

## ARTICLE OPEN

## Discovering the building blocks of atomic systems using machine learning: application to grain boundaries

Conrad W. Rosenbrock<sup>1</sup>, Eric R. Homer<sup>2</sup>, Gábor Csányi<sup>3</sup> and Gus L. W. Hart<sup>1</sup>

Machine learning has proven to be a valuable tool to approximate functions in high-dimensional spaces. Unfortunately, analysis of these models to extract the relevant physics is never as easy as applying machine learning to a large data set in the first place. Here we present a description of atomic systems that generates machine learning representations with a direct path to physical interpretation. As an example, we demonstrate its usefulness as a universal descriptor of grain boundary systems. Grain boundaries in crystalline materials are a quintessential example of a complex, high-dimensional system with broad impact on many physical properties including strength, ductility, corrosion resistance, crack resistance, and conductivity. In addition to modeling such properties, the method also provides insight into the physical “building blocks” that influence them. This opens the way to discover the underlying physics behind behaviors by understanding which building blocks map to particular properties. Once the structures are understood, they can then be optimized for desirable behaviors.

*npj Computational Materials* (2017)3:29; doi:10.1038/s41524-017-0027-x

## INTRODUCTION

Although interactions between small, isolated atomic systems can be studied experimentally and then modeled, real-world systems are exponentially more complex because of multi-scale, many-body interactions between all the atoms. Approximate, statistical methods are then necessary in the quest for deeper understanding. Machine learning is a powerful statistical tool for extracting correlations from high-dimensional data sets; unfortunately, it often suffers from a lack of interpretability. Researchers can create models that approximate the physics well enough, but the physical intuition usually provided by models may be hidden within the complexity of the model (the black-box problem). Here, we present a general method for representing atomic systems for machine learning so that there is a clear path to physical interpretation, or the discovery of those “building blocks” that govern the properties of these systems.

We choose to demonstrate the method for crystalline interfaces because of their inherent complexity, high-dimensionality, and broad impact on many physical properties. Crystalline building blocks are well known and can be classified by a finite set of possible structures. Disordered atomic structures on the other hand are difficult to classify and there is no well-defined set of possible structures or building blocks. Furthermore, these disordered atomic structures often exhibit an oversized influence on material properties because they break the symmetry of the crystals. Crystalline interfaces, more commonly called grain boundaries (GBs), are excellent examples of disordered atomic structures that exert significant influence on a variety of material properties including strength, ductility, corrosion resistance, crack resistance, and conductivity.<sup>1–9</sup> They have macroscopic, crystallographic degrees of freedom that constrain the configuration between the two adjoining crystals.<sup>10, 11</sup> GBs also have microscopic degrees of freedom that define the atomic structure

of the GB.<sup>12–15</sup> While often classified experimentally using the crystallography, the crystallography is only a constraint, and it is the atomic structure that controls the GB properties.

In this article, we examine the local atomic environments of GBs in an effort to discover their building blocks and influence on material properties. This is achieved by machine learning on the space of the atomic environments to make property predictions of GB energy, temperature-dependent mobility trends, and shear coupling. The implications of the work are significant; despite the immense number of degrees of freedom, it appears that GBs in face-centered cubic (FCC) nickel are constructed with a relatively small set of local atomic environments. This means that the space of possible GB structures is not only searchable, but that it is possible to find the atomic environments that give desired properties and behaviors. We emphasize that in addition to being successful for modeling GBs, the methodology presented here could be applied generally to many atomic systems.

Atomic structures in GBs have been examined for decades using a variety of structural metrics<sup>12, 16–26</sup> with the goal of obtaining structure-property relationships.<sup>10, 11, 27–32</sup> Each of the efforts has given unique insight, but none has provided a universally applicable method to find relationships between atomic structure and specific material properties.

Large databases of GB structures have produced property trends<sup>12, 33–35</sup> and macroscopic *crystallographic* structure-property relationships,<sup>36, 37</sup> but no *atomic* structure-property relationships. Machine learning of GBs by Kiyohara et al.<sup>38</sup> has been used to make predictions of GB energy from atomic structures, but we are still left without an understanding of what is important in making the predictions, and how that affects our understanding of the underlying physics and the building blocks that control properties and behaviors. We now present a method to address these limitations.

<sup>1</sup>Department of Physics and Astronomy, Brigham Young University, Provo, UT 84602, USA; <sup>2</sup>Department of Mechanical Engineering, Brigham Young University, Provo, UT 84602, USA and <sup>3</sup>Engineering Laboratory, University of Cambridge, Trumpington Street, Cambridge CB2 1PZ, UK  
Correspondence: Conrad W. Rosenbrock (rosenbrockc@gmail.com) or Eric R. Homer (eric.homer@byu.edu)

Received: 5 April 2017 Revised: 23 May 2017 Accepted: 11 June 2017

Published online: 03 August 2017

## METHODS

To examine atomic structures, we adopt a descriptor for single-species GBs based on the Smooth Overlap of Atomic Positions (SOAP) descriptor.<sup>39, 40</sup> The SOAP descriptor uses a combination of radial and spherical spectral bases, including spherical harmonics. It places Gaussian density distributions at the location of each atom, and forms the spherical power spectrum corresponding to the neighbor density. The descriptor can be expanded to any accuracy desired and goes smoothly to zero at a finite distance, so that it has compact support.

The SOAP descriptor has the following qualities that make it ideal for Local Atomic Environment (LAE) characterization. Specifically, within GBs, the SOAP descriptor (1) is agnostic to the grains' specific underlying lattices (including the loss of periodicity at the GB); (2) has invariance to global translation, global rotation, and permutations of identical atoms; (3) leads to a metric that is smooth and stable against deformations. SOAP vectors are part of a normed vector space so that similarity uses a simple dot product. This dot product can be used to produce a symmetric dissimilarity  $s$ , defined as

$$s = \left| \frac{\|\vec{a}\| + \|\vec{b}\|}{2} - \vec{a} \cdot \vec{b} \right|, \quad (1)$$

that is sensitive to the norm of each SOAP vector. Normally, SOAP similarity uses a dot product on normalized SOAP vectors; however, in our experience this reduces the discriminative ability of the representation.

In GBs, the SOAP descriptor has advantages over other structural metrics in that it requires no predefined set of structures, and a small change in atomic positions produces a correspondingly small (and smooth) change in the SOAP dissimilarity  $s$  (see Eq. 1).<sup>17, 18, 20, 23, 24</sup> Moreover, the SOAP vector is complete in the sense that any given LAE can be reconstructed from its SOAP descriptor.

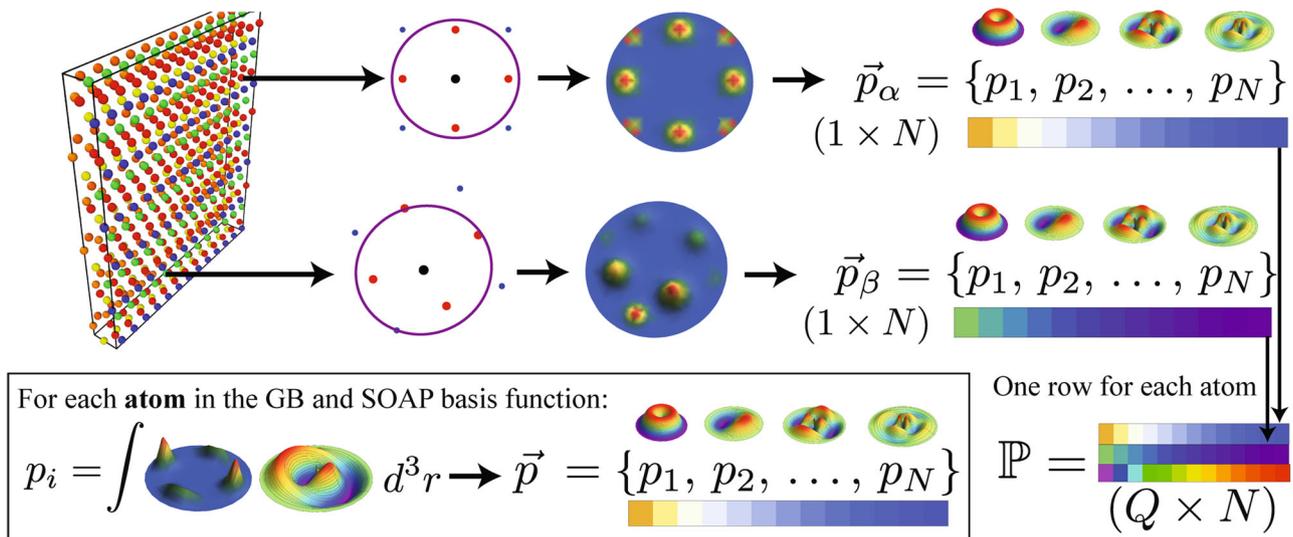
Figure 1 illustrates the process for determining the SOAP descriptor for a GB. First, GB atoms and some surrounding bulk atoms are isolated from their surroundings; a SOAP descriptor for each atom in the set is calculated and represented as a vector of

coefficients. The matrix of these vectors, one for each LAE, is the full SOAP representation for each GB. The SOAP vector can be expanded to resolve any desired features by increasing the number of terms in the basis expansion of the neighbor density at fixed cutoff. For the present work, a cutoff distance of 5 Å ( $\approx 1.4$  lattice parameters) and vector of length 3250 elements produced good results; the selection of SOAP parameters is discussed in Section I of the [Supplementary Information](#). The computed GBs studied in this work are the 388 Ni GBs created by Olmsted, Foiles, and Holm,<sup>12</sup> using the Foiles-Hoyt embedded atom method potential.<sup>41</sup>

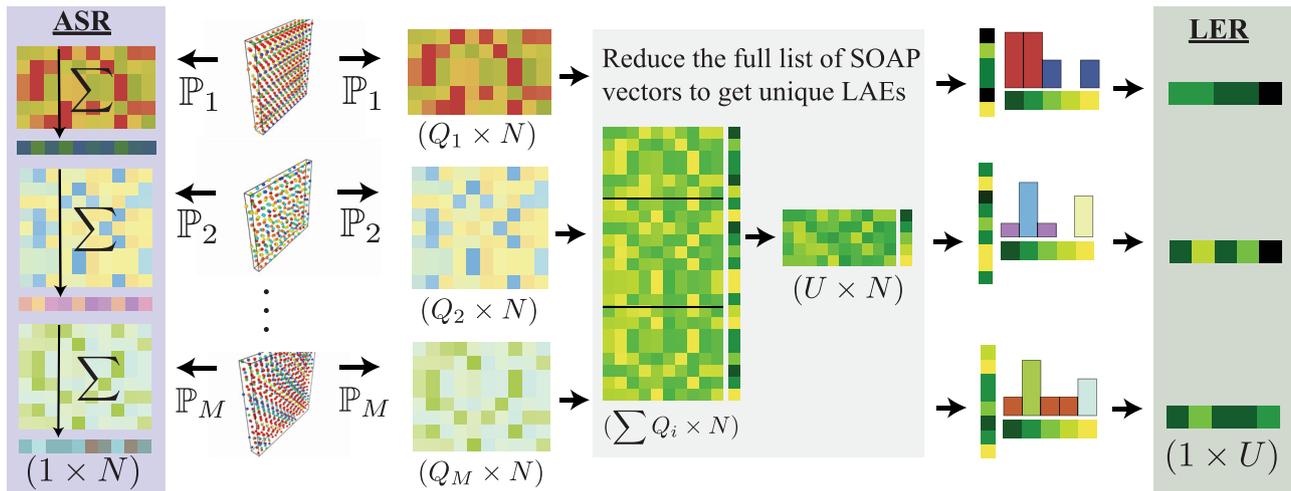
We investigate two approaches for applying machine learning to the GB SOAP matrices. For the first option, we average the SOAP vectors, or coefficients, of all the atoms in a single GB to obtain one averaged SOAP vector that is a measure of the whole GB as shown in Fig. 2. In other words, it is a single description of the average LAE for the whole GB structure. We refer to this single averaged vector representation as the Averaged SOAP Representation (ASR). The ASR for a collection of GBs becomes the feature matrix for machine learning.

Alternatively, we can compile an exhaustive set of *unique* LAEs by comparing the environment of every atom in every GB to all other environments using the dissimilarity metric  $s$  (from Eq. (1)) and a numerical similarity parameter  $\epsilon$  (see Fig. 2). Two LAEs are considered to belong to the same, unique class of LAEs if  $s < \epsilon$ . A SOAP vector will produce a value  $s = 0$  when compared with itself.

Using an  $n^2$  search over all LAEs in all GBs produces the set  $U$  of unique LAE classes, each with a representative LAE, for the GB system. For a sufficiently small  $\epsilon$  each GB will be characterized by a unique fingerprint in terms of the LAEs it contains. As  $\epsilon$  gets smaller, the number of unique LAEs that characterize a GB increases exponentially. When an LAE is sufficiently dissimilar to all others in the set, it is added and becomes the representative LAE for the class of all other LAEs that are similar to it. Any of the LAEs in the class could be the representative LAE since they are all similar. As additional data becomes available, this set of  $U$  LAEs may increase in size if new LAE classes are discovered. Section III in the [Supplementary Information](#) presents additional details.



**Fig. 1** Illustration of the process for extracting a SOAP matrix  $\mathbb{P}$  for a single GB. Given a single atom in the GB, we place a Gaussian particle density function at the location of each atom within a local environment sphere around the atom. Next, the total density function produced by the neighbors is projected into a spectral basis consisting of radial basis functions and the spherical harmonics, as shown in the boxed region. Each basis function produces a single coefficient  $p_i$  in the SOAP vector  $\vec{p}$  for the atom, the magnitude of which is represented in the figure by the colors of the arrays. Once a SOAP vector is available for all  $Q$  atoms in the GB, we collect them into a single matrix  $\mathbb{P}$  that represents the GB. A value of  $N = 3250$  components in  $\vec{p}$  is representative for the present work



**Fig. 2** Illustration of the process for construction of the ASR and LER for a collection of GBs. First, a SOAP matrix  $\mathbb{P}$  is formed (as shown in Fig. 1). ASR: A sum down each of the  $Q$  columns in the matrix produces an averaged SOAP vector that is representative of the whole GB. The ASR feature matrix is then the collection of averaged SOAP vectors for all  $M$  GBs of interest ( $M \times N$ ). LER: The SOAP vectors from all  $M$  GBs in the collection are grouped together and reduced to a set  $U$  of unique vectors using the SOAP similarity metric, of which each unique vector represents a unique LAE. A histogram can then be constructed for each GB counting how many examples of each unique vector are present in the GB. This histogram produces a new vector (the LER) of fractional abundances, whose components sum to 1. The LER feature matrix is then the collection of histograms of unique LEA for the  $M$  GBs in the collection ( $M \times U$ )

**Table 1.** Predictive performance of the machine learning models trained on the ASR and LER representation, respectively

Property	ASR (ML model)	LER (ML model)	Random
GB energy	$89.2 \pm 0.7\%$ (RBF SVM)	$88.5 \pm 0.9\%$ (GBT)	$70.4 \pm 1.6\%$
Temperature-dependent mobility	$77.4 \pm 2.5\%$ (linear SVM)	$74.3 \pm 2.7\%$ (GBT)	$38.5 \pm 2.0\%$
Shear coupling	$61.3 \pm 0.6\%$ (linear SVM)	$61.4 \pm 0\%$ (GBT)	$52.0 \pm 2.5\%$

The models were trained on 50% (194) of the available 388 GBs and then validated on the remaining 194 GBs that the model had never seen. Percent error is relative to the mean. Error bars represent the standard deviation over 50 independent, random samplings (including different combinations of the 50% split, and re-fits of the data set). For the random column, energies were guessed by drawing values from a normal distribution that had the same mean and standard deviation as the 50% training data, and then compared to the actual energies in the validation data. For the classification problems, random choices from the 50% training data class labels were compared to the validation data. The machine learning models used were (1) support vector machine (SVM) with either a linear or radial basis function (RBF) kernel; or (2) gradient-boosted decision tree (GBT). Parameters for each model are discussed in the [Supplementary Information](#)

In the present work, 800,000 LAEs from the atoms in 388 GBs are reduced to 145 unique LAEs. This is a considerable reduction in dimensionality for a machine learning approach. More importantly, these 145 unique LAEs mean that there may be a relatively small, finite set of LAEs used to construct every possible GB in Ni. Using the reduced set of unique LAEs, we represent each GB as a vector whose components are the fraction of each globally unique LAE in that GB. This GB representation is referred to as the Local Environment Representation (LER), and the matrix of LER vectors representing a collection of GBs is also a feature matrix for machine learning. The 145 unique LAEs give a bounded 145-dimensional space, which is a significant improvement over the  $3 \times 800,000$ -dimensional space of the GB data set.

These two approaches are used because they are complementary: physical quantities such as energy, mobility, and shear coupling are best learned from the ASR, while physical interpretability is accessible using the LER, with only marginal loss in predictive power. Because we desire to discover the underlying physics and not just provide a black-box for property prediction, we use the LER to deepen our understanding of which LAEs are most important in predicting material properties such as mobility and shear coupling.

GB energy is measured as the excess energy of a grain boundary relative to the bulk energy as a result of the irregular structure of the atoms in the GB.<sup>12, 42</sup> GB energy is a static property of the system measured at 0 K, and all atomistic structures examined in the machine learning are the 0 K structures associated with this calculation.

Temperature-dependent mobility and shear coupled GB migration are two dynamic properties related to the behavior of a GB when it migrates. The temperature-dependent mobility trend classifies each GB as having (i) *thermally activated*, (ii) *athermal*, or (iii) *thermally damped* mobility, depending on whether the mobility of the GB (related to the migration rate) increases, is constant, or decreases with increasing temperature.<sup>35</sup> GBs that do not move under any of these conditions are classified as being (iv) *immobile*. In addition, when GBs migrate, they can also exhibit a coupled shear motion, in which the motion of a GB normal to its surface couples with lateral motion of one of the two crystals.<sup>34, 43</sup> GBs are then classified as either exhibiting shear coupling or not.

GB energy is a continuous quantity, while temperature-dependent mobility trend and shear coupling are classification properties. Additional details regarding these properties are available in the publications pertaining to their measurements<sup>12,</sup>

4<sup>34, 35</sup> and in Section IV of the [Supplementary Information](#). For the mobility and shear coupling classification, the data set suffered from imbalanced classes; we used standard machine learning resampling techniques to help mitigate the problem.<sup>44–46</sup>

## RESULTS

A summary of the machine learning predictions by the various methods is provided in Table 1. Machine learning was performed using the ASR and LER descriptions of the GBs and the properties of interest for the learning and prediction are GB energy, temperature-dependent mobility, and shear coupled GB migration (obtained from the computed Ni GBs). Table 1 also includes the results of attempting to predict these properties by “educated” random guessing using knowledge of the statistical behavior of the training set. For example, GB energies were guessed by drawing values from a normal distribution that had the same mean and standard deviation as the 50% training data; for the classification problems, random choices from the class labels in the training data were used. In all cases, the machine learning predictions are significantly better than random draws from distributions.

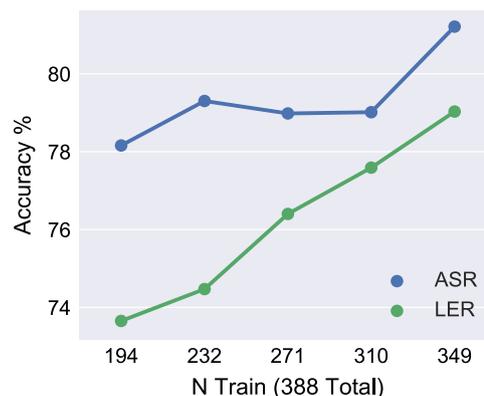
At first glance, the performance for mobility trend and shear-coupling classification (reported in Table 1) may seem mediocre. The results are significant, however, because mobility trend and shear-coupling are *dynamic* quantities, but they were predicted using a representation based on the *static*, 0 K GB structures. The mobility trend results are exceptional because the authors are unaware of any other models that can predict mobility using only knowledge of the atomic positions at the GB.

Shear coupling predictions are a little disappointing, but show some important limitations of the approach and suggest possible physical insights. Since little correlation was found between local environment descriptions and shear coupling, it may imply that the physical phenomenon must be multi-scale. Both the ASR and LER use knowledge of the *local* environments around atoms, but do not consider longer-range interactions between LAEs. Thus, only physical information within the cutoff (5 Å in this case) is considered. A future avenue of research could investigate whether connectivity of LAEs at multiple length scales or the full GB network are responsible for shear coupling.

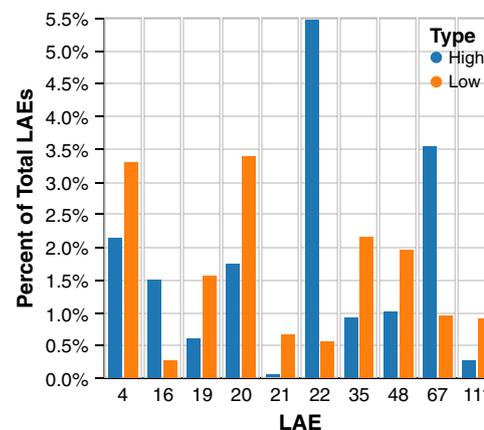
Unfortunately, the size of the data set is a limiting factor in the performance of the machine learning models. In Table 1, we used only half of the available 388 GBs for training. As we increase the amount of training data given to the machine, the learning rates change as shown in Fig. 3. Although it is common practice to use up to 90% of the available data in a small data set for training (with suitable cross validation), we chose to use a lower (pessimistic) split to guarantee that we are not overfitting to non-physical features. Larger data sets would certainly improve the models and our confidence in the physics they illuminate.

## DISCUSSION

For small data sets, ASR does slightly better in predicting energy and temperature-dependent mobility trend; ASR and LER are essentially equivalent for shear coupling. However, the ASR methodology suffers from a lack of interpretability because (1) its vectors and similarity metric live in the abstract SOAP space, which is large and less intuitive; (2) the results reported for ASR were obtained using a support vector machine (SVM), which is not easily interpretable. Details on the algorithm types and interpretation are included in the [Supplementary Information](#). The LER, on the other hand, has direct analogues in LAEs that can be analyzed in their original physical context. The best-performing algorithms for the LER are gradient-boosted decision trees, which lend themselves to easy interpretation. The fitted, gradient-boosted decision trees can be analyzed to determine which of the



**Fig. 3** Learning rate of ASR vs. LER for mobility classification. The x-axis is the number of GBs used in the training set, with the remaining GBs held out for validation. The accuracy was calculated over 25 independent fits. It appears that the LER accuracy increases slightly faster with more data, though a larger data set is necessary to confidently establish this point

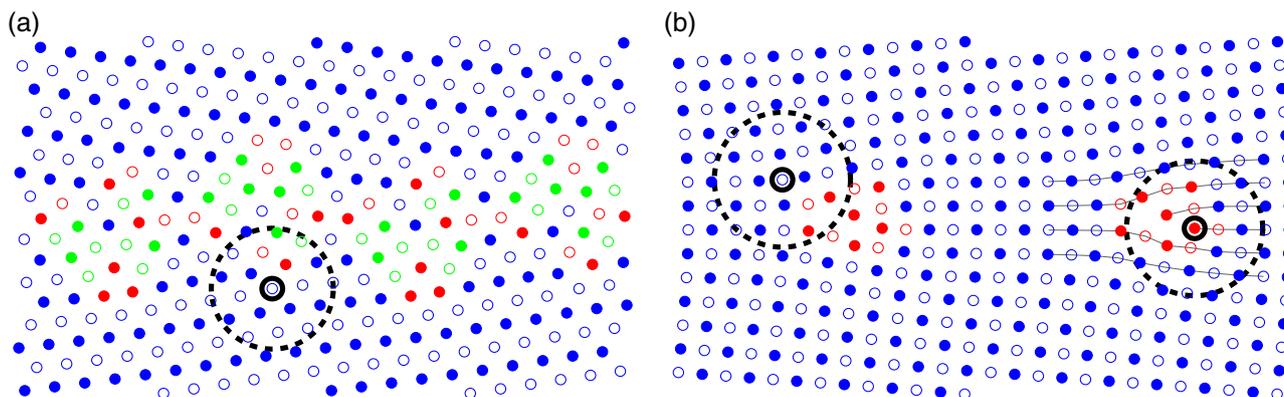


**Fig. 4** Histogram comparing the fraction of total LAEs for high and low GB energy. The 15 highest-energy GBs are compared against the 15 lowest-energy GBs using the 10 most-important LAEs for energy prediction. There are clear differences between the relative abundances of these LAEs in deciding whether a GB will have high or low energy

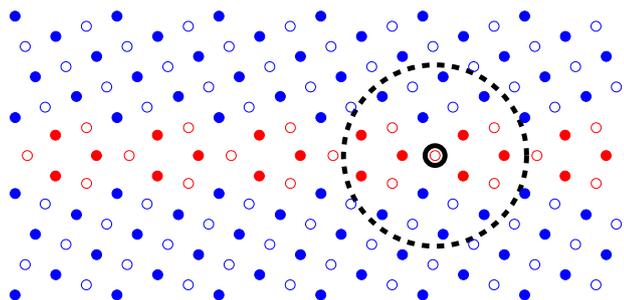
LAEs in the LER are most important. We used information gain as the metric for determining LAE importance, and an example is discussed in Section II of the [Supplementary Information](#). Thus, even at slightly lower accuracy, the physical insights generated by the LER make it the superior choice.

In Fig. 4, we compare the relative abundances of the most important LAEs for high and low energy classification in GBs. The 15 highest- and lowest-energy GBs are compared by calculating the fraction of their LAEs which are in the same class as the 10 most important LAEs for energy prediction. The most important LAEs selected by the machine learning algorithm are good at distinguishing between high and low energy GBs.

In Figs. 5 and 6, we plot some of the most important environments for determining whether a grain boundary will exhibit thermally activated mobility or not (Fig. 5) or thermally damped mobility or not (Fig. 6). These most important LAEs are classified as such because their presence or absence in any of the GBs in the entire data set is highly correlated with the decision to classify them as thermally activated or not, or thermally damped or not. Since such global correlations must be true for all GBs in



**Fig. 5** Illustration of important LAEs for classifying thermally activated GB mobility, as identified in two different GBs. The GB shown in **a** is a  $\Sigma 51a$  ( $16.1^\circ$  symmetric tilt about the  $[110]$  axis,  $\{1\bar{1}10\}$  boundary planes) GB, and has one LAE identified. The LAE shown in **a** has a relative importance of 3% over the entire system and includes a leading partial dislocation that originates from the GB. The GB shown in **b** is a  $\Sigma 85a$  ( $8.8^\circ$  symmetric tilt about the  $[100]$  axis,  $\{0\bar{1}13\}$  boundary planes) GB, and has two LAEs identified. The *leftmost* LAE has a relative importance of 9% (for all GBs in the data set) but its structural importance is not immediately clear, offering an exciting opportunity to discover new physics. The second LAE in **b** encloses edge dislocations, which are regularly spaced to form a tilt GB, (relative importance of 2.7% across all GBs). The *open* and *filled* circles denote atoms on the two unique stacking planes along the  $[100]$  or  $[110]$  direction. The atoms are colored according to common neighbor analysis (CNA) such that *blue*, *green*, and *red* atoms have a local environment that is FCC, HCP, or unclassifiable



**Fig. 6** Illustration of the most important LAE for classifying thermally damped GB mobility, as identified in a  $\Sigma 5$  ( $36.9^\circ$  symmetric tilt about the  $[100]$  axis,  $\{0\bar{1}3\}$  boundary planes) GB. The LAE shown has a relative importance of 6.8% and is centered at the point of a kite structure but includes parts of the kites on either side. These kite structures are “C” structural units that are regularly observed in  $[100]$  axis symmetric tilt GBs. The *open* and *filled* circles denote atoms on the two unique stacking planes along the  $[100]$  direction. The atoms are colored according to common neighbor analysis (CNA) such that *blue*, and *red* atoms have a local environment that is FCC, or unclassifiable

the system, we assume that they are tied to underlying physical processes.

Figure 5a shows a LAE centered around a leading partial dislocation. GBs with partial dislocations emerging from the structure have been associated with thermally activated mobility and immobility, depending upon their presence in simple or complex GB structures;<sup>34</sup> in addition, these structures have also been associated with shear coupled motion or the lack thereof. We now know that there is a strong correlation between the presence of these LAEs and their mobility type, though the presence of other structures is also important in the determination of the exact mobility type. This LAE was presented on equal footing with all others in the feature matrix that trained the machine. In the training, it was selected as important and we can easily see that it has relevant physical meaning.

In Fig. 5b, another LAE has obvious physical meaning as it captures edge dislocations in the environment of the selected atom. Interestingly, arrays of these edge dislocations, as in Fig. 5b,

are the basis for the energetic structure-property relationship of the Read-Shockley model.<sup>27</sup>

Thus, in these first two cases, we see that the LER approach discovers well-known, and physically important structures or defects that are commonly identified in metallic structures. Perhaps even more interesting is the second LAE in Fig. 5b, which has the highest relative importance of all ( $\approx 9\%$ ). The centrosymmetry parameter (CSP) for the atom at the center of the LAE is 0.125, or close to a perfectly structured FCC lattice, as visual inspection of the LAE would suggest. However, the CSP cannot be directly compared with the LAE because CSP examines only nearest neighbors while the LAE encompasses a larger environment, including the defect at the edge of the LAE.<sup>47</sup> Most importantly, this structure may not be immediately identified with any known metallic defect, but we know that it is highly correlated with thermally activated mobility across *all* the GBs in the data set.

In Fig. 6, the most important LAE for predicting thermally damped mobility is shown. Interestingly, it has found the “C” structural unit that is readily found in  $[100]$  axis symmetric tilt GBs,<sup>26</sup> though the LAE spans multiple kite structures. More important to note, however, is the fact that most of the important LAEs for predicting thermally damped mobility, are LAEs that are *not* present in thermally damped GBs. In other words, the machine learning algorithm is able to determine which structures will exhibit thermally damped mobility by the lack of certain LAEs in those structures.

The machine can determine some LAEs that are associated with well-known structures and properties, while also finding other LAEs that are not readily recognizable but are apparently important. This fact offers an exciting avenue to discover new mechanisms and structures governing physical properties. The physical nature of those LAEs that we already understand suggests that these are the building blocks underlying important physical properties and that we may be on the precipice of understanding the atomic building blocks of GBs.

Despite the formidable dimensionality of a raw grain boundary system, machine learning using SOAP-based representations makes the problem tractable. In addition to learning useful physical properties, the models provide access to a finite set of physical building blocks that are correlated with those properties throughout the high-dimensional GB space. Thus, the machine learning is not just a black box for predictions that we do not understand. The work shows that analyzing big data regarding materials science problems can provide insight into physical

structures that are likely associated with specific mechanisms, processes, and properties but which would otherwise be difficult to identify. Accessing these building blocks opens a broad spectrum of possibilities. For example, the reduced space can now be searched for extremal properties that are unique (i.e., special GBs). Poor behavior in certain properties can be compensated for by searching for combinations of other properties. In short, a path is now available to develop methods that optimize GBs (at least theoretically) at the atomic-structure scale. These methods may also provide a route to connect the crystallographic and atomic structure spaces so that existing expertise in the crystallographic space can be further optimized atomistically or vice versa.

While this is exciting within grain boundary science, the methodology presented here (and the SOAP descriptor in particular) has general applicability for building order parameters while studying changes that involve local structure. For example, it can be applied in studying phase change materials, point defects in solids, amorphous materials, cheminformatics, and drug binding. The physical interpretability of the machine learning representations, in terms of atomic environments, will also transfer well to new applications. This can lead to increased physical intuition across many fields of research that are confronted with the same, formidable complexity as seen in grain boundary science.

#### Data availability

Additional details about the machine learning models and data are described in the accompanying [Supplementary Information](#). The feature matrices and code to generate them are available on request.

#### ACKNOWLEDGEMENTS

C.W.R. and G.L.W.H. were supported under ONR (MURI N00014-13-1-0635). E.R.H. is supported by the U.S. Department of Energy, Office of Science, Basic Energy Sciences under Award #DE-SC0016441.

#### AUTHOR CONTRIBUTIONS

C.W.R. conceived the idea, performed all the calculations, and wrote a significant portion of the paper. E.R.H. was responsible for interpretation of the results and guidance of the project, and also wrote a significant portion of the paper. G.C. provided code, guidance and expertise in applying SOAP to the GBs. G.L.W.H. contributed many ideas and critique to help guide the project, and helped write the paper.

#### ADDITIONAL INFORMATION

**Supplementary Information** accompanies the paper on the *npj Computational Materials* website (doi:10.1038/s41524-017-0027-x).

**Competing interests:** The authors declare that they have no competing financial interests.

**Publisher's note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

#### REFERENCES

- Hall, E. O. The deformation and ageing of mild steel: III discussion of results. *Proc. Phys. Soc. B* **64**, 747–753 (1951).
- Petch, N. J. The cleavage strength of polycrystals. *J. Iron Steel Inst.* **174**, 25–28 (1953).
- Hansen, N. Hall–Petch relation and boundary strengthening. *Scripta Mater.* **51**, 801–806 (2004).
- Chiba, A., Hanada, S., Watanabe, S., Abe, T. & T, Obana Relation between ductility and grain-boundary character distributions in  $\text{Ni}_3\text{Al}$ . *Acta Metall. Mater.* **42**, 1733–1738 (1994).
- Fang, T. H., Li, W. L., Tao, N. R. & Lu, K. Revealing extraordinary intrinsic tensile plasticity in gradient nano-grained copper. *Science* **331**, 1587–1590 (2011).
- Shimada, M., Kokawa, H., Wang, Z. J., Sato, Y. S. & Karibe, I. Optimization of grain boundary character distribution for intergranular corrosion resistant 304 stainless steel by twin-induced grain boundary engineering. *Acta Mater.* **50**, 2331–2341 (2002).
- Lu, L. Ultrahigh strength and high electrical conductivity in copper. *Science* **304**, 422–426 (2004).
- Bagri, A., Kim, S.-P., Ruoff, R. S. & Shenoy, V. B. Thermal transport across twin grain boundaries in polycrystalline graphene from nonequilibrium molecular dynamics simulations. *Nano Lett.* **11**, 3917–3921 (2011).
- Meyers, M. A., Mishra, A. & Benson, D. J. Mechanical properties of nanocrystalline materials. *Prog. Mat. Sci.* **51**, 427–556 (2006).
- Wolf, D. & Yip, S. (eds.) *Materials Interfaces: Atomic-Level Structure and Properties* (Chapman & Hall, London, 1992).
- Sutton, A. & Balluffi, R. *Interfaces in Crystalline Materials* (Oxford University Press, 1995).
- Olmsted, D. L., Foiles, S. M. & Holm, E. A. Survey of computed grain boundary properties in face-centered cubic metals: I. Grain boundary energy. *Acta Mater.* **57**, 3694–3703 (2009).
- Cantwell, P. R. et al. Grain boundary complexions. *Acta Mater.* **62**, 1–48 (2014).
- The interplay between grain boundary structure and defect sink/annealing behavior *IOP Conference Series: Materials Science and Engineering* **89**, 012004 (2015).
- Dillon, S. J., Tai, K. & Chen, S. The importance of grain boundary complexions in affecting physical properties of polycrystals. *Curr. Opin. Solid State Mater. Sci.* **20**, 324–335 (2016).
- Weins, M., Chalmers, B., Gleiter, H. & ASHBY, M. Structure of high angle grain boundaries. *Scripta Metall. Mater.* **3**, 601–603 (1969).
- Ashby, M. F., Spaepen, F. & Williams, S. Structure of grain boundaries described as a packing of polyhedra. *Acta Metall. Mater.* **26**, 1647–1663 (1978).
- Gleiter, H. On the structure of grain boundaries in metals. *Mater. Sci. Eng.* **52**, 91–131 (1982).
- Frost, H. J., Ashby, M. F. & Spaepen, F. A catalogue of [100], [110], and [111] symmetric tilt boundaries in face-centered cubic hard sphere crystals. *Harvard Div. Appl. Sci.* 1–216 (1982).
- Sutton, A. P. On the structural unit model of grain boundary structure. *Phil. Mag. Lett.* **59**, 53–59 (1989).
- Wolf, D. Structure-energy correlation for grain boundaries in FCC metals—III. Symmetrical tilt boundaries. *Acta Metall. Mater.* **38**, 781–790 (1990).
- Tschopp, M. A., Tucker, G. J. & McDowell, D. L. Structure and free volume of symmetric tilt grain boundaries with the E structural unit. *Acta Mater.* **55**, 3959–3969 (2007).
- Tschopp, M. A. & McDowell, D. L. Structural unit and faceting description of Sigma 3 asymmetric tilt grain boundaries. *J. Mater. Sci.* **42**, 7806–7811 (2007).
- Spearot, D. E. Evolution of the E structural unit during uniaxial and constrained tensile deformation. *Acta Mater.* **35**, 81–88 (2008).
- Bandaki, A. D. & Patala, S. A three-dimensional polyhedral unit model for grain boundary structure in fcc metals. *Npj Comput. Mater.* **3**, 13 (2017).
- Han, J., Vitek, V. & Srolovitz, D. J. The grain-boundary structural unit model redux. *Acta Mater.* **133**, 186–199 (2017).
- Read, W. & Shockley, W. Dislocation models of crystal grain boundaries. *Phys. Rev.* **78**, 275–289 (1950).
- Frank, F. C. Martensite. *Acta Metall. Mater.* **1**, 15–21 (1953).
- Bilby, B. A., Bullough, R. & Smith, E. Continuous distributions of dislocations: a new application of the methods of non-riemannian geometry. *Proc. Roy. Soc. A-Math. Phys.* **231**, 263–273 (1955).
- Wolf, D. A broken-bond model for grain boundaries in face-centered cubic metals. *J. Appl. Phys.* **68**, 3221–3236 (1990).
- Wolf, D. Correlation between structure, energy, and ideal cleavage fracture for symmetrical grain boundaries in fcc metals. *J. Mater. Res.* **5**, 1708–1730 (1990).
- Yang, J. B., Nagai, Y. & Hasegawa, M. Use of the Frank–Bilby equation for calculating misfit dislocation arrays in interfaces. *Scripta Mater.* **62**, 458–461 (2010).
- Olmsted, D. L., Holm, E. A. & Foiles, S. M. Survey of computed grain boundary properties in face-centered cubic metals-II: Grain boundary mobility. *Acta Mater.* **57**, 3704–3713 (2009).
- Homer, E. R., Foiles, S. M., Holm, E. A. & Olmsted, D. L. Phenomenology of shear-coupled grain boundary motion in symmetric tilt and general grain boundaries. *Acta Mater.* **61**, 1048–1060 (2013).
- Homer, E. R., Holm, E. A., Foiles, S. M. & Olmsted, D. L. Trends in grain boundary mobility: survey of motion mechanisms. *JOM* **66**, 114–120 (2014).
- Bulatov, V. V., Reed, B. W. & Kumar, M. Grain boundary energy function for fcc metals. *Acta Mater.* **65**, 161–175 (2014).

37. Homer, E. R., Patala, S. & Priedeman, J. L. Grain boundary plane orientation fundamental zones and structure-property relationships. *Sci. Rep.* **5**, 15476 (2015).
38. Kiyohara, S., Miyata, T. & Mizoguchi, T. Prediction of grain boundary structure and energy by machine learning arXiv:1512.03502 (2015).
39. Bartók, A. P., Payne, M. C., Kondor, R. & Csányi, G. Gaussian approximation potentials: the accuracy of quantum mechanics, without the electrons. *Phys. Rev. Lett.* **104**, 136403 (2010).
40. Bartók, A. P., Kondor, R. & Csányi, G. On representing chemical environments. *Phys. Rev. B* **87**, 184115 (2013).
41. Foiles, S. M. & Hoyt, J. J. Computation of grain boundary stiffness and mobility from boundary fluctuations. *Acta Mater.* **54**, 3351–3357 (2006).
42. Tadmor, E. B. & Miller, R. E. *Modeling Materials: Continuum, Atomistic and Multi-scale Techniques* (Cambridge University Press, 2011).
43. Cahn, J. W., Mishin, Y. & Suzuki, A. Coupling grain boundary motion to shear deformation. *Acta Mater.* **54**, 4953–4975 (2006).
44. Han, H., Wang, W. -Y. & Mao, B. -H. Borderline-smote: a new over-sampling method in imbalanced data sets learning. In *Proceedings of the 2005 International Conference on Advances in Intelligent Computing - Volume Part I, ICIC'05*, 878–887 (Springer-Verlag, Berlin, Heidelberg, 2005).
45. Nguyen, H. M., Cooper, E. W. & Kamei, K. Borderline over-sampling for imbalanced data classification. *Int. J. Knowl. Eng. Soft Data Paradigm* **3**, 4–21 (2011).
46. Lemaître, G., Nogueira, F. & Aridas, C. K. Imbalanced-learn: a python toolbox to tackle the curse of imbalanced datasets in machine learning. *CoRR* abs/1609.06570 (2016).
47. Kelchner, C. L., Plimpton, S. J. & Hamilton, J. C. Dislocation nucleation and defect structure during surface indentation. *Phys. Rev. B* **58**, 11085–11088 (1998).



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2017