De novo exploration and self-guided learning of potential-energy surfaces

Noam Bernstein1, Gábor Csányi2 and Volker L. Deringer1,2

Interatomic potential models based on machine learning (ML) are rapidly developing as tools for material simulations. However, because of their flexibility, they require large fitting databases that are normally created with substantial manual selection and tuning of reference configurations. Here, we show that ML potentials can be built in a largely automated fashion, exploring and fitting potential-energy surfaces from the beginning (de novo) within one and the same protocol. The key enabling step is the use of a configuration-averaged kernel metric that allows one to select the few most relevant and diverse structures at each step. The resulting potentials are accurate and robust for the wide range of configurations that occur during structure searching, despite only requiring a relatively small number of single-point DFT calculations on small unit cells. We apply the method to materials with diverse chemical nature and coordination environments, marking an important step toward the more routine application of ML potentials in physics, chemistry, and materials science.

Published in partnership with the Shanghai Institute of Ceramics of the Chinese Academy of Sciences

INTRODUCTION

Atomic-scale modeling has become a cornerstone of scientific research. Quantum-mechanical methods, most prominently based on density-functional theory (DFT), describe the atomistic structures and physical properties of materials with high confidence;1 increasingly, they also make it possible to discover previously unknown crystal structures and synthesis targets.2 Still, quantum-mechanical material simulations are severely limited by their high computational cost.

Machine learning (ML) has emerged as a promising approach to tackle this long-standing problem.3–12 ML-based interatomic potentials approximate the high-dimensional potential-energy surface (PES) by fitting to a reference database, which is usually computed at the DFT level. Once generated, ML potentials enable accurate simulations that are orders of magnitude faster than the reference method. They can solve challenging structural problems, as has been demonstrated for the atomic-scale deposition and growth of amorphous carbon films,13 for proton-transfer mechanisms,14 or dislocations in materials,15,16 involving thousands of atoms in the simulation. More recently, it was shown that ML potentials can be suitable tools for global structure searches targeting crystalline phases,17–20 clusters,21–24 and nanostructures.25

Assembling the reference databases to which ML potentials are fitted is currently mostly a manual and laborious process, guided by the physical problem under study. The first artificial neural network (NN)-type potential for materials3 was made by enumerating known crystal structures for silicon and used to describe high-pressure phase transitions.26,27 To incorporate vacancies, surfaces, and so on, hierarchical databases for transition metals have been built that start with simple unit cells and gradually add relevant defect structures;28,29 liquid and amorphous materials can be described by iteratively grown databases that contain relatively small-sized MD snapshots.30–33 A “general-purpose” Gaussian approximation potential (GAP) ML model for elemental silicon was recently developed34 that can describe crystalline phases with meV-per-atom accuracy, treat defects, cracks, and surfaces,35 and generate amorphous silicon structures in excellent agreement with experiments.36 Despite their success in achieving their stated goals, none of these potentials are expected to be even reasonable for crystal structures not included in their databases, say, hitherto undiscovered phases that only become stable at very high pressures.

In contrast, structure searching (i.e., a global exploration of the PES) can be a suitable approach for finding structures to be included in the training databases in the first place.18–20,37 The principal idea to explore configuration space with preliminary ML potentials is well established: since the first high-dimensional ML potentials have been made, it was shown how they can be refined by exploring unknown structures3,30,31 and “on the fly” schemes were proposed to add required data while an MD simulation is being run.3,38–40 We have previously shown that the PES of boron can be iteratively sampled without prior knowledge of any crystal structure involved; we called the method “GAP-driven random structure searching” (GAP-RSS),18 reminiscent of the successful ab initio random structure-searching (AIRSS) approach.41,42 Subsequently, we demonstrated, by way of an example, that the crystal structure of black phosphorus can be discovered by GAP-RSS within a few iterations, and we identified several previously unknown hypothetical allotropes of phosphorus.19

In the context of ML potential fitting, the so-called “active learning” schemes that detect extrapolation (indicating when the potential moves away from known configurations) are currently receiving much attention. A query-by-committee active-learning approach was suggested in 2012 by Artrith and Behler: two NN potential fits are made to the same database, and if their prediction differs for a given (new) structure, this structure needs to be added to the database.43 More recently, Jinnouchi et al. demonstrated how ab initio molecular dynamics (AIMD) simulations of specific systems can be sped up by active learning of the computed forces (in a modified GAP framework), using the predicted error of the Gaussian Process to select new data points.
and to improve the speed of AIMD.\textsuperscript{38,40} In the context of structure exploration, Shapeev and coworkers employed moment tensor potentials\textsuperscript{14} with active learning\textsuperscript{15} to generate ML potentials,\textsuperscript{20,46} and E and coworkers described a generalized active-learning scheme for deep NN potentials.\textsuperscript{47} So far, these studies mainly focused on specific intermetallic systems, namely, Al–Mo\textsuperscript{47} and Cu–Pd, Co–Nb–V, and Al–Ni–Ti.\textsuperscript{46} Furthermore, Podryabinkin et al.\textsuperscript{20} showed that their approach can identify existing and hypothetical boron allotropes.

In this work, we present an efficient and unified approach for generating reference databases for fitting ML potentials, exploring structural space from the beginning (de novo) by ML-driven searching and similarity measures, all without any prior knowledge of what structures are or are not relevant. In contrast with continuous active learning, our aim is to converge to a potential that can describe a wide range of configurations without the need for additional fitting. We demonstrate the ability to cover a broad range of structures and chemistries, from graphite sheets to a densely packed transition metal. Our work provides conceptual insight into how computers can discover structural chemistry based on data and similarity measures alone, and it paves the way for a more routine application of ML potentials in material discoveries.

RESULTS

A unified framework for exploring and fitting structural space

The overarching aim is to construct a ML potential with minimal effort, both in terms of computational resources and in terms of input required from the user. In regard to the former, we use only single-point DFT computations to generate the fitting database.\textsuperscript{18} In regard to the latter, we define general heuristics wherever possible, such that neither the protocol nor its parameters need to be manually tuned for a specific system. The ML architecture to which we couple our method is based on a hierarchical combination of two-, three-, and many-body descriptors,\textsuperscript{32} and it uses GAP as the regressor.\textsuperscript{5} The remaining two parameters that need to be set by the user are a “characteristic” distance and whether the material is primarily covalent or metallic. For the distance, we choose tabulated covalent (for C, B, and Si)\textsuperscript{18} or metallic (for Ti) radii, depending on the nature of the system. These define the volume of the initial structures and the cutoffs for the ML descriptors (Methods section).

Our approach is based on an iterative cycle, as shown in the diagram in Fig. 1a. We generate ensembles of randomized structures as in the AIRSS framework,\textsuperscript{41,42} a structure-searching approach that is widely used in physics, chemistry, and materials science.\textsuperscript{49–51} In the first iteration, we generate 10,000 initial structures, from which we select the N most diverse ones using the leverage-score CUR algorithm.\textsuperscript{52} In the context of PES models, the CUR algorithm was proposed\textsuperscript{53} and then used\textsuperscript{29,32,34} for selection of sparse (representative) points for Gaussian process regression, and also proposed for selection of training configurations.\textsuperscript{54} The distance between candidate structures is quantified by the Smooth Overlap of Atomic Positions (SOAP) descriptor,\textsuperscript{55} which has been widely used in GAP fitting\textsuperscript{32,34} and in structural analysis.\textsuperscript{56–58} While SOAP is normally used to discriminate between pairs of environments of individual atoms, we here use a configuration-averaged SOAP descriptor\textsuperscript{57} that compares entire unit cells to one another (Methods section). We find that selecting the most representative structures is critical, because we can only evaluate a small number (\(\leq 10,000\)) with DFT. In addition, the starting configurations include dimers in vacuum at a wide range of bond lengths; this serves to capture the exchange repulsion at very short interatomic distances, and thereby to make the potentials more robust.\textsuperscript{52}

With the starting configurations in hand, we perform single-point DFT computations and fit an initial potential to the resulting data; in subsequent iterations, we extend the database and thereby refine the potential.\textsuperscript{18} In each iteration, we start from the same number of new random initial structures, and minimize their enthalpy by using the GAP from the previous iteration. We then select the N most relevant and diverse configurations from the full set of configurations seen throughout the minimization trajectories, for which we employ a combination of Boltzmann-probability biased flat histogram sampling (to focus on low-energy structures) and leverage-score CUR (to select the most diverse structures among those), as illustrated in Fig. 1b. These selected configurations are evaluated by using single-point DFT calculations and added to the fitting database.

The iterative procedure runs until the results are satisfactory. Here, we terminate our searches after 2500 DFT data points have been collected, and our results show this to be sufficient to discover and describe all structures discussed in the present work. Other quality criteria, such as those based on the distribution of energies in the database,\textsuperscript{18} might be defined as well; the generality of our approach is not affected by this choice.

Diversity-based selection

We demonstrate the method for boron, one of the most structurally complex elements.\textsuperscript{59} With the exception of a high-pressure α-Ga-type phase, all relevant boron allotropes contain \(\text{B}_{12}\) icosahedra as the defining structural unit.\textsuperscript{59} Boron has been the topic of structure searches with DFT\textsuperscript{60–63} and more recently, with ML potentials for bulk allotropes\textsuperscript{18,20} and gas-phase clusters.\textsuperscript{22} Our previous work showed how the PES for boron can be fitted in a ML framework,\textsuperscript{18} leading to an interatomic potential able to describe the different allotropes. However, at that time, we generated and fed back 2500 cells per iteration (without further selection), and added the structure of \(\alpha\text{-B}_{12}\) manually at a later stage.\textsuperscript{18}

Our new protocol “discovers” the structure of \(\alpha\text{-B}_{12}\) in a self-guided way, as shown in Fig. 2. The figure compares the performance of our selection procedure with alternatives: (i) random selection and (ii) using CUR but on the matrix of SOAP vectors rather than similarity kernels (see Methods section for details). The first of these, random selection, improves the database much less after the first few iterations, and ends up with the highest error (gray in Fig. 2). The second, which uses CUR but neglects the nonlinear aspects of the similarity kernel, initially performs well, but soon stops reducing the error (green). Note that this algorithm is exactly the same as the one used in potential fitting to select representative environments (in that case, even computing the complete similarity kernel matrix quickly becomes impractical). The use of CUR on the similarity kernel for selecting structures to be included in the next iteration is shown to be the most efficient (purple in Fig. 2).

The increasingly accurate description of the \(\text{B}_{12}\) icosahedron is reflected in a gradually lowered energy error, falling below the 10 meV/atom threshold with fewer than 2000 DFT evaluations, and below 4 meV/atom once the cycle is completed. This improvement is best understood by inspecting the respective lowest-energy structures that enter the database in a given iteration (Fig. 2).\textsuperscript{64} The lowest-energy structure at point A already contains several three-membered rings, but no \(\text{B}_{12}\) icosahedra yet. With one more iteration, there is a sharp drop in the GAP error (from 175 to 51 meV/atom), concomitant with the first appearance of a rather distorted \(\alpha\text{-B}_{12}\) structure (B). The final database has seen several instances of the correctly ordered structure (C).

Learning diverse crystal structures

Our method is not restricted to a particular chemical system. To demonstrate this, we now apply it to three prototypical materials
The tetrahedral network of diamond are correctly side by side: carbon, silicon, and titanium, which all exhibit multiple crystal structures.

In carbon (Fig. 3a), both the layered structure of graphite and the tetrahedral network of diamond are correctly "learned" during our iterations. For graphite, the energy error reaches a plateau after only a few hundred DFT evaluations; for diamond, the initial error is very large, and after a dozen or so iterations, we observe a rapid drop—concomitant with a drop in the error for the structurally very similar lonsdaleite ("hexagonal diamond"). The final prediction error is well below 1 meV/atom for the sp³-bonded allotropes, and on the order of 4 meV/atom for graphite. We have previously shown that the forces in diamond show higher locality than those in graphite, making their description by approximated potentials (GAP) (orange). Then, this potential is used to relax a new ensemble of randomized cells (green), selecting again the most relevant snapshots, and repeating the cycle.

In silicon (Fig. 3b), the ground-state (diamond-type) structure is rapidly dropped to the potential has not been under debate: depending on the DFT method, the β-Sn-type allotrope (space group I4₁/amd), which is easily discovered; the larger residual error for β-Sn than for diamond-type Si is consistent with previous studies during synthesis open-framework structure with 24 atoms in the unit cell (oS24), which consists of distorted tetrahedral building units that are linked in different ways, which the potential has not "seen". Still, a good description is achieved after a few iterations.

In titanium (Fig. 3c), a hexagonal close-packed (hcp) structure is observed at ambient conditions; however, the zero-Kelvin ground state has been under debate: depending on the DFT method, either hcp or the so-called ω phase is obtained as the minimum. Our method clearly reproduces the qualitative and quantitative difference between the two allotropes (22 meV/atom with the final GAP-RSS iteration vs. 24 meV/atom with DFT) at the computational level we use, namely PBEsol.

Looking beyond the minimum structures, the DFT energy-volume curves are, by and large, well reproduced by GAP-RSS; see Fig. 3d–f. There is some deviation at large volumes
DFT level) from this iteration is shown, as visualized by using α.

While the most relevant crystal structures for materials are usually well known and available from databases, we show that our approach is more general. In Fig. 4, we use kernel principal component analysis with a SOAP kernel to represent the structures in a 2D plane. Figure 5 shows the resulting plot, in which we have encoded the species by symbols and the average coordination number by color (coordination numbers are determined by counting the nearest neighbors up to 1.2 r0).

The results fall within four groups, moving from the left to the right through Fig. 5. The first group is given by graphite-like examples; they are threefold coordinated and only carbon structures (circles) are found there. Numerical examples in Fig. 5 indicate examples, and in this first group, we observe flat (i) and buckled (ii) graphite sheets. In the second group, we have fourfold coordinated (“diamond-like”) networks, made up of both carbon and silicon (recall that we are using a normalized bond length, so diamond-type carbon, and diamond-type silicon will fall on the same position in the plot). The structures that are shown as insets are characteristic examples; from left to right, there is a distorted lonsdaleite-type structure (iii), the well-known \textit{unij} framework (also referred to as the “chiral framework structure” in group-14 elements (iv)), and a more complex \textit{sp}3-bonded allotrope (v).

While the axis values in our plot are arbitrary, they naturally reflect the structural evolution toward higher coordination numbers, and therefore we next observe a set of high-pressure silicon structures (squares), such as the simple-hexagonal one (vi), with an additional contribution from lower-coordinated titanium structures (circles). Finally, there is a set of densely packed structures, all clustered closely together; these are titanium structures including hcp (vii) and the \textit{ω} type (viii). In the center of the plot, there is a structure that bears resemblance to none of the previously mentioned ones (ix), an energetically high-lying and strongly disordered intermediate from a relaxation trajectory that was added to the reference database, rather than a local minimum (see also Supplementary Tables 1–3). This dissimilarity is reflected in relatively large distances from other entries in the SOAP-based similarity map.

**DISCUSSION**

We have shown that automated protocols can be designed for generating structural databases and fitting PESs of materials in a self-guided way. This allows for the generation of ML-based interatomic potentials with minimal effort, both in terms of computational and user time, when combined with a suitable fitting framework, of which many are presently available. Formalizing the protocols for database construction is an important step toward further methodological developments, and ultimately, toward wide applicability of these techniques in computational materials science.
and a metal (titanium). GAP-RSS models that deviate from the DFT result by ≤ 0.01 meV/atom are considered to be fully optimized with the respective method. GAP-RSS in the context of organic solids. Finally, organic (molecular) materials are also beginning to be described very reliably with ML potentials, and an interesting open question is how to use the structural diversity inherent in RSS in the context of organic solids.

**METHODS**

Interatomic potential fitting

To fit interatomic potentials, we use the established GAP ML framework and the associated computer code, which is freely available for noncommercial research at http://www.libatoms.org. Compared with previous work, we here use suitable heuristics to automate and generalize the choice of fitting parameters where possible. We stress again, however, that the main development in the present work is in the automated generation of databases, not the descriptors or the regressor.

We use a linear combination of 2-, 3-, and many-body terms following refs. 12,24, with defining parameters given in Table 1. The 2-body ("2b") and 3b descriptors are scalar distances and symmetrized three-component vectors, respectively. For the many-body term, we use the SOAP kernel, which has been used to fit GAPs for diverse systems. The overall energy scale of each descriptor’s contribution to the predicted energy (controlled by the parameter δ) is set automatically in our protocol. The 2b value is set from the variance of energies in the fitting database, the 3b value is set from the energy error between a 2b-only fit and the fitting database, and the SOAP value is set from the energy error for a 2b + 3b-only fit.

**Fig. 3** “Learning” diverse crystal structures without prior knowledge, including textbook examples of an insulator (carbon), a semiconductor (silicon), and a metal (titanium). a–c Energy error, defined as the difference between DFT- and GAP-computed energies for structures optimized with the respective method. GAP-RSS models that deviate from the DFT result by <1 meV/atom are considered to be fully converged and therefore their errors are drawn as a constant minimum value to ease visualization.

**Fig. 4** Scatter plots of predicted versus DFT energies for ensembles of structures added to the reference databases in the final iteration. Note that the energy scale is continuous, but it changes from linear to logarithmic scaling at 0.1 eV/at, allowing us to visualize both low- and higher-energy regions.

Our RSS-based reference databases efficiently cover structural space up to a given system size (here, 24 atoms in the unit cell). Once a core database has been constructed in this way, it may be readily improved by adding defect, surface, and liquid/amorphous structural models in much larger simulation cells, while at the same time being sufficiently robust to avoid unphysical behavior—even when taken to the more extreme regions of configuration space that are explored early on during RSS.

We targeted here the space of three-dimensional inorganic crystal structures, but conceptually similar approaches may be useful for nanoparticles and other lower-dimensional systems. Finally, organic (molecular) materials are also beginning to be described very reliably with ML potentials, and an interesting open question is how to use the structural diversity inherent in RSS in the context of organic solids.
diagonal noise terms in Gaussian process regression. We set these expected to affect the conclusions of the present work.

Fig. 5 Visualizing the highly diverse structures, both at low and relatively high energies above the global minimum, which have been explored by GAP-RSS and added to the reference database in the last iteration. A similarity map compares three systems side by side (carbon, graphite, fourfold-coordinated (sp3-like) allotropes of C and Si, onward to high-pressure Si structures and finally densely packed variants of Ti. A higher-energy structure (=0.6 eV/at above diamond-type silicon) from an earlier step in a minimization trajectory is included as an example, as enclosed by a dashed line

Table 1. Hyperparameters for descriptors that we use in GAP fitting

| Descriptor | σcut (Å) | Nsp | lmax | ζ | fcutoff (Å) |
|------------|---------|-----|------|---|-------------|
| Covalent   | 0.5     | 30  | 8    | 8 | 4.5         |
| Metallic   | 1.0     | 100 | 8    | 8 | 2.665r      |
| SOAP       | 0.75    | 2000| 8    | 8 | 4.5         |

For all descriptors: Gaussian width σcut (squared-exponential kernel for 2- and 3-body; atomic density width for SOAP); number of sparse points Nsp. For SOAP only: number of radial functions nmax and angular momenta lmax and kernel exponent ζ. Cutoffs fcutoff are expressed in terms of the characteristic radius r, listed for each material in the interatomic potential fitting subsection.

The cutoffs for the three types of descriptors are expressed in terms of the characteristic radius r (Table 1): that for 2b is the longest (intended to capture only the nearest neighbors), and the SOAP is intermediate in range. The resulting cutoff settings are listed in Table 1, the characteristic radii r for the systems studied here being 0.84, 0.76, 1.11, and 1.47 Å for B, C, Si, and Ti, respectively. An ad hoc choice is made here between predominantly covalent (B, C, and Si) or metallic (Ti) materials for selecting the appropriate tabulated radii; however, settings based on the covalent radius for silicon also produce a satisfactory fit for the metallic (β-tin type) modification (residual error <10 meV/atom; Fig. 3b). Future work might explore more automated ways of extracting optimal atomic radii from datasets, and suitable definitions for multicomponent systems (we stress that the latter, in principle, can be routinely treated by present-day ML potentials). None of this is expected to affect the conclusions of the present work.

The weights on the energies, forces, and stresses that are fit are set by diagonal noise terms in Gaussian process regression. We set these according to the reference energy of a given structure, to make the fit more accurate for relatively low-energy structures at each volume while providing flexibility for the higher-energy regions. The values are piecewise-linear functions in ΔE, which is the per-atom reference energy difference relative to the same volume on the convex hull bounding the set of (V, E) points from below (in energy). For the energy, the error σE is 1 meV/atom for ΔE ≤ 0.1 eV, 100 meV/atom for ΔE ≥ 1 eV, and linearly interpolated in-between. For forces, the corresponding σf values are 31.6 and 316 meV Å⁻¹, and for virials the σv values are 63.2 and 632 meV/atom.

Comparing structures

The same mathematical tools that are used to compare atomic environments for the purpose of constructing potentials can also be used to compare atomic configurations. As for the regression, for these similarity kernels, we also use SOAP, although with different parameters (nmax = lmax = 12, σcut = 0.0875 Å, and fcutoff = 10.5 Å), to compare the similarity of environments in selecting from which data to train (in the CUR step). For the kernel PCA used to generate the map in Fig. 5, we use nmax = lmax = 16, σ = 0.1 r0, and fcutoff = 2.5 r0, where r0 is the shortest bond length, as described in the Results section. We obtain what we call a “configuration-averaged” SOAP by averaging over all atoms in the cell. In the SOAP framework, the neighbor density of a given atom i is expanded using a local basis set of radial basis functions g_i and spherical harmonics Y_{lm}:

\[ \rho_i(r) = \sum_j \left( \frac{\exp(-|r-r_j|^2/2\sigma^2)}{\sqrt{2\pi\sigma^2}} \right) \]

and construct the rotationally invariant power spectrum for the entire unit cell:

\[ \tilde{p}_{nm} = \sum_l \tilde{c}_{nlm} \tilde{C}_{nlm} \]

Note that this is not equal to the average of the usual atomic SOAP power spectra used to describe the atomic neighbor environments.
final kernel to compare two cells, A and B, is then
\[ k_{\text{ps}} = \sum_{\text{cell A}} p_{\text{cell A}}(r_{\text{cell A}}) \sum_{\text{cell B}} p_{\text{cell B}}(r_{\text{cell B}})^\zeta, \]
where \( \zeta \) is a small integer number (here, \( \zeta = 4 \)).

For our main results, our diverse structure selection uses leverage-score CUR approximation potential. For every structure, we employ a random external pressure (and constructing the SOAP vectors for structural selection required \( \approx 1 \) min). For the computational cost of potential cores (and generating 10,000 candidate structures required \( \approx 5 \) min on 16 CPUs/A). The probability of selecting each configuration is approximated by a Boltzmann biasing factor. The biasing factor is exponential in the distribution of enthalpies (each computed using the SOAP descriptors (with the dot-product kernel and exponentiation by \( \zeta \), Eq. (4)). Applying the same algorithm to the rectangular matrix of similarity kernels between atomic positions is repeated by 1 \( \beta \) for covalent, and \( \zeta = 5.5 \) for metallic systems. In the initial iteration, half of the structures are generated from the buildcell-default narrow range of volumes, and half from a wider range, \( \pm 25\% \) from the heuristic value. In all later iterations, only the default narrow range is used.

The wide volume-range configurations are meant to simply span a wide range of structures, and use only even numbers of atoms. The narrow volume-range configurations are meant to be good initial conditions for RSS, and so for \( 80\% \) (20\%) of the seed structures, we choose even (odd) numbers of atoms, respectively. This is because for most known structures, the number of atoms in the conventional unit cell is even (eight for diamond and rocksalt, for example), although for some it is odd, including the \( \mu \) phase.\( ^{19} \) Biasing initial seeds toward distributions that occur in nature is a central idea within the AIRSS formalism.\( ^{76} \) The setup of these cells, in itself, has negligible computational cost compared with the relaxations: generating 10,000 candidate structures required \( \approx 5 \) min on 16 cores (and constructing the SOAP vectors for structural selection required on the order of 1 min). For the computational cost of potential fitting, see Supplementary Fig. 3.

With the initial potential available, we then run structural optimizations by relaxing the candidate configurations with a preconditioned LBFGS algorithm\( ^{72} \) to minimize the enthalpy until residual forces fall below 0.01 eV/Å. As in ref.,\( ^{19} \) we employ a random external pressure \( p \) with probability density
\[ P(p/p_0) = \frac{1}{\beta} \exp \left( -\frac{1}{\beta} p/p_0 \right), \]
where \( p_0 = 1 \) GPa, and \( \beta = 0.2 \). This protocol ensures that there is a small but finite external pressure, and also some smaller-volume structures are included in the fit.\( ^{16,18} \) We choose the same pressure range for all materials, for simplicity, although this value could be adjusted depending on the pressure region of interest.\( ^{19} \)

The selection of configurations for DFT evaluation and fitting at each iteration involves a Boltzmann-biased flat histogram and leverage-score CUR, as illustrated in Fig. 1. To compute the selection probabilities for the flat-histogram stage, the distribution of enthalpies (each computed using the pressure at which the corresponding RSS minimization was done) is approximated by the numpy\( ^{17} \) histogram function, with default parameters. The probability of selecting each configuration is inversely proportional to the density of the corresponding histogram bin, multiplied by a Boltzmann biasing factor. The biasing factor is exponential in the enthalpy per atom relative to the lowest enthalpy configuration, divided by a temperature of 0.3 eV for the first iteration, 0.2 eV for the second, and 0.1 eV for all remaining iterations. The leverage-score CUR selection is based on the singular-value decomposition of the square kernel matrix by using the \( \text{SOAP}\) descriptors (with the dot-product kernel and exponentiation by \( \zeta \), Eq. (4)). Applying the same algorithm to the rectangular matrix of SOAP descriptor vectors was significantly less effective (Fig. 2).

Computational details

Reference energies and forces were obtained by using DFT, with projector-augmented waves\( ^{78,79} \) as implemented in the Vienna Ab Initio Simulation package\( ^{80} \). Valence electrons were described by plane-wave basis sets with cutoff energies of 500 (B), 800 (C), 400 (Si), and 285 eV (Ti), respectively. Reciprocal space was sampled and used a fixed \( \text{KSPACING} \) parameter in VASP, amounting to 0.25 for B, Si, and Ti, and 0.35 for C (in units of Å\(^{-1} \) along the reciprocal lattice vectors which include the 2\( n \) factor). Exchange and correlation were treated by using the PBEsol functional\( ^{80} \) for all materials except carbon, where the opt-B88-wd functional\( ^{81,82} \) was chosen to properly account for the van der Waals interactions in graphitic structures. Benchmark data for energy-volume curves were obtained by scaling selected unit cells within given volume increments and optimizing while constraining the volume and symmetry of the cell.

**DATA AVAILABILITY**

Data supporting this publication are available at https://doi.org/10.17863/CAM.43407.

**CODE AVAILABILITY**

A Python implementation for the protocol developed in this publication is available at https://doi.org/10.17863/CAM.43407.

Received: 24 May 2019; Accepted: 9 September 2019; Published online: 11 October 2019

**REFERENCES**

1. Lejaeghere, K. et al. Reproducibility in density functional theory calculations of solids. Science 351, aad3000 (2016).
2. Oganov, A. R., Pickard, C. J., Zhu, Q. & Needs, R. J. Structure prediction drives materials discovery. Nat. Rev. Mater. 4, 331–348 (2019).
3. Behler, J. & Parrinello, M. Generalized neural-network representation of high-dimensional potential-energy surfaces. Phys. Rev. Lett. 98, 146401 (2007).
4. Baró, 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).
5. Li, Z., Kermode, J. R. & De Vita, A. Molecular dynamics with on-the-fly machine learning of quantum-mechanical forces. Phys. Rev. Lett. 114, 096405 (2015).
6. Arthritus, N. & Urban, A. An implementation of artificial neural-network potentials for atomistic materials simulations: performance for TiO\(_2\). Comput. Mater. Sci. 114, 135–150 (2016).
7. Smith, J. S., Isaev, O. & Roitberg, A. E. ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost. Chem. Sci. 8, 3192–3203 (2017).
8. Chmiela, S. et al. Machine learning of accurate energy-conserving molecular force fields. Sci. Adv. 3, e1603015 (2017).
9. Behler, J. First principles neural network potentials for reactive simulations of large molecular and condensed systems. Angew. Chem. Int. Ed. 56, 12828–12840 (2017).
10. Huan, T. D. A universal strategy for the creation of machine-learning-atomistic force fields. npj Comput. Mater. 3, 37 (2017).
11. Chmiela, S., Sauceda, H. E., Müller, K.-R. & Tkatchenko, A. Towards exact molecular dynamics simulations with machine-learned force fields. Nat. Commun. 9, 3887 (2018).
12. 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).
13. Caro, M. A., Deringer, V. L., Koskinen, J., Laurila, T. & Csányi, G. Growth mechanism of mixed dislocations in bcc Fe from first-principles calculations. Phys. Rev. Lett. 120, 166101 (2018).
14. Hellström, M., Quaranta, V. & Behler, J. One-dimensional vs. two-dimensional proton transport processes at solid-liquid zinc-oxide–water interfaces. Chem. Sci. 10, 1232–1243 (2019).
15. Fellinger, M. R., Tan, A. M. Z., Hector, L. G. & Trinkle, D. R. Geometries of edge and screw dislocations in amorphous carbon structures. ChemPhysChem 18, 873–877 (2017).
16. 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).
ACKNOWLEDGEMENTS
We thank Profs. C.J. Pickard and D.M. Proserpio for ongoing valuable discussions. N.B. acknowledges support from the Office of Naval Research through the U.S. Naval Research Laboratory’s core basic research program. G.C. acknowledges EPSRC grants EP/P022596/1 and EP/L014742/1. V.L.D. acknowledges a Leverhulme Early Career Fellowship and support from the Isaac Newton Trust.

AUTHOR CONTRIBUTIONS
N.B., G.C., and V.L.D. jointly designed the research, developed the approach, and analyzed the data. N.B. developed the computational framework and performed the computations with input from all authors. V.L.D. wrote the paper with input from all authors. All authors revised the paper and approved its final version.

COMPETING INTERESTS
The authors declare no competing interests.

ADDITIONAL INFORMATION
Supplementary information is available for this paper at https://doi.org/10.1038/s41524-019-0236-6.

Correspondence and requests for materials should be addressed to V.L.D.

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

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

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2019