Computational Materials Science 242 (2024) 113073

Contents lists available at ScienceDirect

Computational Materials Science

journal homepage: www.elsevier.com/locate/commatsci

Full length article

Cluster expansion by transfer learning for phase stability predictions

A. Dana [a,b,*], L. Mu [a], S. Gelin [a,b], S.B. Sinnott [a,b,c,d], I. Dabo [a,b,e,**]

[a] Department of Materials Science and Engineering, The Pennsylvania State University, University Park, PA 16802, USA
[b] Materials Research Institute, The Pennsylvania State University, University Park, PA 16802, USA
[c] Department of Chemistry, The Pennsylvania State University, University Park, PA 16802, USA
[d] Institute for Computational and Data Sciences, The Pennsylvania State University, University Park, PA 16802, USA
[e] Department of Physics, The Pennsylvania State University, University Park, PA 16802, USA

ARTICLE INFO

Keywords:
Bayesian sampling
Density-functional theory
Graph neural networks
Reactive potentials
Embedded-atom potentials

ABSTRACT

Recent progress towards universal machine-learned interatomic potentials holds considerable promise for materials discovery. Yet the accuracy of these potentials for predicting phase stability may still be limited. In contrast, cluster expansions provide accurate phase stability predictions but are computationally demanding to parameterize from first principles, especially for structures of low dimension or with a large number of components, such as interfaces or multimetal catalysts. We overcome this trade-off *via* transfer learning. Using Bayesian inference, we incorporate prior statistical knowledge from machine-learned and physics-based potentials, enabling us to sample the most informative configurations and to efficiently fit first-principles cluster expansions. This algorithm is tested on Pt:Ni, showing robust convergence of the mixing energies as a function of sample size with reduced statistical fluctuations.

1. Introduction

Accurate and efficient predictions of phase stability are critical to materials discovery. While machine-learned potentials can efficiently explore the configurations of a phase, their precision is limited when these configurations are not captured by the training dataset. To quantify this limitation, Fig. 1 compares the accuracy of select interatomic potentials, including pre-fitted many-body potentials (charge-optimized many-body potential, COMB3 [1]; reactive force field, ReaxFF [2]; embedded-atom method, EAM [3]; modified embedded-atom method, MEAM [4]) and off-the-shelf machine-learning models (crystal Hamiltonian graph neural network, CHGNet [5]; graph neural network with three-body interactions, M3GNet [6]; atomistic line graph neural network, ALIGNN [7]; message passing multilayer atomic cluster expansion, MACE [8,9]) in predicting the stability of the face-centered cubic Pt:Ni binary [Fig. 1(a)]. Reactive and embedded-atom physics-based potentials (PBPs) rely on predetermined, physically formulated functions, often tailored to specific chemical compositions; as such, they may inherit some transferability beyond the limits of the training data. In contrast, machine-learned potentials (MLPs) do not typically depend on physical approximations; they utilize highly adaptable analytical formulations to predict potential energies and should generally be restricted to the regions of the configurational space that are covered by the training dataset [10].

As shown in Fig. 1(b), for this prototypical bimetallic alloy, MLP and PBP energies can deviate considerably from density-functional theory (DFT) calculations (as MLPs and PBPs may not adequately extrapolate DFT predictions). Although the results for MEAM, EAM, ALIGNN, and M3GNet appear relatively close to the DFT reference for Pt:Ni, discrepancies of up to 40 meV per atom are still observed. While MLPs and PBPs do not yet achieve the precision of DFT models, they still carry important information about the relative energies of the different configurations, as illustrated in Fig. 1(c), which presents rescaled formation energies (the calculation of the scaling factor is explained in Section 3). Convex hull diagrams are also presented in Fig. 1(d), demonstrating that MLPs and PBPs capture the ordering of formation energies to a reasonable extent, although they do not reliably predict the convex hull of the stable configurations for the face-centered cubic phase of Pt:Ni.

We circumvent this limitation by accelerating the parameterization of cluster expansions (CEs), exploiting the latent information contained in MLP and PBP data. CEs evaluate the energy of a lattice by summing energy contributions from finite-size clusters across lattice sites [11,12]. These models have been widely used to study crystalline order and phase stability at reduced computational cost relative to DFT calculations [13–17], and are useful for predicting free energies [13–15,18], magnetic states [19], phase transitions [11,18,20], and defect

* Corresponding author.
** Corresponding author at: Department of Materials Science and Engineering, The Pennsylvania State University, University Park, PA 16802, USA.
E-mail addresses: a.dana@psu.edu (A. Dana), dabo@psu.edu (I. Dabo).

https://doi.org/10.1016/j.commatsci.2024.113073
Received 10 November 2023; Received in revised form 30 April 2024; Accepted 1 May 2024
Available online 15 May 2024
0927-0256/© 2024 Elsevier B.V. All rights reserved.

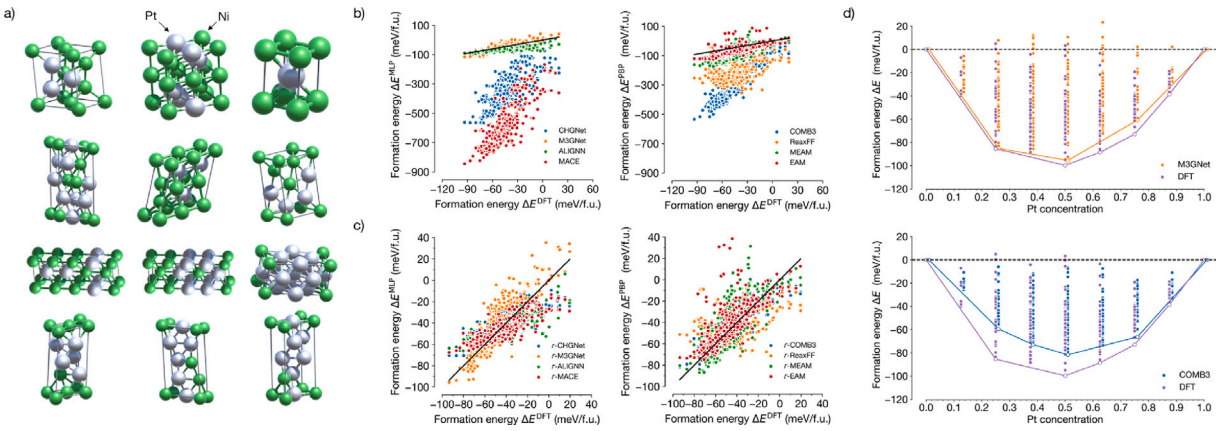

Fig. 1. Accuracy of machine-learned and physics-based (reactive and embedded-atom) potentials in reproducing DFT energies [within the Perdew–Burke–Ernzerhof (PBE) generalized-gradient approximation]. (a) 12 representative supercells (out of a dataset of 413 symmetrically inequivalent structures) for Pt:Ni. (b) Parity plots of mixing energies for machine-learned and physics-based potentials relative to the DFT reference and (c) parity plots after optimal rescaling of the energies with respect to DFT (cf. Section 3 for a detailed description of the rescaling method). (The prefix '*r*-' indicates that the potential is rescaled.) (d) Convex hull plots for M3GNet and COMB3 using the scaled energies. (The energy points are slightly shifted with respect to the actual concentrations for ease of comparison.)

stability [21]. The central complication in constructing CEs is to generate a dataset of DFT energies. This constraint is especially problematic for low-dimension systems, as the absence of full translation symmetry implies that a large number of configurations is needed to capture the interatomic interactions along the nonperiodic direction(s) [12,22]. Considerable effort has been dedicated to generating cluster expansions that minimize prediction errors for a given training set size. While various machine-learning techniques, including active learning [23–25], cross validation [23], regularization [26,27], and feature selection [26–29], are commonly employed to circumvent this bottleneck, significant improvements in computational efficiency may be achieved by leveraging statistical correlations extracted from MLPs and PBPs.

In what follows, we present and validate an algorithm to expedite the fitting of cluster expansions by transfer learning. This approach exploits Bayesian inference to extract prior knowledge from MLPs/PBPs, enabling one to identify the most informative configurations in a given pool [27,30,31]. We show the efficacy of this method by examining Pt:Ni intermetallics.

2. Methodology

2.1. Cluster expansion

Within cluster expansions, the formation energy ΔE of a configuration σ of a system is expressed as the sum of energy contributions associated with symmetrically inequivalent clusters that make up that configuration [17,32]:

$$\Delta E(\boldsymbol{\sigma}) = \sum_{\alpha} m_{\alpha} J_{\alpha} \pi_{\alpha}(\boldsymbol{\sigma}), \tag{1}$$

where $\pi_{\alpha}(\boldsymbol{\sigma}) = \left\langle \prod_i \varphi_{\alpha,i}(\sigma_i) \right\rangle_{\alpha}$ represents a cluster product averaged over a collection of symmetrically inequivalent clusters labeled by the index α with i being the site index and $(\varphi_{\alpha,i})_{\alpha}$ being a basis of orthogonal functions of the site-dependent occupation σ_i. Multiplicity factors (m_{α}) quantify how many times a symmetrically equivalent cluster appears throughout the lattice and J_{α} is the effective cluster interaction (ECI) corresponding to the energy contribution of a cluster to the total energy.

To derive a cluster expansion, it is necessary to determine the ECIs. This process involves acquiring reference data, typically in the form of a set of configurations, along with an associated vector of target energies, which is usually obtained from first-principles calculations. Eq. (1) can be expressed in a simplified vectorial form as [33]

$$\boldsymbol{\Delta E} = \boldsymbol{\Pi}\, \boldsymbol{J}, \tag{2}$$

where the vector $\boldsymbol{\Delta E}$ encodes the energies of the configurations, $\boldsymbol{J}$ represents the ECIs, and $\boldsymbol{\Pi}$ is the matrix of cluster products. The ECIs can be estimated as

$$\boldsymbol{J} = \boldsymbol{\Pi}^{+} \boldsymbol{\Delta E}. \tag{3}$$

where $\boldsymbol{\Pi}^{+} \equiv (\boldsymbol{\Pi}^{\top}\boldsymbol{\Pi})^{-1}\boldsymbol{\Pi}^{\top}$ denotes the pseudoinverse of $\boldsymbol{\Pi}$.

2.2. Bayesian sampling

The Bayesian approach consists of specifying a prior distribution over hypotheses or parameters. Using Bayes' theorem, as new data becomes available, the prior is combined with the likelihood to compute the posterior distribution. Implicitly, Bayes' theorem can be expressed as [34]

$$\big(\text{posterior}\big) = \big(\text{likelihood}\big) \cdot \big(\text{prior}\big) / \big(\text{marginal likelihood}\big). \tag{4}$$

This approach not only enables for parameter estimation but also offers the ability to account for uncertainty and incorporate domain/empirical knowledge using Gaussian statistical distributions $\mathcal{G}$ [35]. An example of energy distribution for a collection of N configurations is shown in Fig. 2. The conventional functional representation of the distribution is illustrated in Fig. 2(a). An equivalent N-dimensional vectorial description is shown in Fig. 2(b). The goal of the Bayesian sampling is to minimize the number of first-principles calculations by identifying a subset of configurations from a larger pool, which most effectively capture the energy trends (the energy covariance).

> The distinct advantage of the proposed method is that the kernel of the prior statistical distribution, which encodes the energy covariance, is directly derived from universal machine-learning potentials and physics-based potentials, rather than being modeled using a chosen metric of structural similarity.

The initial step of the sampling consists of generating the Gaussian distribution prior

$$\mathcal{G}(\boldsymbol{\Delta E}) = \left|2\pi \boldsymbol{A}^{-1}\right|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\boldsymbol{\Delta E} - \boldsymbol{\mu})^{\top} \boldsymbol{A} (\boldsymbol{\Delta E} - \boldsymbol{\mu})\right), \tag{5}$$

where $\boldsymbol{\Delta E}$ is a N-dimensional vector representing the energies of the N configurations, $\boldsymbol{\mu} = \mathbb{E}[\boldsymbol{\Delta E}]$ is the expectation value of $\boldsymbol{\Delta E}$, $\boldsymbol{A} = \boldsymbol{K}^{-1}$ is the inverse of the covariance matrix $\boldsymbol{K}$, which describes the correlations between the energies ($\boldsymbol{K} = \mathbb{E}[(\boldsymbol{\Delta E} - \boldsymbol{\mu})^{\top}(\boldsymbol{\Delta E} - \boldsymbol{\mu})]$), and $|\cdot|$ denotes the determinant. To describe the sampling method, we rewrite $\mathcal{G}(\boldsymbol{\Delta E})$ as

$$\mathcal{G}(\boldsymbol{\Delta E}) = \left|2\pi \boldsymbol{A}^{-1}\right|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\begin{pmatrix} \boldsymbol{\Delta E}_{\perp} - \boldsymbol{\mu}_{\perp} \\ \boldsymbol{\Delta E}_{\parallel} - \boldsymbol{\mu}_{\parallel} \end{pmatrix}^{\top}\right.$$

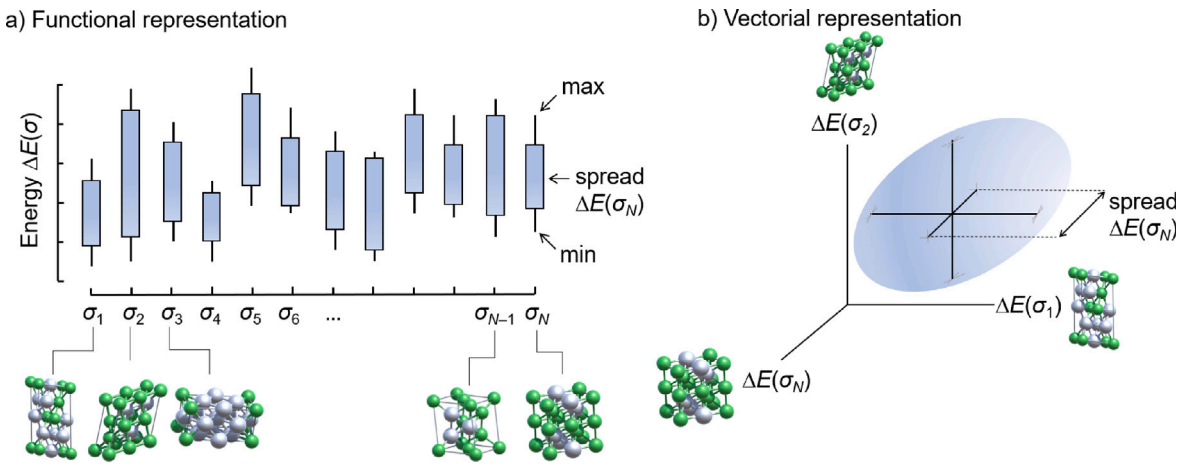

Fig. 2. Functional and vectorial representations of the statistical distribution of the energies of N configurations. (a) By calculating the average and spread of the energy of each cluster across selected empirical potentials, one obtains a statistical energy distribution. (b) This distribution can be represented as a Gaussian probability in N-dimensional vector space of the configurational energies.

$$\times \begin{pmatrix} \boldsymbol{A}_{\perp\perp} & \boldsymbol{A}_{\perp\parallel} \\ \boldsymbol{A}_{\parallel\perp} & \boldsymbol{A}_{\parallel\parallel} \end{pmatrix} \begin{pmatrix} \boldsymbol{\Delta E}_\perp - \boldsymbol{\mu}_\perp \\ \boldsymbol{\Delta E}_\parallel - \boldsymbol{\mu}_\parallel \end{pmatrix} \Bigg), \tag{6}$$

where $\parallel$ indicates the projection on the subspace of the sampled configuration and $\perp$ indicates the projection out of this subspace. The prior can then be refined by Bayesian inference using the configurations that have been sampled at the previous iterations. Using this information, Eq. (4) can be rewritten as

$$\mathcal{G}(\boldsymbol{\Delta E}_\perp | \boldsymbol{\Delta E}_\parallel = \boldsymbol{\Delta E}_0) = \left(\int_\perp \mathcal{G}(\boldsymbol{\Delta E}_\perp, \boldsymbol{\Delta E}_0) d\boldsymbol{\Delta E}_\perp \right)^{-1} \mathcal{G}(\boldsymbol{\Delta E}_\perp, \boldsymbol{\Delta E}_0), \tag{7}$$

where $\mathcal{G}(\boldsymbol{\Delta E}_\perp | \boldsymbol{\Delta E}_\parallel = \boldsymbol{\Delta E}_0)$ denotes the posterior distribution obtained by replacing $\boldsymbol{\Delta E}_\parallel$ with $\boldsymbol{\Delta E}_0$, which represents the energies of the already sampled configurations, $\int_\perp \mathcal{G}(\boldsymbol{\Delta E}_\perp, \boldsymbol{\Delta E}_0) d\boldsymbol{\Delta E}_\perp$ is the marginal likelihood and $\mathcal{G}(\boldsymbol{\Delta E}_\perp, \boldsymbol{\Delta E}_0)$ is the likelihood-weighted prior. Eq. (7) yields

$$\begin{aligned} \mathcal{G}(\boldsymbol{\Delta E}_\perp | \boldsymbol{\Delta E}_\parallel = \boldsymbol{\Delta E}_0) = \left| 2\pi \boldsymbol{A}_{\perp\perp}^{-1} \right|^{-\frac{1}{2}} \exp \Bigg(-\frac{1}{2} \Big[\boldsymbol{\Delta E}_\perp - \boldsymbol{\mu}_\perp \\ - \boldsymbol{A}_{\perp\perp}^{-1} \boldsymbol{A}_{\perp\parallel} (\boldsymbol{\Delta E}_0 - \boldsymbol{\mu}_\parallel) \Big]^\top \boldsymbol{A}_{\perp\perp} \Big[\boldsymbol{\Delta E}_\perp - \boldsymbol{\mu}_\perp \\ - \boldsymbol{A}_{\perp\perp}^{-1} \boldsymbol{A}_{\perp\parallel} (\boldsymbol{\Delta E}_0 - \boldsymbol{\mu}_\parallel) \Big] \Bigg), \end{aligned} \tag{8}$$

which can be further simplified into

$$\begin{aligned} \mathcal{G}(\boldsymbol{\Delta E}_\perp | \boldsymbol{\Delta E}_\parallel = \boldsymbol{\Delta E}_0) = \left| 2\pi \boldsymbol{K}'_{\perp\perp} \right|^{-\frac{1}{2}} \exp \Big(-\frac{1}{2} (\boldsymbol{\Delta E}_\perp - \boldsymbol{\mu}'_\perp)^\top \\ \cdot (\boldsymbol{K}'_{\perp\perp})^{-1} (\boldsymbol{\Delta E}_\perp - \boldsymbol{\mu}'_\perp) \Big), \end{aligned} \tag{9}$$

where $\boldsymbol{\mu}'_\perp$ and $\boldsymbol{K}'_{\perp\perp}$ stand for the mean and the covariance of the posterior, respectively. Thus, by comparing Eqs. (8) and (9), the mean and the covariance of the posterior can be determined iteratively as

$$\boldsymbol{K}_{n+1} = \left[(\boldsymbol{K}_n^{-1})_{\perp\perp} \right]^{-1} \tag{10}$$

$$\boldsymbol{\mu}_{n+1} = (\boldsymbol{\mu}_n)_\perp - \boldsymbol{K}_{n+1} (\boldsymbol{K}_n^{-1})_{\perp\parallel} (E_\parallel \boldsymbol{\sigma}_\parallel - (\boldsymbol{\mu}_n)_\parallel), \tag{11}$$

where $\boldsymbol{K}_n$ and $\boldsymbol{\mu}_n$ are the covariance matrix and mean vector of the prior after n iterations, $E_\parallel$ is the energy of the configuration $\boldsymbol{\sigma}_\parallel$ that has been newly sampled at the current (nth) iteration.

In Eqs. (10) and (11), the configuration $\boldsymbol{\sigma}_\parallel$ that is sampled (that is, the configuration whose energy will be calculated at the DFT level at the next step of the iterative process) is the one that results in the largest reduction in the uncertainty of the posterior, which is represented as the area $S_{n\perp}$ of the $\boldsymbol{\mu}$-centered projection of the prior along the direction of the configuration $\boldsymbol{\sigma}_n$, as illustrated in Fig. 3. In this example, because $S_{2\perp}$ is the lowest cross-section area, the configuration $\boldsymbol{\sigma}_\parallel = \boldsymbol{\sigma}_2$ will be selected, as this choice will result in maximal reduction of the posterior uncertainty. Analytically, $S_{n\perp}$ is calculated as the determinant of the covariance matrix $\boldsymbol{K}_{\perp\perp} = \boldsymbol{A}_{\perp\perp}^{-1}$ after removing the row and column corresponding to that configuration from the inverse covariance matrix $\boldsymbol{A}$. One of the benefits of the Bayesian approach is the ability to quantify uncertainties $\Delta E(\boldsymbol{\sigma}) = 2\left(\boldsymbol{\sigma}^\top \boldsymbol{K}_{n+1} \boldsymbol{\sigma}\right)^{\frac{1}{2}}$ and $\Delta J(\boldsymbol{\alpha}) = 2\left(\boldsymbol{\alpha}^\top \boldsymbol{K}_{n+1} \boldsymbol{\alpha}\right)^{\frac{1}{2}}$ associated to energy predictions and cluster-expansion parameters, where $\boldsymbol{\alpha} = (\pi_\alpha(\boldsymbol{\sigma}_n))_n$ is the vector representing cluster α across the configurational space. The detailed implementation of the Bayesian sampling approach is described in the supplementary information (SI). The next section (Section 3) presents its application and validation in predicting the stability of Pt:Ni binaries.

2.3. Simulations

The Quantum ESPRESSO suite for plane-wave materials simulations was used to perform the DFT calculations [36,37]. Projector-augmented-wave pseudopotentials from the library PseudoDojo were selected to represent the ionic cores [38] and the Perdew–Burke–Ernzerhof (PBE) [39] exchange–correlation functional was used to calculate the energies. The kinetic energy cutoffs for the plane waves expansion of wavefunctions and electronic charge density were set to 80 Ry and 320 Ry, respectively. To sample the Brillouin zone in reciprocal space, the k-point density was set to 0.025 Å^{-1}. Electronic occupations were smoothened using the Marzari–Vanderbilt cold smearing [40], with a smearing width of 0.01 Ry. These kinetic energy cutoffs, k-points, and smearing width were found to be sufficient to converge the total energies within 1 meV per atom and the forces within 1 meV/Å. Classical simulations were performed in the LAMMPS (large-scale atomic/molecular massively parallel simulator) software program [41]. The interatomic potentials follow the parameterization described in Refs. [2–4,42].

3. Results and discussion

To validate the Bayesian sampling, a database of 413 symmetrically unique configurations of Pt:Ni mixtures was produced, corresponding to all distinct supercells with up to eight atoms using the ICET software package [33,43]. As previously stated, our objective is to decrease the number of first-principles calculations required within an extensive training set. This approach utilizes Bayesian analysis to obtain a prior from MLPs and PBPs, facilitating the recognition of the most relevant configurations in the training set. In specific

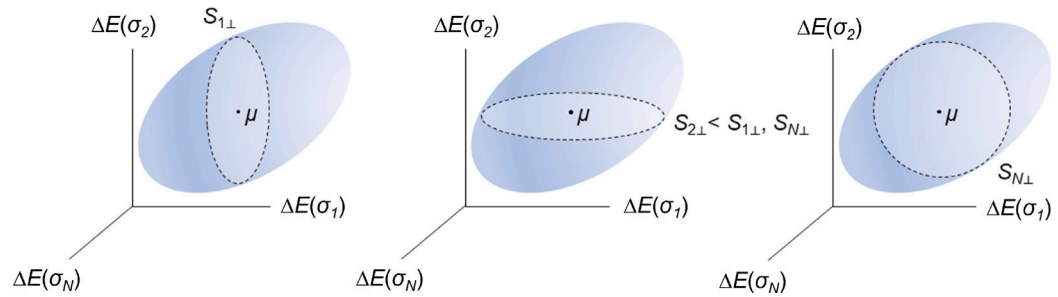

Fig. 3. Evaluation of the posterior uncertainty by examining cross sections (representing the marginal probability distributions) of the prior Gaussian distribution. The estimated uncertainty of the posterior obtained by sampling configuration σ_n equals the area $S_{n\perp}$ of the associated cross section going through the mean μ of the Gaussian distribution. The configuration to be sampled is the one minimizing the cross-sectional area (the marginal uncertainty); here, the configuration σ_2.

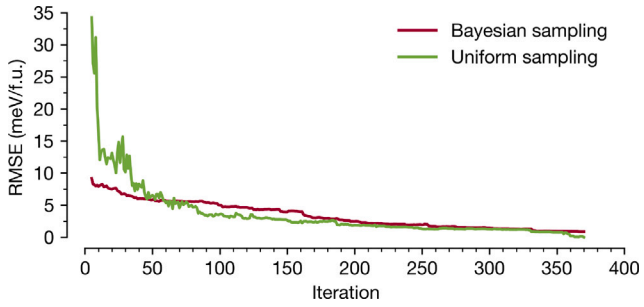

Fig. 4. Root mean squared error (RMSE) of the cluster expansion using Bayesian and uniform sampling as a function of the number of sampling iterations (the number of DFT calculations). RMSE is estimated against a cluster expansion derived from the complete dataset of fully converged DFT energies.

terms, we employed a total of eight interatomic potentials to compute the energy of all structures in our dataset. These encompassed embedded-atom potentials (EAM [3], MEAM [4]), reactive many-body potentials (COMB3 [1], ReaxFF [2]), and machine-learned potentials (CHGNet [5], M3GNet [6], ALIGNN [7], MACE [8,9]).

Our approach relies on the premise that the prior captures correlations between the formation energies of the configurations. We systematically tested this hypothesis by comparing the formation energies calculated from DFT to those computed with the interatomic potentials mentioned in the preceding paragraph. As shown in Fig. 1(a), there exist significant discrepancies between the DFT and MLP/PBP energies. However, these discrepancies do not imply that these potentials cannot provide exploitable information. In fact, upon renormalizing the empirical energies from MLPs or PBPs by the scaling factor

$$\alpha = \left[(\boldsymbol{\Delta}\tilde{E})^{\top}\boldsymbol{\Delta}\tilde{E}\right]^{-1}\left[(\boldsymbol{\Delta}\tilde{E})^{\top}\boldsymbol{\Delta}E\right] \tag{12}$$

(where $\boldsymbol{\Delta}E$ represents the DFT energies and $\boldsymbol{\Delta}\tilde{E}$ is the energy calculated using the MLP/PBP empirical potential), a close correspondence is found between the DFT and empirical trends, suggesting that the ordering of the calculated empirical energies is qualitatively consistent with its DFT counterpart [Fig. 1(b)]. In practice, the calculation of the rescaling factor is repeated for all the interatomic potentials at each iteration, allowing for the gradual improvement of the empirical trends with the progressive incorporation of new DFT energies. After few iterations, the rescaled potentials closely capture energy correlations.

Next, a cluster expansion was parameterized, incorporating clusters up to the fourth order, with cutoff distances of 10 Å for pairs, 7.5 Å for triplets, and 5 Å for quadruplets. This cluster space was composed of a total of 130 parameters, distributed as follows: 1 zerolet, 1 singlet, 17 pairs, 76 triplets, and 35 quadruplets. To analyze how the performance of the approach is affected by increasing the number of DFT calculations in each iteration, we generated a learning curve by assessing the root mean squared error (RMSE) against a cluster expansion derived solely from DFT calculations. It should be mentioned that 370 out of the initial 413 structures were successfully converged during the DFT calculations. Structures that did not converge were excluded from our interatomic potentials database.

We derived the prior distribution by utilizing eight different interatomic potentials with statistical weights representing the amount of information contained in each of them, as explained in Sec. S4 of the SI. The prior was employed in the iterative process of minimizing uncertainty to select the optimal configurations. The performance of the resulting Bayesian sampling method is compared to randomly sampling (uniform sampling) the structures in Fig. 4, which depicts the root mean squared error of the Bayesian and random cluster expansion models with respect to a cluster expansion model solely derived from DFT calculations. A notable difference in the convergence of the RMSE is observed, especially at the initial stages of iterations where Bayesian sampling leads to an immediate decrease of the RMSE. After the 60th iteration, uniform sampling seems to outperform Bayesian sampling because it may better capture configurations away from the convex hull where the accuracy of MLPs/PBPs is expected to deteriorate (as these high-energy configurations are generally less represented in MLP/PBP training). However, as shown in Fig. 5(a), the convex hull generated using uniform sampling at iteration 60 noticeably differs from the convex hull created using all available training structures, while the Bayesian convex hull is already very close to the DFT target.

To correctly assess the convergence of the convex hull, we introduce a direct metric of convex hull accuracy, the areal convex hull error (ACHE), obtained by calculating the area between the convex hulls, as depicted in Fig. 5. Changes in ACHE along the iterative cycle are reported in Fig. 5(d). A noteworthy observation is the close alignment between the Bayesian and DFT curves after 20–40 iterations, while uniform sampling requires 150–170 iterations to reach an ACHE accuracy of 3 meV. Additionally, with Bayesian sampling, the correct prediction of the convex hull is achieved after 100 iterations, whereas uniform sampling provides consistent predictions only after 250 iterations. These observations demonstrate that Bayesian sampling significantly reduces the number of DFT steps in building well-converged cluster expansions.

It is worth noting that further computational acceleration would be achieved by opting for a batch selection strategy at each iteration (as opposed to processing individual structures) and by conducting parallel DFT calculations for the selected batch. To assess the effectiveness of this approach, we conducted a test by calculating the DFT energy for 5 or 10 structures in each iteration. The results demonstrated a marginal increase in the number of DFT calculations required. With a batch of 5, the model achieved convergence after 105 DFT calculations, while for a batch of 10, convergence was attained after 110 DFT calculations. In contrast, the single-structure selection approach reached convergence after 100 DFT calculations. Therefore, it is advisable to employ batch selection to minimize computational time in generating accurate cluster expansions.

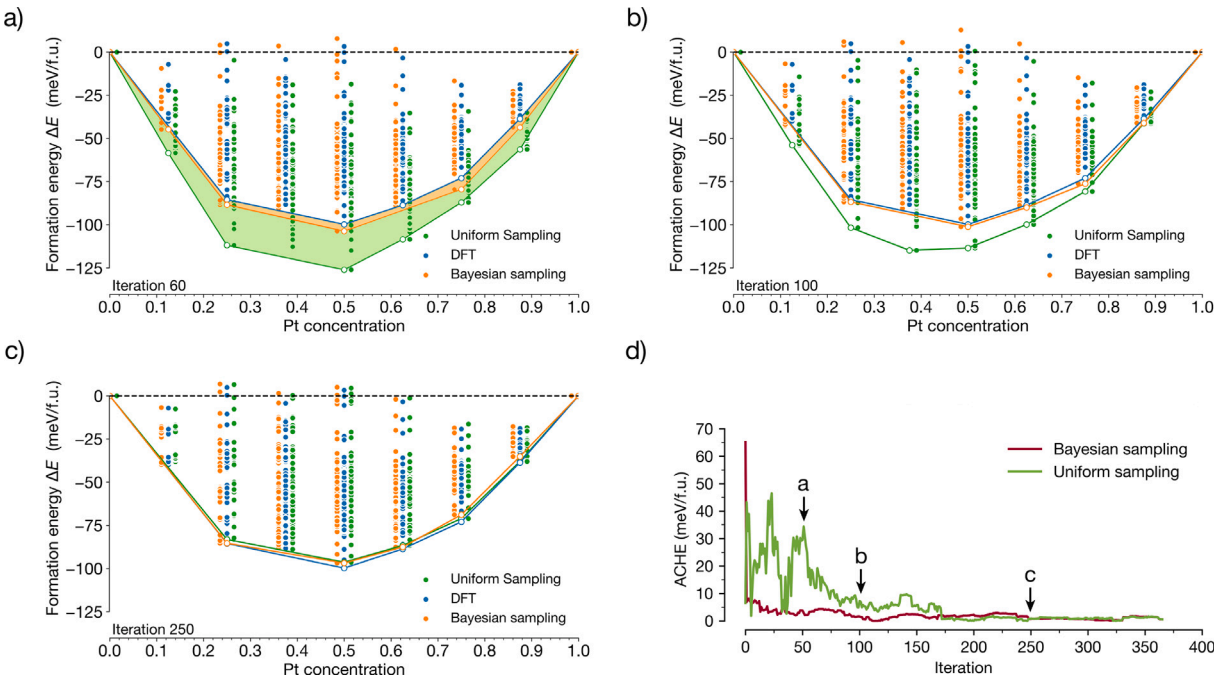

Fig. 5. Cluster expansion performance as a function of the number of sampling iterations. The energy points associated with uniform (Bayesian) sampling are slightly shifted to the right (left) of the actual concentrations for ease of comparison. Convex hull diagrams after (a) 60 iterations, (b) 100 iterations, and (c) 250 iterations. (d) Areal convex hull error (ACHE) between the Bayesian sampling and DFT, and uniform sampling and DFT as function of the number of iterations. The areal convex hull errors (ACHEs) are shown as the colored (orange and green) transparent regions in panel (a) for Bayesian and uniform sampling, respectively.

4. Conclusion

We introduced a Bayesian selection algorithm to expedite the robust parameterization of accurate cluster expansions using covariance information extracted from machine-learned and physics-based interatomic potentials. This prior enables one to identify the most informative structures within a training set for model construction. The energies of the selected structures are calculated at the DFT level. The prior is then updated by incorporating the computed DFT energies. Applying this iterative approach to a prototypical Pt:Ni alloy provided well-converged CE at a fraction of the computational cost of uniform sampling. Importantly, much lower statistical fluctuations were observed using Bayesian inference. Further acceleration was attained by selecting a batch of structures at each iteration rather than performing DFT calculations for single structures. This algorithm provides a powerful approach for future studies of multicomponent interfaces and materials at finite temperature.

CRediT authorship contribution statement

A. Dana: Formal analysis, Investigation, Methodology, Writing – original draft, Data curation, Validation, Visualization. **L. Mu:** Data curation, Formal analysis, Investigation, Validation, Writing – review & editing. **S. Gelin:** Formal analysis, Investigation, Methodology, Project administration, Validation, Writing – review & editing, Supervision. **S.B. Sinnott:** Funding acquisition, Supervision, Writing – review & editing. **I. Dabo:** Conceptualization, Formal analysis, Methodology, Validation, Visualization, Writing – original draft, Funding acquisition, Investigation, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This work was primarily supported by the U.S. Department of Energy, Office of Science, Basic Energy Sciences, CPIMS (Condensed Phase and Interfacial Molecular Science) Program, under Award No. DE-SC0018646. S.G. acknowledges support from the Center for Nanoscale Science under Grant No. DMR-2011839 (code implementation, code validation, and development of accuracy metrics). I.D. and A.D. are thankful to Paul E. Lammert for fruitful discussions on the analytical foundations of the cluster expansion method.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.commatsci.2024.113073.

References

[1] T. Liang, T.-R. Shan, Y.-T. Cheng, B.D. Devine, M. Noordhoek, Y. Li, Z. Lu, S.R. Phillpot, S.B. Sinnott, Classical atomistic simulations of surfaces and heterogeneous interfaces with the charge-optimized many body (COMB) potentials, Mater. Sci. Eng. R 74 (9) (2013) 255–279.

[2] Y.K. Shin, L. Gai, S. Raman, A.C. Van Duin, Development of a ReaxFF reactive force field for the Pt-Ni alloy catalyst, J. Phys. Chem. A 120 (41) (2016) 8044–8055, http://dx.doi.org/10.1021/acs.jpca.6b06770.

[3] S.M. Foiles, M.I. Baskes, M.S. Daw, Embedded-atom-method functions for the fcc metals Cu, Ag, Au, Ni, Pd, Pt, and their alloys, Phys. Rev. B 33 (12) (1986) 7983–7991, http://dx.doi.org/10.1103/PhysRevB.33.7983, URL https://link.aps.org/doi/10.1103/PhysRevB.33.7983.

[4] J.-S. Kim, D. Seol, J. Ji, H.-S. Jang, Y. Kim, B.-J. Lee, Second nearest-neighbor modified embedded-atom method interatomic potentials for the Pt-M (M=Al, Co, Cu, Mo, Ni, Ti, V) binary systems, CALPHAD 59 (2017) 131–141, http://dx.doi.org/10.1016/j.calphad.2017.09.005, URL https://linkinghub.elsevier.com/retrieve/pii/S0364591617301463.

[5] B. Deng, P. Zhong, K. Jun, J. Riebesell, K. Han, C.J. Bartel, G. Ceder, CHGNet as a pretrained universal neural network potential for charge-informed atomistic modelling, Nat. Mach. Intell. 5 (9) (2023) 1031–1041.

[6] C. Chen, S.P. Ong, A universal graph deep learning interatomic potential for the periodic table, Nat. Comput. Sci. 2 (11) (2022) 718–728.

[7] K. Choudhary, B. DeCost, Atomistic line graph neural network for improved materials property predictions, NPJ Comput. Mater. 7 (1) (2021) 185.

[8] I. Batatia, D.P. Kovacs, G. Simm, C. Ortner, G. Csányi, MACE: Higher order equivariant message passing neural networks for fast and accurate force fields, Adv. Neural Inf. Process. Syst. 35 (2022) 11423–11436.

[9] I. Batatia, S. Batzner, D.P. Kovács, A. Musaelian, G.N.C. Simm, R. Drautz, C. Ortner, B. Kozinsky, G. Csányi, The design space of e(3)-equivariant atom-centered interatomic potentials, 2022, http://dx.doi.org/10.48550/arXiv.2205.06643, arXiv:2205.06643.

[10] M. Eckhoff, M. Reiher, Lifelong machine learning potentials, J. Chem. Theory Comput. (2023).

[11] S. Kadkhodaei, J.A. Muñoz, Cluster expansion of alloy theory: a review of historical development and modern innovations, JOM 73 (11) (2021) 3326–3346.

[12] L.J. Nelson, V. Ozoliņš, C.S. Reese, F. Zhou, G.L. Hart, Cluster expansion made easy with Bayesian compressive sensing, Phys. Rev. B 88 (15) (2013) 155105.

[13] C. Wolverton, A. Zunger, Prediction of Li intercalation and battery voltages in layered vs. cubic Li_xCoO2, J. Electrochem. Soc. 145 (7) (1998) 2424.

[14] A. Seko, K. Yuge, F. Oba, A. Kuwabara, I. Tanaka, Prediction of ground-state structures and order-disorder phase transitions in II-III spinel oxides: A combined cluster-expansion method and first-principles study, Phys. Rev. B 73 (18) (2006) 184117.

[15] B. Kolb, G.L. Hart, Nonmetal ordering in TiC1-xNx: Ground-state structure and the effects of finite temperature, Phys. Rev. B 72 (22) (2005) 224207.

[16] A. Carlsson, J. Rosen, M. Dahlqvist, Finding stable multi-component materials by combining cluster expansion and crystal structure predictions, NPJ Comput. Mater. 9 (1) (2023) 21.

[17] Q. Wu, B. He, T. Song, J. Gao, S. Shi, Cluster expansion method and its application in computational materials science, Comput. Mater. Sci. 125 (2016) 243–254.

[18] V. Ozoliņš, C. Wolverton, A. Zunger, Cu-Au, Ag-Au, Cu-Ag, and Ni-Au intermetallics: First-principles study of temperature-composition phase diagrams and structures, Phys. Rev. B 57 (11) (1998) 6427.

[19] J. Sanchez, J. Moran-Lopez, C. Leroux, M. Cadeville, Magnetic properties and chemical ordering in Co-Pt, J. Phys.: Condens. Matter. 1 (2) (1989) 491.

[20] M. Asta, R. McCormack, D. de Fontaine, Theoretical study of alloy phase stability in the Cd-Mg system, Phys. Rev. B 48 (2) (1993) 748.

[21] A. Van der Ven, G. Ceder, Vacancies in ordered and disordered binary alloys treated with the cluster expansion, Phys. Rev. B 71 (5) (2005) 054102.

[22] L. Cao, C. Li, T. Mueller, The use of cluster expansions to predict the structures and properties of surfaces and nanostructured materials, J. Chem. Inf. Model. 58 (12) (2018) 2401–2413.

[23] A. van de Walle, G. Ceder, Automating first-principles phase diagram calculations, J. Phase Equilib. 23 (4) (2002) 348.

[24] A. Seko, Y. Koyama, I. Tanaka, Cluster expansion method for multicomponent systems based on optimal selection of structures for density-functional theory calculations, Phys. Rev. B 80 (16) (2009) 165122.

[25] T. Mueller, G. Ceder, Exact expressions for structure selection in cluster expansions, Phys. Rev. B 82 (18) (2010) 184107.

[26] E. Cockayne, A. van de Walle, Building effective models from sparse but precise data: Application to an alloy cluster expansion model, Phys. Rev. B 81 (1) (2010) 012104.

[27] T. Mueller, G. Ceder, Bayesian approach to cluster expansions, Phys. Rev. B 80 (2) (2009) 024103.

[28] L.J. Nelson, G.L. Hart, F. Zhou, V. Ozoliņš, et al., Compressive sensing as a paradigm for building physics models, Phys. Rev. B 87 (3) (2013) 035125.

[29] R. Tibshirani, Regression shrinkage and selection via the lasso, J. R. Stat. Soc. Ser. B Stat. Methodol. 58 (1) (1996) 267–288.

[30] D. Packwood, et al., Bayesian Optimization for Materials Science, Springer, 2017.

[31] M. Todorović, M.U. Gutmann, J. Corander, P. Rinke, Bayesian inference of atomistic structure in functional materials, NPJ Comput. Mater. 5 (1) (2019) 35.

[32] J. Sanchez, T. Mohri, Approximate solutions to the cluster variation free energies by the variable basis cluster expansion, Comput. Mater. Sci. 122 (2016) 301–306.

[33] M. Ångqvist, W.A. Muñoz, J.M. Rahm, E. Fransson, C. Durniak, P. Rozyczko, T.H. Rod, P. Erhart, ICET–a Python library for constructing and sampling alloy cluster expansions, Adv. Theory Simul. 2 (7) (2019) 1900015.

[34] C.E. Rasmussen, C.K. Williams, Gaussian Processes for Machine Learning, Vol. 1, Springer, 2006.

[35] A. Gelman, J.B. Carlin, H.S. Stern, D.B. Rubin, Bayesian Data Analysis, Chapman and Hall/CRC, 1995.

[36] P. Giannozzi, S. Baroni, N. Bonini, M. Calandra, R. Car, C. Cavazzoni, D. Ceresoli, G.L. Chiarotti, M. Cococcioni, I. Dabo, et al., QUANTUM ESPRESSO: a modular and open-source software project for quantum simulations of materials, J. Phys.: Condens. Matter. 21 (39) (2009) 395502.

[37] P. Giannozzi, O. Andreussi, T. Brumme, O. Bunau, M.B. Nardelli, M. Calandra, R. Car, C. Cavazzoni, D. Ceresoli, M. Cococcioni, et al., Advanced capabilities for materials modelling with Quantum ESPRESSO, J. Phys.: Condens. Matter. 29 (46) (2017) 465901.

[38] F. Jollet, M. Torrent, N. Holzwarth, Generation of Projector Augmented-Wave atomic data: A 71 element validated table in the XML format, Comput. Phys. Comm. 185 (4) (2014) 1246–1254.

[39] J.P. Perdew, K. Burke, M. Ernzerhof, Generalized gradient approximation made simple, Phys. Rev. Lett. 77 (18) (1996) 3865.

[40] N. Marzari, D. Vanderbilt, A. De Vita, M. Payne, Thermal contraction and disordering of the Al (110) surface, Phys. Rev. Lett. 82 (16) (1999) 3296.

[41] S. Plimpton, Fast parallel algorithms for short-range molecular dynamics, J. Comput. Phys. 117 (1) (1995) 1–19, http://dx.doi.org/10.1006/JCPH.1995.1039.

[42] J.A. Martinez, A. Chernatynskiy, D.E. Yilmaz, T. Liang, S.B. Sinnott, S.R. Phillpot, Potential optimization software for materials (POSMat), Comput. Phys. Comm. 203 (2016) 201–211, http://dx.doi.org/10.1016/j.cpc.2016.01.015.

[43] G.L. Hart, R.W. Forcade, Algorithm for generating derivative structures, Phys. Rev. B 77 (22) (2008) 224115.