

## ARTICLE OPEN

## Crowd-sourcing materials-science challenges with the NOMAD 2018 Kaggle competition

Christopher Sutton<sup>1\*</sup>, Luca M. Ghiringhelli<sup>1\*</sup>, Takenori Yamamoto<sup>2</sup>, Yury Lysogorskiy<sup>3</sup>, Lars Blumenthal<sup>4,5</sup>, Thomas Hammerschmidt<sup>3</sup>, Jacek R. Golebiowski<sup>4,5</sup>, Xiangyue Liu<sup>1</sup>, Angelo Ziletti<sup>1</sup> and Matthias Scheffler<sup>1</sup>

A public data-analytics competition was organized by the Novel Materials Discovery (NOMAD) Centre of Excellence and hosted by the online platform Kaggle by using a dataset of 3,000  $(\text{Al}_x\text{Ga}_y\text{In}_{1-x-y})_2\text{O}_3$  compounds. Its aim was to identify the best machine-learning (ML) model for the prediction of two key physical properties that are relevant for optoelectronic applications: the electronic bandgap energy and the crystalline formation energy. Here, we present a summary of the top-three ranked ML approaches. The first-place solution was based on a crystal-graph representation that is novel for the ML of properties of materials. The second-place model combined many candidate descriptors from a set of compositional, atomic-environment-based, and average structural properties with the light gradient-boosting machine regression model. The third-place model employed the smooth overlap of atomic position representation with a neural network. The Pearson correlation among the prediction errors of nine ML models (obtained by combining the top-three ranked representations with all three employed regression models) was examined by using the Pearson correlation to gain insight into whether the representation or the regression model determines the overall model performance. Ensembling relatively decorrelated models (based on the Pearson correlation) leads to an even higher prediction accuracy.

npj Computational Materials (2019)5:111

; <https://doi.org/10.1038/s41524-019-0239-3>

## INTRODUCTION

Computational approaches have become a powerful tool for guided design of new compounds to potentially aid the development of advanced technologies. However, the identification and discovery of new materials that are ideal for targeted applications is a nontrivial task that requires examining enormous compositional and configurational degrees of freedom. For example, an alloy with two substitutional atoms in the unit cell and with  $M$  sites displays about  $2^M$  possible configurational states (neglecting symmetry) for each lattice, and most often several polymorphs have to be examined.

Density-functional theory (DFT) typically provides the best compromise between accuracy and cost; nevertheless, we used a single energy evaluation by using DFT scaling as a high-order polynomial with system size. As a result of the high computational demand, DFT-based exploration of configurational spaces of alloys is only feasible for unit cells with a relatively small number of atoms. To efficiently search this vast chemical space, methods that allow for fast and accurate estimates of material properties have to be developed.

Machine learning (ML) promises to accelerate the discovery of novel materials by allowing to rapidly screen candidate compounds at significantly lower computational cost than traditional electronic structure approaches.<sup>1–7</sup> A key consideration for an ML model of material properties is how to include atomic and structural information as a fixed-length feature vector to enable regression, which is referred to as the representation or descriptor. Given that knowledge of the atomic positions and chemical species (e.g., the atomic number) for a given system is sufficient to construct the Hamiltonian, a ML descriptor should include the

geometrical and chemical information in a convenient way. A considerable amount of work has been devoted to defining suitable ML descriptors of molecules or materials by encoding the chemical and geometrical information in various ways such as Coulomb matrices,<sup>8,9</sup> scattering transforms,<sup>10</sup> diffraction patterns,<sup>11</sup> bags of bonds,<sup>12</sup> many-body tensor representation,<sup>13</sup> smooth overlap of atomic positions (SOAP),<sup>14,15</sup> and several symmetry-invariant transformations of atomic coordinates.<sup>16–18</sup> All of these approaches represent the training or test samples and are typically combined with kernel ridge regression (KRR) or Gaussian process regression (GPR)<sup>19</sup> methods to effectively identify differences in the structures of the dataset. In addition, generalized atom-centered symmetry functions have also been developed to be combined with a neural network (NN).<sup>20</sup> Other approaches such as a modified Least Absolute Shrinkage and Selection Operator (LASSO)<sup>21</sup> and the Sure Independence Screening and Sparsifying Operator (SISSO)<sup>22</sup> have focused on identifying the best descriptor out of a large space of mathematical combinations of simple features that represent the chemical information and structural information.<sup>23–25</sup> Of particular importance for the efficient modeling of the large configurational space of substitutional alloys, the cluster expansion (CE) method<sup>26–32</sup> is an ML representation by using only an occupational variable for each substitutional lattice site. However, the lack of explicit local atomic information (e.g., bond distances and angles) of the crystalline systems prevents a broad and transferable application of this approach. Along these same lines, semiempirical interatomic potentials or force field-based approaches use parameterized models based on classical mechanics (e.g., short-

<sup>1</sup>Fritz Haber Institute of the Max Planck Society, Berlin, Germany. <sup>2</sup>Research Institute for Mathematical and Computational Sciences (RIMCS), Yokohama, Japan. <sup>3</sup>ICAMS, Ruhr-Universität Bochum, Bochum, Germany. <sup>4</sup>EPSRC Centre for Doctoral Training on Theory and Simulation of Materials, Department of Physics, Imperial College London, London, UK. <sup>5</sup>Thomas Young Centre for Theory and Simulation of Materials, Department of Materials, Imperial College London, London, UK. \*email: [sutton@fhi-berlin.mpg.de](mailto:sutton@fhi-berlin.mpg.de); [ghiringhelli@fhi-berlin.mpg.de](mailto:ghiringhelli@fhi-berlin.mpg.de)

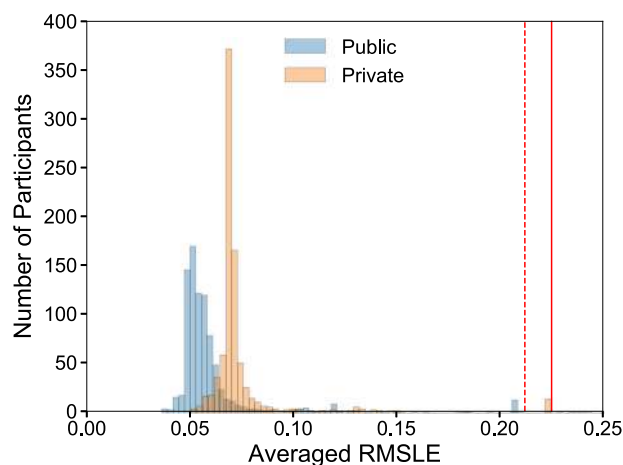
range two-body and three-body interactions, long-range Coulomb interactions) to approximate quantum mechanical properties.

With so many choices of the various structural representations, it is often unclear which will be the most insightful or accurate for a given problem. Furthermore, optimizing an ML model for a particular application can be a time-consuming endeavor: a given representation is combined with a specific regression model (i.e., a model class and an induction algorithm) whose hyperparameters are tuned subsequently. Therefore, typically, only a few combinations of representation and regression algorithms are carefully tested for a specific application, which limits the understanding of how well various ML models perform. Crowd sourcing offers an alternative approach for examining several ML models by identifying a key problem and challenging the community to solve it by proposing solutions that are ranked in an unbiased way. To this end, the Novel Materials Discovery (NOMAD)<sup>33</sup> Centre of Excellence organized a data-analytics competition for predicting the key properties of transparent conducting oxides (TCOs) with Kaggle, which is one of the most recognized online platforms for hosting data-science competitions.

TCOs are an important class of well-developed and commercialized wide bandgap materials that have been employed in a variety of (opto)electronic devices such as solar cells, light-emitting diodes, field-effect transistors, touch screens, sensors, and lasers.<sup>34–44</sup> However, only a small number of compounds display both transparency and electronic conductivity suitable enough for these applications. For example, tin-doped indium oxide ( $\text{In}_2\text{O}_3:\text{Sn}$ ) serves as the primary transparent electrode material for (opto)electronic devices because of its high transparency over the visible range, resulting from an electronic bandgap energy of 2.7 eV,<sup>45,46</sup> and its high electrical conductivity,<sup>47–49</sup> which are typically competing properties. A wide range of experimental bandgap energies from 3.6 to 7.5 eV have been reported from alloying  $\text{In}_2\text{O}_3/\text{Ga}_2\text{O}_3$  or  $\text{Ga}_2\text{O}_3/\text{Al}_2\text{O}_3$ ,<sup>50–56</sup> which suggest that alloying of group-III oxides is a viable strategy for designing new wide-bandgap semiconductors. However,  $\text{Al}_2\text{O}_3$ ,  $\text{Ga}_2\text{O}_3$ , and  $\text{In}_2\text{O}_3$  all display very different ground-state structures. Therefore, it is unclear which structure will be stable for various compositions.

The goal of the competition was to identify the best ML model for both the formation energy (an indication of the stability) and the bandgap energy (an indication of transparency) by using a dataset that contained 3,000  $(\text{Al}_x\text{Ga}_y\text{In}_{1-x-y})_2\text{O}_3$  compounds (with  $x + y + z = 1$ ) spanning six preassigned lattice symmetries ( $C2/m$ ,  $Pna2_1$ ,  $R\bar{3}c$ ,  $P6_3/mmc$ , and  $la\bar{3}$ , and  $Fd\bar{3}m$ ), which are some of the phases that have been experimentally reported for the binary or ternary compounds.<sup>57–67</sup> In terms of the dataset, 2,400 structures were used for the training set, with the remaining 600 structures included in the test set that was kept secret throughout the entire competition. A similar distribution of the two target properties is observed for the training and test sets (Fig. S1); a boxplot of the two target properties for each of the six lattices is provided in the Supporting Information (see Fig. S2). A ternary diagram of the unique compositions used in the dataset is provided in Fig. S3.

The competition was launched on 18 December, 2017 and ended on 15 February, 2018, attracting 883 participants. Figure 1 shows the distribution of the so-called public and private leaderboard scores for all the participants of the competition. The public score was calculated for only 100 fixed samples from the test set in order to quickly assess the performance of the submitted models, with the two target properties of these samples still kept secret. The remaining 500 samples of the test set were used to determine the winner of the competition, which is displayed in the private leaderboard. The scoring metric used in the competition was the root mean square logarithmic



**Fig. 1** Histogram of averaged RMSLE of the bandgap and formation energies for all of the 883 models submitted in the NOMAD 2018 Kaggle competition. The scores are shown for the Kaggle public and private scoreboards of the test containing 600 samples with the values of these two target properties withheld for the entire competition. The public score was calculated for 100 fixed samples; the private score was calculated for 500 samples and was used to determine the winner of the competition. The vertical red lines correspond to the predictions from taking the average value of the training set to predict the public (dashed line) and private (line) datasets

error (RMSLE)

$$\text{RMSLE} = \sqrt{\frac{1}{N} \sum_{i=1}^N \left( \log \left( \frac{\hat{y}_i + 1}{y_i + 1} \right) \right)^2}$$

where  $N$  is the total number of samples. The error is calculated as the log ratio of the predicted target property  $\hat{y}_i$  and the corresponding reference value  $y_i$ . The error for both the formation energy and bandgap energy is then averaged for a final assessment of the model performance. The log ratio of the errors is a convenient choice because it prevents the bandgap, which is an order of magnitude larger than the formation energy (see Fig. S1), from dominating an analysis of the predictive capability of each model.

For the practical application of ML models for high-throughput screening, it is of particular importance to have a model that inputs structural features based on a generalized unrelaxed geometry because the relaxed structures are not readily available. If the relaxed geometry needed to be calculated to obtain the input features for an ML model, then all of the quantities of interest would already be available. In this competition, the structures were provided by the linear combination of the stoichiometric amounts of the  $\text{Al}_2\text{O}_3$ ,  $\text{Ga}_2\text{O}_3$ , and  $\text{In}_2\text{O}_3$  geometries at the same lattice symmetry (i.e., obtained by applying Vegard's law<sup>68,69</sup> for the lattice vectors to generate the input structures); however, the target properties are computed from the fully relaxed geometries.

In this paper, we first describe the performance of the three ML approaches on the original dataset provided in the NOMAD 2018 Kaggle competition. We then provide a comparison in the performance of these three representations with various regression methods to gain an understanding of the key determining factors for the high performance of the winning models. A comparison between the errors of the fully optimized geometries and those obtained by using the starting structures generated by using Vegard's law is also provided. Finally, we examine the generalization error of the ML models for lattice symmetries outside of the training set. We only briefly describe the models in

**Table 1.** A summary of the three winning models of the competition with the test-set root mean square log error (RMSLE) and mean absolute error (MAE) of the formation energy and bandgap energy

Ranking	ML representation + regression method	Formation energy		Bandgap energy	
		Root mean square log error	Mean absolute error (meV/cation)	Root mean square log error	Mean absolute error (meV/cation)
First	<i>n</i> -gram + KRR	0.020 (0.021 <sup>a</sup> )	14 (15 <sup>a</sup> )	0.078 (0.077 <sup>a</sup> )	106 (114 <sup>a</sup> )
Second	<i>c</i> /BOP + LGBM	0.022	15	0.081	93
Third	SOAP + NN	0.021	13	0.083	98

<sup>a</sup>Determined by using the 4-gram representation instead of an ensemble of 3-gram and 4-gram that was used in the actual competition

the main text; a detailed description of each of the three winning ML models from the competition is provided in the “Methods” section.

## RESULTS

Performance of the three winning approaches from the NOMAD 2018 Kaggle competition

As already mentioned in the introduction, the errors in both the bandgap and formation energy of the crystalline system differ by about an order of magnitude in their mean and standard deviations. Thus, simply averaging the two absolute errors would result in an error metric that is dominated by the bandgap energy because of its larger magnitude. This is why the RMSLE was the performance metric used in the competition. However, for this discussion, we use the mean absolute errors (MAE) of the bandgap and formation energies separately because they allow for a more intuitive quantification of model performance from a physical point of view:

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^N |\hat{y}_i - y_i|$$

Table 1 compares the RMSLE and MAE for the top-three models. The first-place model employed a crystal-graph representation composed of histograms of unique coordination environments and unique edge sequences (called *n*-grams<sup>70</sup>) extracted from a multigraph representing a bonding network in the crystal structure, which was combined with kernel ridge regression (KRR). An ensemble of *n*-gram feature lengths was used in the model submission because it gave the lowest cross-validation (CV)-score RMSLE; however, in this paper, the results from the concatenated list of 1-gram, 2-gram, 3-gram, and 4-gram features (which is referred to as the 4-gram herein) are discussed to facilitate a comparison between each of the different regression methods without the need for additional tuning of the mixing coefficient in the ensemble model.

The second-place model (*c*/BOP+LGBM) starts from a large set of candidate features (i.e., weighted chemical properties as well as atomic-environment representations based on analytic bond-order potentials (BOP)<sup>71–74</sup> and basic geometric measures), which was then optimized and combined with the light gradient-boosting machine (LGBM) regression model.<sup>75</sup> The third-place solution used the SOAP representation developed by Bartók et al.<sup>14,15</sup> that incorporates information on the local atomic environment through a rotationally integrated overlap of the Gaussian-shaped densities centered at the neighbor atoms, which was combined with a three-layer feed-forward NN (SOAP+NN).

The top-three models have a test-set MAE for the formation energy within 2 meV/cation, whereas a larger range of 21 meV is observed for the predictions of the bandgap energy (Table 1). Based on the learning curves provided in Fig. S4, the formation energy MAEs of all the three methods converge to  $\leq$ ca. 2.4 meV/cation for training-set sizes  $\geq$ 960 samples relative to the error

obtained when training on the full dataset of 2,400 structures. For the bandgap energies, a test-set MAE  $\leq$ 16 meV relative to the error obtained when training on the full 2400 samples is achieved for 960, 1440, and 1920 training samples for SOAP + NN, *c*/BOP + LGBM, and 4-gram + KRR, respectively. Furthermore, for all three models, these errors vary by  $<$ ca. 3 meV/cation and 20 meV for the formation energy and bandgap energy, respectively, for five additional random 80/20% splits of the entire 3000-compound dataset (Table S1).

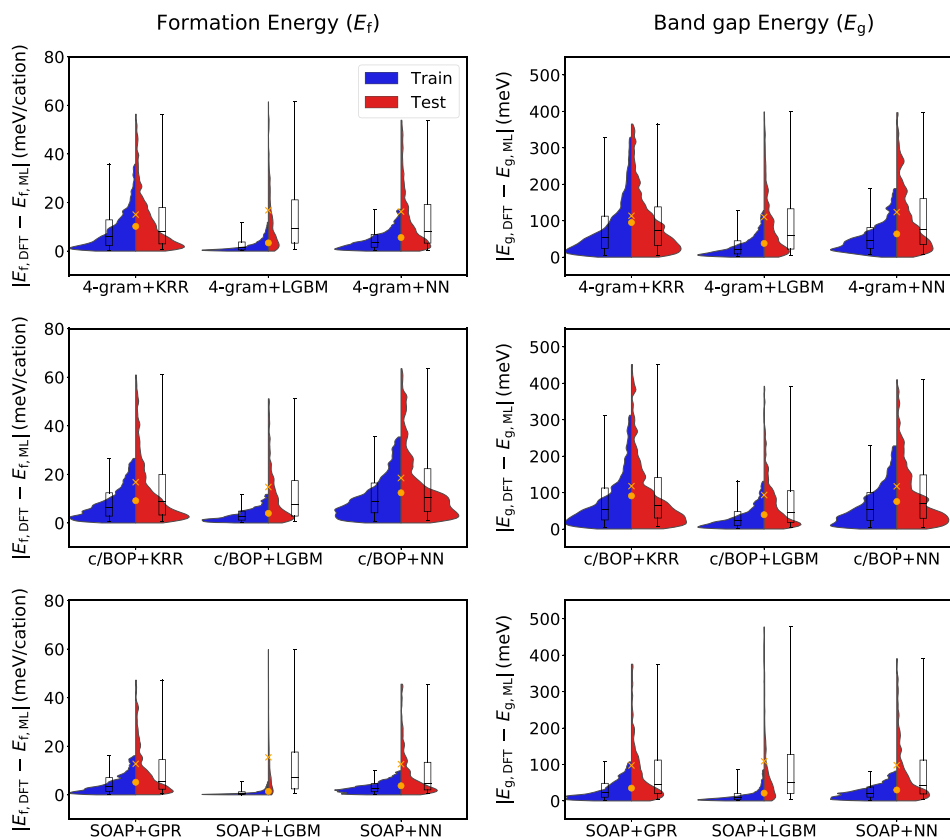
Overall, the higher accuracy in the formation energy for all three approaches is attributed to the inclusion of the local atomic topography in each model. The lower accuracy for the bandgap energy is attributed partly to the fact that the valence band is determined by hybridization of oxygen atoms, whereas the conduction band is described by the metal–metal interactions. Therefore, an accurate description of this property most likely requires additional information to be included in the representation beyond the local structure.

Three winning representations combined with all three regression methods

To understand the effect of the choice of representation vs. regression model on the overall error, we now examine the performance of each representation combined with KRR/GPR, NN, and LGBM, with the hyperparameters optimized for each representation and regression method combination. A detailed description of each of the nine models is provided in the Methods section.

The primary goal for training an ML model is to accurately generalize the rules learned on the training set to make predictions on unseen data. Overfitting describes the propensity of an ML model to give a higher accuracy on the training set compared with the test set, which is an indication of poorly generalizable predictions of the model. To evaluate the generalizable error, we investigate the difference between the 95% percentiles of the MAE for the training and test sets for each of the nine ML models ( $\Delta$ 95%). The 95% percentiles for the training set and test set are given by the upper edges of the boxplots in Fig. 2 (the explicit values for the MAE and 95% percentile are provided in Table S2).

Beginning with a discussion of the errors in the formation energy, a practically identical error is observed among the predictions from all the three regression models (KRR/GPR, NN, and LGBM) by using the same *c*/BOP, SOAP, and 4-gram representations, with a maximum difference of 4 meV/cation, 2 meV/cation, and 2 meV/cation, respectively (Fig. 2). However, a large variation of the  $\Delta$ 95% value between the training and test predictions is observed. For example, a consistently larger  $\Delta$ 95% value is calculated when the NN and LGBM regression methods are used, irrespective of the three representations. This is apparent in Fig. 2 with a much narrower distribution of the training-set absolute errors (blue) compared with the test-set absolute errors (red). More specifically and focusing only on the formation energy,



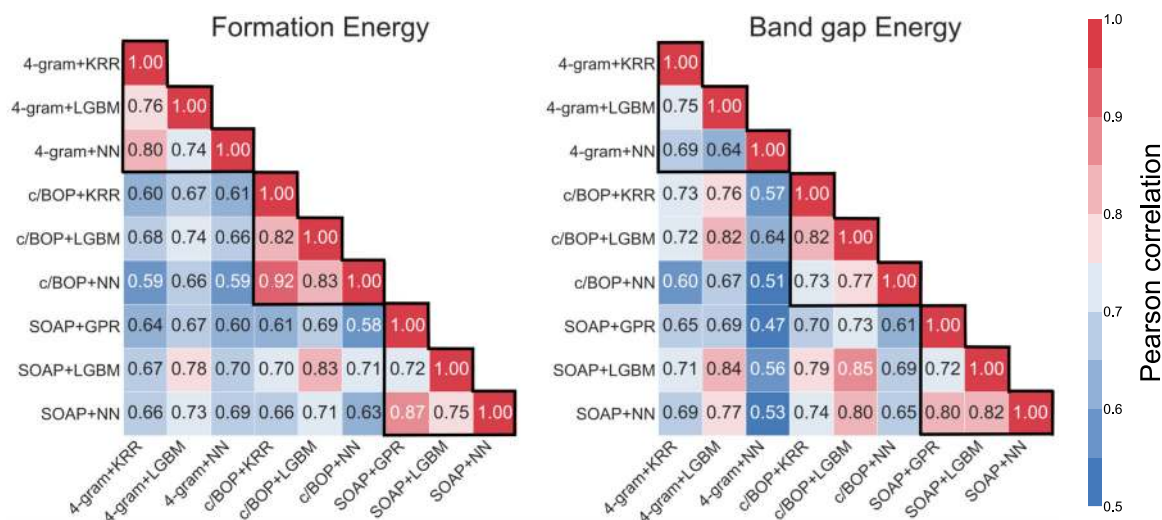
**Fig. 2** A comparison of the distribution of the absolute errors for the training set (blue) and test set (red) of the formation energy ( $E_f$ , left) and bandgap energy ( $E_g$ , right) from the three winning representations (4-gram, c/BOP, and SOAP) of the competition combined with the KRR/GPR, NN, and LGBM regression models. The mean absolute errors (MAE) of the test set (orange cross) and training set (orange filled circle) are provided. Boxplots are included for each training and test-set distribution to indicate the 25, 50, and 75% percentiles of the absolute errors. The box and violin plots only extend to the 95% percentile. For the training-set predictions, the maximum absolute error in the formation (bandgap) energy for 4-gram + KRR, c/BOP + LGBM, and SOAP + NN is 103 meV/cation (1047 meV), 185 meV/cation (606 meV), and 376 meV/cation (497 meV), respectively. The corresponding maximum absolute test errors are 282 meV/cation (1112 meV), 276 meV/cation (1680 meV), and 286 meV/cation (1198 meV) for the 4-gram + KRR, c/BOP + LGBM, and SOAP + NN models, respectively

a markedly larger  $\Delta 95\%$  is observed for 4-gram + NN ( $\Delta 95\% = 95\%_{\text{train}} - 95\%_{\text{test}} = 37$  meV/cation) and LGBM ( $\Delta 95\% = 49$  meV/cation) compared with KRR ( $\Delta 95\% = 20$  meV/cation). A similar trend is found for SOAP representation combined with NN ( $\Delta 95\% = 36$  meV/cation), LGBM ( $\Delta 95\% = 55$  meV/cation), and GPR ( $\Delta 95\% = 31$  meV/cation). In comparison, only a slightly larger difference between the 95% confidence thresholds of the training and test sets is computed for c/BOP with LGBM ( $\Delta 95\% = 39$  meV/cation), c/BOP + NN ( $\Delta 95\% = 28$  meV/cation), and c/BOP + KRR ( $\Delta 95\% = 34$  meV/cation). These results indicate a consistently larger  $\Delta 95\%$  when the NN and LGBM regression models are used because these regressors are potentially more prone to overfitting in this application. This observation is consistent with the expectation that overfitting is more likely with highly nonlinear functions that have more flexibility when learning a target. However, this might be resolved by a more careful hyperparameter optimization.

The Pearson correlation ( $r$ ) between signed errors in the test-set predictions is used to quantify correlations between test-set errors for all combinations of representation (c/BOP, SOAP, and 4-gram) and regression models (LGBM, NN, or KRR/GPR) to elucidate the dominant factors of the model performance through a comparison of the nine models (Fig. 3). The Pearson correlation is chosen for this analysis because it is a simple parameter-free measure of the linear correlation between two variables (i.e., the residuals between two models) to indicate where two ML models have similar predictions for the test set. As we noted above, a practically identical error in the formation energy is observed among the predictions from all the three regression models (KRR/GPR, NN,

and LGBM) by using the c/BOP, SOAP, and 4-gram representations, with a maximum difference of 4 meV/cation, 2 meV/cation, and 2 meV/cation, respectively. The minor variation in the average error is attributed to the dominant effect of the representation in the overall accuracy. However, the range of  $r$  values between errors of the three models by using these representations combined with the three different regression models ( $r = 0.74$ – $0.80$ ,  $r = 0.72$ – $0.87$ , and  $r = 0.82$ – $0.92$  for 4-gram, SOAP, and c/BOP models, respectively) indicates that the accuracy of the three ML models is correlated but not identical. In addition, the highest Pearson correlations in the formation energy errors is observed for the predictions obtained with the c/BOP representation, indicating that these models have a strongly correlated description of the test set. Furthermore, among all three representations, the highest Pearson coefficients are consistently obtained for the residuals in the formation energy between predictions by using KRR/GPR and NN, with 4-gram + KRR vs. 4-gram + NN ( $r = 0.80$ ), SOAP + GPR vs. SOAP + NN ( $r = 0.87$ ), and c/BOP + KRR vs. c/BOP + NN ( $r = 0.92$ ). In general, the high Pearson correlation among errors of the same representation indicates that the choice of the representation is a determining factor in the performance of these approaches.

In contrast to what is observed for the formation energy where the predictions made from the same representations are the most correlated and largely independent of the regression model, the bandgap energies are less correlated overall. Compared with what is found in the errors in the formation energy, an overall decrease in the correlation in the residuals of the models by utilizing the



**Fig. 3** Pearson correlation in the test-set errors of the formation energy (left) and bandgap energy (right) between each of the nine combinations of representation and regression models examined in this study. The predictions obtained for the same representation are outlined in black

KRR/GPR and NN regressors is observed: 4-gram + KRR vs. 4-gram + NN ( $r = 0.69$ ), c/BOP + KRR vs. c/BOP + NN ( $r = 0.73$ ), and SOAP + GPR vs. SOAP + NN ( $r = 0.80$ ). These lower Pearson correlation scores for the bandgap errors both indicate that even with the same representation, the three respective ML models perform differently for the bandgap predictions, which is potentially a result of the larger errors in this target property.

With an understanding of the correlation for each representation but using different regressors, a key question becomes how correlated the prediction errors are between all nine ML models. The highest correlation is observed when the LGBM regression model is used with the three representations. For the formation energy residuals, the correlations between 4-gram + LGBM vs. SOAP + LGBM ( $r = 0.78$ ), and c/BOP + LGBM vs. SOAP + LGBM ( $r = 0.83$ ) show a higher correlation compared with the predictions with 4-gram + LGBM vs. c/BOP + LGBM ( $r = 0.74$ ). For the errors in the bandgap energies, the highest correlations ( $r = 0.82$ – $0.85$ ) are observed between each representation and combined with LGBM, which further demonstrates that this regression model dominates the prediction of this target property. This is rationalized to occur because the LGBM algorithm builds an accurate ML model by ensembling weak learners, which are flowchart-like structures that allow for input data points to be classified based on questions learned from the data.<sup>76</sup> To improve the model predictions, gradient boosting is used to iteratively train additional models on the error. This process specifically addresses the weak points of the previous models, and therefore, the improved correlation indicates that the larger errors are more consistently described by these regression models.

A new model obtained by averaging two models with uncorrelated errors (i.e., small  $r$  values) can perform better than individual ML models, which is the basic idea behind the so-called ensembling.<sup>77–79</sup> To demonstrate that this idea holds for the present dataset and set of learners, we have combined various ML models with both small and large Pearson correlations. More specifically, an equivalent test-set error in the formation energy of the 4-gram + KRR model (MAE = 15 meV/cation) can be achieved by averaging the predictions from the 4-gram + NN (MAE = 16 meV/cation) and c/BOP + NN (MAE = 19 meV/cation) models, which have an  $r = 0.59$ . In contrast, averaging two models with a high correlation such as c/BOP + KRR model (MAE = 17 meV/cation) with c/BOP + NN (MAE = 19 meV/cation), which have an  $r = 0.92$ , gives an MAE = 17 meV/cation. This result indicates that the ensembling two correlated models cannot lower the

prediction errors. Furthermore, a test-set MAE = 12 meV/cation – i.e., less than the error of the winning model in the actual competition – can be obtained by averaging the predictions from 4-gram + KRR model (MAE = 15 meV/cation) with SOAP + GPR (MAE = 13 meV/cation), which have  $r = 0.64$ . For the bandgap energy, averaging the 4-gram + NN (MAE = 124 meV) and SOAP + GPR (MAE = 97 meV) models, ( $r = 0.47$ ) yields an MAE = 96 meV, which is lower than the first-place 4-gram + KRR model (MAE = 114 meV). Overall, these results demonstrate that the Pearson correlation allows for an identification of models with weakly correlated predictions, which can be combined to obtain even lower errors.

#### Training and test-set errors using features derived from relaxed structures

For the purposes of efficient predictions in high-throughput screening, it is important to incorporate structural features without performing a geometry optimization. If atomic structural information were required from optimized geometries, then most other quantities would be known as well, and no predictions were necessary. The discussion has so far been limited to a dataset constructed by using geometries generated from the weighted average of the optimized pure binary crystalline systems (i.e., applying Vegard's law<sup>68,69</sup> to generate the input structures). However, the target formation and bandgap energies correspond to the fully optimized structures with the lattice vectors and atomic positions allowed to relax self-consistently. Therefore, to examine the additional challenge for the ML description by using this structure generation procedure, the performance of the top-three ML approaches by using the fully relaxed geometries is also examined.

Differences in the test-set MAEs of 2, 1, and 2 meV/cation for the formation energy are calculated for the 4-gram + KRR, c/BOP + LGBM, and SOAP + NN models, respectively, by using features generated from the relaxed structures compared with the Vegard's law starting structures (Table 2). For the bandgap energy, there exist differences of 1, 3, and 14 meV, respectively, between the MAE values by using the two sets of geometries. The increasing error of predictions from the 4-gram model when features are built from the relaxed geometry is attributed to the rigid definition of the coordination environment, which is determined as the number of interatomic distances within predetermined cutoff values based on the sum of ionic radii. In

**Table 2.** Comparison of test-set MAEs for the different regression methods retrained using fully relaxed geometries for the NOMAD 2018 Kaggle dataset compared with idealized geometries

Representation	Regression method	Formation energy		Bandgap energy	
		Mean absolute error (meV/cation)		Mean absolute error (meV)	
4-gram	KRR	17 <sup>a</sup>	15 <sup>b</sup>	113 <sup>a</sup>	114 <sup>b</sup>
c/BOP	LGBM	14 <sup>a</sup>	15 <sup>b</sup>	96 <sup>a</sup>	93 <sup>b</sup>
SOAP	NN	11 <sup>a</sup>	13 <sup>b</sup>	84 <sup>a</sup>	98 <sup>b</sup>

<sup>a</sup>Calculated by using features generated from the fully optimized structure  
<sup>b</sup>Calculated by using features generated from the Vegard's law structure

**Table 3.** Comparison of test-set MAE values for the three winning models trained for a re-partitioned dataset containing five of six lattice symmetries (2384 samples) and a test set (616 samples) comprising only one lattice symmetry ( $la\bar{3}$ )

Representation	Regressor	Formation energy (meV/cation)	Bandgap energy (meV)
4-gram	KRR	53	179
c/BOP	LGBM	40 <sup>a</sup> /36 <sup>b</sup>	180 <sup>a</sup> /111 <sup>b</sup>
SOAP	NN	132	527

<sup>a</sup>Feature selection and model hyperoptimization according to fivefold CV with splits generated randomly

<sup>b</sup>Feature selection and model hyperoptimization according to fivefold CV with splits generated based on the spacegroup number

the 4-gram representation, the parameterization of the coordination environment for each of the lattice symmetries augments the additional challenge of the Vegard's law starting structure by inputting bias into the model; however, this then leads to a representation that is less flexible to different input structures. In contrast, the SOAP representation is strongly dependent on the geometry used for building the descriptor, which leads to a larger difference in errors between the two sets of structures.

Examining the model generalizability to lattices outside of the training set

Each model was retrained on a dataset that contained only five out of six lattice structures and then tested on a dataset containing only the lattice excluded from the training set. The  $la\bar{3}$  lattice was chosen as the test set in this investigation because it displays one of the largest differences in the bandgap minimum and maximum values of all the lattices (4.42 eV) with a standard deviation of 0.98 eV (for the combined set of values in the training and test set). The model performance for this re-partitioned training set (2384 structures encompassing five lattice symmetries) and the test set (616 structures of the  $la\bar{3}$  symmetry) results in significantly larger MAE values for the formation energy of 53 meV/cation, 40 meV/cation, and 132 meV/cation in the 4-gram + KRR, c/BOP + LGBM, and SOAP + NN models, respectively (Table 3). A large increase in the bandgap energies is also observed for 4-gram + KRR (MAE = 179 meV), c/BOP + LGBM (MAE = 180 meV), and SOAP + NN (MAE = 527 meV), respectively. The significant increase in the errors compared with the original dataset is attributed to the absence of common local atomic-environment descriptors between the training and test sets.

To examine if an improved generalizability of each model can be obtained by training a model for each lattice type separately, the c/BOP + LGBM model is retrained by performing the feature selection and hyperparameter optimization procedure for each spacegroup separately and then tested on the left-out  $la\bar{3}$  lattice.

This procedure results in a test-set MAE score of 36 meV/cation (111 meV) for the formation (bandgap) energy, which is improved compared with the MAE of 40 meV/cation (180 meV) when training the model to the entire training set.

To give an indication of the prediction quality of these three ML models for the left-out lattice, a CE model was trained by using a random training/test 75%/25% split of the 616 structures with the  $la\bar{3}$  lattice symmetry using the software CELL [<https://sol.physik.hu-berlin.de/cell/>]. For a CE model that includes two-point clusters up to six angstroms, a test-set MAE of 23 meV/cation for the formation energy is obtained. A saturation in the learning curve with a training-set size of only 50 samples for the CE approach (Fig. S5), indicates that this approach is largely incapable of achieving a higher accuracy with more data. In comparison with the CE test-set accuracy, the 4-gram + KRR and c/BOP + LGBM have about twice the error when the  $la\bar{3}$  lattice is completely left out of the training set. For the bandgap energy, the 4-gram + KRR and c/BOP + LGBM models are much more accurate compared with what is achieved with CE (229 meV). These results indicate that the simple CE representation still provides a reasonable accuracy when examining only one lattice; however, the disadvantage of the CE approach is that the model is not transferable, and therefore, a new model must be trained for each symmetry.

## DISCUSSION

We have presented the three top-performing machine-learning models for the prediction of two key properties of transparent conducting oxides during a public crowd-sourced data-analytics competition organized by NOMAD and hosted by the online platform Kaggle. One key outcome of this competition was the development of a new representation for materials science based on the  $n$ -gram model. Because of the diverse set of methods and regression techniques, the interplay between the combination of the representation and regression methods was also analyzed. In particular, consistently large differences between the mean absolute errors and the 95 percentile distributions of the training and test-set errors are consistently observed when a neural network and light gradient-boosting machine is used at the regression models, which indicates a higher potential for overfitting for these methods. The Pearson correlation was used to investigate correlations between the estimates of the test-set values among the various ML models to give additional insight into the model performance. By using this analysis, the largest Pearson correlations were observed for predictions from the same representations combined with different regressors for the formation energy. In particular, the highest predictions were observed for the same representations by using neural network and kernel ridge regression (Gaussian process regression). The Pearson correlation allows for an identification of models with

weakly correlated predictions to obtain even lower errors through ensembling.

## METHODS

The two target properties of the NOMAD 2018 Kaggle competition were the formation energy and bandgap energy. The bandgap energy was taken to be the minimum value (either direct or indirect). The formation energy was calculated relative to pure  $\text{In}_2\text{O}_3$ ,  $\text{Al}_2\text{O}_3$ , and  $\text{Ga}_2\text{O}_3$  phases and were normalized per number of cations according to

$$E_f = E[(\text{Al}_x\text{Ga}_y\text{In}_z)_2\text{O}_3] - xE[\text{Al}_2\text{O}_3] - yE[\text{Ga}_2\text{O}_3] - zE[\text{In}_2\text{O}_3]$$

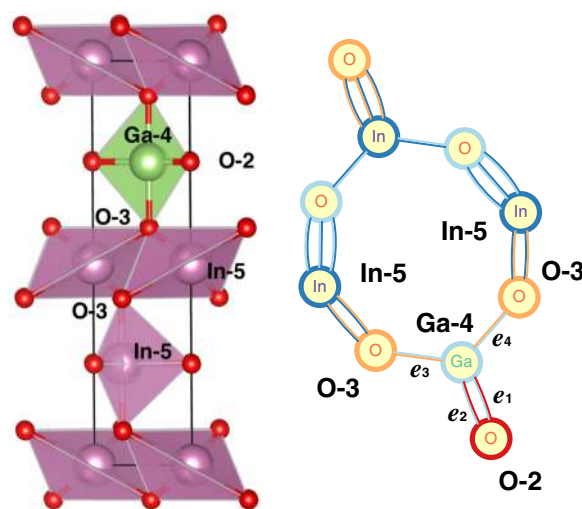
where,  $x$ ,  $y$ , and  $z$  are the corresponding relative concentrations of Al, Ga, and In, respectively defined as  $x = \frac{N_{\text{Al}}}{N_{\text{Al}}+N_{\text{Ga}}+N_{\text{In}}}$ ,  $y = \frac{N_{\text{Ga}}}{N_{\text{Al}}+N_{\text{Ga}}+N_{\text{In}}}$ ,  $z = \frac{N_{\text{In}}}{N_{\text{Al}}+N_{\text{Ga}}+N_{\text{In}}}$ .

$E[(\text{Al}_x\text{Ga}_y\text{In}_z)_2\text{O}_3]$  is the energy of the mixed system,  $E[\text{Al}_2\text{O}_3]$ ,  $E[\text{Ga}_2\text{O}_3]$ , and  $E[\text{In}_2\text{O}_3]$  are the energies of the pure binary crystalline systems in their thermodynamic ground-state lattices  $\alpha\text{-Al}_2\text{O}_3$  ( $R\bar{3}c$ ),  $\beta\text{-Ga}_2\text{O}_3$  ( $C2/m$ ), and  $c\text{-In}_2\text{O}_3$  ( $la\bar{3}$ ), respectively. Thus, this relative formation energy provides an estimate of the stability of the mixed system with respect to the stable ground state of the binary components and differs from the usual definition that instead uses the atomic energies for reference values. By using the bulk energies of the pure binary components as reference values (instead of the atomic energy references that are typically used), the relative formation energy is a more difficult property to learn.

The formation energy and bandgap energy were computed by using the PBE exchange-correlation DFT functional with the all-electron electronic structure code FHI-aims with tight settings.<sup>80</sup> In a separate study, we have carefully compared the performance of different exchange-correlation functionals (e.g., PBE, PBEsol, and HSE06) for group-III oxides and observed that the PBE functional provides a qualitatively correct prediction of the minimum energy structures for  $\text{Al}_2\text{O}_3$ ,  $\text{Ga}_2\text{O}_3$ , and  $\text{In}_2\text{O}_3$  as  $R\bar{3}c$ ,  $C2/m$ , and  $la\bar{3}$ , respectively out of five polymorphs included in this dataset (Fig. S6). In addition, a comparison of the computed and experimental lattice parameters for seven experimentally reported  $\text{Al}_2\text{O}_3$ ,  $\text{Ga}_2\text{O}_3$ , and  $\text{In}_2\text{O}_3$  structures shows that the PBE-computed volumes overestimate the experimental values by an average error of 4.2% (see Fig. S7). In terms of the bandgap energies computed by using different exchange-correlation functionals, we note that the differences between the PBE and correct bandgaps can indeed be significant. In the absence of reliable experimental values, we have instead compared PBE with HSE06 and found that the PBE bandgap energies systematically underestimate the HSE06 values by 1.5–2 eV. We emphasize that the differences between PBE and HSE06 (or even *GW* bandgaps) display regularities. For example, a largely systematic shift was reported for a dataset of 250 epasolites<sup>81</sup> and for a broad class of inorganic systems.<sup>6</sup> Therefore, PBE values are still useful to screen materials because they provide a first indication of the bandgap energies (but not much more). Clearly, a few of the possibly interesting materials will then need to be studied with more accurate methods (e.g., *GW* approach). Furthermore, it is possible to use machine learning for describing the differences (see, e.g., ref.<sup>81</sup>), but this is beyond the scope of the present study.

## *n*-gram model

The first step to construct an *n*-gram representation of a crystalline system is to map the real-space 3-dimensional (3D) periodic structure onto a crystal graph (see Fig. 4 for an illustration), where the nodes represent atoms in a given lattice position and the edges represent bonds. In the present implementation, a node is labeled by the atomic species and the number of bonded neighbors (e.g., Ga-4 indicates indium in a four-coordinate environment for gallium). For this material class, the cations (i.e., Al, Ga, and In) are only bonded to oxygen atoms and vice versa; therefore, an edge between nodes in the crystalline graph occurs if the cation-oxygen distance in the 3D crystal is less than a lattice-specific and linearly scaled (discussed in more detail below) threshold that is based on the sum of the ionic radii of the two species. The empirically tabulated Shannon ionic radii were used in this study, which are oxidation-state and coordination-environment specific values, and thus, have to be defined beforehand.<sup>82</sup> The two-coordinate  $\text{O}^{2-}$  radius (1.35 Å) and the six-coordinate radii for  $\text{Al}^{3+}$  (0.535 Å),  $\text{Ga}^{3+}$  (0.62 Å), and  $\text{In}^{3+}$  (0.8 Å) were used. Because the lattice structures in the dataset have coordination environments that vary from these values, lattice-specific scaling factors of the radii of 1.3 [ $P6_3/mmc$ ], 1.4 [ $C2/m$ ,  $Pna2_1$ , and  $Fd\bar{3}m$  (for the subset of lattices within this spacegroup with  $\gamma < 60$ )], and 1.5 [ $R\bar{3}c$ ,  $la\bar{3}$ , and  $Fd\bar{3}m$



**Fig. 4** Unit cell of the  $\text{In}_3\text{Ga}_1\text{O}_6$  structure (left) depicted as a crystal-graph representation (right), which shows the connections between each node that are defined by the chemical bonds

(for the subset of lattices within this spacegroup with  $\gamma > 60$ ) were used. The optimal scaling factors used in the generation of the crystal graph were found for each lattice symmetry through trial and error. We note that previously a similar crystal-graph representation of solid-state lattices was already introduced,<sup>83</sup> but the graph itself was used as a descriptor and combined with a convolutional neural network to learn several properties of materials. In our case, the information contained in the crystal graph is transformed into features through histograms, as explained below.

The 1-gram features are generated from counting the unique coordination environments in the graph (i.e., one four-coordinate gallium [Ga-4], three five-coordinate indium atoms [In-5], one two-coordinate oxygen [O-2], three three-coordinate oxygen atoms [O-3], and two four-coordinate oxygen atoms [O-4]). The higher-order *n*-grams are the contiguous sequences of the cation-oxygen-directed edges in the crystal graph varying from 1 (2-gram), 2 (3-gram), and 3 (4-gram). We note that a directed-graph notation is used here, where parallel edges indicate equivalent sites because of the symmetry present in crystalline systems.

To illustrate how the higher-order *n*-gram features are generated, we focus on only the 2-grams, 3-grams, and 4-grams associated with a single node Ga-4 in the example presented in Fig. 4. The 2-gram set is determined by counting the edges of the Ga-4 node: Ga-4 to O-2 (labeled as e1), Ga-4 to O-2 (e2), Ga-4 to O-3 (e3), and Ga-4 to O-3 (e4), resulting in a histogram of two Ga-4/O-2 and two Ga-4/O-3 2-grams. The 3-gram set for the Ga-4 node is obtained from the six unique combinations of edges [(e1, e2), (e2, e3), (e2, e4), (e1, e4), (e1, e3), and (e3, e4)], giving histograms of the following 3-grams: O-2/Ga-4/O-2, O-2/Ga-4/O-3, and O-3/Ga-4/O-3. The 4-gram can be built by combining three contiguous edges. For example, the edge labeled e4 (Ga-4/O-3) can be combined with e3 (O-3/Ga-4) on the left, and on the right, an edge connecting O-3 to In-5 to create the 4-gram: O-3/Ga-4/O-3/In-5. An additional 4-gram (O-2/Ga-4/O-3/In-5) can be built from e4 by combining either e1 or e2 (i.e., the left edge O-2/Ga-4) and the right edge O-3/In-5.

For each structure in the dataset, this procedure is applied to generate histograms of all 1-grams and 3-grams for every node and all 2-grams and 4-grams for every edge in the crystal graph. The resulting histograms are then normalized by the unit cell volume to account for the varying unit cell sizes in the dataset.

For the NOMAD 2018 Kaggle dataset, a total of 13 unique 1-grams were used that range from 4 to 6 and unique oxygen coordination numbers that range from 2 to 5. To illustrate the histogram features generated from the *n*-gram model by using the 1-gram for two 80-atom structures with the formula  $(\text{Al}_{0.25}\text{Ga}_{0.28}\text{In}_{0.47})_2\text{O}_3$  and  $(\text{Al}_{0.63}\text{Ga}_{0.34}\text{In}_{0.03})_2\text{O}_3$  and  $C2/m$  and  $P6_3/mmc$  symmetry types are shown in Fig. 5.

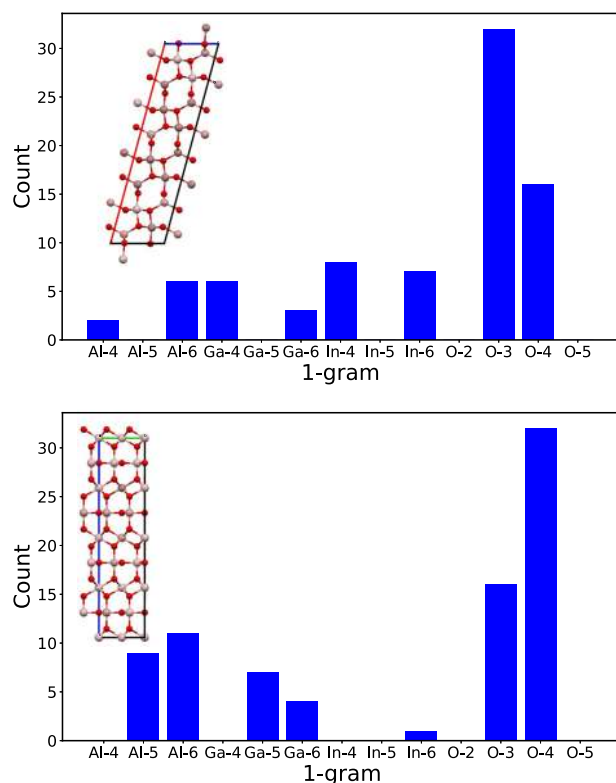
The *n*-gram features are combined with a KRR model by using the Gaussian radial basis function kernel. The values of the two hyperparameters (the regularization constant  $\lambda$  and the length scale of the Gaussian  $\gamma$ , which controls the degree of correlation between training points) were determined by performing grid searches with fivefold CV and compare well with the private leaderboard score (Table 4). Similar to what was

discussed in the context of ensembling different models with low correlation, here too the highest accuracies are obtained from an ensemble score of the 3-gram and 4-gram predictions:  $P_{\text{mix}} = \alpha_{\text{mix}}P(3\text{-gram}) + (1 - \alpha_{\text{mix}})P(4\text{-gram})$ , where a mixing parameter of 0.64 and 0.69 for the formation and bandgap energies was used, respectively. Although such an ensemble gives the lowest CV-score RMSLE, the concatenated list of 1-gram, 2-gram, 3-gram, and 4-gram features was used throughout this paper to facilitate a comparison between each of the different regression methods. This is a convenient choice to avoid having to retrain the mixing parameter for each analysis.

The  $n$ -gram features were combined with a NN architecture consisting of 10 dense layers with 100, 50, 50, 20, 20, 20, 10, 10, and 10 neurons, respectively, and LeakyReLU activation functions. The NN was implemented in PyTorch<sup>82</sup> and optimized by using Adam<sup>84</sup> with a learning rate of 0.005. The  $n$ -gram features were combined with LGBM with the model hyperparameter optimization performed as described in the “Atomic and bond-order-potential derived features” section of the Methods.

### Atomic and bond-order-potential derived features

For the second-place model, many descriptor candidates are examined from a set of compositional, atomic-environment-based, and average



**Fig. 5** Histogram of the complete set of 13 1-gram features formed from the total list of the unique coordination environment for each atom type for two training-set structures  $(\text{Al}_{0.25}\text{Ga}_{0.28}\text{In}_{0.47})_2\text{O}_3$  and  $(\text{Al}_{0.63}\text{Ga}_{0.34}\text{In}_{0.03})_2\text{O}_3$

structural properties (Fig. 6). Of this list, the optimal 175 (212) features are selected for the prediction of the bandgap (formation) energy based on an iterative procedure by using the auxiliary gradient-boosting regression tree (XGBoost) and used with the LGBM learning algorithm.

The weighted chemical properties are computed from reference data by using either the overall stoichiometry or the nearest neighbors. This approach is motivated by the concepts of structure maps that chart the structural stability of compounds in terms of chemical properties of the constituent atoms and the overall chemical composition.<sup>85–87</sup> For generating per-structure features, the weighted arithmetic mean of bandgap and of formation energy is computed from the stoichiometry by using the respective values for  $\text{In}_2\text{O}_3$  ( $R\bar{3}c$ ,  $1a\bar{3}$ , and  $Pnma$ ),  $\text{Ga}_2\text{O}_3$  ( $C/2m$  and  $R\bar{3}c$ ), and  $\text{Al}_2\text{O}_3$  ( $C/2m$ ,  $Pna2_1$ ,  $R\bar{3}c$ , and  $P4_232$ ) from the Materials Project.<sup>88</sup> In addition, the average and difference of several free-atom properties such as the electronic affinity, ionization potential, atomic volume, and covalent radius (all values were obtained from mendeleev [<https://bitbucket.org/lukaszmentel/mendeleev>]) are computed between each atom and each of its nearest neighbors to generate per-atom features. The list of nearest atomic neighbors is generated by using the ASE package<sup>89</sup> and determined based on the distance between two atoms being smaller than the sum of the computed free-atom radii.

The representations of the atomic environment are incorporated by using BOP-based properties and simple geometric measures. The latter comprise averaged atomic bond distances, averaged cation-oxygen nearest-neighbor bond distances, centrosymmetric parameters (determined from a sum of the vectors formed between atom  $i$  and its nearest neighbors), and the volume per atom. The characterization of atomic environments by the BOP methodology relies on moments and the closely related recursion coefficients that connect the local atomic environment and local electronic structure (DOS) by the moments theorem.<sup>90</sup> Within the analytic BOP formalism, these properties can be computed efficiently in an approximate way<sup>71,73</sup> and used as per-atom features that discriminate and classify the local atomic environment.<sup>74,91</sup> For each atom  $j$ , the  $n$ th moment  $\mu_n^{(j)}$  is computed by multiplying pairwise model Hamiltonians along self-returning paths (i.e., start and end at the same atom) up to length  $n$ . The representation of local atomic environments uses scaled recursion coefficients  $a_i^{(j)}$  and  $b_i^{(j)}$  obtained from  $\mu_n^{(j)}$  with scaled volumes  $v_j$  as described in refs.<sup>74,91</sup> In this work, a total of 12 moments corresponding to the atomic environment up to the sixth nearest-neighbor shell were used. This procedure is to some degree comparable to the  $n$ -gram approach of the first-place solution with regard to sampling the environment. More specifically, a 4-gram feature would correspond to one-half of a self-returning path in an eighth-moment calculation. One of the differences in the two methodologies is that all path segments are used explicitly in the  $n$ -gram approach, whereas only the individual self-returning paths are subsumed in the moments of the  $c$ /BOP approach.

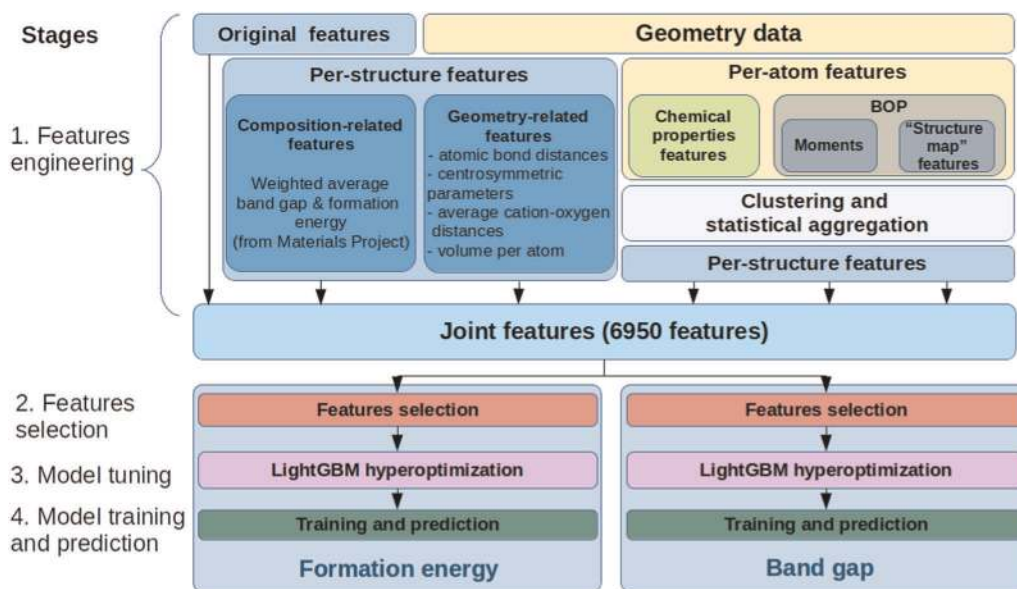
For each atom in the structure, this procedure generates a list with a length that is equivalent to the number of neighbors. A clustering scheme is then applied to the average, and standard deviation of these features is used to generate a fixed-length representation. These properties were clustered into seven groups based on the atomic environments described by  $a_1^{(j)}$ ,  $b_2^{(j)}$ , and  $v_j$  by using the k-means clustering algorithm applied separately to O and Al, Ga, and In for each structure in the dataset. These clusters of varying lengths were then projected into a fixed-length vector by taking only the mean and standard deviation. If one of the seven atomic-environment groups is not present in a given structure, then the corresponding feature is set to zero.

In total, this approach resulted in a set of 6,950 features (ca. 120 atomic properties per atom  $\times$  7 atomic-environment groups  $\times$  4 element types  $\times$  2 statistical aggregation measures), which were reduced to a set of 175

**Table 4.** Fivefold CV RMSLE values of the formation energies and bandgap energies and public and private leaderboards for  $n$ -grams of various lengths (normalized by unit cell volume)

$n$ -gram lengths	Formation energy RMSLE	Bandgap RMSLE	Public RMSLE	Private RMSLE
1-gram	0.0229	0.0817	0.0518	0.0560
2-gram	0.0230	0.0811	0.0472	0.0540
3-gram	0.0223	0.0814	0.0381	0.0514
Ensemble of 3- and 4-gram	0.0222	0.0806	0.0514	0.0510
4-gram	0.0237	0.0829	0.0394	0.0506





**Fig. 6** Illustration of feature engineering and subsequent stages for the construction of the second-place *c*/BOP + LGBM model

and 212 features for the prediction of the bandgap and formation energies that produced the highest accuracy based on an iterative procedure by using XGBoost.<sup>92</sup> The final set of features were then combined with LGBM<sup>75</sup> for the final model with the hyperparameters tuned by using 10-fold CV within the hyperopt package<sup>93</sup> and a suggestion algorithm by using tree-structured Parzen estimators,<sup>94</sup> which resulted in RMSLE values of 0.0462 and 0.0521 for the public and private leaderboards.

The selection of the optimal set of features requires attributing an importance to each of the 6,950 features. However, recently, popular feature attribution methods were shown to have a lower assigned importance relative to the true impact of that feature in modeling the target property.<sup>95</sup> The SHapley Additive exPlanations (SHAP) method<sup>96</sup> was proposed to give more accurate relative feature importances and was calculated here as a normalized mean absolute value of the SHAP values for each feature (see Fig. S8). For prediction of the bandgap energy, the features with the largest relative importance (ca. 17% each) are the weighted bandgap of  $\text{Al}_2\text{O}_3$ ,  $\text{Ga}_2\text{O}_3$ , and  $\text{In}_2\text{O}_3$  and the volume per atom. In contrast, all features have a relatively small importance for the prediction of the formation energy; only geometrical information describing the environment of indium and the length centrosymmetric parameter has the highest importance. The per-atom features have a total relative importance of 40 and 33% for formation energy and bandgap energy, respectively, including ca. 20 and 15% of the relative feature importance for the BOP-related features.

The same set of top features used with LGBM to achieve the second-place score were also combined with KRR and NN. The features used with the KRR and NN regressors were rescaled to have a zero mean and unit variance. The KRR model employed a Gaussian radial basis function kernel with the  $\lambda$  and  $\gamma$  hyperparameters tuned by using a 5-fold CV grid search. The Keras package [<https://keras.io>] with the Tensorflow backend<sup>97</sup> was used to generate a three-layer NN containing 1,024, 256, and 256 neurons with batch normalization, hyperbolic tangent activation function, and 20% dropout in each layer. The output layer contained one neuron only, had no batch normalization, and used an ReLU activation function.<sup>98</sup> The NN was trained for 500 epochs.

#### Smooth overlap of atomic positions

The third-place solution used the smooth overlap of atomic positions (SOAP) kernel developed by Bartók et al. that incorporates information on the local atomic environment through a rotationally integrated overlap of Gaussian densities of the neighboring atoms.<sup>14,15</sup> The SOAP kernel describes the local environment for a given atom through the sum of Gaussians centered on each of the atomic neighbors within a specific cutoff radius. The SOAP vector was computed with the QUIPPY package

[<https://libatoms.github.io/QUIP/index.html>] by using a real-space radial cutoff  $f_{\text{cut}}$  of 10 Å, the smoothing parameter  $\sigma_{\text{atom}} = 0.5$  Å, and the basis set expansion values of  $l = 4$  and  $n = 4$ . For each structure, a single-feature vector was used by averaging the per-atom SOAP vector for each atom in the unit cell, which resulted in a vector with a length of 681 values. These aggregated mean feature vectors for the dataset were then scaled so that each dimension has a mean equal to zero and variance equal to one.

The average SOAP features were used in a three-layer feed-forward NN by using PyTorch<sup>99</sup> [<https://pytorch.org/>] with batch normalization and 20% dropout in each layer. For predicting the bandgap energies and the formation energies, the initial layer had 1024 and 512 neurons, respectively. In both cases, the remaining two layers had 256 neurons each. The neural networks were trained for 200 or 250 epochs, and the final predictions were based on 200 independently trained NNs by using the same architecture but with different initial weights.

The average SOAP vector of each structure was combined with Gaussian Process Regression (GPR),<sup>19</sup> where the covariance function between two structures was defined as a polynomial kernel

$$k(\mathbf{R}_i, \mathbf{R}_j) = (a\mathbf{R}_i \cdot \mathbf{R}_j + b)^c$$

where  $\mathbf{R}_i$  and  $\mathbf{R}_j$  are descriptor vectors for structure  $i$  and  $j$ ;  $a$ ,  $b$ , and  $c$  are kernel coefficients. Several values for the Polynomial kernel degree  $c$  (ranging from 1 to 6) with  $a = 1.0$  and  $b = 0.0$  were examined until the lowest RMSLE was obtained. This resulted in two hyperparameters for the model construction: regularization term and the degree of the kernel. Optimal hyperparameters were identified by using repeated random subsampling CV for 100 training and validation splits. Finally, the final GPR model was averaged over all 100 splits. The SOAP vector was also combined with LGBM regression with the model hyperparameter optimization performed as described in the section “Atomic and bond-order-potential derived features” of the Methods.

#### DATA AND CODE AVAILABILITY

The dataset used in the Kaggle competition are publicly available in the NOMAD Repository (<http://dx.doi.org/10.17172/NOMAD/2019.06.14-1>), the Kaggle competition website (<https://www.kaggle.com/c/nomad2018-predict-transparent-conductors>), and the labeled training and test set can be found on github ([https://github.com/csutton7/nomad\\_2018\\_kaggle\\_dataset](https://github.com/csutton7/nomad_2018_kaggle_dataset)). The three winning models are available at <https://analytics-toolkit.nomad-coe.eu>.

Received: 30 November 2018; Accepted: 12 September 2019;  
Published online: 18 November 2019

## REFERENCES

1. Meredig, B. et al. Combinatorial screening for new materials in unconstrained composition space with machine learning. *Phys. Rev. B* **89**, 094104 (2014).
2. Isayev, O. et al. Materials cartography: representing and mining materials space using structural and electronic fingerprints. *Chem. Mater.* **27**, 735–743 (2015).
3. Olynyk, A. O. et al. High-throughput machine-learning-driven synthesis of full-heusler compounds. *Chem. Mater.* **28**, 7324–7331 (2016).
4. Schmidt, J. et al. Predicting the thermodynamic stability of solids combining density functional theory and machine learning. *Chem. Mater.* **29**, 5090–5103 (2017).
5. Pilania, G. et al. Machine learning bandgaps of double perovskites. *Sci. Rep.* **6**, 19375 (2016).
6. Lee, J., Seko, A., Shitara, K., Nakayama, K. & Tanaka, I. Prediction model of band gap for inorganic compounds by combination of density functional theory calculations and machine learning techniques. *Phys. Rev. B* **93**, 115104 (2016).
7. Goldsmith, B. R., Esterhuizen, J., Liu, J.-X., Bartel, C. J. & Sutton, C. Machine learning for heterogeneous catalyst design and discovery. *AIChE J.* **64**, 2311–2323 (2018).
8. Rupp, M., Tkatchenko, A., Müller, K.-R. & von Lilienfeld, O. A. Fast and accurate modeling of molecular atomization energies with machine learning. *Phys. Rev. Lett.* **108**, 058301 (2012).
9. Hansen, K. et al. Assessment and validation of machine learning methods for predicting molecular atomization energies. *J. Chem. Theory Comput.* **9**, 3404–3419 (2013).
10. Hirn, M., Poilvert, N. & Mallat, S. Quantum energy regression using scattering transforms. Preprint at <https://arxiv.org/abs/1502.02077> (2015).
11. Ziletti, A., Kumar, D., Scheffler, M. & Ghiringhelli, L. M. Insightful classification of crystal structures using deep learning. *Nat. Commun.* **9**, 2775 (2018).
12. Hansen, K. et al. Machine learning predictions of molecular properties: accurate many-body potentials and nonlocality in chemical space. *J. Phys. Chem. Lett.* **6**, 2326–2331 (2015).
13. Huo, H. & Rupp, M. Unified representation for machine learning of molecules and crystals. Preprint at <https://arxiv.org/abs/1704.06439> (2017).
14. 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).
15. Bartók, A. P., Kondor, R. & Csányi, G. On representing chemical environments. *Phys. Rev. B* **87**, 184115 (2013).
16. Seko, A., Hayashi, H., Nakayama, K., Takahashi, A. & Tanaka, I. Representation of compounds for machine-learning prediction of physical properties. *Phys. Rev. B* **95**, 144110 (2017).
17. Schütt, K. T. et al. How to represent crystal structures for machine learning: towards fast prediction of electronic properties. *Phys. Rev. B* **89**, 205118 (2014).
18. Faber, F., Lindmaa, A., von Lilienfeld, O. A. & Armiento, R. Crystal structure representations for machine learning models of formation energies. *Int. J. Quantum Chem.* **115**, 1094–1101 (2015).
19. Rasmussen, C. E. & Williams, C. K. I. in *Adaptive Computation and Machine Learning* (The MIT Press, 2005).
20. Behler, J. Atom-centered symmetry functions for constructing high-dimensional neural network potentials. *J. Chem. Phys.* **134**, 074106 (2011).
21. Ghiringhelli, L. M., Vybiral, J., Levchenko, S. V., Draxl, C. & Scheffler, M. Big data of materials science: critical role of the descriptor. *Phys. Rev. Lett.* **114**, 105503 (2015).
22. Ouyang, R., Curtarolo, S., Ahmetcik, E., Scheffler, M. & Ghiringhelli, L. M. SISSO: A compressed-sensing method for identifying the best low-dimensional descriptor in an immensity of offered candidates. *Phys. Rev. Mater.* **2**, 083802. <https://doi.org/10.1103/PhysRevMaterials.2.083802> (2018).
23. Kim, C., Pilania, G. & Ramprasad, R. From organized high-throughput data to phenomenological theory using machine learning: the example of dielectric breakdown. *Chem. Mater.* **28**, 1304–1311 (2016).
24. Bartel, C. J. et al. Physical descriptor for the Gibbs energy of inorganic crystalline solids and temperature-dependent materials chemistry. *Nat. Commun.* **9**, 4168 (2018).
25. Bartel, C. J. et al. New tolerance factor to predict the stability of perovskite oxides and halides. *Sci. Adv.* **5**, eaav0693 (2019).
26. Nelson, L. J., Ozoliņš, V., Reese, C. S., Zhou, F. & Hart, G. L. W. Cluster expansion made easy with Bayesian compressive sensing. *Phys. Rev. B* **88**, 155105 (2013).
27. De Fontaine, D. in *Solid State Physics* Vol. Volume 47 (eds Ehrenreich Henry & Turnbull David) 33–176 (Academic Press, 1994).
28. Müller, S. Bulk and surface ordering phenomena in binary metal alloys. *J. Phys.: Condens. Matter* **15**, R1429 (2003).
29. Sanchez, J. M., Ducastelle, F. & Gratias, D. Generalized cluster description of multicomponent systems. *Phys. A: Stat. Mech. its Appl.* **128**, 334–350 (1984).
30. van de Walle, A., Asta, M. & Ceder, G. The alloy theoretic automated toolkit: a user guide. *Calphad* **26**, 539–553 (2002).
31. Zunger, A., Wang, L. G., Hart, G. L. W. & Sanati, M. Obtaining Ising-like expansions for binary alloys from first principles. *Model. Simul. Mater. Sci. Eng.* **10**, 685 (2002).
32. Laks, D. B., Ferreira, L. G., Froyen, S. & Zunger, A. Efficient cluster expansion for substitutional systems. *Phys. Rev. B* **46**, 12587–12605 (1992).
33. Ghiringhelli, L. M. et al. Towards efficient data exchange and sharing for big-data driven materials science: metadata and data formats. *npj Comput. Mater.* **3**, 46 (2017).
34. Kinoshita, A., Hirayama, H., Aino, M., Aoyagi, Y. & Hirata, A. Room-temperature operation at 333 nm of Al<sub>0.03</sub>Ga<sub>0.97</sub>N/Al<sub>0.25</sub>Ga<sub>0.75</sub>N quantum-well light-emitting diodes with Mg-doped superlattice layers. *Appl. Phys. Lett.* **77**, 175–177 (2000).
35. Ohta, H. et al. Current injection emission from a transparent p–n junction composed of p-SrCu<sub>2</sub>O<sub>2</sub>/n-ZnO. *Appl. Phys. Lett.* **77**, 475–477 (2000).
36. Tsukazaki, A. et al. Repeated temperature modulation epitaxy for p-type doping and light-emitting diode based on ZnO. *Nat. Mater.* **4**, 42–46 (2005).
37. Nakamura, S., Mukai, T. & Senoh, M. Candela-class high-brightness InGaN/AlGaN double-heterostructure blue-light-emitting diodes. *Appl. Phys. Lett.* **64**, 1687–1689 (1994).
38. Arulkumaran, S. et al. Improved dc characteristics of AlGaN/GaN high-electron-mobility transistors on AlN/sapphire templates. *Appl. Phys. Lett.* **81**, 1131–1133 (2002).
39. Kubovic, M. et al. Microwave performance evaluation of diamond surface channel FETs. *Diam. Relat. Mater.* **13**, 802–807 (2004).
40. Hoffman, R., Norris, B. J. & Wager, J. ZnO-based transparent thin-film transistors. *Appl. Phys. Lett.* **82**, 733–735 (2003).
41. Nishii, J. et al. High mobility thin film transistors with transparent ZnO channels. *Jpn. J. Appl. Phys.* **42**, L347 (2003).
42. Nomura, K. et al. Thin-film transistor fabricated in single-crystalline transparent oxide semiconductor. *Science* **300**, 1269–1272 (2003).
43. Nomura, K. et al. Room-temperature fabrication of transparent flexible thin-film transistors using amorphous oxide semiconductors. *Nature* **432**, 488–492 (2004).
44. Dehuff, N. et al. Transparent thin-film transistors with zinc indium oxide channel layer. *J. Appl. Phys.* **97**, 064505 (2005).
45. Irmischer, K. et al. On the nature and temperature dependence of the fundamental band gap of In<sub>2</sub>O<sub>3</sub>. *Phys. status solidi (a)* **211**, 54–58 (2014).
46. Oliver, B. Indium oxide—a transparent, wide-band gap semiconductor for (opto) electronic applications. *Semiconductor Sci. Technol.* **30**, 024001 (2015).
47. Christoph, J. et al. Experimental electronic structure of In<sub>2</sub>O<sub>3</sub> and Ga<sub>2</sub>O<sub>3</sub>. *New J. Phys.* **13**, 085014 (2011).
48. Preissler, N., Bierwagen, O., Ramu, A. T. & Speck, J. S. Electrical transport, electrothermal transport, and effective electron mass in single-crystalline In<sub>2</sub>O<sub>3</sub> films. *Phys. Rev. B* **88**, 085305 (2013).
49. Gupta, R. K., Ghosh, K., Mishra, S. R. & Kahol, P. K. High mobility, transparent, conducting Gd-doped In<sub>2</sub>O<sub>3</sub> thin films by pulsed laser deposition. *Thin Solid Films* **516**, 3204–3209 (2008).
50. Yang, F., Ma, J., Luan, C. & Kong, L. Structural and optical properties of Ga<sub>2(1-x)</sub>In<sub>2x</sub>O<sub>3</sub> films prepared on α-Al<sub>2</sub>O<sub>3</sub> (0001) by MOCVD. *Appl. Surf. Sci.* **255**, 4401–4404 (2009).
51. Kong, L., Ma, J., Luan, C. & Zhu, Z. Structural and optical properties of Ga<sub>2</sub>O<sub>3</sub>:In films deposited on MgO (100) substrates by MOCVD. *J. Solid State Chem.* **184**, 1946–1950 (2011).
52. Kong, L., Ma, J., Yang, F., Luan, C. & Zhu, Z. Preparation and characterization of Ga<sub>2-x</sub>In<sub>2(1-x)</sub>O<sub>3</sub> films deposited on ZrO<sub>2</sub> (100) substrates by MOCVD. *J. Alloy. Compd.* **499**, 75–79 (2010).
53. Oshima, T. & Fujita, S. Properties of Ga<sub>2</sub>O<sub>3</sub>-based (In<sub>x</sub>Ga<sub>1-x</sub>)<sub>2</sub>O<sub>3</sub> alloy thin films grown by molecular beam epitaxy. *Phys. Status Solidi (c)* **5**, 3113–3115 (2008).
54. Kokubun, Y., Abe, T. & Nakagomi, S. Sol–gel prepared (Ga<sub>1-x</sub>In<sub>x</sub>)<sub>2</sub>O<sub>3</sub> thin films for solar-blind ultraviolet photodetectors. *Phys. Status Solidi (a)* **207**, 1741–1745 (2010).
55. Zhang, F. et al. Wide bandgap engineering of (AlGa)<sub>2</sub>O<sub>3</sub> films. *Appl. Phys. Lett.* **105**, 162107 (2014).
56. Ito, H., Kaneko, K. & Fujita, S. Growth and band gap control of corundum-structured α-(AlGa)<sub>2</sub>O<sub>3</sub> thin films on sapphire by spray-assisted mist chemical vapor deposition. *Jpn. J. Appl. Phys.* **51**, 100207 (2012).
57. Kondo, S., Tateishi, K. & Ishizawa, N. Structural evolution of corundum at high temperatures. *Jpn. J. Appl. Phys.* **47**, 616 (2008).
58. Favaro, L. et al. Experimental and ab initio infrared study of χ<sub>r</sub>, κ- and α- aluminas formed from gibbsite. *J. Solid State Chem.* **183**, 901–908 (2010).
59. Halvarsson, M., Langer, V. & Vuorinen, S. Determination of the thermal expansion of κ-Al<sub>2</sub>O<sub>3</sub> by high temperature XRD. *Surf. Coat. Technol.* **76**, 358–362 (1995).
60. Hussain, E. & Repelin, Y. Structural studies of transition aluminas. Theta alumina. *Eur. J. Solid State Inorg. Chem.* **33**, 1223–1231 (1996).
61. He, H. et al. First-principles study of the structural, electronic, and optical properties of Ga<sub>2</sub>O<sub>3</sub> in its monoclinic and hexagonal phases. *Phys. Rev. B* **74**, 195123 (2006).

62. Lorenz, H. et al. Methanol steam reforming: CO<sub>2</sub>-selective Pd<sub>2</sub>Ga phases supported on  $\alpha$ - and  $\gamma$ -Ga<sub>2</sub>O<sub>3</sub>. *Appl. Catal. A: Gen.* **453**, 34–44 (2013).
63. Playford, H. Y., Hannon, A. C., Barney, E. R. & Walton, R. I. Structures of uncharacterised polymorphs of gallium oxide from total neutron diffraction. *Chem. – A Eur. J.* **19**, 2803–2813 (2013).
64. Goldschmidt, V. M., Barth, T. F. W. & Lunde, G. *Geochemische Verteilungsgesetze der Elemente Skr. Norske*, Vol. I–VII. (Ved. Akad., 1925).
65. Liu, D. et al. Large-scale synthesis of hexagonal corundum-type In<sub>2</sub>O<sub>3</sub> by ball milling with enhanced lithium storage capabilities. *J. Mater. Chem. A* **1**, 5274–5278 (2013).
66. Garcia-Domene, B. et al. High-pressure lattice dynamical study of bulk and nanocrystalline In<sub>2</sub>O<sub>3</sub>. *J. Appl. Phys.* **112**, 123511 (2012).
67. Shannon, R. D. & Prewitt, C. T. Synthesis and structure of phases in the In<sub>2</sub>O<sub>3</sub>-Ga<sub>2</sub>O<sub>3</sub> system. *J. Inorg. Nucl. Chem.* **30**, 1389–1398 (1968).
68. Vegard, L. Die Konstitution der Mischkristalle und die Raumfüllung der Atome. *Physik* **5**, 17–26 (1921).
69. Denton, A. R. & Ashcroft, N. W. Vegards Law. *Phys. Rev. A* **43**, 3161–3164 (1991).
70. Broder, A. Z., Glassman, S. C., Manasse, M. S. & Zweig, G. Syntactic clustering of the Web. *Computer Netw. ISDN Syst.* **29**, 1157–1166 (1997).
71. Hammerschmidt, T. et al. BOPfox program for tight-binding and analytic bond-order potential calculations. *Comp. Phys. Comm.* **235**, 221–233 (2019).
72. Drautz, R. & Pettifor, D. G. Valence-dependent analytic bond-order potential for transition metals. *Phys. Rev. B* **74**, 174117 (2006).
73. Drautz, R., Hammerschmidt, T., Čák, M. & Pettifor, D. G. Bond-order potentials: derivation and parameterization for refractory elements. *Model. Simul. Mater. Sci. Eng.* **23**, 074004 (2015).
74. Jenke, J. et al. Electronic structure based descriptor for characterizing local atomic environments. *Phys. Rev. B* **98**, 144102 (2018).
75. Ke G. et al. LightGBM: a highly efficient gradient boosting decision tree. *Adv. Neural Inf. Process Syst.* 3146–3154 (2017).
76. Mason, L., Baxter, J., Bartlett, P. L. & Frean, M. R. Boosting algorithms as gradient descent. *Adv. Neural Inf. Process Syst.* 512–518 (2000).
77. Breiman, L. Stacked regressions. *Mach. Learn.* **24**, 49–64 (1996).
78. Shapire, R. The strength of weak learnability. *Mach. Learn.* **5**, 197–227 (1990).
79. Opitz, D. & Maclin, R. Popular ensemble methods: an empirical study. *J. Artif. Intell. Res.* **11**, 169–198 (1999).
80. Blum, V. et al. Ab initio molecular simulations with numeric atom-centered orbitals. *Computer Phys. Commun.* **180**, 2175–2196 (2009).
81. Pilania, G., Gubernatis, J. E. & Lookman, T. Multi-fidelity machine learning models for accurate bandgap predictions of solids. *Computational Mater. Sci.* **129**, 156–163 (2017).
82. Shannon, R. Revised effective ionic radii and systematic studies of interatomic distances in halides and chalcogenides. *Acta Crystallographica Section A* **32**, 751 (1976).
83. Xie, T. & Grossman, J. C. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Phys. Rev. Lett.* **120**, 145301 (2018).
84. Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. Pre-print at <https://arxiv.org/abs/1412.6980> (2014).
85. Pettifor, D. G. The structures of binary compounds. I. *Phenomenological Struct. maps. J. Phys. C: Solid State Phys.* **19**, 285 (1986).
86. Seiser, B., Drautz, R. & Pettifor, D. G. TCP phase predictions in Ni-based superalloys: structure maps revisited. *Acta Materialia* **59**, 749–763 (2011).
87. Bialon, A. F., Hammerschmidt, T. & Drautz, R. Three-parameter crystal-structure prediction for sp-d-valent compounds. *Chem. Mater.* **28**, 2550–2556 (2016).
88. Jain, A. et al. Commentary: the materials project: a materials genome approach to accelerating materials innovation. *APL Mater.* **1**, 011002 (2013).
89. Bahn, S. R. & Jacobsen, K. W. An object-oriented scripting interface to a legacy electronic structure code. *Comput. Sci. Eng.* **4**, 56–66 (2002).
90. Cryot-Lackmann, F. On the electronic structure of liquid transition metals. *Adv. Phys.* **16**, 393–400 (1967).
91. Hammerschmidt, T., Ladines, A. N., Koßmann, J. & Drautz, R. Crystal-structure analysis with moments of the density-of-states: application to intermetallic topologically close-packed phases. *Crystals* **6**, 18 (2016).
92. Chen, T. & Guestrri, C. Xgboost: A scalable tree boosting system *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 785–794 (2016).
93. Bergstra, J., Yamins, D. & Cox, D. D. in *Proc. of the 30th International Conference on Machine Learning (ICML 2013)*.
94. Bergstra, J. S., Bardenet, R., Bengio, Y. & Kégl, B. Algorithms for hyper-parameter optimization. *Adv. Neural Inf. Process Syst.* 2546–2554 (2011).
95. Lundberg, S. M. & Lee, S. I. in *Advances in Neural Information Processing Systems* (eds I. Guyon et al.) 4765–4774 (Long Beach, CA, USA, 2017).
96. Lundberg, S. M., Erion, G. G. & Lee, S. I. Consistent individualized feature attribution for tree ensembles. Preprint at <http://arxiv.org/abs/1802.03888> (2018).
97. Abadi, M. et al. Tensorflow: a system for large-scale machine learning. *OSDI* **16**, 265–283 (2016).
98. Vinod, N. & Hinton, G. E. in *Proceedings of the 27th international conference on machine learning* 807–814 (OmniPress, Haifa, Israel, 2010).
99. Paszke, A. et al. Automatic differentiation in PyTorch. NIPS-W (2017).

## ACKNOWLEDGEMENTS

The idea of organizing a competition was given to us by Bernhard Schölkopf and Samuel Kaski, who are on the scientific advisory committee of NOMAD. We also gratefully acknowledge the help of Will Cukierski from Kaggle for assistance in launching the competition. The NOMAD 2018 Kaggle competition award committee members are Claudia Draxl, Daan Frenkel, Kristian Thygesen, Samuel Kaski, and Bernhard Schölkopf. In addition, we thank Gabor Csanyi, Matthias Rupp, Mario Boley, Christopher Bartel, and Marcel Langer for the helpful discussions and providing feedback on this paper. The project received funding from the European Union's Horizon 2020 research and innovation program (grant agreement no. 676580) and the Molecular Simulations from First Principles (MS1P). C.S. gratefully acknowledges funding by the Alexander von Humboldt Foundation.

## AUTHOR CONTRIBUTIONS

C.S. performed all calculations. C.S., L.M.G., A.Z. and M.S. designed the studies. T.Y., Y.L./T.H. and L.B./J.R.G. obtained the highest scores of the competition. C.S., T.Y., Y.L./T.H., L.B./J.R.G. and X.L. trained the additional models discussed herein. C.S., T.Y., Y.L., T.H. and L.B. wrote the paper. C.S., L.M.G. and M.S. supervised the project. All the authors discussed the results and implications and edited the paper.

## COMPETING INTERESTS

The authors declare no competing interests.

## ADDITIONAL INFORMATION

**Supplementary information** is available for this paper at <https://doi.org/10.1038/s41524-019-0239-3>.

**Correspondence** and requests for materials should be addressed to C.S. or L.M.G.

**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.



**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) 2019