Acta Mat A-21-0401 Response to Reviewers

Reviewer #1: The authors develop a framework for approximating grain boundary geodesic distance kernels that is more computationally efficient than prior work without significantly impacting the accuracy of property regression on reference datasets. Two ideas central to the efficiency of the framework are 1) the approximation of geodesic GB distances with a Euclidean metric and 2) the construction of a fundamental zone for GBs based on Voronoi tessellations on the 7-sphere. It is argued that the distance approximation errors arising from (2) are small and insignificant for property prediction. Out of several regression methods explored in the paper, the authors recommend Gaussian Process Regression (GPR) for GB property interpolation because of its relative accuracy and built-in uncertainty quantification.   
  
The main impact of this work, in my view, is that it provides a method for GB property interpolation that scales to large GB datasets. The structure of the GB character space, and the way in which properties vary over that space, are complex and not well understood yet, and methods which help advance scientific efforts in this direction are welcome.

We appreciate the positive feedback.  
  
I do have several comments and concerns. My primary concern about this paper is its atypical and somewhat confusing organization, addressed in points 1-4 below. One theoretical comment is given in point 5. Finally, several typos are identified in points 6.   
  
1. Several sections seem to overlap in content and should be merged or reorganized, such as "3.1.3 Our Four Interpolation Methods" and "3.2.1 Comparison of Models".

References to accuracy in 3.2.1 were removed (which is in the efficiency section), and interpolation visualization was moved to its own section (3.3) so that the “four methods” accuracy and efficiency sections are close together.

2. The narrative is difficult to read sequentially: sections do not clearly transition from one to another, sections do not have topic sentences summarizing main points.

We agree that some sections lacked topic sentences; we have added topic sentences/summaries for these sections and reorganized some sections in order to improve the reading experience. In particular, the VFZO Framework (Section 2.1) and its subsections.

3. The Introduction starts with an outline, where claims about a "highly efficient" model seem premature. Consider moving the outline to the end of the Introduction.

This paragraph has been moved to the end of the Introduction.

4. Consider relegating details of the codebase to the Supplementary Information. This is a suggestion meant to lead to a better reading experience.

We appreciate this suggestion. While we consider the development of our approach and the results of this work meaningful on their own, we believe that the full value is realized through use of the extensive codebase developed around this metric. We wish to emphasize that not only are the distance calculations and interpolation methods efficient, but that these and other methods are available in functions with well-defined purposes. That said, we have moved a section that focuses specifically on the implementation to the supplementary information and removed other sentences that may be considered distracting.

5. In Section 2.1.1, an example is given for VFZ construction on the 2-sphere. It seems that this construction should be related to the standard stereographic triangle, a space with which materials scientists are already familiar. Is there a relationship between these constructions?

The primary purpose of the 2-sphere construction was to help with visualization of high-dimensional hyperspheres and illustrate key similarities (e.g. crossing the border of a VFZ). However, based on this comment, we noticed that double cover or inversion symmetry was not present for the 2-sphere VFZ construction. After applying double cover operations, the standard stereographic triangle (SST) for Oh symmetry is produced. We have updated the text to mention this relationship.

6. Typos:   
Figure 1 caption, "fixed reference point": no comma needed  
Figure 1 caption, insert comma before "demonstrating"

This has been updated in the manuscript.

Reviewer #2:

See *Summary of Comments on A-21-0401-Review 2 Attachment.pdf*

Reviewer #3: This paper was interesting for this reviewer to read, but I am not sure it will find its readership if published in Acta Mat. It is rather atypical in its focus on algorithmic and computational aspects whereas Acta papers are traditionally more focused on physical aspects of material behavior. I understand quite well where the authors are coming from but, really, will a typical Acta reader go into all or any of the technical details presented here? I am not necessarily recommending doing it, but if I were its author, I'd rather submit a brief but logical summary of this work to Scripta and a full article to some other journal more directly focused on data sciences.

We agree that this paper has a large computational focus. The paper was lacking in focus on physical aspects of material behavior, and we have updated the paper to focus on answering the following questions:

* Can a 5DOF FZ be defined with continuous coordinates?
* When working with randomly generated, *symmetrized* GBs, how does the # GBs affect the density and distribution of points?
* What are the dimensions of a 5DOF fundamental zone?
* How correlated are nearby GBs w.r.t. crystallography and GBE?
* Are grain boundaries of interest (i.e. low-Sigma GB cusps) connected or separated by activation energy barriers?

This involved inclusion or better emphasis of the following fundamental insights:

* an estimate of the intrinsic error of an Fe simulation dataset (Section 3.1.1. Experimental and Simulation Error, see also Figure S11). The data is estimated to have an intrinsic error on the order of 0.065 J m^-2.
* 1NN distribution and kNN distances for 50000 random cubochorically sampled points in a VFZ (Figure 2). If you take 50000 random points, the average NN distance is 2.87 +/- 0.69 deg and the distribution is Gaussian. Typically, the first 10 NNs have average NN distances less than 5 deg.
* average 1NN distance as a function of set size (Figure 5)
* the maximum dimension of an O\_h VFZ is ~60 deg (2.1.3. Distance Calculations in the Voronoi Fundamental Zone, see also Figure S1), i.e. the largest minimum distance path for O\_h point group symmetry is ~60 deg. Puts correlation lengths into perspective (10 deg is pretty large)
* large amounts of noise in the input data make it difficult to resolve deep cusps accurately (sort of a negative for large, noisy datasets --> smaller, low-noise datasets are generally to be preferred, but somewhere in-between is probably better)
* a GB correlation length of 10 or 15 degrees is overestimated. The GPR model gives a correlation length of 7.4 degrees for the Olmsted Ni GBE data. If you assume low noise, the correlation length drops to ~2 degrees. If you take 50,000 points from the BRK model, the fitted correlation length is 10.5 degrees -> indicating the BRK model is "smoothed" more than the Olmsted data warrants.
* Sigma 3 and Sigma 7 are connected, but Sigma 3 is not connected to Sigma 5, 9, or 11
* we have an estimate for the most likely function value (1.16 J/m^2) in a random cubochoric sense for the BRK function, which is biased towards larger GBE (min, midpoint, and max GBE are ~0.1, 0.8, and 1.4 J/m^2)
* the first 5DOF FZ defined by continuous coordinates

Having wrote all that above, I find the logic and the methods presented here elegant and potentially quite useful. Especially nice is the idea to map all symmetry equivalent representations of a given GB to a single point inside the Voronoi volume. Then, the approximate euclidization of the resulting space is making interpolations so much faster. On its technical content, it is a high quality paper. And very well written. The authors even post their user friendly [matlab] functions that can be used by others to much benefit.

We appreciate the positive feedback.  
  
And a couple of miscellaneous comments:  
  
1. Unless I missed it somehow, the authors are not giving a comparison of their, say, GPR predictions for Ni against the original computed data of Olmstead. Instead, in Fig. 6 they only compare their four interpolation functions with the BRK function. But the BRK validation function is an interpolation function itself so, while it is obviously useful for generating synthetic training data, it is smoother by construction than the atomistic data set to which it was fitted. It would be of interest to compare BRK and GPR interpolations directly to Olmstead's atomistic data.

Figure S9 contains a parity plot of a GPR model trained on the atomistic data and tested using leave-one-out cross-validation, but to the reviewer’s point, even this does not show a direct comparison of BRK and GPR interpolation. We added Figure 9 which makes this direct comparison for this interesting suggestion. No cross-validation was used for this plot because cross-validation will be difficult to implement for the BRK function. If we set the input noise to close to zero (0.0001 J m-2), the data passes almost exactly through the points (see Figure 9c and discussion below).  
  
2. Is it really important to be able to fit to 50,000 boundaries? Given uncertainties in experimental and computational determination of the GB energy, how much really gained by sampling so many more GBs than 388 in Olmsted's dataset?

Yes and no. First, we answer for the case when the 388 GBs are randomly chosen. Based on Figure 7, we really start to see diminishing accuracy returns at around 5000 or 10000 GBs. At only 388 GBs; however, the performance is much worse. Next, we answer for the Olmsted dataset specifically, many of these 388 GBs were handpicked based on extensive prior knowledge relating to high-symmetry FCC GBs. As frequent users of the BRK function which built a scaffolding out of this dataset, we certainly appreciate its existence. If, however, we move to less-studied point groups, multiphase systems, properties other than energy, etc. is the same approach possible, feasible, and/or effective? We believe the current approach is more extensible; indeed, the idea of a scaffolding-like implementation in the VFZ framework is something of interest. In its current implementation, setting the correlation length of the GPR method close to 0 would mean that the model passes (at least very close) through the training data for all data points, but with 388 points, an ensemble or data augmentation scheme is likely necessary to mitigate issues with distance overestimation.

In terms of uncertainty, we offer the counterpoint that *because* experimental and computational data has uncertainty, we need more data, and GPR is well-equipped to handle the input uncertainty and provide measures of output uncertainty. This may in turn explain characterization error, simulation stochasticity, and/or metastable states, for example.