npj | Computational Materials

www.nature.com/npjcompumats

ARTICLE OPEN

Check for updates

Thermodynamics of order and randomness in dopant distributions inferred from atomically resolved imaging

Lukas Vlcek[1,2]✉, Shize Yang[3], Yongji Gong[4], Pulickel Ajayan[5], Wu Zhou[1,7], Matthew F. Chisholm[6], Maxim Ziatdinov[6], Rama K. Vasudevan[6]✉ and Sergei V. Kalinin[6]✉

Exploration of structure-property relationships as a function of dopant concentration is commonly based on mean field theories for solid solutions. However, such theories that work well for semiconductors tend to fail in materials with strong correlations, either in electronic behavior or chemical segregation. In these cases, the details of atomic arrangements are generally not explored and analyzed. The knowledge of the generative physics and chemistry of the material can obviate this problem, since defect configuration libraries as stochastic representation of atomic level structures can be generated, or parameters of mesoscopic thermodynamic models can be derived. To obtain such information for improved predictions, we use data from atomically resolved microscopic images that visualize complex structural correlations within the system and translate them into statistical mechanical models of structure formation. Given the significant uncertainties about the microscopic aspects of the material's processing history along with the limited number of available images, we combine model optimization techniques with the principles of statistical hypothesis testing. We demonstrate the approach on data from a series of atomically-resolved scanning transmission electron microscopy images of $Mo_xRe_{1-x}S_2$ at varying ratios of Mo/Re stoichiometries, for which we propose an effective interaction model that is then used to generate atomic configurations and make testable predictions at a range of concentrations and formation temperatures.

npj Computational Materials (2021)7:42 ; https://doi.org/10.1038/s41524-021-00507-7

INTRODUCTION

Condensed matter physics and materials science are both predicated on tuning physical and chemical functionalities via changes in chemical composition. Paradigmatic examples of this approach are the doping of silicon and other semiconductors that underpins virtually all aspects of the semiconductor industry and electronics[1], compositional tuning of oxides that underpin catalysis, energy technologies, and electroceramics[2–4], alloying of metals, and many others. From a fundamental perspective, most physical studies are performed (and hence functionalities defined) for single crystal solid solutions, a fact which propelled single crystal growth to be a key enabling component of modern research.

The relationship between the atomistic mechanisms of materials doping and emerging functionalities is highly non-trivial. For many materials, such as metals and silicon, the electron wavefunctions are sufficiently delocalized that the doping effects can be interpreted within effective mean-field models, e.g., via the shift of Fermi level or chemical potential of corresponding mobile species. The residual effects of chemical inhomogeneities can then be described via increased scattering rates and reduced mean free paths for electrons and phonons, or effective resistance, whereas exact positions of dopant species are less relevant. Overall, in these cases, doping effects are well-described through an effective change of bulk material parameters[5].

This approach however does not hold for materials with higher levels of disorder, giving rise to intriguing physical behaviors such as Anderson localization[6]. The latter is associated with macroscopically disordered ground states resulting in the localization of electronic wavefunctions. Similarly, in systems with localized interactions such as strongly correlated materials[7–11], complex behaviors emerge that are dependent on the strength and directionality of local interactions[12]. Correspondingly, electronic and functional properties will depend not only on average dopant concentrations but also on the exact configurations of dopant atoms[13]. For phenomena such as phase transformations, including the nucleation and transformation of domains and associated movement of interfaces during the transformation, the details of local atomic arrangements also become important—here, they determine the magnitude of the pinning of the interface, affect the transformation front geometry and account for roughness, and can thus greatly affect other relevant behaviors[14,15].

Notably, the statistics of atomic configurations of dopant atoms in real space, and hence the effects of the doping on materials behaviors strongly depend on the interactions between the dopant atoms. The effective attractive interactions between the same type of solid solution components can lead to dopant clustering and, above a certain threshold, to segregation of the second phase below the spinodal line. Similarly, repulsive interactions can lead to the formation of additional periodicity on the length scales determined by dopant concentration. These atomic configurations will correspondingly affect the electron, phonon, ferroelectric, or quantum behaviors of the material. The dopants interactions are strongly temperature-dependent as determined by the entropic term of free energy. Hence, in realistic

[1]Materials Science and Technology Division, Oak Ridge National Laboratory, Oak Ridge, TN, USA. [2]Joint Institute for Computational Sciences, University of Tennessee, Knoxville, Oak Ridge, TN, USA. [3]Center for Functional Nanomaterials, Brookhaven National Laboratory, Brookhaven, NY, USA. [4]School of Materials Science and Engineering, Beihang University, Beijing, China. [5]Department of Materials Science and NanoEngineering, Rice University, Houston, TX, USA. [6]Center for Nanophase Materials Sciences, Oak Ridge National Laboratory, Oak Ridge, TN, USA. [7]Present address: School of Physical Sciences and CAS Center for Excellence in Topological Quantum Computation, University of Chinese Academy of Sciences, Beijing, China. ✉email: lvlcek@utk.edu; vasudevanrk@ornl.gov; sergei2@ornl.gov

materials dopant distributions can be significantly different from the thermodynamic minimum and determined by the preparation history. Furthermore, nanoscale confinement effects can significantly affect even the equilibrium thermodynamics, leading to stabilization of higher-energy phases, the emergence of new phases, broadening the regions of solid solution, and other changes.

These considerations necessitate understanding the thermodynamics and effective dopant interactions in real materials. Advances in atomically resolved techniques such as scanning transmission electron microscopy (STEM)[16–18], scanning tunneling microscopy (STM)[19,20] and non-contact AFM (NC-AFM)[21], and atom probe tomography (APT) have allowed insight into atomic configurations on an atom by atom level. However, quantitative information extracted from these observations has been limited, usually because it is difficult to perform appropriate theory-experiment matching at the length scales of both simulation and experiment. Furthermore, while these experiments can in principle produce libraries of possible atomic configurations and structures, the throughput and hence the statistics of these experiments is generally limited.

Here, we analyze the structure of solid solutions from the series of atomically resolved images in scanning transmission electron microscopy and infer the microscopic thermodynamic interactions at the formation temperature. This approach allows us to avoid the statistical bottleneck and develop microscopic and thermodynamic generative models for the solid solution formation that can be used to test the alternative hypothesis about the formation of the observed structures and provide extrapolations to multiple concentrations and temperatures.

RESULTS

Modeling approach

Atomically resolved images provide a wealth of information about the interactions and history of the investigated material. In principle, each atom's chemical identity and position within the structure contains a piece of useful information about the system's physics. However, it is not immediately clear what this information may be and how we can use it. We approach this problem, which requires dealing with potentially large and noisy imaging data sets, by applying statistical and machine learning (ML) techniques to develop physically interpretable statistical mechanical models. Specifically, we use model selection and optimization methods that operate in the space of measurement outcomes.

As an illustration relevant for the current task, we consider the solid solution (exemplified here by $\mathrm{Mo}_x\mathrm{Re}_{1-x}\mathrm{S}_2$) extending over N metal atoms on a regular lattice, where $N \sim 400$. Ignoring structural defects, there are $k = 2^N$ possible elementary outcomes corresponding to different atomic configurations, where each can be represented as a unit basis vector in a k-dimensional real-valued Hilbert space, H^k [22]. We note that for classical systems and in the absence of experiment-specific errors (e.g., misidentified atoms), the space of measurement outcomes is equivalent to the space of the system's coarse-grained states corresponding to all distinguishable lattice configurations of Mo/Re metal atoms. For large sample numbers, the relative frequency of different configurations collected from repeated measurements converges to the probability distribution of the system surface configurations, where each distribution can be represented as a unit vector on the probability space of all possible distributions.

As shown by Wootters for pure quantum states and by Braunstein and Caves for density matrices[23,24], the angle between probability vectors, typically referred to as statistical distance, presents the natural metric for quantifying distinguishability of physical systems. It is defined as,

$$s^2 = \arccos^2\left(\sum_{i=1}^{k} \sqrt{p_i q_i}\right) \tag{1}$$

where p_i and q_i are the probabilities of states i in systems P and Q, and the argument represents a scalar product between k-dimensional probability vectors. We have recently proposed to use this metric to measure model quality and used it as an optimization loss function that avoids the pitfalls of other commonly used functions, such as the Kullback–Leibler divergence, simple least squares, or various energy and force matching methods for force field optimization[25].

We have shown earlier that a convenient loss function for D independent data sets in the form of histograms collected from multiple sources, such as images at different conditions, can be written as[22],

$$S^2 = \frac{1}{n_{\mathrm{Tot}}} \sum_{d=1}^{D} n_d s_d^2 \tag{2}$$

where s_d^2 is squared statistical distance for data set d, n_d is the number of samples in data set d, and n_{Tot} is the total number of samples in all data sets.

The practical challenge in dealing with microscopic imaging data using the outlined formalism is the enormous dimension of the Hilbert space and a limited number of samples (individual images), which may often amount to just one. In this situation, it is impossible to obtain an accurate estimate of the limiting probability distribution P that should be matched by a model. According to the maximum likelihood approach, the probability distribution estimate is equal to the distribution of relative frequencies, which would imply zeros for nearly all states[26]. Consequently, there is virtually no chance of a model matching the particular observed configuration.

An alternative estimate of P more suited for dealing with zero counts is to use a non-informative Jeffreys prior over the states, which is a uniform distribution on the probability space and whose effect is equivalent to assigning an extra 1/2 of a sample to each state. The estimate of the system's probability distribution P is then[26],

$$p_i = \frac{x_i + 1/2}{n + k/2} \tag{3}$$

where p_i is the estimated probability of state i of a k-state system, x_i is the number of counts in the histogram bin corresponding to i, and $n = \sum_{i=1}^{k} x_i$ is the total number of samples. It is easy to see that in the case of large k and small n the estimated P will be nearly uniform for any measurement, and the optimal model will be therefore random with not enough data to support a more complex model.

To overcome this obstacle and obtain more discriminative information from an image, we can first consider the crystalline system as composed of a large number m of subsystems, each with l dimensions, $l \ll k$. The original Hilbert space can be then expressed as a direct product of the subsystem spaces, $H^k = H^{l^m} = \otimes_{i=1}^{m} H_i^l$. In case the subsystems are uncorrelated because of their spatial separation, a lower-dimensional space obtained as the direct sum of subsystem spaces can be formed, $H^{m\times l} = \oplus_{i=1}^{m} H_i^l$, which can represent the full physically relevant information. If we further assume that the subsystems are statistically identical as a result of translational symmetry, we can collect all relevant statistics in a single l-dimensional space spanning only the states of the subsystem. For a single image, we obtain a larger number of samples, equal to the number m of subsystems, and lower dimension l of the subsystem state space. The maximum likelihood or Eq. (3) will therefore provide a much more accurate estimate of the limiting probabilities that still capture the full relevant information.

We note that this approach is equivalent to the presence of translational statistical invariance in the system and assumes the absence of long-range fields (such as depolarizing field in ferroelectrics). A similar approach was used in the statistical analysis of structural and electronic order parameters using sliding transforms, as reported by Vasudevan et al.[27].

The optimal choice of the subsystems is a feature selection problem. In the limit of large subsystems, we end up with a single sample per image, as discussed above. In the opposite limit of subsystems of the size of a single atom, we can collect a large number of samples, but the two-state (Mo/Re) subsystems will provide only a minimal amount of information to discriminate between candidate models because many plausible models can easily fit a binomial distribution (i.e., average concentration). The ideal subsystems that balance the number of samples and the number of distinguishable states l (resolution) will therefore lie in between these extremes and depend on the amount of data. The choice of the most discriminative features will also influence the maximum model complexity that can be supported by the data. As a general rule, when developing models based on microscopic images, we select features that can support the most complex models. Physical considerations of the locality of interactions may guide us to consider features (subsystems) in the form of local configurations that contain information about the direct correlations between atoms that roughly span the range of direct atom–atom interactions[28]. Typically, these may contain the nearest and next-nearest metal atom neighbors. The statistics of such configurations in the form of histograms represent a natural signature, or fingerprint, of the observed structure, which the model should reproduce. We note that this approach to feature selection is a variation of the bag-of-visual-words ML method used for image classification[29–31].

Statistical distance, as the geodesic on the probability space, is directly related to the statistical hypothesis testing. In this interpretation, a model of structure formation can be considered a testable hypothesis about the origin of the observed data. While we cannot prove the correctness of the model, we can rule out possible scenarios that are not compatible with the experimental data. For instance, it may not be clear whether configurations observed in microscopic images result from an equilibrium process and can be therefore directly related to interatomic interactions, or whether they represent history-dependent samples from a non-equilibrium distribution.

The target and model distributions of repeated measurement outcomes form multinomial distributions centered around the limiting probability distributions P and Q, defined on the probability space. In the large sample limit, these distributions are well approximated by normal distributions with variance equal to ¼. In this setting, statistical distance can be considered an instance of a Mahalanobis distance M defined on the $k-1$-dimensional probability space. We can then use the relation of M^2 to p-value[32], which quantifies the probability that the model generates a distribution that is at least as different as the target distribution. Since s^2 follows the χ^2_{k-1} distribution for $k-1$ the degrees of freedom, p-value can be determined as,

$$p = 1 - \mathrm{CDF}\left(\chi^2_{k-1}, 4ns^2\right) \tag{4}$$

where CDF denotes the cumulative distribution function of χ^2_{k-1} evaluated at $4ns^2$. Minimizing s^2 then results in a model representing a hypothesis that is most difficult to reject using the significance test, i.e., the model distribution is the most difficult to distinguish from the experimental one.

As an alternative to classical statistical significance testing, which evaluates individual models, we can also employ relative model selection criteria. Ideally, we would want to employ the minimum description length (MDL) criterion[33], which can be interpreted as penalizing model complexity based on the number of distinguishable configurations the model can generate[34]. This criterion is fully consistent with the ideas of the statistical distance framework utilized here. However, for practical reasons we use the simplified version valid in the large sample limit, which coincides with the Bayesian information criterion (BIC)[35], defined here as,

$$\mathrm{BIC} = 2ns^2 + \frac{r}{2}\ln n \tag{5}$$

where the first term is the negative log-likelihood of the model generating the observed distribution, r is the number of model parameters, and the rest of the symbols have the same meaning as before.

Imaging segregation and phase transition in $\mathrm{Re}_x\mathrm{Mo}_{1-x}\mathrm{S}_2$

As a model system, we have chosen the $\mathrm{Re}_x\mathrm{Mo}_{1-x}\mathrm{S}_2$ solid solutions for varying Re concentrations synthesized as described in "Methods" section[36,37]. The atomically resolved images across the composition series for $x = 0.05$, 0.55, 0.78, and 0.95 were acquired on the Nion UltraSTEM100 microscope and are shown in Fig. 1. The Re atoms are clearly visible as bright dots, as expected given the higher atomic number of Re.

To analyze the images, we adopt the atom finding algorithm based on the procedure outlined by Somnath et al.[38]. Briefly, this involves the first image denoising step via a sliding window reconstruction with principal components, followed by motif-matching and thresholding to find sub-lattices of distinct types and isolate the individual atoms. This functionality is available through the open-source python package PyCroscopy[39,40]. Subsequent Gaussian fitting enables sub-pixel accuracy of the atomic coordinates to be determined. Notably, this approach allows not only positional identification of all the atoms in the image, but also classifies them as Mo or Re based on simple thresholding given the change in contrast expected due to higher Z number of Re (details are included in SI document). The identified atom types are shown superimposed on the atomic contrast in Fig. 1. Thus, obtained data sets contain the information on the atomic configuration of cations in the 2D triangular lattice, i.e., compositional fluctuations. The latter, in turn, can be related to the thermodynamics of the solid solution via the formalism described above.

Models of dopant segregation

Here, we restrict our modeling to Mo/Re atom distribution on an idealized hexagonal lattice and ignore defects such as sulfur vacancies. As the first step, we select structural descriptors on this lattice, whose statistics will serve as the target structural fingerprint for model optimization and statistical significance testing. Given the limited amount of data, we constrain our analysis to the statistics of local configurations consisting of an atom and its six nearest neighbors (Fig. 2a). Assuming the translational symmetry of the sample, the seven atoms of two possible types can result in 2^7 configurations. Taking further into account rotational and reflective symmetries, the total number of distinct configurations reduces[26]. The statistics of these configurations in the form of relative frequencies collected from four images at different stoichiometries are shown in Fig. 3.

The complexity of the models reproducing the statistics of local configurations can theoretically range from a null model with zero adjustable parameters and single probability distribution to a model with 4 × 25 parameters, each of which controls the statistics of individual histogram bins collected from the four images. Such a model, which is in effect equivalent to that described by Eq. (3), achieves the maximum complexity with possible probability distributions spanning the entire probability space. Clearly, such a model will overfit and therefore possesses limited predictive power. Physically motivated constraints are thus needed to select a lower-dimensional subspace of possible distributions.

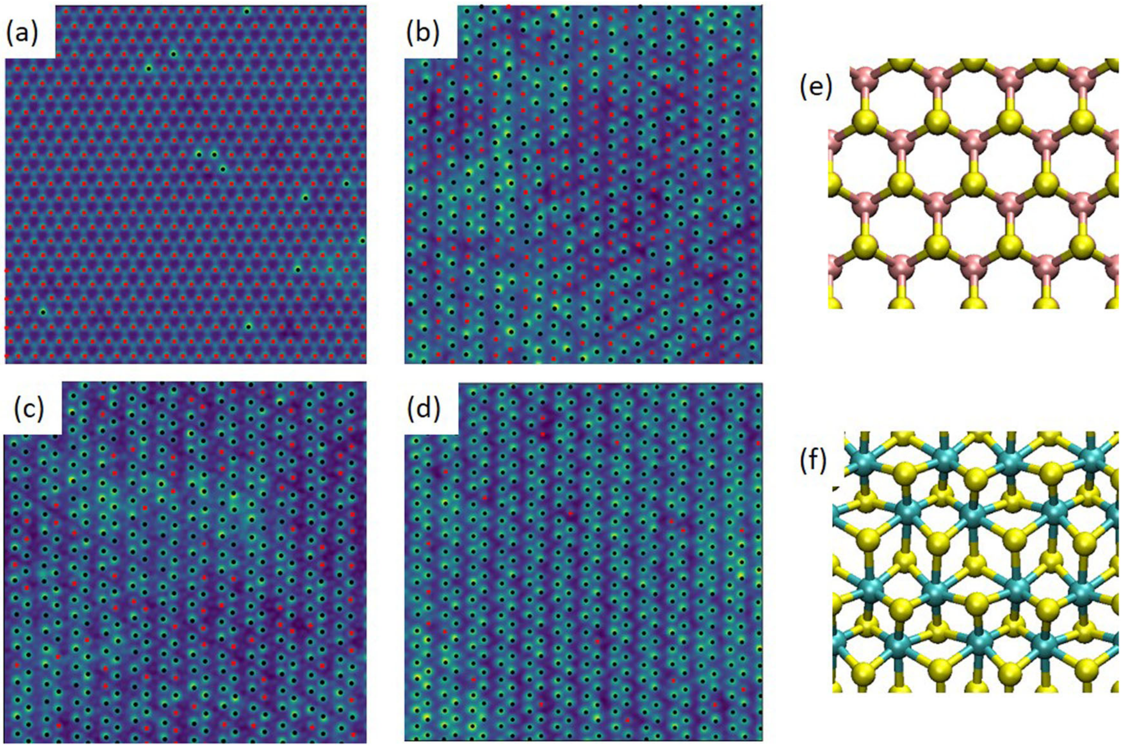

Fig. 1 STEM images of $Re_xMo_{1-x}S_2$ at different values of Re concentration. **a** $x = 0.05$, **b** $x = 0.55$, **c** $x = 0.78$, and **d** $x = 0.95$. Identified Mo and Re atoms are indicated by red and black dots, respectively. At $x = 0.05$, the material adopts the MoS_2 lattice structure (**e**), while for higher Re ratios it adopts the ReS_2 lattice (**f**). Color code: Mo (pink), Re (cyan), and S (yellow).

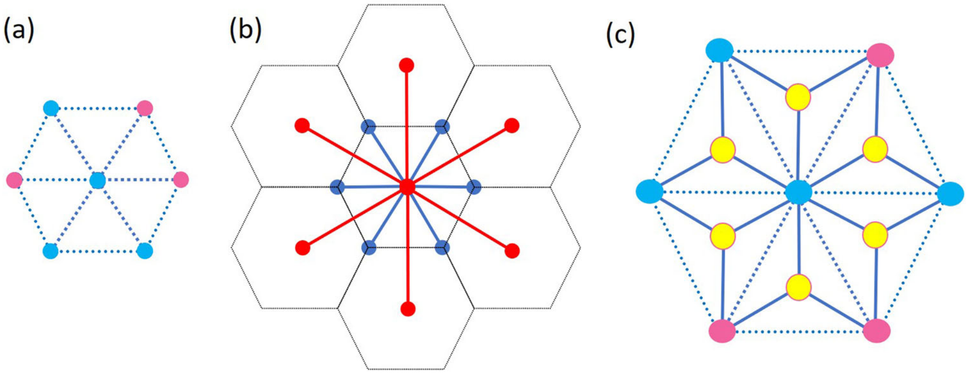

Fig. 2 Local atom configurations and the topology of direct atom–atom interactions. **a** An example from a set of 26 local surface configurations whose statistics are to be matched by a model; Mo (pink), Re (cyan). **b** Nearest (blue) and next-nearest (red) neighbor metal atom pairs considered in the lattice Hamiltonian of Eq. (6). **c** The triplets of Mo and Re atoms connected to individual sulfur (yellow) atoms define the many-body Hamiltonian of Eq. (7).

As the simplest possible model, the null hypothesis for the observed statistics, we assume that the $Mo_xRe_{1-x}S_2$ configurations collected from the four images are completely random. Physically, such a distribution of metal atoms may result from random deposition of Mo and Re atoms without subsequent thermal equilibration. Alternatively, a random distribution of metal atoms could be formed in an equilibrium system in which the differences in the effective energetics of Mo–Mo, Mo–Re, and Re–Re interactions are very weak.

The random model statistics are compared with the target data in Fig. 3. A quick visual comparison of the two histogram sets suggests that most of the variation in the configuration probabilities can be attributed to their symmetry numbers. To make this comparison more quantitative, we calculated the statistical distances between the target and model distributions and the corresponding p-values for data based on individual images as well as for the combined data sets. The results, summarized in Table 1 under model R, show that while the random model would pass the significance tests at the typical levels of $\alpha = 0.01$ or 0.05 for the images with very low and very high Re concentrations, we can reject it for the intermediate concentrations, as well as for the combined data set. It does appear that the distributions are non-random, and detectable ordering happens at the intermediate concentrations. The BIC criterion, Eq. (5), with $r = 0$, attains the value of 73.4.

To probe the segregation hypothesis further, we test a model assuming that the images present equilibrium structures that can be described by a class of models with a simple pair-additive

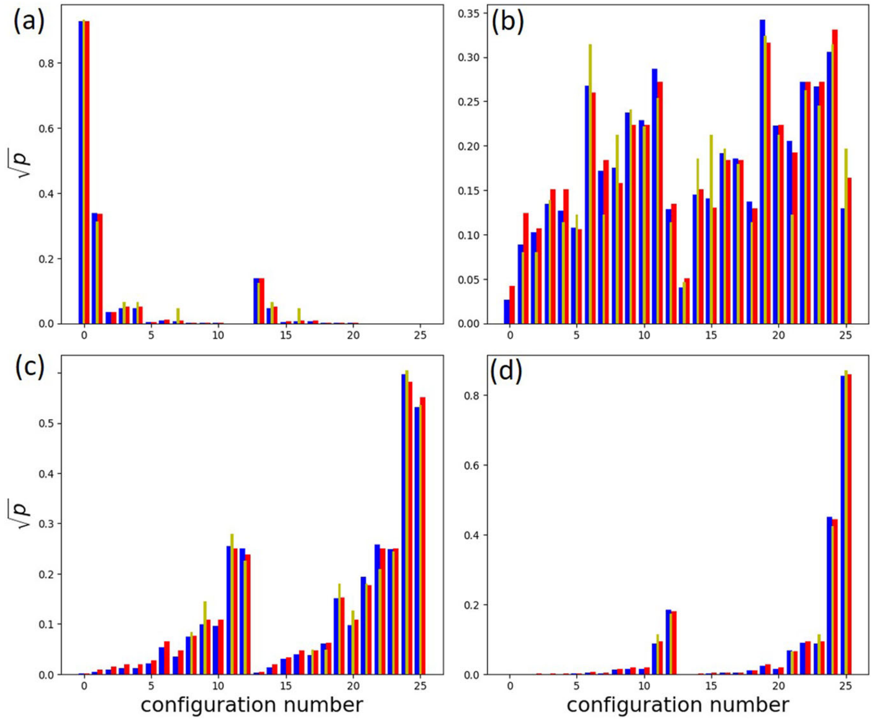

Fig. 3 Comparison of experimental and model-generated statistics of local configurations. Square roots of relative frequencies, $\sqrt{p_i}$, of unique local configurations in the target images (yellow), random model (red), and equilibrium model (blue) for four compositions studied in the present work. Plots for different values of *x*: **a** 0.05, **b** 0.55, **c** 0.78, and **d** 0.95. The configuration numbers are assigned identification numbers in Supplementary Fig. 1.

Table 1. Statistical significance tests and BIC scores of the different models: random (R) and pair additive with one (P1) and two (P2) parameters, pair additive with one parameter (P1), and many-body (M).

Data set	*N*	S^2 (R)	PV (R)	S^2 (P1)	PV (P1)	S^2 (P2)	PV (P2)	S^2 (M)	PV (M)
$x = 0.05$	464	0.0062	0.9899	0.0067	0.9829	0.0070	0.9771	0.0073	0.9680
$x = 0.55$	466	0.0325	0.0001	0.0283	0.0009	0.0288	0.0007	0.0284	0.0009
$x = 0.78$	434	0.0298	0.0013	0.0263	0.0069	0.0247	0.0143	0.0275	0.0040
$x = 0.95$	471	0.0124	0.5602	0.0129	0.4976	0.0121	0.5852	0.0121	0.5864
Total	1835	0.0200	0.0015	0.0184	0.0107	0.0180	0.0168	0.0187	0.0079
BIC		73.4		71.4		73.6		76.1	

The columns list the values of sample numbers (*N*), statistical distance (S^2), and *p*-value (PV) for individual and combined data sets.

Hamiltonian that includes the first- and second nearest-neighbor interactions (Fig. 2b). Both of these interactions effectively account for bonds between Mo and Re atoms mediated by sulfur bridges. The energy of configuration *i* can be written as,

$$u_i = w_1 \sum\nolimits_{\{NN\}} \delta_{MoRe} + w_2 \sum\nolimits_{\{NNN\}} \delta_{MoRe} \quad (6)$$

where w_1 and w_2 are interaction energies between Mo and Re atoms in the nearest (NN) and next-nearest neighbor (NNN) positions, respectively; the summation runs over all nearest and next-nearest atom pairs with $\delta_{MoRe} = 1$ a for Mo–Re pairs and $\delta_{MoRe} = 0$ otherwise. This class contains our null hypothesis as a special case with the interaction parameters set to zero, and also a subclass of nearest neighbor models with $w_2 = 0$.

The interaction parameters were optimized to minimize the statistical distance between the target histograms and those collected from equilibrium Monte Carlo simulations with the model. As described in "Methods" section, we combined five reference simulations with tentative models to construct the profile of the combined squared statistical distance S^2 as a function of interaction parameters (Fig. 4a). The minimum of this profile was found at $w_1 = -0.1$ and $w_2 = -0.06$. Examples of configurations generated by the equilibrium model at different stoichiometries are presented in Fig. 5. While at the low and high Re ratios *x* the configurations appear random, ordering of like atoms into smaller clusters seems present at the intermediate concentrations. Even though the profiles of Helmholtz free energy and excess entropy in Fig. 6 indicate increased order at *x* ~ 0.5 (negative excess entropy), they are essentially featureless and do not indicate any phase separation, as can be expected from the attractive effective interactions between Mo and Re atoms (or, equivalently, the repulsion between like atoms).

To quantify the agreement between these structures and the target images, we performed hypothesis testing. The *p*-values summarized in Table 1 (under model P2) show that the equilibrium model is more difficult to reject using standard hypothesis testing. Similar to the random model, it would also

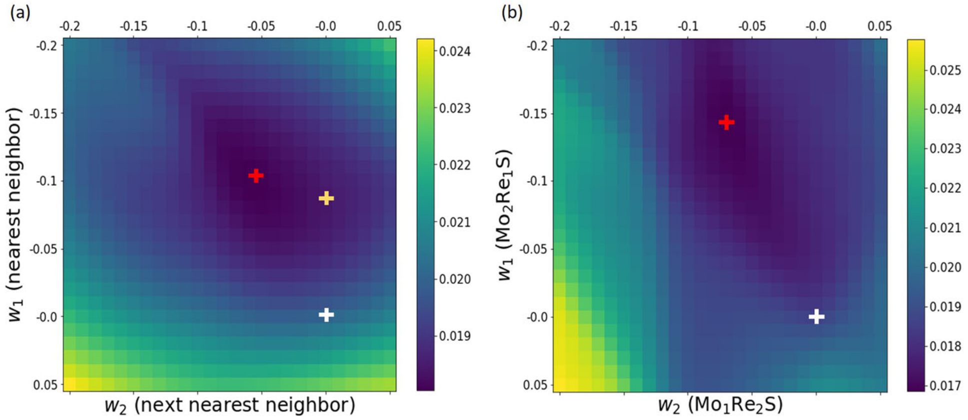

Fig. 4 Statistical distance profiles as a function of interaction parameters. The profiles of statistical distance based on a data set combining all four images, Eq. (2), as a function of parameters w_1 and w_2 of the effective Hamiltonian defined by Eq. (6) on the left (**a**), and Eq. (7) on the right (**b**). Darker colors denote lower values of the loss function, with the minimum for each of the two-parameter models indicated by a red cross, for the single parameter model by a yellow cross, and the random model by a white cross.

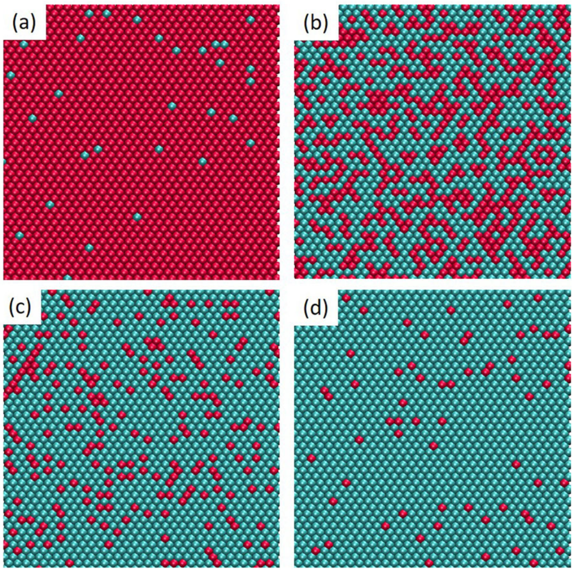

Fig. 5 Simulated Mo and Re atom distributions for different stoichiometries. Configurations generated by two-parameter next-nearest neighbor model, Eq. (6), for different Re fractions *x*. **a** x = 0.05, **b** x = 0.55, **c** x = 0.78, and **d** x = 0.95. Mo (red), Re (cyan).

pass as a generator of configurations at the two extreme concentrations, $x = 0.05$ and 0.95, but performs better for the intermediate concentrations with p-values approximately an order of magnitude larger. This improvement means that the model would pass the test for $x = 0.78$ at the significance level of $\alpha = 0.01$. However, it would still fail to explain the configuration statistics at $x = 0.55$. Given the closeness of the random and equilibrium models, we can be also certain that any transition between these two that would represent partial equilibration from the random state would not pass the significance tests. The BIC for the equilibrium model, Eq. (6), with $r = 2$, attains the value of 73.6, which is nearly identical to the random model. Therefore, the improvement in the statistical distance (Table 1) does not fully justify the two-parameter pair-additive equilibrium model. A simpler pair-additive model can be easily obtained by restricting the interaction parameter to the nearest neighbors by setting $w_2 = 0$ and optimizing only w_1. The optimum of s^2 is then found at $w_1 = -0.08$, as indicated in Fig. 4a. While this choice slightly deteriorates the p-value and s^2, the BIC for this lower-complexity model with $r = 1$ is found to be 71.4, which is more favorable than both the two-parameter and random models. Therefore, accepting this criterion, the amount of available data can justify the choice of the simple nearest-neighbor model.

We may speculate that the overall poor agreement of our pair-additive models stems from their inability to capture the correct form of physical interactions across the range of stoichiometries. In particular, they do not explicitly account for the different bonding topologies of the MoS_2 and ReS_2 lattices identified at low and high x values, respectively. To test an alternative model of bonding interactions within the system, we constructed a model with a simple many-body Hamiltonian that reflects bonding between triplets of metal atoms sharing the same sulfur atom. Since we are using simulations in the canonical ensemble, which keeps the number of particles of each type constant, we can set the pure-phase energies to zero and only optimize interactions responsible for the mixing of Mo–Re atoms. Within this model, the energy of configuration i can be written as,

$$u_i = w_1 \sum\nolimits_{\{S\}} \delta_{\mathrm{MoMoRe}} + w_2 \sum\nolimits_{\{S\}} \delta_{\mathrm{MoReRe}} \qquad (7)$$

where w_1 and w_2 are interaction energies of sulfur with Mo_2Re and $MoRe_2$ neighbors; the summation runs over all S atoms with $\delta_{\mathrm{MoMoRe}} = 1$ for S with two Mo and one Re bonds, and $\delta_{\mathrm{MoMoRe}} = 0$ otherwise; similarly for δ_{MoReRe} with two Re and one Mo. As in the pair-additive model, this model class contains the null hypothesis as a special case with the interaction parameters set to zero.

We followed the same optimization procedure as in the pair-additive model to find the two interaction parameters. The profile of combined squared statistical distance S^2 as a function of interaction parameters is shown in Fig. 4b, with the minimum found at $w_1 = -0.14$ and $w_2 = -0.07$. As in the previous cases, the negative interaction coefficients indicate favorable mixing of Mo and Re. The statistical distances and *p*-values summarized in Table 1 show that the many-body model is more difficult to reject than the random model based on standard hypothesis testing but performs worse than the simple nearest-neighbor model. Taking model complexity into account, the BIC criterion for the equilibrium model, Eq. (6), with $r = 2$, attains the value of 76.1, which is slightly worse than even the random model.

While we were able to find a simple pair-additive model of elemental segregation in $Mo_xRe_{1-x}S_2$, the overall agreement with the imaging data is not completely satisfactory. This indicates that not all physically important effects are captured by the current equilibrium and random models. One possibility to further improve the equilibrium models is to include elastic contributions in the Hamiltonian. A more likely explanation of the discrepancies

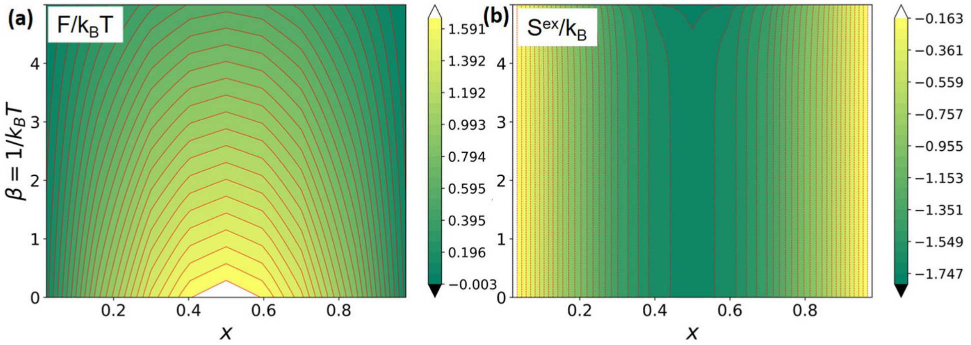

Fig. 6 Thermodynamic properties predicted by the next-nearest neighbor model. a Helmholtz free energy, F, and **b** excess entropy, S^{ex}, of the equilibrium model as a function of Re fraction x and inverse reduced temperature β.

Table 2. Adjusted p-values based on the effective number of pseudo-experimental samples, N_{eff}, from conditional simulations.

Data set	N^{eff}	PV* (R)	PV* (P1)	PV* (P2)	PV* (M)
$x = 0.05$	148	0.9999	0.9999	0.9999	0.9999
$x = 0.55$	365	0.0043	0.0211	0.0177	0.0204
$x = 0.78$	300	0.0746	0.1700	0.2367	0.1301
$x = 0.95$	248	0.9839	0.9788	0.9865	0.9865

The columns list the adjusted p-values (PV*) obtained from Eq. (4) using N_{eff} for each data set (notation follows Table 1).

seems to be the presence of structures created by non-equilibrium processes, whose reproduction would require adequate models. For instance, a model of spinodal decomposition could be tested in a similar manner. However, more data in the form of additional images would be needed to justify selecting a more complex model (equilibrium or dynamic) capable of explaining the observed structures.

Data limitations and uncertainty estimates

Since we only have a single STEM image available for each concentration, it is possible that the visible area is not fully representative of the overall atom distribution. One way to estimate the effect of concentration heterogeneity and its effect on model uncertainty is to evaluate atom–atom correlations within an observed area and derive a corresponding Gaussian process model. Subsequently, analyzing the model's uncertainties and generating synthetic images using conditional simulations can give us an idea about the error caused by spatial correlations. As described in more detail in the Supplementary Information document, we have performed such an analysis and simulations, finding only a very limited range of correlations not extending beyond the nearest neighbors, which is consistent with the absence of clear large-scale patterns. Nevertheless, the effective number of samples, estimated from the variance of local configuration histogram values collected from 100 conditional simulations, suggests that the models with optimized interactions (P1, P2, and M) would pass the significance test at the level of $\alpha = 0.01$ or better (Table 2).

DISCUSSION

We have used atomically resolved STEM images of compositional fluctuations in $Mo_xRe_{1-x}S_2$ to develop statistical models of elemental distribution at different stoichiometries. Using these thermodynamic models, we tested alternative hypotheses about the origins of observed structures. While the random model, which ignores any interaction effects (ideal solid solution), appears sufficient to explain the structures observed at low and high ends of the Re relative concentrations, it does not pass the commonly accepted significance levels for the intermediate concentrations, at which the Mo and Re atoms appear to be structured. Alternative equilibrium models with a simple effective pair-additive and many-body Hamiltonians improved the agreement with the observed data relative to the random model, even though only passing the statistical significance test at 0.01 level for the medium range of Re and Mo concentrations (x ~ 0.5). Based on this analysis, we conclude that the investigated material is close to an ideal solution at forming temperature with weak attractive interactions between the Mo and Re atoms, i.e., the tendency for chemical mixing.

We note that while it is difficult to prove that the observed sets of configurations are samples from an equilibrium distribution, and in fact, the results indicate that it is unlikely that the structures are not influenced by the material's history, it is possible to test different statistical mechanical models that incorporate both equilibrium and non-equilibrium effects, with the complexity of these models only limited by the amount of available data.

Overall, this approach greatly increases the value of STEM data by allowing it to be connected to the thermodynamic or more complex properties of the system. By the same token, it necessitates the acquisition of much larger data volumes[41,42]. While previously a single image provided qualitative information on the system properties, the use of more data enables more statistics, which in turn facilitates improved understanding and discrimination ability between competing models. Furthermore, this approach can be used with data from other experimental tools, including atomic probe tomography, etc., and necessitates the development of automated workflows for data analysis and extraction.

The presented analysis, which integrates statistical mechanics principles with statistical learning methods and statistical hypothesis testing can be easily incorporated into materials science workflows for materials design. In general, the presented work follows the path towards seamless integration of physical theory, machine learning, and experiments. Future work will focus on the further development of the unsupervised learning methods for automated feature selection and structure analysis, as well as on expanding the approach to dynamic data and kinetic Monte Carlo modeling.

METHODS

Sample growth

Molybdenum oxide powder (99%, Sigma Aldrich), sulfur powder (99.5%, Sigma Aldrich), and ammonium perrhenate (99%, Sigma Aldrich) were used as precursors for CVD growth. A selected ratio of molybdenum oxide and ammonium perrhenate was added to an alumina boat with a Si/SiO$_2$ (285 nm) wafer cover. The furnace temperature was ramped to 550 °C in 15 min and then kept at 550 °C for another 15 min for the growth of the $Re_xMo_{1-x}S_2$ alloy materials. Sulfur powder in another alumina boat was placed upstream where the temperature was roughly 200 °C. After growth, the furnace was cooled to room temperature using natural convection. The growth process was carried out with 50 SCCM argon at atmospheric pressure.

Electron microscopy characterization

The $Re_xMo_{1-x}S_2$ flakes were transferred to TEM grids by spin-coating PMMA to support the flakes and etching with KOH to release them from the substrates (by dissolving the SiO_2). The annular dark-field images (ADF) were collected using a Nion UltraSTEM100 microscope operated at 60 kV. The as-recorded images were filtered using a Gaussian function (full width half maximum = 0.12 nm) to remove high-frequency noise. The convergence half angle of the electron beam was set to 30 mrad and the collection inner half-angle of the ADF detector was 51 mrad. The samples were baked in a vacuum at 140 °C overnight before STEM observation. During STEM observation, the probe current was controlled between 10 and 60 pA to reduce beam damage.

Monte Carlo simulations and model optimization

Simulations with the effective interaction models were performed on a 2-dimensional hexagonal lattice with periodic boundary conditions along with the $Mo_xRe_{1-x}S_2$ plane directions. The simulation cell contained $N = 2048$ metal atoms which were equilibrated at reduced temperature $T^* = 1$. After equilibration, a total of $10^5 \times N$ individual MC steps consisting of swaps of Mo and Re atoms were performed in each simulation. The search over the model parameter space to minimize the statistical distance loss function was accomplished with the perturbation technique[25,43], which allowed us to minimize the number of MC simulations in the optimization process and reduce thus the computational cost of the inverse problem solution. In the present case of target data with poor statistics, the basic version of the technique based on reweighting the results of a single MC simulation provided inaccurate estimates. Therefore, we used the multistate Bennett acceptance ratio (MBAR) method[44] to combine the results of 5 reference system simulations performed with models with interaction parameters (w_1, w_2) set to (0, 0), (0.2, 0.0), (−0.2, 0.0), (0.0, 0.2), and (0.0, −2.0).

DATA AVAILABILITY

The data used for the analysis of atomic configurations and Gaussian process simulations is is located along with the processing code on materialscloud.org, https://doi.org/10.24435/materialscloud:w8-k3. The rest of the data that support the findings of this study are available from the corresponding author upon request.

CODE AVAILABILITY

The code used to perform the analysis of atomic configurations is available on materialscloud.org, https://doi.org/10.24435/materialscloud:w8-k3.

Received: 17 October 2019; Accepted: 25 January 2021;
Published online: 25 March 2021

REFERENCES

1. Riordan, M., & Hoddeson, L. E. *Crystal Fire: The Invention of the Transistor and the Birth of the Information Age* (1998).
2. Bagotsky, V. S. *Fuel Cells: Problems and Solutions* (2009).
3. Winter, M., Besenhard, J. O., Spahr, M. E. & Novak, P. Insertion electrode materials for rechargeable lithium batteries. *Adv. Mater.* **10**, 725–763 (1998).
4. Chung, S. Y., Kim, I. D. & Kang, S. J. L. Strong nonlinear current-voltage behaviour in perovskite-derivative calcium copper titanate. *Nat. Mater.* **3**, 774–778 (2004).
5. Kittel, C., McEuen, P. & McEuen, P. *Introduction to Solid State Physics* Vol. 8 (1976).
6. Manley, M. E. et al. Phonon localization drives polar nanoregions in a relaxor ferroelectric. *Nat. Commun.* **5**, 3683 (2014).
7. Dagotto, E. Complexity in strongly correlated electronic systems. *Science* **309**, 257–262 (2005).
8. Dagotto, E., Hotta, T. & Moreo, A. Colossal magnetoresistant materials: the key role of phase separation. *Phys. Rep. Rev. Sec. Phys. Lett.* **344**, 1–153 (2001).
9. Tokura, Y. & Nagaosa, N. Orbital physics in transition-metal oxides. *Science* **288**, 462–468 (2000).
10. Imada, M., Fujimori, A. & Tokura, Y. Metal-insulator transitions. *Rev. Mod. Phys.* **70**, 1039–1263 (1998).
11. Vugmeister, B. E. & Rabitz, H. Kinetics of electric-field-induced ferroelectric phase transitions in relaxor ferroelectrics. *Phys. Rev. B* **65**, 024111 (2001).
12. Ma, E. Y. et al. Charge-order domain walls with enhanced conductivity in a layered manganite. *Nat. Commun.* **6**, 7595 (2015).
13. Vasudevan, R. K. et al. Surface reconstructions and modified surface states in La1-xCaxMnO3. *Phys. Rev. Mater.* **2**, 104418 (2018).
14. Chang, H. et al. Watching domains grow: in-situ studies of polarization switching by combined scanning probe and scanning transmission electron microscopy. *J. Appl. Phys.* **110**, 052014 (2011).
15. Nelson, C. T. et al. Domain dynamics during ferroelectric switching. *Science* **334**, 968–971 (2011).
16. Pennycook, S. J. & Nellist, P. D. *Scanning Transmission Electron Microscopy: Imaging and Analysis* (2011).
17. Krivanek, O. L. et al. Atom-by-atom structural and chemical analysis by annular dark-field electron microscopy. *Nature* **464**, 571–574 (2010).
18. Huang, P. Y. et al. Imaging atomic rearrangements in two-dimensional silica glass: watching silica's dance. *Science* **342**, 224–227 (2013).
19. Stroscio, J. A., Feenstra, R. M. & Fein, A. P. Electronic structure of the Si(111)2 × 1 surface by scanning-tunneling microscopy. *Phys. Rev. Lett.* **57**, 2579–2582 (1986).
20. Bonnell, D. A. & Garra, J. Scanning probe microscopy of oxide surfaces: atomic structure and properties. *Rep. Prog. Phys.* **71**, 044501 (2008).
21. Sugimoto, Y. et al. Chemical identification of individual surface atoms by atomic force microscopy. *Nature* **446**, 64–67 (2007).
22. Vlcek, L., Vasudevan, R. K., Jesse, S. & Kalinin, S. V. Consistent integration of experimental and ab initio data into effective physical models. *J. Chem. Theory Comput.* **13**, 5179–5194 (2017).
23. Wootters, W. K. Statistical distance and Hilbert space. *Phys. Rev. D.* **23**, 357–362 (1981).
24. Braunstein, S. L. & Caves, C. M. Statistical distance and the geometry of quantum states. *Phys. Rev. Lett.* **72**, 3439–3443 (1994).
25. Vlcek, L. & Chialvo, A. A. Rigorous force field optimization principles based on statistical distance minimization. *J. Chem. Phys.* **143**, 144110 (2015).
26. Tuyl, F. A note on priors for the multinomial model. *Am. Stat.* **71**, 298–301 (2017).
27. Vasudevan, R. K., Ziatdinov, M., Jesse, S. & Kalinin, S. V. Phases and interfaces from real space atomically resolved data: Physics-based deep data image analysis. *Nano Lett.* **16**, 5574–5581 (2016).
28. Hansen, J. P. & McDonald, I. R. *Theory of Simple Liquids* 3rd edn (2006).
29. Yang, J., Jiang, Y.-G., Hauptmann, A. G. & Ngo, C.-W. Evaluating bag-of-visual-words representations in scene classification. In *Proc. International Workshop on Workshop on Multimedia Information Retrieval* 197–206 (2007).
30. Vlcek, L., Maksov, A., Pan, M., Vasudevan, R. K. & Kalinin, S. V. Knowledge extraction from atomically resolved images. *ACS Nano* **11**, 10313–10320 (2017).
31. Vlcek, L. et al. Learning from imperfections: predicting structure and thermodynamics from atomic imaging of fluctuations. *ACS Nano* **13**, 718–727 (2019).
32. Elfadaly, F. G., Garthwaite, P. H. & Crawford, J. R. On point estimation of the abnormality of a Mahalanobis index. *Comput. Stat. data Anal.* **99**, 115–130 (2016).
33. Rissanen, J. J. Fisher information and stochastic complexity. *IEEE Trans. Inf. Theory* **42**, 40–47 (1996).
34. Grünwald, P. Model selection based on minimum description length. *J. Math. Psychol.* **44**, 133–152 (2000).
35. Schwarz, G. Estimating the dimension of a model. *Ann. Stat.* **6**, 461–464 (1978).
36. Yang, S. Z. et al. Rhenium-doped and stabilized MoS2 atomic layers with basal-plane catalytic activity. *Adv. Mater.* **30**, 1803477 (2018).
37. Yang, S.-Z. et al. Direct cation exchange in monolayer MoS 2 via recombination-enhanced migration. *Phys. Rev. Lett.* **122**, 106101 (2019).
38. Somnath, S. et al. Feature extraction via similarity search: application to atom finding and denoising in electron and scanning probe microscopy imaging. *Adv. Struct. Chem. Imaging* **4**, 3 (2018).
39. Somnath, S., Smith, C., Laanait, N., Vasudevan, R. & Jesse, S. USID and pycroscopy—open source frameworks for storing and analyzing imaging and spectroscopy data. *Microsc. Microanal.* **25**, 220–221 (2019).
40. Somnath, S., Smith, C. R., Laanait, N. & Jesse, S. Pycroscopy. *Comput. Softw.* https://pycroscopy.github.io/pycroscopy/ (2019).
41. Kalinin, S. V. et al. Big, deep, and smart data in scanning probe microscopy. *ACS Nano* **10**, 9068–9086 (2016).
42. Kalinin, S. V., Sumpter, B. G. & Archibald, R. K. Big-deep-smart data in imaging for guiding materials design. *Nat. Mater.* **14**, 973–980 (2015).
43. Chialvo, A. A. Excess properties of liquid-mixtures from computer simulation—a coupling parameter approach to the determination of their dependence on molecular asymmetry. *Mol. Phys.* **73**, 127–140 (1991).
44. Shirts, M. R. & Chodera, J. D. Statistically optimal analysis of samples from multiple equilibrium states. *J. Chem. Phys.* **129**, 124105 (2008).

ACKNOWLEDGEMENTS

The effort at ORNL including electron microscopy (S.Y., M.F.C., W.Z.) and image analytics and statistical distance minimization (R.K.V., S.V.K., L.V.) was supported by the U.S. Department of Energy, Office of Science, Basic Energy Sciences, Materials

Sciences, and Engineering Division and performed at the Center for Nanophase Materials Sciences, which is a US DOE Office of Science User Facility.

AUTHOR CONTRIBUTIONS

S.Y., Y.G., P.A., M.F.C., and W.Z. performed electron microscopy; R.K.V., S.V.K., and M.Z. performed image processing and analytics and co-wrote the manuscript. L.V. has performed the analysis of atomic configurations, run Monte Carlo simulations, optimized models, and co-wrote the manuscript.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41524-021-00507-7.

Correspondence and requests for materials should be addressed to L.V., R.K.V. or S.V.K.

Reprints and permission information is available at http://www.nature.com/reprints

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