in Figure S2 for mean and median homogenizations. This is to be expected because the original octonion distances [1] are well-approximated using an ensemble size of 10 (Figure 3c and Figure 4).

S2.3. Possibility: Combining Ensemble with Gaussian Process Regression Mixture

A scheme which preferentially favors the ensemble minimum for low GBE predictions and defaults to ensemble mean or median for all other GBEs may produce even better results across the full range of GBEs. For example, this could be accomplished by combining the ensemble scheme described here with the GPR mixture model described in Section S4.1.

S3. Barycentric Interpolation

S3.1. High-Aspect Ratios

An artifact of the barycentric interpolation method which occurs due to the presence of high-aspect ratio facets is shown in Figure S4. As the dimensionality increases for a constant number of points and

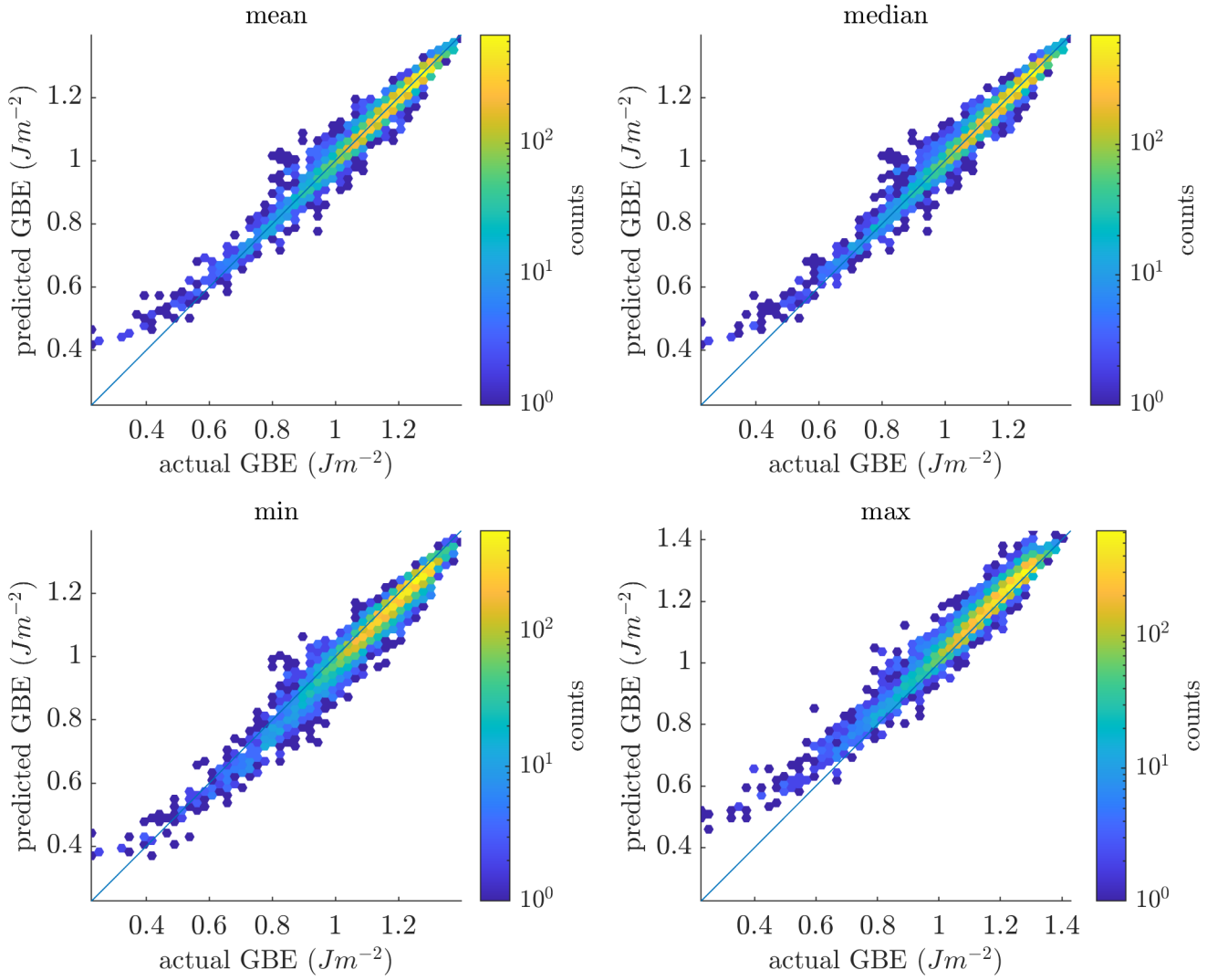

Figure S3: Hexagonally binned parity plots for (a) mean, (b) median, (c) minimum, and (d) maximum ensemble homogenization functions. A GPR model with 50 000 input and 10 000 prediction VFZOs was used.

from our numerical tests, the rate of missed facet intersections increases. This artifact and our method for addressing it are discussed in Appendix B.2 of the main text.

S4. Kim Interpolation

A GPR mixing model is developed to accommodate the non-uniformly distributed, noisy Fe simulation data [3] and better predict low GBE. Details of the method (Section S4.1) and an analysis of the input data quality (Section S4.2) are given. The code implementation is given in `gprmix.m` and `gprmix_test.m` of the VFZO repository [4].

S4.1. Details of Gaussian Process Regression Mixture

As shown in Figure S5a, prediction using the standard approach of the main document (termed the ϵ_1 model) overestimates low GBEs for this dataset. By training the model on only GBs with a GBE less than $1.2\,\mathrm{J\,m^{-2}}$ (termed the ϵ_2 model) and by using an exponential (`KernelFunction='exponential'`)

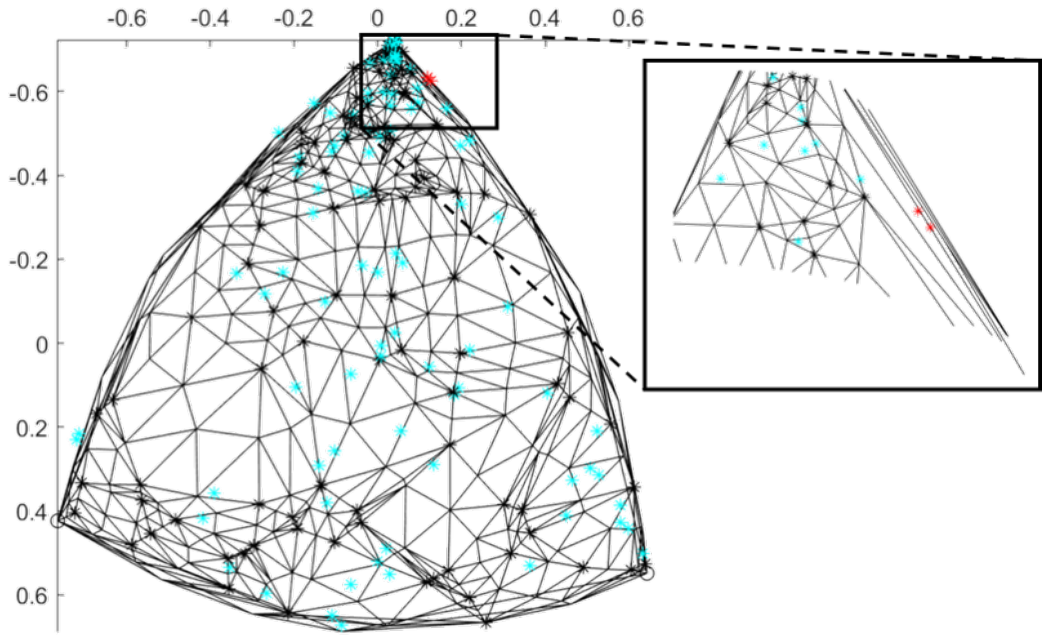

Figure S4: Illustration of two prediction points (red) for which no intersecting facet is found due to being positioned within a high-aspect ratio facet. The inset shows that facets connected to the nearest neighbor (NN) do not contain the prediction point. Many NNs would need to be considered before an intersection is found. Additionally, it is expected that if found, the interpolation will suffer from higher error due to use of facet vertices far from the interpolation point. Proper intersections of prediction points with the mesh are shown in blue.

rather than a squared exponential kernel, prediction of low GBEs improves, but naturally underestimation occurs for higher GBEs (Figure S5b).

A combined, disjoint model (Figure S5c) is taken (ϵ_3) by replacing ϵ_1 GBE predictions for GBs with GBE less than $1.2\,\mathrm{J\,m^{-2}}$ with the corresponding ϵ_2 predictions. Finally, a weighted average (Eq. (S1)) is taken according to:

$$\epsilon_{mix} = f\epsilon_1 + (f-1)\epsilon_2 \tag{S1}$$

where ϵ_1 and ϵ_2 represent the standard GPR model and the GPR model trained on the subset of GBs with a GBE less than $1.2\,\mathrm{J\,m^{-2}}$, respectively, and f is the sigmoid mixing fraction given by:

$$f = \frac{1}{e^{-m(\epsilon_3 - b)} + 1} \tag{S2}$$

and shown in Figure S6 with $m = 30$ and $b = 1.1\,\mathrm{J\,m^{-2}}$, as used in this work. Larger values of m yield a steeper sigmoid function and larger values of b shift the sigmoid function further to the right. Specific values for m and b were chosen by visual inspection and trial and error. This results in a GPR mixing model which better predicts low GBEs while retaining overall predictive accuracy (Figure S5d).

Uncertainty of the GPR mixing model is similarly obtained by taking a weighted average of the uncertainties of each model according to:

$$\sigma_{mix} = f\sigma_1 + (f-1)\sigma_2 \tag{S3}$$

where σ_1 and σ_2 are the corresponding uncertainties of ϵ_1 and ϵ_2, respectively, and f is given by Eq. (S2).

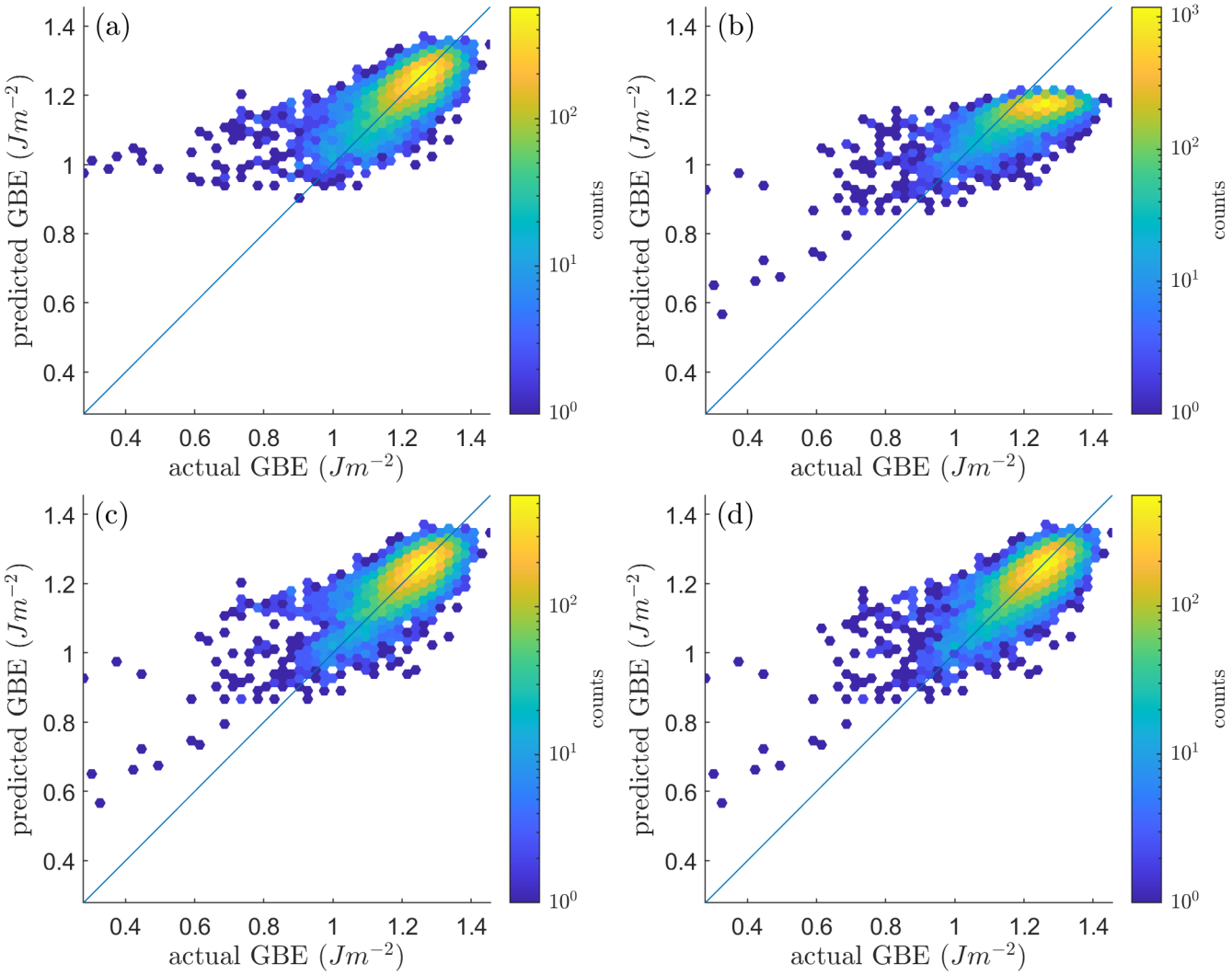

Figure S5: (a) Hexagonally binned parity plot of the standard GPR model. (b) All prediction GBs based on the model using only training GBs with a GBE less than 1.2 J m^{-2}. (c) Combined disjoint model as explained in the text. (d) Hexagonally binned parity plots of the final GPR mixing model. Points in (c) are produced by splitting the prediction data into less than and greater than 1.2 J m^{-2}. A sigmoid mixing function (Figure S6) is then applied where the predicted GBEs shown in (c) determines the mixing fraction (f) to produce a weighted average of models (a) and (b). A large Fe simulation database [3] using 46 883 training datapoints and 11 721 validation datapoints in an 80%/20% split. The GPR mixture model decreases error for low GBE and changes overall RMSE and MAE from 0.057 932 J m^{-2} and 0.039 857 J m^{-2} in the original model (shown in (a)) to 0.057 502 J m^{-2} and 0.041 272 J m^{-2} (shown in (d)), respectively.

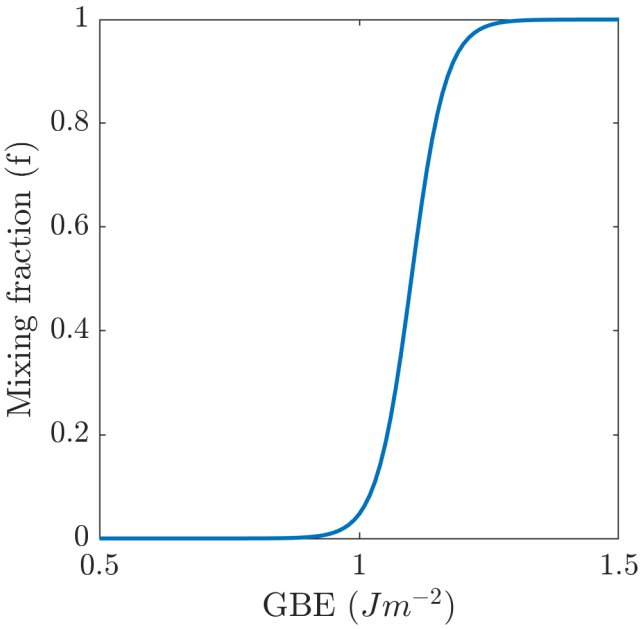

Figure S6: Sigmoid mixing function used in the GPR mixing model with $m = 30$ and $b = 1.1\,\mathrm{J\,m^{-2}}$ (Eq. (S2)).

S4.2. Input Data Quality

Of the ∼60 000[2] GBs in [3], ∼10 000 GBs were repeats that were identified by converting to VFZOs and applying VFZO repository function `avg_repeats.m`. In [3], mechanically selected GBs were those which involved sampling in equally spaced increments[3] for each five degree-of-freedom (5DOF) parameter, and a few thousand intentionally selected GBs (i.e. special GBs) were also considered. Of mechanically and intentionally selected GBs, 9170 and 112 are repeats, respectively, with a total of 2496 degenerate sets[4] (see Figure S7 for a degeneracy histogram). Thus, on average there is a degeneracy of approximately four per set of degenerate GBs.

By comparing GBE values of (unintentionally[5]) repeated GBs in the Fe simulation dataset [3], we can estimate the intrinsic error of the input data. For example, minimum and maximum deviations from the average value of a degenerate set are $-0.2625\,\mathrm{J\,m^{-2}}$ and $0.2625\,\mathrm{J\,m^{-2}}$, respectively, indicating that a repeated Fe GB simulation from [3] can vary by as much as $0.525\,\mathrm{J\,m^{-2}}$, though rare. Additionally, RMSE and MAE values can be obtained within each degenerate set by comparing against the set mean. Overall RMSE and MAE are then obtained by averaging and weighting by the number of GBs in each degenerate set. Following this procedure, we obtain an average set-wise RMSE and MAE of $0.065\,29\,\mathrm{J\,m^{-2}}$ and $0.061\,90\,\mathrm{J\,m^{-2}}$, respectively, which is an approximate measure of the intrinsic error of the data. Figure S8 shows histograms and parity plots of the intrinsic error. The overestimation of intrinsic error mentioned in the main text (Section 3.3) could stem from bias as to what type of GBs exhibit repeats based on the sampling scheme used in [3] and/or that many of the degenerate sets contain a low number of repeats (Figure S7).

Next, we see that by binning GBs into degenerate sets, most degenerate sets have a degeneracy of

[2]The "no-boundary" GBs (i.e. GBs with close to $0\,\mathrm{J\,m^{-2}}$ GBE) were removed before testing for degeneracy.
[3]In some cases, this was equally spaced increments of the argument of a trigonometric function.
[4]A degenerate "set" is distinct from a VFZO "set", the latter of which is often used in the main text.
[5]To our knowledge, the presence of repeat GBs were not mentioned in [3] or [5]

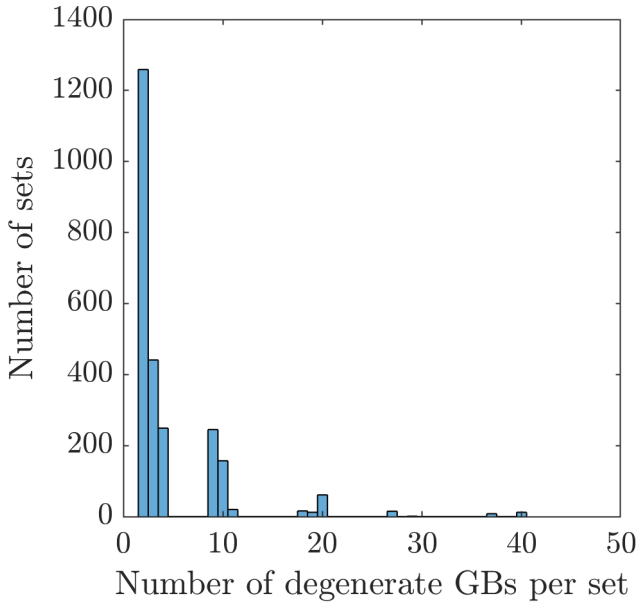

Figure S7: Histogram of number of sets vs. number of degenerate GBs per set for the Fe simulation dataset [3]. Most sets have a degeneracy of fewer than 5.

fewer than 5 Figure S7. We split the repeated data into sets with a degeneracy of fewer than 5 and greater than or equal to 5 and plot the errors (relative to the respective set mean) in both histogram form (Figure S8a and Figure S8c, respectively) and as hexagonally-binned parity plots (Figure S8b and Figure S8d, respectively). While heavily repeated GBs tend to give similar results, occasionally repeated GBs often have larger GBE variability. This could have physical meaning: Certain types of (e.g. high-symmetry) GBs tend to have less variation (i.e. fewer and/or more tightly distributed metastable states). However, it could also be an artifact of the simulation setup that produced this data (e.g. deterministic simulation output for certain types of GBs).

S5. Olmsted Interpolation

As illustrated in Figure S9, leave-one-out cross-validation (LOOCV) interpolation results for 0 K molecular statics (MS) low-noise Ni simulations using the GPR method are similar to Laplacian kernel regression (LKR) results reported in Figure 6a of Chesser et al. [2] (reproduced on the right of Figure S9 for convenience).

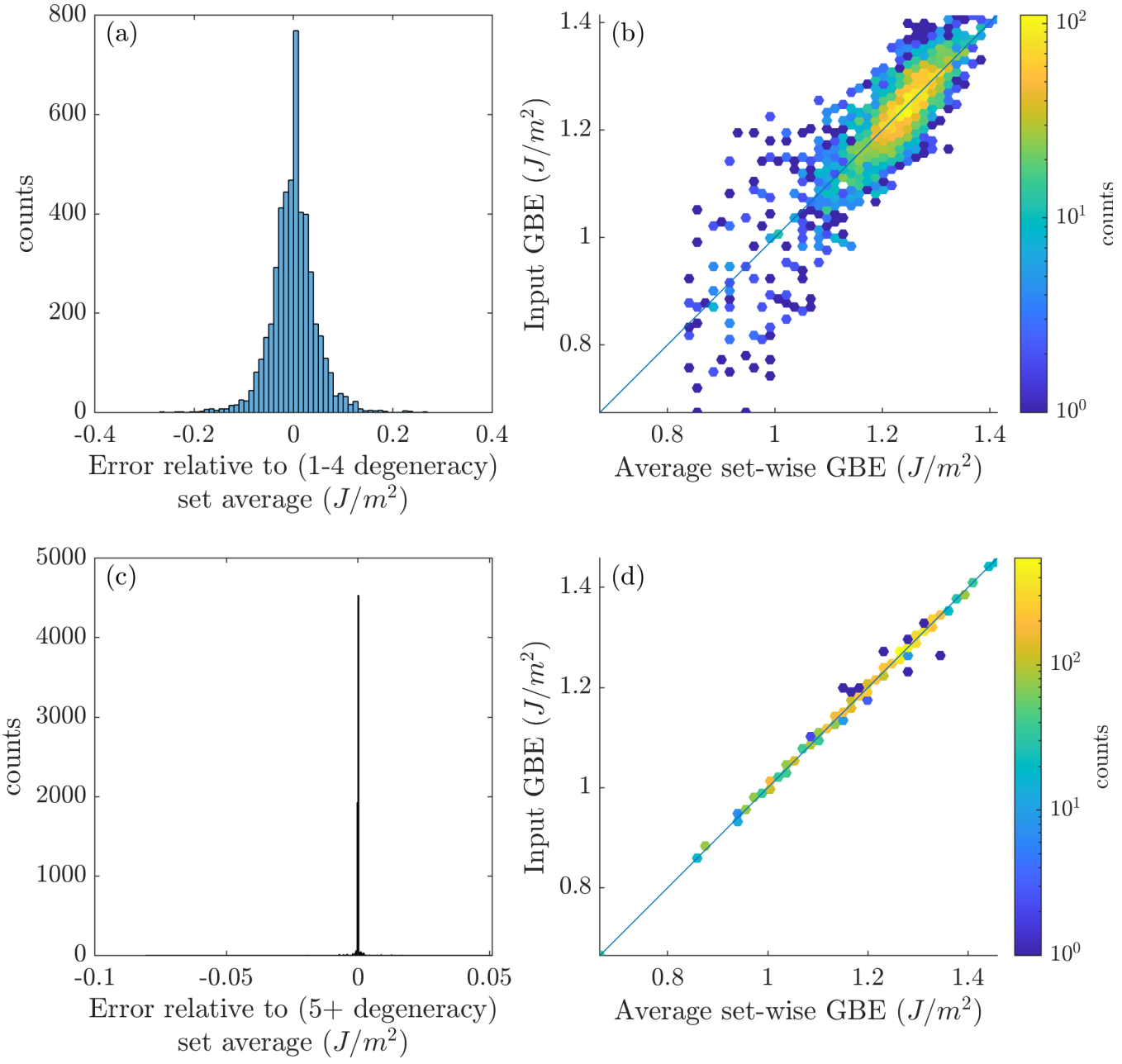

Figure S8: Degenerate GBs sets are split into those with a degeneracy of fewer than 5 and greater than or equal to 5 and plotted as ((a) and (c), respectively) error histograms and ((b) and (d), respectively) hexagonally-binned parity plots. Large degenerate sets tend to have very low error, whereas small degenerate sets tend to have higher error. In other words, GBs that are more likely to be repeated many times based on the sampling scheme in [3] tend to give similar results, whereas GBs that are less likely to be repeated often have larger variability in the simulation output. We do not know if this has physical meaning or is an artifact of the simulation setup.

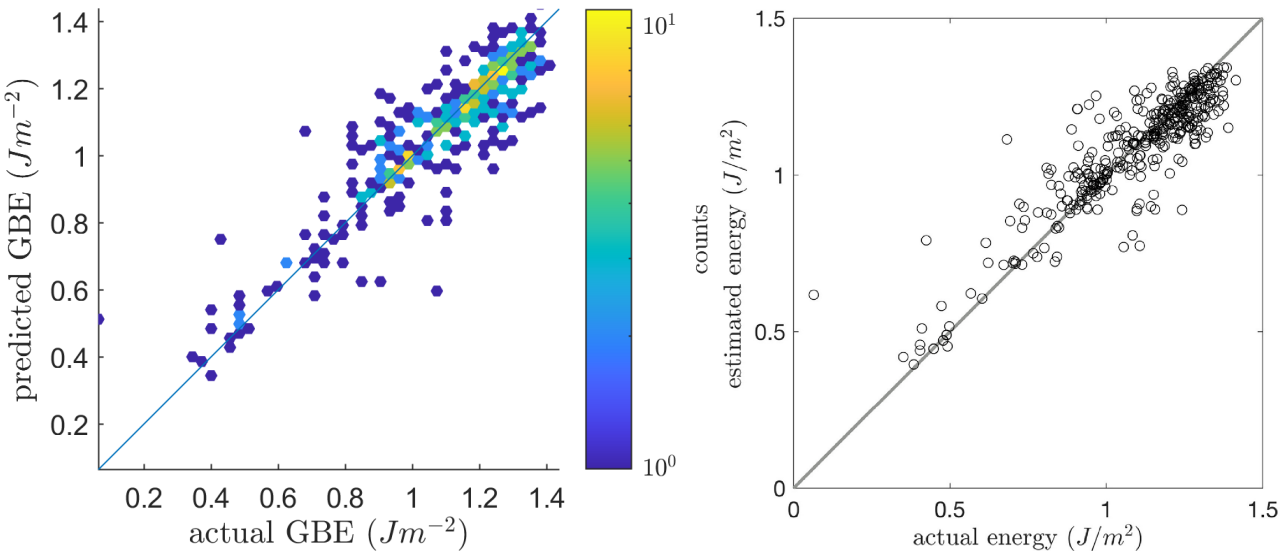

Figure S9: (left) Hexagonally binned parity plot for Ni simulation grain boundary energy (GBE) interpolation using LOOCV. (right) Parity plot for leave-one-out cross-validation (LOOCV) interpolation results reproduced from Figure 6a of Chesser et al. [2] under CC-BY Creative Commons license.

Acronyms

5DOF five degree-of-freedom 5

BRK Bulatov Reed Kumar 8

GB grain boundary 1

GBE grain boundary energy 2

GPR Gaussian process regression 1

GPRM Gaussian process regression mixture 1

kNN k-nearest neighbor 8

LOOCV leave-one-out cross-validation 11

MAE mean absolute error 2

MS molecular statics 7

NN nearest neighbor 5

RMSE root mean square error 2

VFZ Voronoi Fundamental Zone 2

VFZO Voronoi Fundamental Zone octonion 1

References

[1] T. Francis, I. Chesser, S. Singh, E. A. Holm, M. De Graef, A geodesic octonion metric for grain boundaries, Acta Materialia 166 (2019) 135–147. doi:10.1016/j.actamat.2018.12.034.

[2] I. Chesser, T. Francis, M. De Graef, E. Holm, Learning the grain boundary manifold: Tools for visualizing and fitting grain boundary properties, Acta Materialia 195 (2020) 209–218. doi:10.1016/j.actamat.2020.05.024.

[3] H.-K. Kim, S. G. Kim, W. Dong, I. Steinbach, B.-J. Lee, Phase-field modeling for 3D grain growth based on a grain boundary energy database, Modelling and Simulation in Materials Science and Engineering 22 (2014) 034004. doi:10.1088/0965-0393/22/3/034004.

[4] S. Baird, O. Johnson, Five Degree-of-Freedom (5DOF) Interpolation, 2020. URL: github.com/sgbaird-5dof/interp.

[5] H. K. Kim, W. S. Ko, H. J. Lee, S. G. Kim, B. J. Lee, An identification scheme of grain boundaries and construction of a grain boundary energy database, Scripta Materialia 64 (2011) 1152–1155. doi:10.1016/j.scriptamat.2011.03.020.