

<https://doi.org/10.1038/s41524-024-01321-7>

Machine learned interatomic potentials for ternary carbides trained on the AFLOW database



Josiah Roberts¹, Biswas Rijal¹, Simon Divilov^{2,3}, Jon-Paul Maria⁴, William G. Fahrenholtz⁵, Douglas E. Wolfe⁴, Donald W. Brenner⁶, Stefano Curtarolo^{2,3} & Eva Zurek¹ ✉

Large-density functional theory (DFT) databases are a treasure trove of energies, forces, and stresses that can be used to train machine-learned interatomic potentials for atomistic modeling. Herein, we employ structural relaxations from the AFLOW database to train moment tensor potentials (MTPs) for four carbide systems: CHfTa, CHfZr, CMoW, and CTaTi. The resulting MTPs are used to relax ~6300 random symmetric structures, and are subsequently improved via active learning to generate robust potentials (RP) that can relax a wide variety of structures, and accurate potentials (AP) designed for the relaxation of low-energy systems. This protocol is shown to yield convex hulls that are indistinguishable from those predicted by AFLOW for the CHfTa, CHfZr, and CTaTi systems, and in the case of the CMoW system to predict thermodynamically stable structures that are not found within AFLOW, highlighting the potential of the employed protocol within crystal structure prediction. Relaxation of over three hundred ($\text{Mo}_{1-x}\text{W}_x$)C stoichiometry crystals first with the RP then with the AP yields formation enthalpies that are in excellent agreement with those obtained via DFT.

Machine-learned interatomic potentials (ML-IAPs), which are trained on density functional theory (DFT) data, have irrevocably changed the way in which computational materials science is carried out. They have increased the complexity of the systems that can be studied via static calculations¹, the duration of molecular dynamics trajectories², and made it possible to routinely model effects, such as anharmonicity³, which are commonly neglected due to the large computational cost associated with the requisite DFT calculations. Moreover, they are becoming increasingly important in the field of crystal structure prediction (CSP) where it may be necessary to relax hundreds or thousands of structures to find the global minimum and a handful of low-lying local minima⁴. As CSP methods tackle systems with increased combinatorial complexity such as ternaries and quaternaries⁵, it is becoming increasingly important to develop protocols that can be used to train and employ ML-IAPs for CSP.

Some of the most well-known ML-IAPs include neural networks (NNs)^{6,7}, the spectral neighbor analysis potential (SNAP)⁸, moment tensor potentials (MTPs)⁹, the Gaussian approximation potential (GAP)¹⁰, and, more recently ultra-fast (UF)¹¹, and ephemeral data-derived (EDD) potentials¹². Provided an ML-IAP can accurately describe the potential energy surface (PES) of a multicomponent system, it can significantly

accelerate a CSP search while at the same time enabling exploration of wider regions of compositional space⁴. Indeed, recent algorithmic developments have shown promising results: CSP methods including random and evolutionary searches, as well as particle-swarm optimization have been coupled with various ML-IAPs and applied to predict crystalline structures of boron^{13–15}, carbon^{14,16}, sodium¹⁴, phosphorus¹⁷, lithium¹⁸ as well as Mg–Ca¹⁹ and various metal–tin^{20,21} alloys at 1 atm and under pressure. For ternary crystalline systems, early work showed that adaptive classical potentials can accelerate CSP in the Mg–Si–O system²², and recently ephemeral deep learning potentials combined with random searching have been used to predict candidate structures for a (disputed) Lu–N–H superconducting phase²³, and $\text{Zn}(\text{CN})_2$ metal–organic-frameworks²⁴.

A challenge in developing ML-IAPs for CSP is that they need to be able to predict the energies of unstable structures, as well as those that are close to the local minima in the PES. Moreover, because ML-IAPs are poor at extrapolation, they cannot accurately predict the energies (or forces or stresses) of configurations that differ significantly from those they have seen before. Traditionally, DFT datasets containing thousands to hundreds of thousands of structures have been generated for the training and testing of ML-IAPs^{14,25}. A wide range of procedures have been employed to create

¹Department of Chemistry, State University of New York at Buffalo, Buffalo, NY, USA. ²Department of Mechanical Engineering and Materials Science, Duke University, Durham, NC, USA. ³Center for Autonomous Materials Design, Duke University, Durham, NC, USA. ⁴Department of Materials Science and Engineering, The Pennsylvania State University, University Park, PA, USA. ⁵Department of Materials Science and Engineering, Missouri University of Science and Technology, Rolla, MO, USA. ⁶Department of Materials Science and Engineering, North Carolina State University, Raleigh, NC, USA. ✉e-mail: ezurek@buffalo.edu

these DFT datasets including generating structures randomly¹⁴, perturbing the geometries of such structures via “shaking”¹², ab initio molecular dynamics runs at various temperatures²⁶, decorating predefined lattices with atoms of different types while simultaneously varying the chemical composition²⁵, relaxation of structures generated via constrained evolutionary searches⁷, straining crystal lattices, creating defect structures, and more^{26,27}.

Unfortunately, even when large DFT datasets are employed for training, it is unlikely that the resulting ML-IAPs can predict, with sufficient accuracy, the energies of the various structures encountered in the course of a CSP run. One strategy that has been proposed for the generation of a multi-purpose ML-IAP, given a limited number of DFT calculations, relies on the automatic iterative building of the fitting database by selecting the most diverse structures²⁸. Additional techniques include various active learning^{29,30} and learning-on-the-fly³¹ methods, where the potential is updated and improved during the course of the search. It has been suggested that active learning could be used to generate two ML-IAPs: a robust one that is able to optimize any structure the CSP algorithm encounters and make rough predictions, and an accurate one trained on, and used for, only the low-energy structures^{14,25}.

ML-IAP-based simulations where the potential is updated on-the-fly may be initialized using either an empty/untrained potential²⁵, or one that has been pre-trained. The former strategy may require just as many, if not more, DFT evaluations than the latter because the likelihood of encountering a configuration that is deemed extrapolative is high, necessitating a retraining of the ML-IAP¹⁴. Thus, there appears to be “no-free-lunch” since the construction of a reliable ML-IAP requires numerous expensive DFT calculations. At the same time, a large number of databases exist—AFLOW (Automatic FLOW)^{32–34}, the Materials Project (MP)³⁵, the Open Quantum Materials Database³⁶, etc.—each containing millions of DFT evaluations of the energies, forces and stresses of extended systems. Furthermore, it is becoming standard practice for researchers to deposit the DFT data generated during the course of a computational project in repositories such as NOMAD³⁷, OCELOT³⁸, and NIST Materials Data³⁹. One way this data is being used is to train ML-IAPs for the computational study and exploration of the vast PES of all possible chemistries. Some of the forefront examples of such “universal” ML-IAPs, which can predict energies, forces, and stresses using equivariant graph neural networks, include M3GNet⁴⁰, CHGNet⁴¹, ALIGNN-FF⁴², MACE-MP-0⁴³, and GNoME⁴⁴.

The training of potentials on already-existing DFT data is illustrated here by combining outputs present within the AFLOW³² database with MTPs⁹, as implemented within the Machine-Learned Interatomic Potentials (MLIP) program⁴⁵. Specifically, chemically sensible structures, which are randomly selected from the relaxation trajectories stored within the AFLOW database, are used to train an MTP that is subsequently employed to relax a large number of random symmetric structures spanning a wide composition range. In contrast to the recently developed universal ML-IAPs^{40–44,46}, here we only train on a subset of the data found within AFLOW, chosen for the application in mind. Therefore, the ML-IAPs we develop are system-specific, and not universal. What distinguishes our study from prior works^{7,12,14,26,27} is that rather than generating our own DFT-training set, we employ already-existing data found within AFLOW. In a further step, the AFLOW-trained potentials are improved via active learning, generating ML-IAPs that can be robust (for rough optimizations of any configuration) or accurate (for more precise optimizations of low-energy structures near the convex hull). Thus, only a small number of supplementary DFT calculations are required to develop system-specific MLIPs that enable the computational exploration of evermore complex PESs toward the discovery of materials.

A utility package that automates this training process, the *Plan for Robust and Accurate Potentials* (PRAPs), is described. The method is used to determine the zero-Kelvin phase diagrams of four ternary metal carbides, chosen because they represent materials with superlative mechanical properties^{47,48}. The PRAPs pre-training on AFLOW improves the robust potential predictions. The convex hulls relaxed with the accurate potential

are generally in good agreement with the hulls found within AFLOW, but relaxation of the low-energy structures from both datasets with DFT improves the agreement. Moreover, in the case of the CMoW system, thermodynamically stable structures are found that are not present in the AFLOW data. Further calculations with the accurate potential find a variety of $(\text{Mo}_{1-x}\text{W}_x)\text{C}$ stoichiometry phases at/near the tie-line indicating the possibility of a solid solution with a very low miscibility gap critical temperature.

Results and discussion

Plan for robust and accurate potentials (PRAPs)

We created MTPs of varying complexity for a number of ternary metal carbides and investigated their capabilities to predict structures outside of their training sets. Our choice of MTPs was motivated by their excellent balance between model accuracy and computational efficiency^{26,49}, their application towards multicomponent systems^{25,27,50–52}, their ability to predict phonons and thermodynamic properties^{3,53,54}, and the availability of a powerful active learning scheme (ALS) interfaced with the MTP method^{25,55}.

The MLIP software package trains MTPs and uses them to relax the geometries and minimize the energies of a wide variety of chemical systems⁴⁵. The simplest form of training, basic training, employs the energy, force and stress (EFS) data of a set of configurations, as obtained from DFT or other quantum chemical calculations, to generate an MTP. The complexity of the MTP is described by a user-selected level, a notation containing information about the number of basis functions and parameters comprising the potential. The ALS employs a D-optimality criterion to calculate the extrapolation degree or grade, γ , for every structure that is generated throughout the course of the simulation (relaxation trajectory or molecular dynamics run)^{25,55}. MLIP automatically selects configurations to be added to the training set if their γ falls within a user-defined range; we choose the default $2 < \gamma \leq 10$. The calculation (relaxation or molecular dynamics run) is terminated if γ exceeds the upper bound, triggering retraining of the MTP, and the procedure repeats until the simulation finishes with $\gamma \leq 10$. Further information about MTPs, including their functional form, the quantities included in the cost function, and details of the active learning procedure are provided in the Supplementary Information Section 2.1.

In what follows we give a brief overview of the PRAPs utility package employed in this study; a forthcoming manuscript will describe its composition and usage. PRAPs interfaces with MLIP and automates the creation of a Robust Potential (RP) and Accurate Potential (AP) using the aforementioned ALS, and employs basic training for supplemental tasks. From a given a set of configurations obtained from AFLOW, a subset of ~800 configurations is chosen randomly and used to train an MTP (Fig. 1, gray box). This procedure is repeated five times, generating five different MTPs (as recommended in the MLIP manual⁴⁵), mimicking a cross-validation procedure. From these PRAPs finds the “best” MTP, defined as the one that identifies the most high-and-low-energy structures (ten of each are compared), or if multiple MTPs fulfill this criteria, the MTP with the lowest training root-mean-squared error (RMSE). This step is intended to mitigate the initial random parameters that MLIP assigns at the beginning of training, but users may bypass it if they so desire. To ensure that training is performed on sensible structures, PRAPs can filter out configurations with undesirable interatomic distances (default), or cell volumes. If relaxation trajectories comprise the dataset, dozens of ionic steps originating from the same system may be present. Though using many similar configurations for training may be beneficial in some cases, in others it may be desirable to exclude the intermediate steps and only include the final relaxed configuration, and PRAPs provides an option for users to select the desired behavior.

The RP training (Fig. 1, blue box) begins using the PRAPs-determined “best” MTP. The set of configurations may be optionally augmented with structures lacking any EFS data (Fig. 1, left-hand-side, no box), which are combined with the initial DFT dataset to form the relaxation set. The relaxation set is used to train the RP by active learning. The active learning

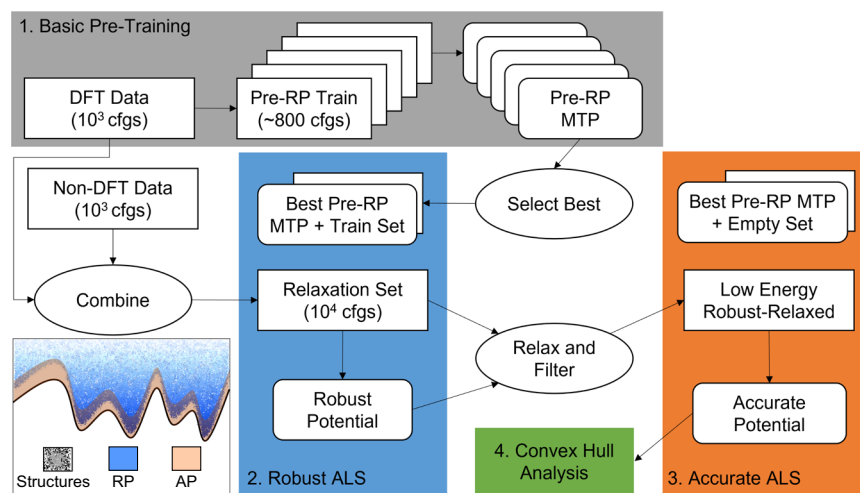


Fig. 1 | Workflow in the Plan for Robust and Accurate Potentials (PRAPs) package. PRAPs automates the generation of a moment tensor potential (MTP), given a set of data generated from quantum-mechanical calculations (here the online AFLOW database³²), and/or a pre-generated set of structures (here RANDSPG⁷⁰). Five MTPs are trained on a set of configurations (cfs) and the best (the pre-Robust Potential, pre-RP) is chosen (gray box). The pre-RP potential and training set are employed to initialize the training of the RP via an active learning scheme (ALS, blue box). This requires iterations of relaxing and retraining on structures in, and chosen from, the Relaxation Set. After training, the RP relaxes

everything in the Relaxation Set, and the results are filtered to retain only the lowest-energy configurations in each composition (Low-Energy Robust Relaxed). This is then used to train the Accurate Potential (AP) via active learning (orange box). The AP-relaxed structures may be sent for relaxation via AFLOW-DFT, followed by subsequent analysis (green box). The inset schematically illustrates the energy distribution of the structures that can be relaxed with the RP and the AP. Rectangles with sharp corners represent structural data, while curved corners represent potentials.

begins with the “best” training set and MTP from the pre-training step (if performed, otherwise PRAPs trains from scratch). During the active learning process, the relaxation set is relaxed using MLIP’s built-in functionality and additional structures are added to the training set. The final output is an RP capable of performing a reasonable relaxation of most any configuration in the relaxation set.

After relaxation with the RP, PRAPs filters out structures with energies above a specified cutoff (default 50 meV/atom) for each composition present. This creates the “Robust Relaxed” set, which is employed to train the AP via active learning (Fig. 1, orange box). The procedure for training the AP is similar to the RP except that MLIP begins with an empty training set instead of the set of structures that were used to train the “best” pre-trained MTP. We find that this results in better predictions as the AP should not see high-energy structures during training, thereby improving the EFS predictions for the low-energy structures. For additional efficiency, users can alter the convergence criteria of the AP training at the cost of a few meV/atom in the training error (as described more thoroughly below).

PRAPs performs a number of analyses throughout and at the end of the training procedure. In addition to calculating the training and prediction root-mean square and mean absolute error (MAE), and comparing the ten high-and-low-energy structures, it can generate a set of convex hull candidates, invoke AFLOW^{32,56} to relax them, and use this data to generate composition-energy convex hull plots (Fig. 1, green box). Finally, PRAPs contains a checkpoint system to allow users to start or re-start a job from a certain step. PRAPs is primarily a project management software that performs and automates many menial tasks, and generates plots that may be desired, thereby reducing the amount of human time required to generate MTPs for a particular system and analyze their performance.

In what follows, we seek to answer two questions using PRAPs as applied to four ternary alloys: (i) Does pre-training a robust MTP on already-existing quantum-mechanical data improve its ability to identify low-energy structures? (ii) And, can we subsequently generate an MTP designed for these low-energy structures via active learning, then use it to discover thermodynamically stable systems comprising the convex hull that are not present in the DFT-training set? In what follows, we show the answers to both questions is “yes”.

Machine-learned interatomic potentials from AFLOW data

To illustrate how already-existing DFT results can be scraped from large databases and repurposed towards the generation of system-specific MLIPs, we chose four ternary alloy systems: CHfTa, CHfZr, CTaTi, and CMoW. TaC, HfC, ZrC, and TiC all adopt the rocksalt structure and are refractory ceramic materials with desirable mechanical properties⁵⁷. Though MoC can adopt the same rocksalt structure under pressure⁵⁸, at ambient conditions MoC and WC prefer a hexagonal arrangement instead⁵⁹. For nearly a century, the propensity for transition metal carbides to form high-melting-point solid solutions with compositions such as $(\text{Hf}_{1-x}\text{Ta}_x)\text{C}$ has been known^{60,61}. The metal formulation can be engineered to contain multiple atoms, and when there are five or more metals, the configurational entropy stabilizes single-phase high entropy carbides (HECs)^{47,62,63} and their thin films⁴⁸. Recently, ML-IAPs have been developed for various HECs including a deep learning potential for $(\text{ZrHfTiNbTa})\text{C}_5$ ⁶⁴ and a low-rank potential for $(\text{TiZrNbHfTa})\text{C}_5$ ⁶⁵. Herein, we illustrate that data found within AFLOW, coupled with active learning, can be used to train MTPs for ternary carbides, with the future goal of HECs in mind.

For training, we employed DFT relaxation trajectories of ~210 structures obtained from the AFLOW database, which have been generated through a combination of structure prototyping of naturally occurring compounds^{66–68} and structure enumeration algorithms⁶⁹, followed by relaxation using DFT. This resulted in ~5500–6000 individual configurations on which the MTP was trained (Table 1). From this training set, ~800 configurations, with a minimum interatomic distance greater than 1.1 Å, were chosen randomly and MTPs of levels 10, 16 and 22 were trained in the “basic” mode. For the CMoW system, no level 22 data is reported due to the excessive computational cost required to obtain well-trained potentials. Five trainings were performed, and the best potential was chosen as the pre-Robust Potential (pre-RP). As expected, the training errors for the pre-RP (Supplementary Table 1) decreased with increasing MTP level, with the average MAE (RMSE) for the energies being 27 (44), 16 (25), and 8 (13) meV/atom, and for the forces being 75 (217), 51 (147), and 26 (78) meV/Å for levels 10, 16, and 22, respectively. A comparison of the ten highest and ten lowest-energy structures as predicted by DFT and the pre-RP revealed that the MTP rarely miscategorized the structures, but had

Table 1 | Number of unique structures for the studied ternary carbides found within the AFLOW database, and the number of individual ionic steps (configurations) from their structural relaxations

System	AFLOW structures	Configurations	RANDSPG structures
CHfTa	209	6064	6346
CHfZr	210	5811	6356
CMoW	210	5524	6338
CTaTi	211	5831	6347

The number of individual structures generated by RANDSPG that were employed to develop the robust and accurate potentials using the MLIP program package are also provided.

Table 2 | Prediction errors for energies (meV/atom) and forces (meV/Å) by system, level, and potential

Alloy	Energy errors (meV/atom)		Force errors (meV/Å)	
Level 10	Robust potential	Accurate potential	Robust potential	Accurate potential
CHfTa	45 (75)	24 (36)	89 (281)	82 (270)
CHfZr	55 (87)	37 (62)	81 (255)	57 (170)
CMoW	53 (105)	88 (155)	99 (318)	88 (273)
CTaTi	51 (85)	53 (70)	99 (303)	76 (220)
Level 16	Robust potential	Accurate potential	Robust potential	Accurate potential
CHfTa	42 (72)	29 (90)	79 (238)	53 (169)
CHfZr	43 (77)	22 (30)	68 (211)	42 (112)
CMoW	46 (96)	76 (129)	88 (278)	55 (167)
CTaTi	41 (60)	45 (59)	90 (285)	56 (161)
Level 22	Robust potential	Accurate potential	Robust potential	Accurate potential
CHfTa	44 (70)	21 (57)	85 (270)	23 (71)
CHfZr	39 (62)	29 (91)	61 (197)	18 (59)
CTaTi	36 (58)	82 (218)	73 (223)	25 (79)
Level 16*	Robust potential	Accurate potential	Robust potential	Accurate potential
CHfTa	126 (156)	16 (19)	129 (381)	50 (150)
CHfZr	78 (107)	24 (37)	93 (278)	40 (114)
CMoW	91 (131)	67 (120)	108 (336)	52 (196)
CTaTi	119 (153)	92 (112)	110 (345)	45 (134)

Each box contains the mean absolute-error (MAE) and, in parentheses, the root-mean-squared error (RMSE). Only the Robust Potential (RP) and the Accurate Potential (AP) developed during the active learning scheme (ALS) as illustrated in Fig. 1 are shown here; training data for the pre-Robust Potential is provided in Supplementary Table 1. The errors for the RP are calculated by comparing the energies of the final relaxed structures found within AFLOW to their RP-predicted energies and forces. The errors for the AP are calculated in a similar fashion, except only the structures whose AFLOW energies are within 50 meV/atom of the most stable configuration for a particular stoichiometry were considered.

*Calculations performed without pre-training the MTP on the AFLOW data (gray box in Fig. 1).

difficulty correctly categorizing more than 7 in the correct highest or lowest-energy set (Supplementary Table 2).

Previous studies have suggested that when used for CSP, MTPs should be trained on a very diverse set of structures^{14,25}, and as Table 1 shows, the original AFLOW data was somewhat limited. To obtain this diversity, we used the RANDSPG⁷⁰ program to create crystal lattices with up to eight atoms in the unit cell, for all possible ternary compositions. Lattice vectors were constrained to fall between 3 and 10 Å, and unit cell volumes between 200 and 600 Å³, with a minimum interatomic distance of 1.1 Å. For each ternary system, ~6300 structures were generated (Table 1) and combined

with the initial AFLOW data to form the relaxation set (Fig. 1) used in training the RP.

For a given stoichiometry RANDSPG determines the compatible spacegroups, based upon the Wyckoff positions, and randomly chooses one prior to decorating its sites with atoms, thereby enabling the creation of random symmetric crystal lattices. Employing symmetric structures in the first generation of an evolutionary or particle-swarm-directed CSP search greatly decreases the number of configurations that need to be optimized to locate the global minimum in the PES. The reason for this is that symmetric structures tend to be either very stable or unstable, spanning a greater amount of the potential energy hypersurface than those generated without symmetry constraints⁵. Indeed, tests have shown that the average energy of random structures that are symmetric is higher (less negative) than those that are purely random⁷⁰. The training errors for the RP and AP, provided in Supplementary Table 1, are significantly larger than the pre-RP likely due to the diversity of the RANDSPG-created structures, and the propensity of the D-optimality criterion to choose the most diverse structures for training.

How do we determine the prediction error for the active learning-derived-potentials since each level of theory (DFT or MTP) encountered different structures during the relaxation process? We initially tried predicting the energy of every structure comprising the AFLOW relaxation trajectories via the RP or the AP. Noise present in the first steps of the relaxation trajectory, which may have erroneous EFS that result from changes in the plane-wave basis during variable cell optimizations, made this problematic. Moreover, the AFLOW data contained a few configurations whose per-atom-energies (magnitude) were substantially larger than others—some surviving the distance-filtering-criteria. Therefore, in Table 2 we have opted to compare the DFT energies and forces of the final AFLOW relaxed structures against their RP and AP predicted values for different MTP levels. Plots comparing the MTP-predicted energies against the DFT data found within AFLOW are provided in Supplementary Figs. 2–16. The prediction errors with the RP generally decreased with increasing MTP level, but the MAE did not fall below 40 meV/atom for energies and 60 meV/Å for forces. While these errors might seem large compared to the < 4 meV/atom and < 160 meV/Å RMSEs computed for single-component systems wherein the testing and training dataset both contained structures that could be derived via perturbations of the ground-state crystal²⁶, they are in-line with the some of the errors presented in ref. 27 where MTPs for Li–Al alloys were developed and applied to a broad range of compositions and lattices.

A common practice in CSP is to use less accurate, but quicker methods to estimate the energies of a large number of structures, and then to optimize those with low energies with progressively more accurate, but costlier, methods^{23,24}. This workflow saves computational expense by filtering out structures that are unlikely to be stable prior to performing calculations that yield a more predictive rank order. A similar strategy was employed in refs. 25 and 14 where both robust and accurate MTPs were trained for CSP. In addition to generating 375,000 binary and ternary bcc, fcc, and hcp-type unit cells, 1463 Al–Ni–Ti ternary structures were created via decorating prototypes²⁵. Because the prototype-derived structures could have large prediction errors arising from geometries with short metal-metal distances and too-small volumes, structures with formation energies (as determined with an RP) that were within 100 meV/atom of the convex hull were chosen for re-relaxation using active learning starting from an empty MTP to generate an AP. The training set MAEs (RMSEs) were 18 (27) meV/atom for the RP and 7 (9) meV/atom for the AP, but prediction errors were not reported. In ref. 14, a RP yielded a training RMSE of 170 meV/atom for allotropes of boron, and an AP yielded errors of 11 meV/atom taking into consideration the 100 lowest-energy structures that were found.

A key difference between our workflow and that of Gubaev and co-workers²⁵ is the pre-training step on the AFLOW data. To test what effect this may have, we used PRAPs to develop level 16 MTPs without performing this pre-training step (denoted schematically in the gray box in Fig. 1). Table 2 illustrates that the RP-energy MAEs were significantly improved by the AFLOW pre-training, and the RMSEs were slightly improved (average difference of 60 and 79 meV/atom). For the RP, the errors obtained for the

forces were smaller when the pre-training was performed. Therefore, the AFLOW pre-training helps develop RPs able to correctly identify most low-energy configurations, and not misidentify them as having high energies, thereby curating a new set of training data that can then be used to create an AP. One way to gauge this is by the MAEs, which frequently fell below 50 meV/atom. A more definitive way, discussed in detail below, is by examining the convex hulls generated during the PRAPs procedure. The AFLOW pre-training does not impact the AP errors directly since the initial training set for the AP is intentionally chosen to be empty, but it can influence which structures are chosen for training.

For an MTP level of 22, the training in some cases took substantially longer than at lower levels. The reason for this is that the ALS, as originally designed⁵⁵, terminates when MLIP does not find any configurations that need to be added to the training set. Here a different protocol was used, where the active learning for the AP was stopped when the number of structures to be added to the training set was less than 1% of its size. This choice was motivated by the observation that at times fewer than ten configurations were being added to a training set of thousands in a single iteration, which took a full day to process, for a gain in RMS training error that was <1 meV/atom. Tests showed that the choice of a variety of early termination criteria (which can be chosen as options in PRAPs) comes at the cost of 2–5 meV/atom in training error.

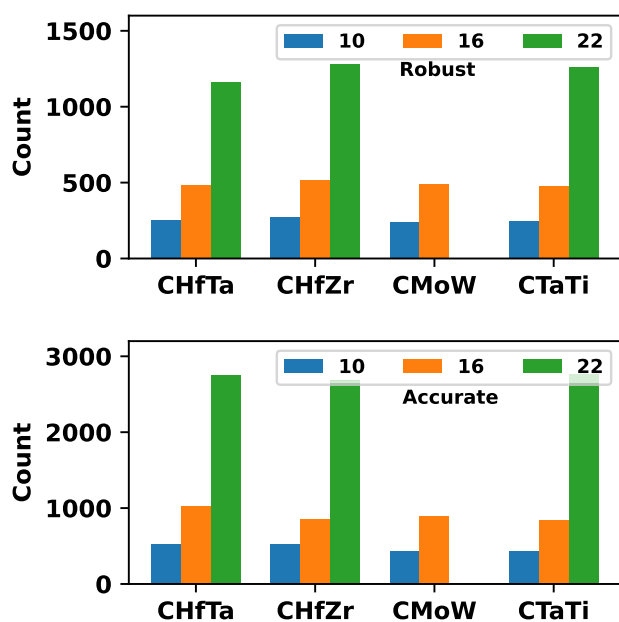


Fig. 2 | Density functional theory (DFT) calculations required for model training. Number of DFT single-point energy evaluations performed during the generation of the robust (top) and accurate (bottom) potentials according to the procedure illustrated in Fig. 1 for each ternary carbide system at a particular MTP level (see legend).

A key question to be answered is: “Does the ML-aided procedure reduce the total computational expense?” For each ternary the AFLOW and RANDSPG datasets contained ~210 and ~6300 individual structures, respectively (Table 1). Therefore, ~6500 geometry optimizations would be needed to relax all of these configurations. Using MTPs, a maximum number of single-point DFT energy evaluations performed during the course of the training was ~3000 (Fig. 2) for the CTaTi system at level 22. Dividing the number of configurations comprising the AFLOW data by the number of structures gives an average estimate of the number of steps required per geometry optimization (~28). Thus, our protocol reduces the number of DFT evaluations that would need to be performed to relax all of the considered structures by a factor of ~30 or more (as high as 180 for the CHfTa system at an MTP level of 10). We note that as the MTP level increases, the number of single-point calculations increases (as does the total training time), as expected (e.g., see Table 3 in ref. 45). The reason for this is that higher-level MTPs contain more parameters, and therefore require more training data to fit these parameters.

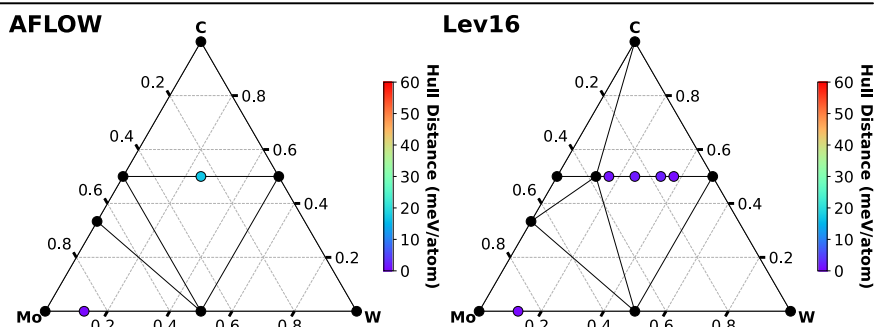
Predicted convex hulls and solid solution-forming ability

Let us now examine the convex hulls and investigate the structures that PRAPs relaxed with the robust and accurate potentials. Since optimization of random symmetric configurations is the first step in CSP, we examined if the aforementioned workflow could discover lattices not found within AFLOW whose energies lie on, or close to the convex hull. We filtered out the AP predicted configurations that were within 50 meV/atom of the lowest-energy structure for each composition and DFT-relaxed them via AFLOW. The resulting geometries were then concatenated with the fully relaxed AFLOW data to produce convex hulls, for different MTP levels (e.g., Fig. 3).

The data found within AFLOW for the CHfTa, CHfZr, and CTaTi systems all contained multiple compounds on or near the convex hull with $(M_{1-x}N_x)_3C$ stoichiometries ($x = 0.25, 0.33, 0.5, 0.67, 0.75$) (Supplementary Figs. 17–27). Examination of these structures suggested that they could all be obtained via relaxation of metal-carbide rocksalt structures whose metal sites were decorated with two different types of atoms, as would be expected for a solid solution. On the other hand, for the CMoW system, AFLOW did not contain any ternary carbide phases with a $(Mo_{1-x}W_x)_3C$ composition that were on the hull (Fig. 3). Unlike the cubic binary carbides comprised of metal atoms from group 5 or 6, isostructural MoC and WC (Fig. 4a) adopt the hexagonal $P6m2$ (#187) spacegroup, suggesting that $(Mo_{1-x}W_x)_3C$ stoichiometry structures would be hexagonal as well. Examination of a $Imm2$ symmetry $Mo_{0.5}W_{0.5}C$ phase that was 17 meV/atom above the convex hull (teal dot in Fig. 3) showed that it could not be derived from a decoration of a hexagonal metal-carbide lattice. Instead, it was related to a high-pressure phase of GaAs (AFLOW prototype [AB_014_44_a_b](#)) where the Ga atoms were replaced by C, half of the As atoms were substituted by Mo, and the other half by W.

In addition to the elemental endpoints, as well as MoC and WC, cubic MoW and rhombohedral C_2Mo_4 (Fig. 3) comprise the AFLOW convex hull for the CMoW system. Tetragonal $Mo_{14}W_2$ (labeled by a purple dot) lies 1 meV/atom above the hull—a value that is within the error of the k -mesh and kinetic energy cutoffs employed in our plane-wave calculations. A

Fig. 3 | A sample of convex hull diagrams produced by PRAPs for CMoW. Convex hulls obtained by DFT-optimizing structures predicted to be within 50 meV/atom of the convex hull obtained from AFLOW (left) and the PRAPs procedure at an MTP level of 16 (right) for the CMoW system. The black dots are structures lying on the convex hull, and phases within 60 meV/atom of the hull are represented by colored dots (see color bar). Purple dots are within 1 meV/atom of the hull.



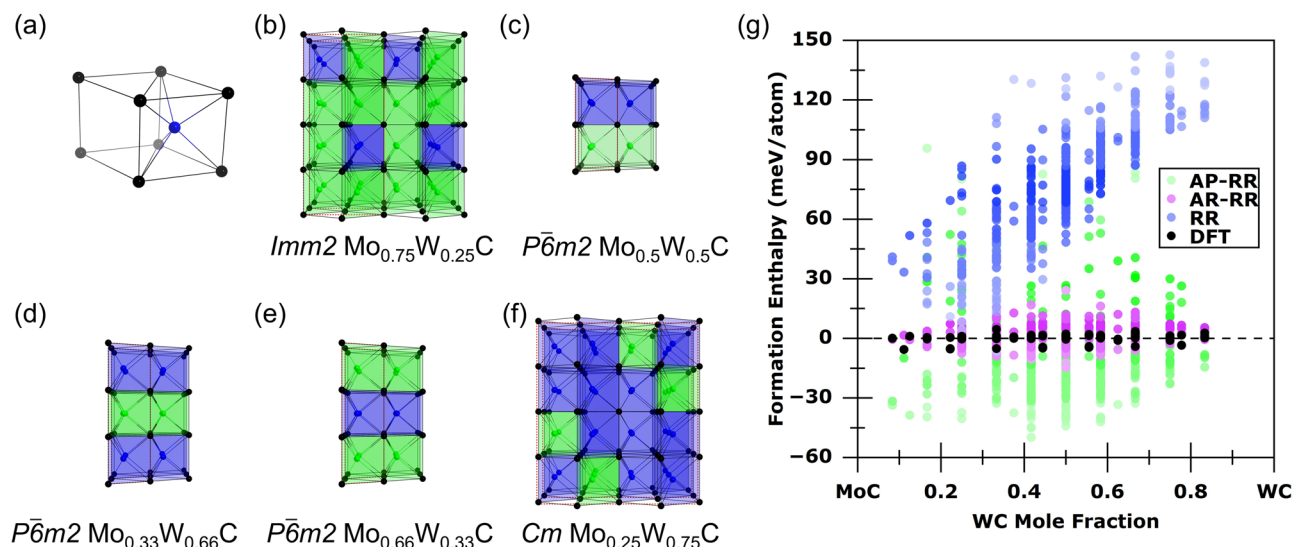


Fig. 4 | Crystal structures of CMoW compounds on or near the convex hull, and formation enthalpies of recolorings of those structures. Crystal structures of (a) parent phases: $P\bar{6}m2$ WC (isostructural with MoC), (b) $Imm2$ $Mo_{0.75}W_{0.25}C$, (c) $P\bar{6}m2$ $Mo_{0.5}W_{0.5}C$, (d) $P\bar{6}m2$ $Mo_{0.33}W_{0.66}C$, (e) $P\bar{6}m2$ $Mo_{0.66}W_{0.33}C$ and (f) Cm $Mo_{0.25}W_{0.75}C$. Carbon atoms are colored black, tungsten atoms are blue and molybdenum atoms are green. Colored polyhedra are employed to emphasize the decoration of the structure by the two types of metal atoms. The

c -axis is oriented perpendicular to the metal/carbon triangular nets. **g** The formation enthalpy, ΔH , in meV/atom for the reaction $(1-x)MoC + x(WC) \rightarrow (Mo_{1-x}W_x)C$ as a function of MoC/WC composition. DFT values (black dots) are provided along with results obtained following relaxation with the robust potential (RR), prediction of the energetics of the RR structures with the accurate potential (AP-RR), and further relaxation of the RR structures with the accurate potential (AR-RR) given by blue, green and pink dots, respectively.

different choice of DFT functional, inclusion of zero point energy or finite temperature contributions may place this structure on the hull. We then examined if the PRAPs procedure, pre-trained on AFLOW data, could relax structures created with RANDSPG and identify other thermodynamically stable compounds not found within AFLOW for the CMoW system.

We compare the AFLOW hull with hulls predicted using the PRAPs procedure (Fig. 3 and Supplementary Figs. 23 and 24). At an MTP level of 10 no additional structures emerged. However, for an MTP level of 16, some thermodynamically stable structures were found, including $Imm2$ $Mo_{0.75}W_{0.25}C$, in addition to the on-and-near-hull compounds present in AFLOW. The $Imm2$ phase, containing two formula units per primitive cell, resembles hexagonal MoC except that in every second layer half of the metal atoms are replaced by W, and the substituted metal-containing triangular nets are arranged in an ...ABAB... stacking sequence with respect to each other (Fig. 4b). This phase likely originated from the RANDSPG set, which was subsequently optimized, in an active learning sense, via the robust and accurate potentials. When an MTP of level 10 was used instead, $Imm2$ $Mo_{0.75}W_{0.25}C$ was not on the convex hull likely because the relaxation process with the robust potential pushed this particular configuration too high in energy. To test this hypothesis the level 16 data for $Imm2$ $Mo_{0.75}W_{0.25}C$ was concatenated with the structures that are present on the level 10 hull, and further analysis revealed that the phase was predicted to be thermodynamically stable.

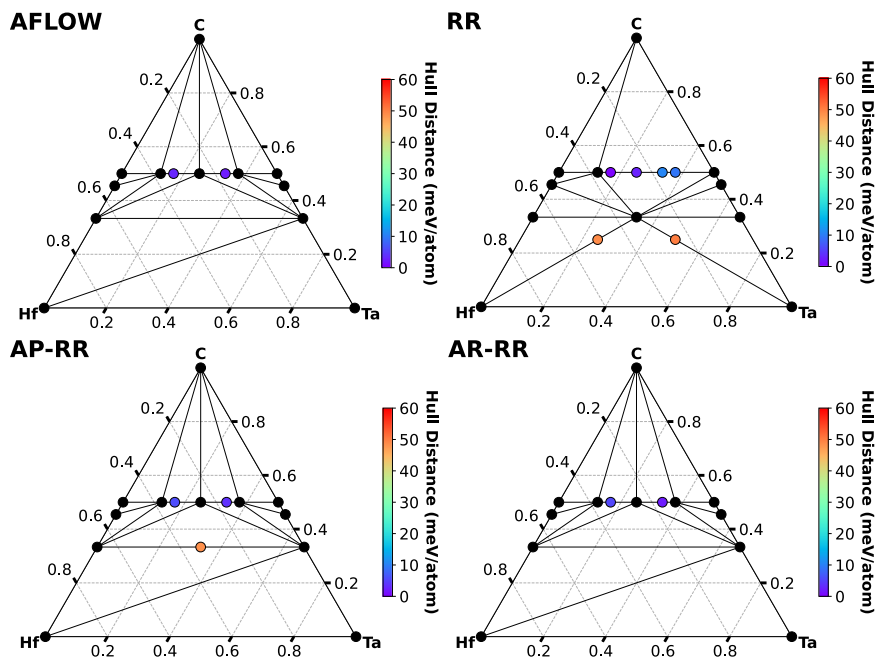
In addition, four more structures, within 1 meV/atom of the convex hull, lay on the level 16 hull: $P\bar{6}m2$ $Mo_{0.5}W_{0.5}C$, $P\bar{6}m2$ $Mo_{0.33}W_{0.66}C$, $P\bar{6}m2$ $Mo_{0.66}W_{0.33}C$ and Cm $Mo_{0.25}W_{0.75}C$ (Fig. 4c–f). Though the first has the same composition as the structure present within AFLOW, it is 17 meV/atom lower in energy. In fact, if we do not distinguish between the identities of the metal atoms, the AFLOW structure can be transformed into $P\bar{6}m2$ $Mo_{0.5}W_{0.5}C$ by doubling it along the b -axis followed by three sets of translations of various subsets of atoms. In both phases the C atoms fall within trigonal-prismatic holes, but in this particular structure the triangular (and square) faces all point along the same crystallographic direction, while in the AFLOW structure half of the prisms are rotated, thereby swapping the axes along which the two sets of faces lie. Importantly, PRAPs-found $P\bar{6}m2$ $Mo_{0.5}W_{0.5}C$ corresponds to a coloring of the hexagonal CMo/CW

prototype structure with an ...ABAB... arrangement for the metal-containing hexagonal nets (Fig. 4c). Similarly, the remaining three PRAPs-found structures can be derived from colorings of the hexagonal parent phase, with $P\bar{6}m2$ $Mo_{0.33}W_{0.66}C$ and $P\bar{6}m2$ $Mo_{0.66}W_{0.33}C$ being inverses of each other, while Cm $Mo_{0.25}W_{0.75}C$ can be described as a W-rich ...ABAB... layered decoration of this same hexagonal prototype.

The identified near-and-on-hull phases lie on a straight line joining the two end-members comprising this CMo/CW series. They represent examples of an ensemble of phases with highly-variable concentrations, suggesting the existence of a solid solution with a very low critical temperature of the miscibility gap. To investigate this, we optimized ~ 366 $(Mo_{1-x}W_x)C$ structures ($x = 0.08\bar{3}, 0.1, 0.125, 0.1\bar{6}, 0.2, 0.25, 0.3, 0.375, 0.41\bar{6}, 0.4, 0.5, 0.5\bar{8}\bar{3}, 0.625, 0.6, 0.75, 0.7, 0.8\bar{3}$) with 4–24 atoms in the unit cell, and between 2 and 86 unique structures were optimized per composition. The previously generated level 16 robust and accurate potentials were used to predict their energies and to relax them. Figure 4g plots the resulting enthalpies of formation, ΔH , from the monocarbide endpoints: relaxed with the robust potential (RR), subsequently predicted by the accurate potential (AP-RR), and finally relaxed with the accurate potential (AR-RR).

All of the DFT-optimized compounds fell on or within 5.7 meV/atom of the line joining the CMo and CW endpoints, suggesting that their ΔH is close to 0 meV/atom. For a given composition, various decorations were computed to be nearly isoenthalpic, suggesting that configurational entropy will play a role in the stability of this family of structures. Turning to the results obtained with the generated MTPs, the computed ΔH , as predicted on structures relaxed by the RP, was largely positive (blue dots) with the deviation from the zero-energy line steadily increasing for larger W concentrations. Whereas the distance from the CMo–CW tie-line, averaged over all structures, was calculated as being 0.8 meV/atom ($\sigma = 1.10$) via DFT, the robust relaxed protocol resulted in an average tie-line distance of 78.4 meV/atom ($\sigma = 28.95$). Prediction of the energies of the robust relaxed structures with the AP (green dots) yielded an average ΔH of 18.8 meV/atom ($\sigma = 14.96$). It is only via relaxation with the AP (purple dots) that we obtain an average tie-line distance of 4.2 meV/atom ($\sigma = 3.58$). This example illustrates that structural relaxation with the AP is key for obtaining energetics that are in good agreement with those derived from DFT calculations.

Fig. 5 | A sample of convex hull diagrams produced by PRAPs for CHfTa. Structures within 60 meV/atom of the AFLOW-derived convex hull (top left), as well as the PRAPs procedure at an MTP level of 22 for the CHfTa system. Structures are colored (see color bar) according to the distances from the hull obtained after relaxing with the robust potential (RR), prediction of the enthalpies of the RR structures with the accurate potential (AP-RR) and relaxing the RR structures with the accurate potential (AR-RR). Black dots are on the hull, and purple dots are within 1 meV/atom of the hull. In contrast to the plots shown in Fig. 3, the data illustrated here has not undergone further DFT relaxations.



The convex hulls discussed and presented above (Fig. 3) were optimized with DFT, and the conclusions regarding thermodynamic stability of particular phases were made based upon these hulls. This procedure is commonplace⁷¹ but it might make one wonder about the limits of the utility of ML-IAPs in CSP. Part of the answer lies above where we show that ML can significantly reduce the number of required DFT calculations. But, the other part of this answer is in the convex hull candidate structures: the output of PRAPs relaxations and predictions before the final DFT step. The analysis of the CMoW system suggested that relaxation with the AP is key for obtaining energetics that are in-line with DFT results. To further study this aspect, in Fig. 5 we plot the convex hulls for the CHfTa system calculated at an MTP level of 22. Comparison of the AFLOW-derived hull with one that is obtained after relaxation with the robust potential (RR) shows that the latter predicts a structure that is not found within AFLOW, with $\text{Hf}_{0.5}\text{Ta}_{0.5}\text{C}$ composition, to lie on the hull (after DFT relaxation, it falls 123 meV/atom above the hull) whereas $(\text{Hf}_{1-x}\text{Ta}_x)\text{C}$ stoichiometries lie around 15 meV/atom above the hull. The rogue $\text{Hf}_{0.5}\text{Ta}_{0.5}\text{C}$ structure disappears after AP prediction, and the energies of the $(\text{Hf}_{1-x}\text{Ta}_x)\text{C}$ species fall onto-and-just-above the hull. Relaxation with the AP yields a hull that is virtually indistinguishable from the one derived from AFLOW, similar to the results obtained for the CMoW system. In Supplementary Figs. 17–27, we provide these same four convex hulls for each carbide system considered and each MTP level, before and after subsequent relaxation with DFT. Comparison of the AR-RR hulls with the hulls constructed from relaxing the AFLOW data with DFT indicates that all of the structures within 1 meV/atom of the latter are predicted to be within 50 meV/atom of the level 16 or 22 MTP-derived hulls. Moreover, for CMoW the structures were within ~35 meV/atom of the level 16 hull. This suggests that the AR-RR protocol is useful for screening a large dataset for structures that may be thermodynamically stable, but further DFT relaxations of this reduced set of structures is important for accurate energy evaluations.

Discussion

The density functional theory (DFT) computed energies, forces and stresses found within the AFLOW database of four ternary carbide systems (HfTaC , HfZrC , MoWC , and TaTiC) were employed to train system-specific machine learning interatomic potentials of the moment tensor potential (MTP) flavor. A utility package that can be used to generate both robust potentials (RP), capable of roughly relaxing any structure, and accurate potentials (AP), tailored towards the relaxation of low-energy structures,

which was employed to automate this training, is described. The AFLOW data was augmented with ~6300 random symmetric structures resembling those that would be created in the first step of a crystal structure prediction (CSP) search, and these were relaxed with MTPs updated via active learning. Pre-training on the AFLOW data was shown to decrease prediction errors with the RP. For the HfTaC system, relaxation with the AP yielded a convex hull that agreed perfectly with the one found within AFLOW.

Moreover, this procedure identified five $(\text{Mo}_{1-x}\text{W}_x)\text{C}$ stoichiometry compounds, not found within AFLOW, that lay on the convex hull and corresponded to colorings of the hexagonal CMo/CW prototypes, illustrating how the described protocol can accelerate CSP. Subsequently, the RP and AP were used to relax hundreds of $(\text{Mo}_{1-x}\text{W}_x)\text{C}$ lattices spanning a broad composition range, and it was shown that relaxation with the AP yielded formation enthalpies that were in excellent agreement with those computed via DFT. The ideas and tools described here may aid in the generation of ML-IAPs from already-existing DFT data, to be used for materials prediction.

Methods

Computational details

The density functional theory (DFT) calculations were performed using the Vienna ab initio Simulation Package version 5.4.12⁷² coupled with the Perdew, Burke, Ernzerhof (PBE) gradient-corrected exchange and correlation functional⁷³ and the projector augmented wave method⁷⁴. During the active learning procedure, the VASP calculations were performed using Γ -centered Monkhorst-Pack k -meshes where the number of divisions along each reciprocal lattice vector was chosen such that the product of this number with the real lattice constant was 30 Å. The carbon $2s^2 2p^2$, Hf $6s^2 5d^4$, Ta $6s^2 5d^4$, Zr $5s^2 4d^4$, Ti $4s^2 3d^3$, Mo $5s^2 4d^5$, and W $6s^2 5d^4$ electrons were treated as valence, and an energy cutoff of 400 eV was employed. After training was complete, the convex hull analysis included a DFT relaxation accomplished by calling AFLOW's management protocol, using the standard settings described in ref. 56; the AFLOW hull data were also re-relaxed using this procedure. Structures from the AFLOW database and those generated by RANDSPG⁷⁰, as described in the main text, comprised the full relaxation set employed for the development of the MTPs. The crystals whose geometries were relaxed to construct Fig. 4g were generated from $P6m2$ $\text{Mo}_{0.5}\text{W}_{0.5}\text{C}$ using the Supercell program, employing the merge option to remove duplicate structures⁷⁵.

PRAPs was run on each ternary carbide using MTP levels 10, 16, and 22 with a MLIP-relaxation-iteration limit of 100, and an

extrapolation grade of $2 < \gamma \leq 10$. The cutoff distances for generating the MTP were $1.1 \text{ \AA} < x < 5 \text{ \AA}$. Active learning was, in most cases, declared to be converged when no new structures were considered for addition to the training set. In the case of the level 22 trainings, the ALS procedure was stopped when the number of structures to be added to the training set was less than 1% of the number already in the training set. PRAPs filtered out configurations with interatomic distances $1.1 \text{ \AA} < d < 3.1 \text{ \AA}$ before beginning the pre-training, and when beginning the AP training removed all structures that were higher than 50 meV/atom of the most stable configuration for each composition.

Data availability

The datasets generated during and/or analyzed during the current study are summarized in the supplementary information, and are available from the corresponding author on reasonable request.

Code availability

The PRAPs code will be released in a subsequent publication, and in the meanwhile, may be obtained from the corresponding authors upon reasonable request.

Received: 22 December 2023; Accepted: 14 June 2024;

Published online: 02 July 2024

References

- Liu, X., Zhang, J. & Pei, Z. Machine learning for high-entropy alloys: progress, challenges and opportunities. *Prog. Mater. Sci.* **131**, 101018 (2023).
- Zhang, L., Han, J., Wang, H., Car, R. & E, W. Deep potential molecular dynamics: a scalable model with the accuracy of quantum mechanics. *Phys. Rev. Lett.* **120**, 143001 (2018).
- Grabowski, B. et al. Ab initio vibrational free energies including anharmonicity for multicomponent alloys. *NPJ Comput. Mater.* **5**, 80 (2019).
- Tong, Q. et al. Combining machine learning potential and structure prediction for accelerated materials design and discovery. *J. Phys. Chem. Lett.* **11**, 8710–8720 (2020).
- Falls, Z., Avery, P., Wang, X., Hilleke, K. P. & Zurek, E. The xtalopt evolutionary algorithm for crystal structure prediction. *J. Phys. Chem. C* **125**, 1601–1620 (2021).
- Behler, J. & Parrinello, M. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Phys. Rev. Lett.* **98**, 146401 (2007).
- Hajinazar, S., Shao, J. & Kolmogorov, A. N. Stratified construction of neural network based interatomic models for multicomponent materials. *Phys. Rev. B* **95**, 014114 (2017).
- Thompson, A. P., Swiler, L. P., Trott, C. R., Foiles, S. M. & Tucker, G. J. Spectral neighbor analysis method for automated generation of quantum-accurate interatomic potentials. *J. Comput. Phys.* **285**, 316–330 (2015).
- Shapeev, A. V. Moment tensor potentials: a class of systematically improvable interatomic potentials. *Multiscale Model. Sim.* **14**, 1153–1173 (2016).
- 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).
- Xie, S. R., Rupp, M. & Hennig, R. G. Ultra-fast interpretable machine-learning potentials. *NPJ Comp. Mat.* **9**, 162 (2023).
- Pickard, C. J. Ephemeral data derived potentials for random structure search. *Phys. Rev. B* **106**, 014102 (2022).
- Yang, Q. et al. Hard and superconducting cubic boron phase via swarm-intelligence structural prediction driven by a machine-learning potential. *Phys. Rev. B* **103**, 024505 (2021).
- Podryabinkin, E. V., Tikhonov, E. V., Shapeev, A. V. & Oganov, A. R. Accelerating crystal structure prediction by machine-learning interatomic potentials with active learning. *Phys. Rev. B* **99**, 064114 (1–7) (2019).
- Deringer, V. L., Pickard, C. J. & Csányi, G. Data-driven learning of total and local energies in elemental boron. *Phys. Rev. Lett.* **120**, 156001 (2018).
- Deringer, V. L. & Csányi, G. Machine learning based interatomic potential for amorphous carbon. *Phys. Rev. B* **95**, 094203 (2017).
- Deringer, V. L., Pickard, C. J. & Proserpio, D. M. Hierarchically structured allotropes of phosphorus from data-driven exploration. *Angew. Chem. Int. Ed.* **59**, 15880–15885 (2020).
- Wang, X. et al. Data-driven prediction of complex crystal structures of dense lithium. *Nat. Commun.* **14**, 2924 (2023).
- Ibarra-Hernandez, W. et al. Structural search for stable mg-ca alloys accelerated with a neural network interatomic model. *Phys. Chem. Chem. Phys.* **20**, 27545–27557 (2018).
- Kharabadze, S., Thorn, A., Koulakova, E. A. & Kolmogorov, A. N. Prediction of stable li-sn compounds: boosting ab initio searches with neural network potentials. *NPJ Comput. Mater.* **8**, 136 (2022).
- Thorn, A., Gochitashvili, D., Kharabadze, S. & Kolmogorov, A. N. Machine learning search for stable binary sn alloys with Na, Ca, Cu, Pd and Ag. *Phys. Chem. Chem. Phys.* **25**, 22415–22436 (2023).
- Wu, S. Q. et al. An adaptive genetic algorithm for crystal structure prediction. *J. Phys.: Condens. Matter* **26**, 035402 (2014).
- Ferreira, P. P. et al. Search for ambient superconductivity in the lu-n-h system. *Nat. Commun.* **14**, 5367 (2023).
- Salzbrenner, P. T. et al. Developments and further applications of ephemeral data derived potentials. *J. Chem. Phys.* **159**, 144801 (2023).
- Gubaev, K., Podryabinkin, E. V., Hart, G. L. W. & Shapeev, A. V. Accelerating high-throughput searches for new alloys with active learning of interatomic potentials. *Comput. Mater. Sci.* **156**, 148–156 (2019).
- Zuo, Y. et al. Performance and cost assessment of machine learning interatomic potentials. *J. Phys. Chem. A* **124**, 731–745 (2020).
- Liu, Y. & Mo, Y. Assessing the accuracy of machine learning interatomic potentials in predicting the elemental orderings: a case study of li-al alloys. *Acta Materialia* **268**, 119742 (2024).
- Bernstein, N., Csányi, G. & Deringer, V. L. De novo exploration and self-guided learning of potential-energy surfaces. *NPJ Comput. Mater.* **5**, 99 (2019).
- Smith, J. S., Nebgen, B., Lubbers, N. & Isayev, O. Less is more: sampling chemical space with active learning. *J. Chem. Phys.* **148**, 241733 (2018).
- Zhang, L., Lin, D. Y., Wang, H., Car, R. & E, W. Active learning of uniformly accurate interatomic potentials for materials simulation. *Phys. Rev. Mater.* **3**, 023804 (2019).
- Jinnouchi, R., Karsai, F. & Kresse, G. On-the-fly machine learning force field generation: application to melting points. *Phys. Rev. B* **100**, 014105 (2019).
- Curtarolo, S. et al. Aflow: an automatic framework for high-throughput materials discovery. *Comp. Mater. Sci.* **58**, 218–226 (2012).
- Esters, M. et al. aflow.org: a web ecosystem of databases, software and tools. *Comput. Mater. Sci.* **216**, 111808 (2023).
- Oses, C. et al. aflow++: A C++ framework for autonomous materials design. *Comput. Mater. Sci.* **217**, 111889 (2023).
- Jain, A. et al. Commentary: The materials project: a materials genome approach to accelerating materials innovation. *APL Mater.* **1**, 011002 (2013).
- Saal, J. E., Kirklin, S., Aykol, M., Meredig, B. & Wolverton, C. Materials design and discovery with high-throughput density functional theory: the open quantum materials database (oqmd). *JOM-J Min Met Mat S* **65**, 1501–1509 (2013).
- Draxl, C. & Scheffler, M. The nomad laboratory: from data sharing to artificial intelligence. *J Phys-Mat.* **2**, 036001 (2019).
- Ai, Q. et al. Ocelot: An infrastructure for data-driven research to discover and design crystalline organic semiconductors. *J Chem. Phys.* **154**, 174705 (2021).

39. NIST-Materials. <https://materialsdata.nist.gov/> (2023).
40. Chen, C. & Ong, S. P. A universal graph deep learning interatomic potential for the periodic table. *Nat. Comp. Sci.* **2**, 178–228 (2022).
41. Deng, B. et al. Chgnet as a pretrained universal neural network potential for charge-informed atomistic modelling. *Nat. Machine Intelligence* **5**, 1031–1041 (2023).
42. Choudhary, K. et al. Unified graph neural network force-field for the periodic table: solid state applications. *Digital Discovery* **2**, 346–355 (2023).
43. Batatia, I. et al. A foundation model for atomistic materials chemistry. Preprint at <https://arxiv.org/abs/2401.00096> (2023).
44. Merchant, A. et al. Scaling deep learning for materials discovery. *Nature* **624**, 80–85 (2023).
45. Novikov, I. S., Gubaev, K., Podryabinkin, E. V. & Shapeev, A. V. The mlip package: moment tensor potentials with mpi and active learning. *Mach. Learn.: Sci. Technol.* **2**, 025002 (2021).
46. Schaarschmidt, M. et al. Learned force fields are ready for ground state catalyst discovery. Preprint at <https://arxiv.org/abs/2209.12466> (2022).
47. Hossain, M. D. et al. Entropy landscaping of high-entropy carbides. *Adv. Mater.* **33**, 2102904 (2021).
48. Hossain, M. D. et al. Carbon stoichiometry and mechanical properties of high entropy carbides. *Acta Mater.* **215**, 117051 (2021).
49. Nyshadham, C. et al. Machine-learned multi-system surrogate models for materials prediction. *NPJ Comput. Mater.* **5**, 75 (2019).
50. Jafary-Zadeh, M., Khoo, K. H., Laskowski, R., Branicio, P. S. & Shapeev, A. V. Applying a machine learning interatomic potential to unravel the effects of local lattice distortion on the elastic properties of multi-principal element alloys. *J. Alloy. Compd.* **803**, 1054–1062 (2019).
51. Gubaev, K. et al. Performance of two complementary machine-learned potentials in modelling chemically complex systems. *NPJ Comput. Mater.* **9**, 129 (2023).
52. Zeng, C., Neils, A., Lesko, J. & Post, N. Machine learning accelerated discovery of corrosion-resistant high-entropy alloys. *Comput. Mater. Sci.* **237**, 112925 (2024).
53. Korotaev, P., Novoselov, I., Yanilkin, A. & Shapeev, A. Accessing thermal conductivity of complex compounds by machine learning interatomic potentials. *Phys. Rev. B* **100**, 144308 (2019).
54. Mortazavi, B. et al. Exploring phononic properties of two-dimensional materials using machine learning interatomic potentials. *Appl. Mater. Today* **20**, 100685 (2020).
55. Podryabinkin, E. V. & Shapeev, A. V. Active learning of linearly parametrized interatomic potentials. *Comp. Mater. Sci.* **140**, 171–180 (2017).
56. Calderon, C. E. et al. The aflow standard for high-throughput materials science calculations. *Comp. Mater. Sci.* **108**, 233–238 (2015).
57. Nakamura, K. & Yashima, M. Crystal structure of nacl-type transition metal monocarbides mc (m = v, ti, nb, ta, hf, zr), a neutron powder diffraction study. *Mater. Sci. Eng.-B Adv.* **148**, 69–72 (2008).
58. Clougherty, E. V., Kafalas, J. A. & Lothrop, K. H. A new phase formed by high-pressure treatment - face-centered cubic molybdenum monocarbide. *Nature* **191**, 1194 (1961).
59. Schuster, J., Rudy, E. & Nowotny, H. Moc-phase with wc structure. *Monatsh. Chem.* **107**, 1167–1176 (1976).
60. Harrington, T. J. et al. Phase stability and mechanical properties of novel high entropy transition metal carbides. *Acta Mater.* **166**, 271–280 (2019).
61. Vorotilo, S. et al. Phase stability and mechanical properties of carbide solid solutions with 2-5 principal metals. *Comp. Mater. Sci.* **201**, 110869 (2022).
62. Sarker, P. et al. High-entropy high-hardness metal carbides discovered by entropy descriptors. *Nat. Commun.* **9**, 4980 (2018).
63. Divilov, S. et al. Disordered enthalpy-entropy descriptor for high-entropy ceramics discovery. *Nature* **625**, 66–73 (2024).
64. Dai, F.-Z., Wen, B., Sun, Y., Xiang, H. & Zhou, Y. Theoretical prediction on thermal and mechanical properties of high entropy (zr_{0.2}hfo_{0.2}ti_{0.2}nb_{0.2}ta_{0.2}) c by deep learning potential. *J. Mater. Sci. Technol.* **43**, 168–174 (2020).
65. Pak, A. Y. et al. Machine learning-driven synthesis of tizmbhftac5 high-entropy carbide. *NPJ Comput. Mater.* **9**, 7 (2023).
66. Mehl, M. J. et al. The aflow library of crystallographic prototypes: part 1. *Comp. Mater. Sci.* **136**, S1–S828 (2017).
67. Hicks, D. et al. The aflow library of crystallographic prototypes: part 2. *Comp. Mater. Sci.* **161**, S1–S1011 (2019).
68. Hicks, D. et al. The aflow library of crystallographic prototypes: part 3. *Comp. Mater. Sci.* **199**, 110450 (2021).
69. Hicks, D. et al. Aflow-xtalfinder: a reliable choice to identify crystalline prototypes. *NPJ Comput. Mater.* **7**, 30 (2021).
70. Avery, P. & Zurek, E. Randspg: an open-source program for generating atomistic crystal structures with specific spacegroups. *Comput. Phys. Commun.* **213**, 208–216 (2017).
71. Roberts, J., Bursten, J. R. & Risko, C. Genetic algorithms and machine learning for predicting surface composition, structure, and chemistry: a historical perspective and assessment. *Chem. Mater.* **33**, 6589–6615 (2021).
72. Kresse, G. & Hafner, J. Ab Initio molecular dynamics for liquid metals. *Phys. Rev. B* **47**, 558 (1993).
73. Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.* **77**, 3865–3868 (1996).
74. Blöchl, P. E. Projector augmented-wave method. *Phys. Rev. B* **50**, 17953 (1994).
75. Okhotnikov, K., Charpentier, T. & Cadars, S. Supercell program: a combinatorial structure-generation approach for the local-level modeling of atomic substitutions and partial occupancies in crystals. *J. Cheminformatics* **8**, 1–15 (2016).

Acknowledgements

The authors would like to gratefully acknowledge the DoD SPICES MURI sponsored by the Office of Naval Research (Naval Research contract N00014-21-1-2515) for their financial support of this work. Calculations were performed at the Center for Computational Research at SUNY Buffalo (<http://hdl.handle.net/10477/79221>). We would like to acknowledge Xiaoyu Wang for artistic help and guidance, Masashi Kimura for assistance with the convex hull diagrams, Hagen Eckert, Xiomara Campilongo, and Corey Osos for fruitful discussions.

Author contributions

E.Z. and S.C. conceived the research and supervised the study. J.R. carried out the method development of the PRAPs code, and performed the calculations and analysis. All authors (J.R., B.R., S.D., J-P.M., W.G.F., D.E.W., D.W.B., S.C., and E.Z.) participated in discussing the results, and commented on the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41524-024-01321-7>.

Correspondence and requests for materials should be addressed to Eva Zurek.

Reprints and permissions 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 licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence 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 licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2024