Training sets based on uncertainty estimates in the cluster-expansion method

IOP Publishing J. Phys. Energy 3 (2021) 034012 https://doi.org/10.1088/2515-7655/abf9ef

Journal of Physics: Energy

OPEN ACCESS

RECEIVED
25 January 2021

REVISED
16 April 2021

ACCEPTED FOR PUBLICATION
20 April 2021

PUBLISHED
19 May 2021

PAPER

Training sets based on uncertainty estimates in the cluster-expansion method

David Kleiven[1,*], **Jaakko Akola**[1,2], **Andrew A Peterson**[3,4], **Tejs Vegge**[4] **and Jin Hyun Chang**[4,*]

[1] Department of Physics, Norwegian University of Science and Technology, NO-7491 Trondheim, Norway
[2] Computational Physics Laboratory, Tampere University, P.O. Box 692, FI-33014 Tampere, Finland
[3] School of Engineering, Brown University, Providence, RI 02912, United States of America
[4] Department of Energy Conversion and Storage, Technical University of Denmark, DK-2800 Kgs. Lyngby, Denmark
[*] Authors to whom any correspondence should be addressed.

E-mail: david.kleiven@ntnu.no **and** jchang@dtu.dk

Keywords: cluster expansion, Monte Carlo, phase transition, bootstrapping, machine learning, energy materials

Abstract

Cluster expansion (CE) has gained an increasing level of popularity in recent years, and its applications go far beyond its original root in binary alloys, reaching even complex crystalline systems often used in energy materials research. Similar to other modern machine learning approaches in materials science, many strategies have been proposed for training and fitting the CE models to first-principles calculation results. Here, we propose a new strategy for constructing a training set based on their relevance in Monte Carlo sampling for statistical analysis and reduction of the expected error. The CE model constructed from the proposed approach has lower dependence on the specific details of the training set, thereby increasing the reproducibility of the model. The same method can be applied to other machine learning approaches where it is desirable to sample relevant configurational space with a small set of training data, which is often the case when they consist of first-principles calculations.

1. Introduction

Cluster expansion (CE) has emerged as a valuable tool in modern computational materials science. The CE technique was first applied to binary alloys [1, 2] to investigate their thermodynamic behavior by conducting Monte Carlo simulation based on the CE model. Later, it has been applied to a large number of alloys [3, 4], oxides [5, 6], and even to thermodynamic studies of crystallographic surfaces [7–9] and nanoparticles [10, 11]. The CE model captures the structure–property relationship of crystalline materials, which is particularly useful for describing the behavior of disordered systems. At the same time, disordered materials are used in many modern energy applications, including next-generation battery materials [12–15], fuel cells catalysts [16–19] and materials for solar cells [20–22]. As the cost of advanced energy materials is typically a bottleneck for the commercialization of sustainable energy technologies, accelerating the development process of low cost alternatives is critical. Consequently, CE is becoming a popular method for analyzing and designing energy materials.

In the CE scheme, the crystalline material is modeled by generalized Ising models, where the various many-body cluster interactions term called effective cluster interaction (ECI) must be determined. The total energy of a binary alloy is given by

$$E = \sum_c V_c \Phi_c(\vec{\sigma}), \tag{1}$$

where c denote a cluster. V_c is an ECI of cluster c, and $\vec{\sigma}$ is an occupation variable which represents the type of element that occupies a specific site ($+1/-1$ or $+1/0$ is commonly used). The correlation function of cluster c, Φ_c, is expressed as

$$\Phi_c = \prod_{i \in c} \sigma_i, \tag{2}$$

where σ_i is the occupation variable at site i. Details of a similar but more complex expression for multi-component systems are explained in [1, 23]. Mathematically, the CE model is represented as a linear regression model where the ECIs are the coefficient of each term. Albeit simple with respect to modern machine learning models based on neural networks, the underlying linear equation to represent the energy of the system makes it very attractive to perform thermodynamic analysis since evaluating energies using the trained CE model takes only few milliseconds on modern computers.

CE models are typically trained using *ab initio* simulation results to map its structure-energy relationship of a given crystalline material. Fitting the CE model to the energy values obtained from *ab initio* simulations is to determine the best set of ECI values for the set of structures chosen for training. Ideally, it is desirable to include all possible configurations in the fitting process to best represent the material's overall behavior. However, the configurational space of even simple binary alloys is too vast to sample manually, and the high computational cost of *ab initio* calculations makes it difficult to cover even a small fraction of this space. Therefore, one is forced to employ a structure selection method to generate structures for the training set, and it is desirable to have a selection scheme such that the prediction errors have a low sensitivity to the details of the training set.

A large number of ECI terms are typically needed for multi-component systems and systems with complex space groups, and the CE models obtained using direct least-square fits are prone to overfitting. Therefore, many different techniques have been used to balance the bias and variance of the resulting CE model. One popular method to avoid overfitting is to use l_2-regularized fits [24]. However, the l_2 regularization does not promote sparsity, which often leads to many ECI terms included in the resulting CE model. It is undesirable to have many ECI terms because calculating many terms in the Monte Carlo simulations leads to a slower speed. Consequently, most of recent developments are aimed at reducing overfitting by selecting only a subset of ECIs, and the examples include l_1-regularized fits [25, 26], Bayesian compressive sensing [27], evolutionary algorithms [28] and automatic relevance detection [29].

The quality of the CE model is typically determined using a cross-validation score (i.e. average prediction error in a validation set). While the CV score gives an estimate of the average predictive power of the CE model, they fall short in two main aspects: (a) it only measures the *average* prediction error without any insight on *where* in the configurational space has the larger error and (b) it fails to provide information about sensitivity to variations in the dataset. The implication of these shortcomings are that the model may have high errors for important configurations (e.g. ground-state structures) in spite of the low cross-validation score, and the addition/removal of a small subset of the training set may dramatically change the model's description of the material due to its high sensitivity to the training set. These two shortcomings are interrelated and depend heavily on the selection of the training set. The strong dependence of the CE model on the training set raises a concern on the reproducibility of reported results as well as a question on how to select a training set that leads to a model that best describes the material behavior.

Commonly employed strategies for selecting training set structures are a full enumeration of all possible structures up to certain cell size, random selection, ground-state search, derivative superstructure [30–32] and probe structures [33]. The results obtained from the model trained using the structure enumeration do not suffer from the low reproducibility since all possible structures on a fixed lattice are included in the training set, albeit a severe constraint is imposed on the cell size. The structure enumeration method is not feasible for complex systems where a primitive unit cell has many atoms. On the other hand, an element of randomness is inherently present in other selection methods, and the dataset will vary even if the same code/algorithm is employed. Our work aims to reduce the dependence of the CE model on the training set by selecting the structures with the highest uncertainty among those encountered in the Monte Carlo sampling trajectory. We argue that the structures that are relevant for capturing the material's physical behavior are likely to be present in the Monte Carlo trajectory and that the inclusion of the structures with the highest uncertainty naturally suppresses the oversampling of a small configurational space. This concept is similar to employing a bootstrapping technique for training machine learning potential [34]. We call our method a *bootstrapping uncertainty structure selection* (BUSS) scheme and investigate its performance on reducing the dependence of the training set by tracking the evolution of the selected clusters as well as the cross-validation score.

2. Method

A set of *ab initio* simulations are needed for fitting the ECIs of a CE model. From the physical point of view, one should choose the training set representing the structures likely to be encountered during

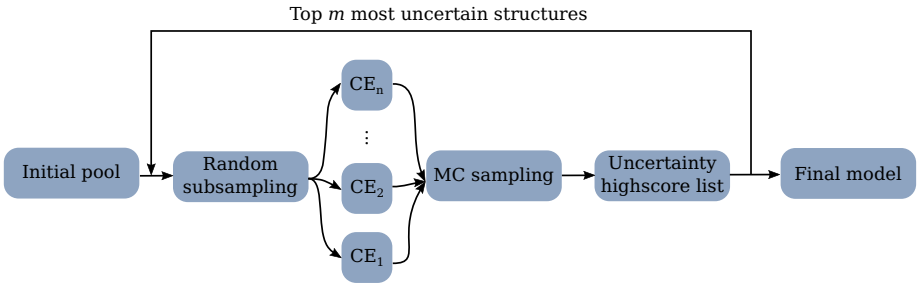

Figure 1. Flow chart of the BUSS scheme.

thermodynamic sampling of the given material. Random selection is an example that is prone to naively selecting the configurations with very high energies of formation, which will not appear unless the temperature or pressure is unrealistically high. Liu *et al* suggested an approach of selecting the training set from the structures encountered during the Monte Carlo sampling [35]. In this work, we present the bootstrapping uncertainty structure selection (BUSS) scheme where an additional criterion is imposed to select the structures with the highest uncertainty from the entire set of structures encountered during the Monte Carlo sampling. The benefit of choosing structures from the Monte Carlo sampling is to avoid the inclusion of highly unstable structures in the training set. The BUSS scheme consists of two stages. The first is a bootstrapping stage where multiple CE models are fitted to different portions of the training set. A Monte Carlo simulation is carried out in the second stage to determine the most uncertain structure in the sampling trajectory, which represents the ensemble of configurations that material is likely to take. The workflow of the BUSS scheme is illustrated in figure 1.

The BUSS scheme starts by populating the pool of structures with q randomly chosen structures ($q = 10$ in this work). The main purpose of having the initial pool of random structure is to have a sufficient number of structures to randomly select in the bootstrapping procedure. Therefore, other structure generation methods such as the enumeration method, derivative superstructure [30–32] and special quasi-random structures [36] can be employed as well. There does not exist a standard way of generating the initial pool of structures, and the BUSS scheme supports any initial pool of structures as long as two or more structures are provided for the subsequent bootstrapping step. It is noted that the pool generated using these approaches can include the structures with unrealistically high energies. However, their impact on the quality of the CE model becomes less relevant as more structures are added to the training in the subsequent steps.

In the subsequent bootstrapping steps, we denote the current number of structures in the training set as N. n new datasets are constructed from the original N structures by randomly sampling them *with* replacement. That is, N structures are randomly chosen from N structures n times to form n datasets (note that the same structure can be chosen multiple times, and each dataset can contain the same structures). One can then fit CE models on each of the n datasets, leading to n different CE models (n is set to 100 in this work). This procedure leads to n individual models, each trained on a different portion of the data. Some data is left out of each set, and some data points are over-represented. The probability of being left out in a dataset is around 36% in the limit of a large number of structures. That is,

$$P_{\text{left}-\text{out}} = \left(1 - \frac{1}{N}\right)^N \xrightarrow[N\to\infty]{} e^{-1} \approx 0.36. \tag{3}$$

In this way, the models will tend to agree when making predictions in regions of training space where there is much data, but will diverge from one another in regions of sparsity. The strategy of using the spread of the ensemble prediction has been applied in the past [37]. However, the BUSS scheme also actively uses the uncertainty to propose the training structures for first-principles calculations.

One can employ many fitting methods, and we used least absolute shrinkage and selection operator (LASSO) regression [25] throughout this study where the hyper parameter is chosen such the five-fold cross-validation score is minimized. The choice of regression technique can impact the cluster selection, especially when a technique that promotes sparsity is employed. Alternative methods such as elastic net [38], automatic relevance determination regression [39] can also be employed. However, a detailed comparison of the effect of the specific choice of regression method is beyond the scope of this work, and we demonstrate the effect of the BUSS scheme using the widely used LASSO regression. We note that one should be careful in interpreting the meaning of the cross-validation score of each individual models; it should not be taken as a measure of the predictive accuracy as the dataset was generated while allowing the same structure to be

selected multiple times. Therefore, a structure can be present in multiple 'folds' during the hyperparameter optimization. The inclusion of multiple instances of the same structure may lead to a model that is insufficiently regularized, and the resulting cross-validation score generally underestimates the model error. However, the objective of the bootstrapping approach is to identify the structures with the highest uncertainty scores, so the precise meaning and value of the cross-validation score in each bootstrapped CE models should not be examined meticulously.

The next part of the BUSS scheme is to search for new candidate structures to be included in the training set. We form a single CE model by averaging all the models (i.e. taking the average of the ECI values), which generally leads to less overfitting. Averaging each model's ECI values is equivalent to averaging the prediction of all models since the CE models are linear. Thus, the unified model can be constructed *a priori*. It is noted that a nonlinear CE scheme has been proposed recently [40]. The BUSS scheme can be applied to the nonlinear models, except that representing the average prediction of all models using their average ECI values is no longer valid. Various types of Monte Carlo simulations based on the CE model can be used to sample microstates of the system where the CE model allows a rapid energy evaluation. In this work, we carried out simulated annealing where the temperatures are gradually lowered from 4000 K to 1 K. The energy of configurations is evaluated for each Monte Carlo step (e.g. swapping two atoms) using each of the n CE models. The energy evaluated from the average CE model is used to determine whether the trial move is accepted based on the acceptance probability, P_{acc}, of the Metropolis algorithm [41],

$$P_{\text{acc}} = \min\left\{1, \exp\left(\frac{-(E_{\text{new}} - E_{\text{old}})}{k_{\text{B}}T}\right)\right\}. \tag{4}$$

E_{new} and E_{old} are the energies of the configuration before and after the trial move, respectively. k_{B} is Boltzmann's constant and T is the temperature in kelvin. For each trial move, the standard deviation of the energy predicted by each of the n CE models is evaluated and taken as a measure of how sensitive the energy prediction is to the variations in the dataset (i.e. uncertainty measure). A list of structures with the highest uncertainty score (highest standard deviation) is kept during the Monte Carlo simulations, and the m structures with the highest uncertainty are added to the training set. It is noted that adding Boltzmann weighting to the uncertainty (i.e. penalize uncertainty in low energy structures) will further promote the inclusion of low-energy structures if the low energy regime is of importance. Each iteration cycle is referred to as a 'generation,' and we selected ten most uncertain structures per generation (i.e. $m = 10$) for *ab initio* simulations for the energy assessment. The procedure is repeated until the uncertainty drops below the predefined threshold. In this work, a total of 100 structures have been evaluated using the BUSS scheme, where the first generation of ten random structures and nine generations of ten most uncertain structures were added in the training set in sequence.

The resulting CE model of each generation is the unified version of all bootstrapped models obtained by taking the average of their ECI values. The aforementioned issue of overfitting in each bootstrapped CE model is alleviated in the unified model by averaging the predictions of all models. It is worth noting that the ECI values obtained by taking the average of the bootstrapped models differ from those obtained using standard regularization methods. The most common approach to obtain the ECI values in the CE community is to tune the hyper parameter in the regularization to minimize the leave-one-out cross-validation score. The quality and robustness of averaging the bootstrapped models are assessed in the current work by comparing its leave-one-out cross-validation scores against the ones obtained in a traditional manner.

Two convergence criteria are used to determine when to stop the training procedure in this work. The first criterion is to achieve the leave-one-out cross-validation score of the model to be less than 5 meV/atom, and the second criterion is to have no abrupt increase/decrease in both uncertainty score and the leave-one-out cross-validation score for three generations in a row. The first criterion ensures the overall prediction accuracy of the model, while the second ensures that a sufficient number of structures are added to the training set, such that further introduction of the least certain structures do not change the model (i.e. reached a convergence).

We have also carried out an additional 90 DFT simulations on the structures generated using a traditional selection scheme where a combination of random structure and ground-state search (hitherto referred to as an RGS scheme). The same initial pool of ten random structures in the BUSS scheme are used for for the RGS scheme. Subsequently, ground-state structures were searched through simulated annealing, and the newly found structures were added to the training set if they differ from the rest of the structures in the training set. We introduced ten structures per generation, and the random structures were added when less than ten new ground-state structures were found. This way, a consistent comparison can be made between RGS and BUSS schemes as they both consists of 100 structures introduced through 10 generations.

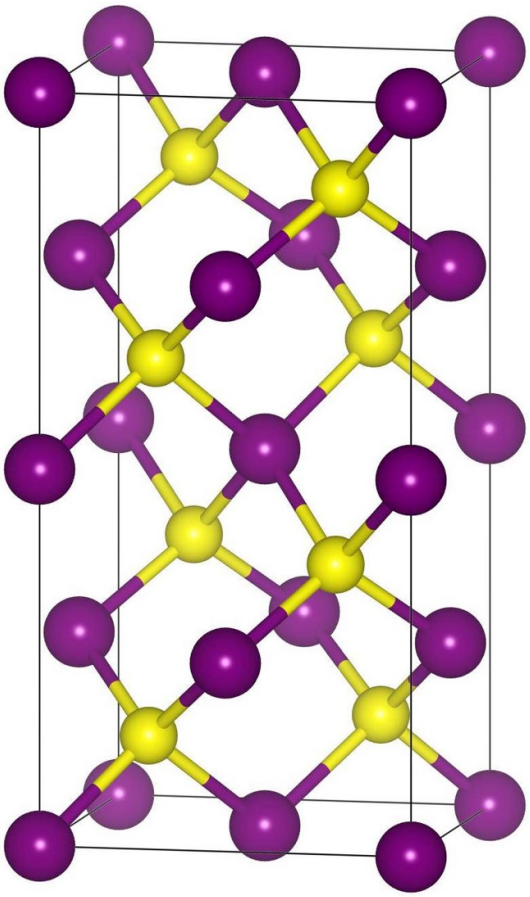

Figure 2. Sketch of disordered Cu_2ZnSnS_4. The mixed cation $Cu_{1/2}Zn_{1/4}Sn_{1/4}$ is shown in purple and the S anion is shown in yellow.

Comparison of the two schemes allows us to evaluate how the BUSS scheme performs in reducing the dependence of the training set with respect to the traditional approach while only introducing the structures that are likely to occur during the statistical sampling at moderate conditions.

We demonstrate the BUSS scheme with disordered Cu_2ZnSnS_4 (CZTS), a promising material for thin-film photovoltaic applications as it consists of earth-abundant and non-toxic elements. Despite its merit, the power conversion efficiency of CZTS needs to improve significantly for commercial adoption, and improving its efficiency is one of the active areas of research. The leading cause of poor efficiency is thought to have originated from the point defects, such as site disorders [42]. A possible route for improving the efficiency of CZTS is to synthesize them in Zn-rich and Cu-poor conditions, which leads to structural disorder [43]. Therefore, an improved structural understanding of CZTS is crucial for further development.

CZTS has a cation sublattice consisting of Cu, Zn and Sn, analogous to metallic ternary alloys. Although CZTS takes on kesterite structure, the significant cationic disorder has been observed at elevated temperatures (see [44] for more detailed discussion). Understanding the substitutional disorder in CZTS is important in gaining further insights into the material and designing a similar class of materials because the disorder can significantly alter the electronic properties of the material. Although the photovoltaic application is used in this work to illustrate the effectiveness of our approach, the same principle can be applied to study other disordered materials other energy applications such as batteries, fuel cells and hydrogen storage. The crystal structure of the cation-disordered CZTS is shown in figure 2.

All of the CE procedures such as generating the training set, Monte Carlo sampling and fitting the CE models were carried out using CL in Atomic Simulation Environment (CLEASE) package [3]. CLEASE does not allow duplicate structures to be in the training set by design, and our algorithm is robust against the unlikely scenarios where the same structure is repeatedly added to the training set. The CE models are fitted to the energies obtained from density functional theory (DFT) simulations. All DFT simulations were carried out on periodic structures using the Vienna *Ab initio* Simulation Package (VASP) [45–48] with a plane-wave cutoff of 400 eV and the k-space was sampled using a Γ-centered Monkhorst–Pack grid [49] with a density of 4.0 points per $Å^{-1}$. Various sizes and shapes of the cells containing up to 56 atoms were used for

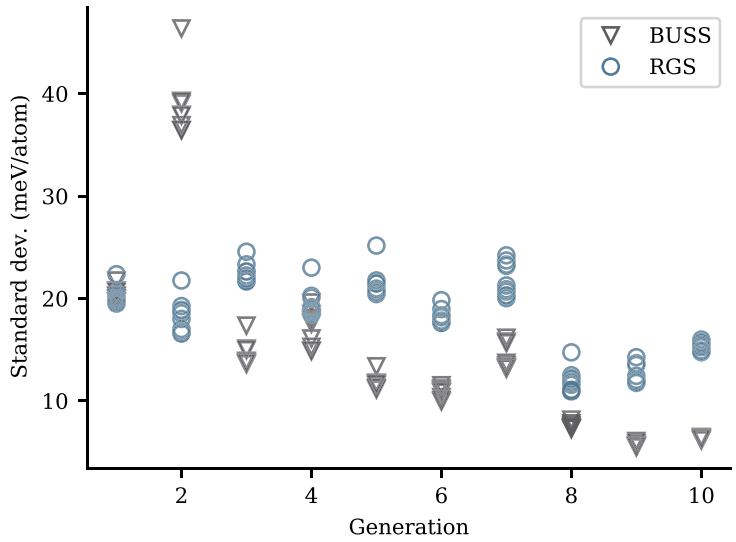

Figure 3. Standard deviation of energies of the ten most uncertain structures observed during the Monte Carlo sampling for the structures generated using BUSS (gray inverted triangle) and RGS (blue circle) schemes.

training structures. The exchange-correlation functional parametrized by Perdew, Burke and Ernzerhof [50] was used. The core electrons were treated in the projector augmented wave (PAW) [51] formalism. The Hubbard U correction [52, 53] was applied to the d orbital of Cu with the U value of 6.0 eV. While the specific choice of U for the study of CZTS can have some implications on the local chemical environment, Ramkumar *et al* reported that the mean absolute error (MAE) of the obtained energies for of PBE+U is 3.3 meV f.u.$^{-1}$ when compared to hybrid functional for 11 different CZTS structures [44]. Furthermore, a specific choice of U value will alter the conclusions of the analysis presented in this work. The performance of PBE+U for CZTS is reported to be comparable with considerably more expensive post-DFT methods such as hybrid functional and self-consistent GW methods for the same material [54, 55]. The PBE+U method was successfully used to study defect formation energy, phase stability and phase transition of CZTS [44, 55]. The atomic positions and cells were fully relaxed until the maximum force on any atom is below 0.02 eV Å^{-1}. The structural relaxation was carried out in the Atomic Simulation Environment [56] using the preconditioned FIRE optimizer [57].

3. Results and discussion

The evolution of the standard deviation (i.e. uncertainty measure) of 10 most uncertain structures based on 100 bootstrapped CE models is shown in figure 3. We also applied the bootstrapping technique on the dataset constructed using the RGS method to assess the effect of choosing structures according to the BUSS scheme by comparing the evolution of uncertainties of the two approaches. The standard deviation for the first generation is the same for both BUSS and RGS schemes as the same initial pool of structures are used. For the BUSS scheme, the standard deviation spikes in the second generation after the first set of structures with the highest uncertainties are introduced. However, the standard deviation quickly decays in the third generation and keeps declining except for a minor increase in the seventh generation. In contrast, there is a lack of reduction in the standard deviation for the first seven generations for the RGS scheme until it finally drops after the addition of 80 structures. The standard deviation slightly increases upon adding more training structure, which shows no sign of convergence. Not only the standard deviation decays faster for the BUSS scheme, but it also is consistently lower for the BUSS scheme from the third generation, resulting in the lower uncertainty in the model using only a small number of training structures. One can observe a clear benefit of the BUSS scheme from figure 3: Selecting training structures specifically aimed at improving the predictive power in the most uncertain regions of the configurational space leads to a faster reduction and convergence of the maximum uncertainty of the model.

The LASSO regularization is used for fitting all the bootstrapped CE models used for determining the uncertainties. While LASSO reduces the number of clusters used in the CE model, both the types of chosen clusters and their ECI values may vary a lot as more training data are included. The large variations of the selected cluster and their ECI values are common to regularization/selection methods that promote sparsity. The BUSS scheme uses a strategy to average many bootstrapped models fitted to different portions of the

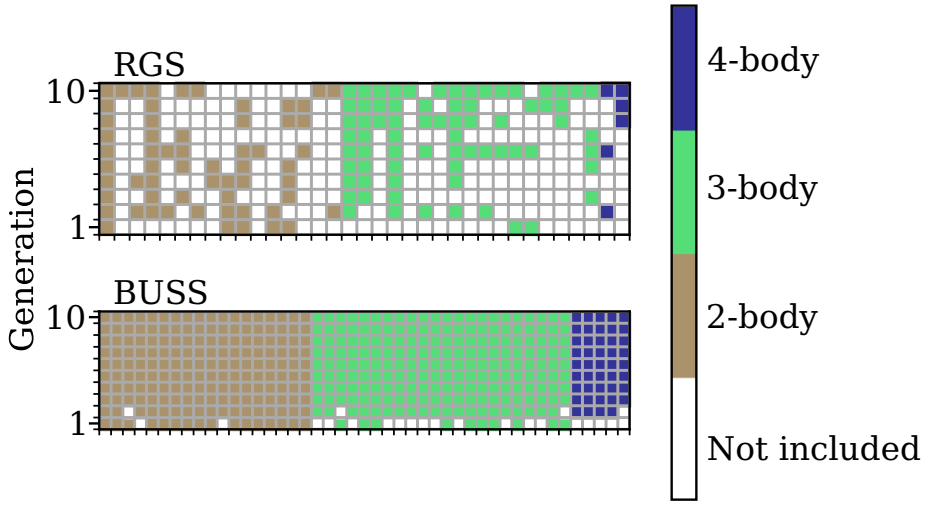

Figure 4. Effect of averaging multiple LASSO fits. A colored square indicates that a feature was included in a particular generation, while a white indicates that a feature was not included. Colors represent different cluster sizes.

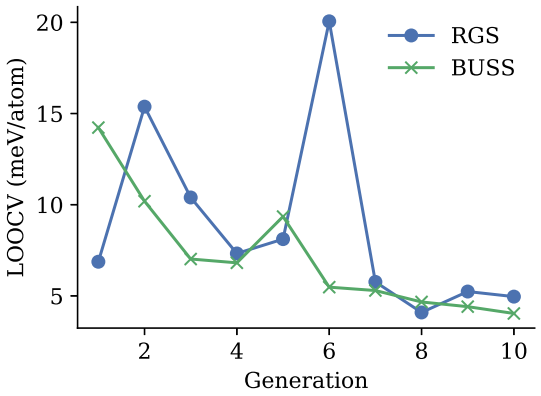

Figure 5. Leave-one-out cross-validation (LOOCV) score as a function of generations where ten additional structures were added to the training set in each generation based on RGS (blue) and BUSS (green) schemes.

dataset, which greatly reduces the variations in the cluster selection as shown in figure 4. The averaging strategy of the BUSS scheme includes a cluster in the model if at least one of the bootstrapped CE models includes it. This approach leads to a less sparse model, although not as dense as l_2-norm based regularization where sparsity is not promoted. The resulting consistency in cluster selection is similar to what can be achieved in the group LASSO approach [26], which imposes a constraint to include a cluster only after all of its subclusters are included.

It is shown in figure 4 that the selection of clusters is inconsistent for the RGS scheme, indicating the large variation in the cluster selection from one generation to another. In fact, only one cluster is included throughout the entire training process while most of the clusters are included in less than half of the generations. In contrast, a majority of clusters are included in > 9 generations with a few exceptions of being included in eight generations for the BUSS scheme, and the cluster selection remains unchanged after the third generation. The clear difference between the two schemes illustrates that the BUSS scheme is much more consistent in cluster selection compared to the standard regularization methods that promote sparsity. Our results indicate that model averaging is an effective approach to improve the inconsistent feature selectors.

The final CE models (the unified CE model in the case of the BUSS scheme) are evaluated after each generation. We used LASSO regularization to find the lowest leave-one-out cross-validation score after each generation for the RGS scheme, a standard practice for obtaining the score. On the other hand, the leave-one-out cross-validation score of the BUSS scheme is obtained without regularization but by taking the average ECI values of the bootstrapped models as described above. The evolution of the leave-one-out cross-validation score for RGS and BUSS schemes are shown in figure 5.

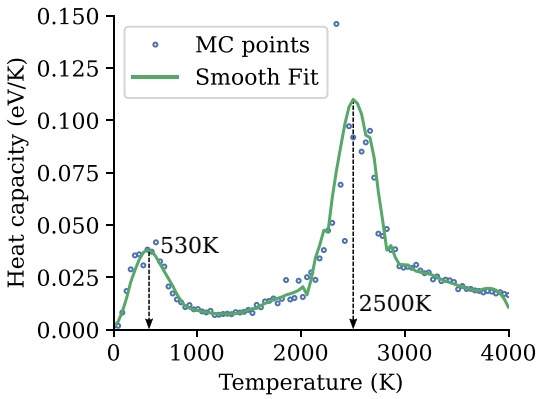

Figure 6. Heat capacity of CZTS as a function of temperature. Blue circles are raw values obtained from Monte Carlo simulations and green line is a smooth curve obtained using smooth fit.

The leave-one-out cross-validation score is quite different for the two schemes for the first generation, although the same initial pool of structures is used. The difference stems from the fact different fitting methods are used for obtaining the ECI values. Although it started from a much higher value, the cross-validation score of the BUSS scheme decays rapidly in a rather prediction manner, except for a minor spike in the fifth generation. On the other hand, the cross-validation score of the RGS scheme fluctuates substantially between generations, albeit exhibiting a general downward trend. The steady and predictable decrease of the cross-validation score for the BUSS scheme indicates that the model is much less sensitive to the variations in the training set even when there are a small number of structures present. Not only that the BUSS scheme has more predictable decay in the cross-validation score, but it also outperforms the RGS scheme during most of the training process with a lower final score. It is interesting to find that the cross-validation score of the BUSS scheme is lower than the traditional approach, although it populates the dataset solely with the most uncertain structures observed in the Monte Carlo sampling.

As a final step, we proceed to investigate the properties of CZTS using the unified CE model obtained after the tenth generation based on the BUSS scheme. The heat capacity of CZTS is traced during the simulated annealing to monitor the order-disorder phase transition temperatures of CZTS on temperature. The heat capacity is shown as a function of temperature in figure 6. There are two peaks that correspond to transition temperatures: 530 K and 2500 K. The obtained transition temperatures are very close to 545 K and 2891 K the values reported in a previous CE study by Ramkumar *et al* [44]. Ramkumar *et al* compared their findings with other experimental reports and found that the lower transition temperature from CE closely resembles that of experimental values of 533 K $\pm$ 10 K and 552 K [58–60]. However, the second transition temperature obtained from the CE model significantly overestimates the experimentally observed temperature of 1150 K [61, 62]. Ramkumar *et al* attributed the source of discrepancy to the lack of vibrational entropy and vacancy contribution in the CE model, as well as the possibility of the material being off-stoichiometric in the experimental studies [44].

The pair correlation functions are also monitored to investigate the structural dependence of CZTS on temperature. The average pair correlation functions up to the sixth nearest neighbors (nn) on the cation sublattice are shown in figure 7 for selected temperatures. The pair correlation functions are reported for Zn and Sn species as the information on the Cu in the same sublattice can be inferred from the presented information. The pair correlation function between elements A and B, Φ_{AB}, is defined as

$$\Phi_{AB} = \frac{1}{2N}\sum_{i=1}^{N} n_{AB,i}, \tag{5}$$

where N is the total number of cations in a cell and $n_{AB,i}$ is the number of A–B pairs that includes site i.

All pair correlation functions have the same value at a high temperature limit, which represents a state where the elements are randomly mixed. A pattern starts to emerge as the temperature is lowered. CZTS remains relatively well mixed at 4000 K, a temperature above the second transition temperature. A sign of short-range order appears at 1536 K, a temperature between the first and second order-disorder transition temperature. Most notably, strong signals of the third and fourth nearest neighbor Zn–Zn pairs while the signals of the first, second and fifth nearest neighbor Zn–Zn disappear. Some ordering patterns also appear for Zn–Sn and Sn–Sn pairs, although they are not as pronounced. A more ordering pattern appears upon further cooling. At 203 K, a temperature below the first transition temperature, the signal for the third nearest

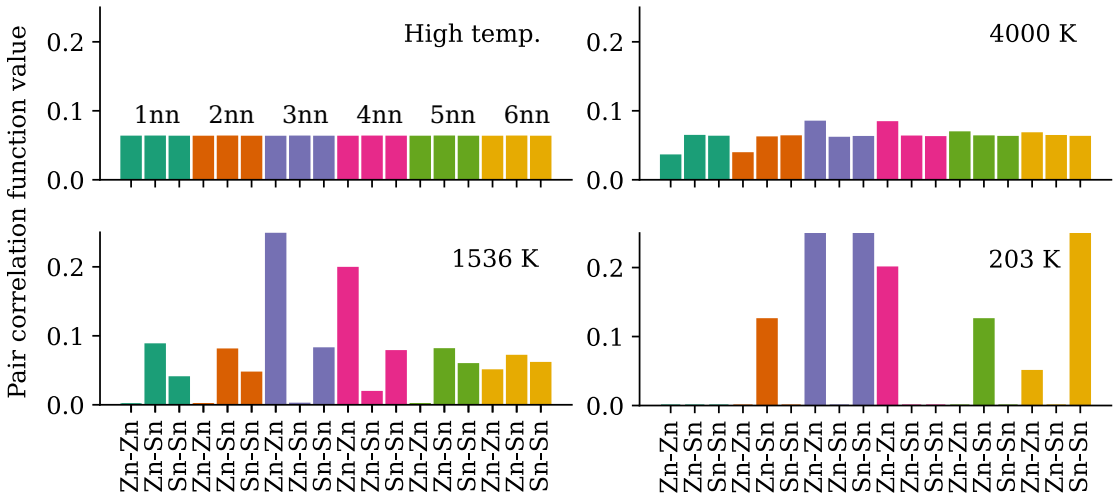

Figure 7. Average of pair correlation functions at different temperatures.

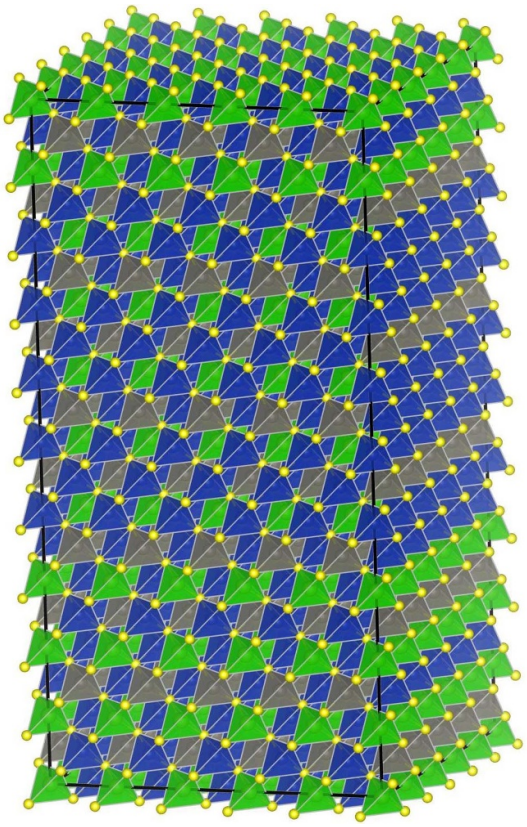

Figure 8. Ground-state structure produced by the CE model after 10 generations of the BUSS scheme. The Cu atoms are shown in blue, the Sn atoms are shown in green, the Zn atoms are shown in gray and the S atoms are shown in yellow.

neighbor Zn–Zn pair remains virtually the same, while the signal for the fourth nearest neighbor Zn–Zn pair is increased slightly. The pair correlation function pattern becomes more distinct at temperatures below the first order-disorder transition. The signal for the first nearest neighbors disappears entirely, indicating that all the nearest neighbor pairs contain Cu atom, which is reminiscent of the kesterite structure. Finally, the final structure obtained from the simulated annealing is shown in figure 8. The structure closely resembles the kesterite structure with alternating Cu/Zn and Cu/Sn planes along the [001] direction, except some discrepancy in the plane ordering. The kesterite structure still is predicted to be the ground-state structure with the energy of −3.64 eV/atom, while the structures shown in figure 8 has the energy of −3.60 eV/atom.

4. Conclusion

A new training set construction technique for CE is presented. The motivation is to train the model using the structures that one is likely to encounter during the statistical sampling process, thereby ensuring that the model has sufficient information on the relevant sampling trajectory. A list of the bootstrapped model is used to determine the ECI values and estimate the uncertainty of the predicted energies quantitatively. A set of structures with the highest uncertainty values is added to the training set to avoid oversampling of small regions in the configurational space. The final model is constructed by unifying the bootstrapped models by taking the average of their ECIs. The model constructed using our approach displays both lower and more predictable decay of the cross-validation score compared to the model constructed in a traditional approach, indicating that the model is less sensitive to the details of the training set while outperforming the traditional approach. The presented approach will be useful for many practitioners of the CE method as it reduces the required number of training data, which consists of the structures that represent the microstates of the material to be sampled in a statistical investigation. Furthermore, the presented method of averaging the multiple bootstrapping models can be combined with any feature selectors to make the model more robust.

Data availability statement

The data that support the findings of this study are available upon reasonable request from the authors.

Acknowledgment

The authors acknowledge support from the European Union's Horizon 2020 research and innovation programme under Grant Agreement No. 957189.

ORCID iDs

Tejs Vegge https://orcid.org/0000-0002-1484-0284
Jin Hyun Chang https://orcid.org/0000-0003-0668-4530

References

[1] Sanchez J M, Ducastelle F and Gratias D 1984 *Physica* A **128** 334–50
[2] Sanchez J M 1993 *Phys. Rev.* B **48** 14013–5
[3] Chang J H, Kleiven D, Melander M, Akola J, Garcia-Lastra J M and Vegge T 2019 *J. Phys.: Condens. Matter* **31** 325901
[4] Kleiven D and Akola J 2020 *Acta Mater.* **195** 123–31
[5] Pedersen C S, Chang J H, Li Y, Pryds N and Garcia Lastra J M 2020 *APL Mater.* **8** 121101
[6] Chable J, Baur C, Chang J H, Wenzel S, García-Lastra J M and Vegge T 2020 *J. Phys. Chem.* C **124** 2229–37
[7] Chen W, Dalach P, Schneider W F and Wolverton C 2012 *Langmuir* **28** 4683–93
[8] Penev E S, Bhowmick S, Sadrzadeh A and Yakobson B I 2012 *Nano Lett.* **12** 2441–5
[9] Sadigh B, Asta M, Ozoliņš V, Schmid A K, Bartelt N C, Quong A A and Hwang R Q 1999 *Phys. Rev. Lett.* **83** 1379–82
[10] Yuge K 2010 *J. Phys.: Condens. Matter* **22** 245401
[11] Wang L L, Tan T L and Johnson D D 2014 *Nano Lett.* **14** 7077–84
[12] Meng Y S and Arroyo-De Dompablo M E 2009 *Energy Environ. Sci.* **2** 589–609
[13] Wang F, Kim S W, Seo D H, Kang K, Wang L, Su D, Vajo J J, Wang J and Graetz J 2015 *Nat. Commun.* **6** 6668
[14] Huang W, Urban A, Rong Z, Ding Z, Luo C and Ceder G 2017 *npj Comput. Mater.* **3** 30
[15] Chang J H *et al* 2020 *J. Mater. Chem.* A **8** 16551–9
[16] Han B C, van der Ven A, Ceder G and Hwang B J 2005 *Phys. Rev.* B **72** 1–9
[17] Stamenkovic V R, Fowler B, Mun B S, Wang G, Ross P N, Lucas C A and Markovic N M 2007 *Science* **315** 493–7
[18] Cao L and Mueller T 2015 *J. Phys. Chem.* C **119** 17735–47
[19] Cao L and Mueller T 2016 *Nano Lett.* **16** 7748–54
[20] Yamamoto K, Iikubo S, Yamasaki J, Ogomi Y and Hayase S 2017 *J. Phys. Chem* C **121** 27797–804
[21] Zheng Y F, Yang J H and Gong X G 2019 *AIP Adv.* **9** 035248
[22] Wong Z M, Deng T, Shi W, Wu G, Tan T L and Yang S W 2020 *Mater. Adv.* **1** 1176–85
[23] Zhang X and Sluiter M H F 2016 *J. Phase Equilibria Diffus.* **37** 44–52
[24] Sanchez J 2019 *Phys. Rev.* B **99** 134206
[25] Tibshirani R 1996 *J. R. Stat. Soc.* B **58** 267–88
[26] Leong Z and Tan T L 2019 *Phys. Rev.* B **100** 134108
[27] Nelson L J, Ozoliņš V, Reese C S, Zhou F and Hart G L 2013 *Phys. Rev.* B **88** 155105
[28] Hart G L, Blum V, Walorski M J and Zunger A 2005 *Nat. Mater.* **4** 391–4
[29] Ångqvist M, Muñoz W A, Rahm J M, Fransson E, Durniak C, Rozyczko P, Rod T H and Erhart P 2019 *Adv. Theory Simul.* **2** 1900015
[30] Hart G L W and Forcade R W 2008 *Phys. Rev.* B **77** 1–12
[31] Hart G L W, Nelson L J and Forcade R W 2012 *Comput. Mater. Sci.* **59** 101–7
[32] Morgan W S, Hart G L and Forcade R W 2017 *Comput. Mater. Sci.* **136** 144–9
[33] Seko A, Koyama Y and Tanaka I 2009 *Phys. Rev.* B **80** 1–7

[34] Peterson A A, Christensen R and Khorshidi A 2017 *Phys. Chem. Chem. Phys.* **19** 10978–85
[35] Liu X, Zhang J, Yin J, Bi S, Eisenbach M and Wang Y 2021 *Comput. Mater. Sci.* **187** 110135
[36] Zunger A, Wei S H, Ferreira L G and Bernard J E 1990 *Phys. Rev. Lett.* **65** 353–6
[37] Kostiuchenko T, Körmann F, Neugebauer J and Shapeev A 2019 *npj Comput. Mater.* **5** 1–7
[38] Friedman J, Hastie T and Tibshirani R 2010 *J. Stat. Softw.* **33** 1–22
[39] Wipf D P, Nagarajan S S, Platt J, Koller D and Singer Y 2007 A new view of automatic relevance determination *NIPS* pp 1625–32
[40] Natarajan A R and van der Ven A 2018 *npj Comput. Mater.* **4** 56
[41] Metropolis N, Rosenbluth A W, Rosenbluth M N, Teller A H and Teller E 1953 *J. Chem. Phys.* **21** 1087–92
[42] Bosson C J, Birch M T, Halliday D P, Knight K S, Gibbs A S and Hatton P D 2017 *J. Mater. Chem.* A **5** 16672–80
[43] Delbos S 2012 *EPJ Photovolt.* **3** 35004
[44] Ramkumar S P, Miglio A, van Setten M J, Waroquiers D, Hautier G and Rignanese G M 2018 *Phys. Rev. Mater.* **2** 085403
[45] Kresse G and Hafner J 1993 *Phys. Rev.* B **47** 558–61
[46] Kresse G and Hafner J 1994 *Phys. Rev.* B **49** 14251–69
[47] Kresse G and Furthmüller J 1996 *Comput. Mater. Sci.* **6** 15–50
[48] Kresse G and Furthmüller J 1996 *Phys. Rev.* B **54** 11169–86
[49] Monkhorst H J and Pack J D 1976 *Phys. Rev.* B **13** 5188–92
[50] Perdew J P, Burke K and Ernzerhof M 1996 *Phys. Rev. Lett.* **77** 3865–8
[51] Blöchl P E 1994 *Phys. Rev.* B **50** 17953–79
[52] Anisimov V I, Zaanen J and Andersen O K 1991 *Phys. Rev.* B **44** 943–54
[53] Cococcioni M and de Gironcoli S 2005 *Phys. Rev.* B **71** 035105
[54] Botti S, Kammerlander D and Marques M A 2011 *Appl. Phys. Lett.* **98** 2–5
[55] Sarker P, Al-Jassim M M and Huda M N 2015 *J. Appl. Phys.* **117** 035702
[56] Larsen A H *et al* 2017 *J. Phys.: Condens. Matter.* **29** 273002
[57] Bitzek E, Koskinen P, Gähler F, Moseler M and Gumbsch P 2006 *Phys. Rev. Lett.* **97** 170201
[58] Ritscher A, Hoelzel M and Lerch M 2016 *J. Solid State Chem.* **238** 68–73
[59] Scragg J J, Choubrac L, Lafond A, Ericson T and Platzer-Björkman C 2014 *Appl. Phys. Lett.* **104** 2–6
[60] Scragg J J, Larsen J K, Kumar M, Persson C, Sendler J, Siebentritt S and Platzer Björkman C 2016 *Phys. Status Solidi* b **253** 247–54
[61] Schorr S 2011 *Sol. Energy Mater. Sol. Cells* **95** 1482–8
[62] Schorr S and Gonzalez-Aviles G 2009 *Phys. Status Solidi* a **206** 1054–8