Ephemeral data derived potentials for random structure search

Chris J. Pickard*
Department of Materials Science & Metallurgy, University of Cambridge, 27 Charles Babbage Road, Cambridge CB3 0FS, United Kingdom and
Advanced Institute for Materials Research, Tohoku University 2-1-1 Katahira, Aoba, Sendai, 980-8577, Japan
(Dated: May 30, 2022)

Structure prediction has become a key task of the modern atomistic sciences, and depends on the rapid and reliable computation of energy landscapes. First principles density functional based calculations are highly reliable, faithfully describing entire energy landscapes. They are, however, computationally intensive and slow compared to interatomic potentials. Great progress has been made in the development of machine learning, or data derived, potentials, which promise to describe entire energy landscapes at first principles quality. Compared to first principles approaches, their preparation can be time consuming and delay searching. Ab initio random structure searching (AIRSS) is a straightforward and powerful approach to structure prediction, based on the stochastic generation of sensible initial structures, and their repeated local optimisation. Here, a scheme, compatible with AIRSS, for the rapid construction of disposable, or ephemeral, data derived potentials (EDDPs) is described. These potentials are constructed using a homogeneous, separable manybody environment vector, and iterative neural network fits, sparsely combined through non-negative least squares. The approach is first tested on methane, boron nitride, elemental boron and urea. In the case of boron, an EDDP generated using data from small unit cells is used to rediscover the complex γ-boron structure without recourse to symmetry or fragments. Finally, an EDDP generated for silane (SiH_4) at 500 GPa enables the discovery of an extremely complex, dense, structure which significantly modifies silane's high pressure phase diagram. This has implications for the theoretical exploration for high temperature superconductivity in the dense hydrides, which have so far largely depended on searches in smaller unit cells.

I. INTRODUCTION

The knowledge of the arrangement, and nature, of atoms in a system is an essential starting point for its theoretical or computational study. First-principles approaches to crystal structure prediction have provided a route to this knowledge which is independent of experiment or intuition.[1] Early approaches were based on evolutionary algorithms,[2] or random search,[3,4] but many related algorithms have been proposed since.[5,6] Over the last decade and a half, first-principles structure prediction has led to a number of computational "discoveries". These include dense transparent sodium,[7] the structure of phase III of hydrogen and its mixed phase IV,[8] and complex host-guest structures in aluminium at terapascal pressures.[9] The first application of random structure search[3] was to testing Ashcroft's prediction[10] that compressed hydrides might offer a route to high temperature superconductivity.[11] This has been dramatically confirmed with the experimental discovery of superconductivity in hydrogen sulphide at 203K[12] and 250K in LaH_{10}.[13] In both cases the structures were predicted from first principles and the superconductivity anticipated computationally.[14–16]

Ab initio random structure searching (AIRSS) is a particularly simple, yet powerful, approach to structure prediction.[4] Random structures are generated and relaxed to nearby local minima of the energy landscape, repeatedly and in parallel. With a focus on exploration rather than exploitation, the initial random structures are generated to broadly sample a sub-volume of the total configuration space, see Figure 1. This sub-volume is defined by the search parameters. These parameters include the range of unit cell volumes and shapes, species dependent minimum distances, structural or molecular units, and symmetry. If these settings are well chosen, the initial random structures are 'sensible' and steer the search to promising regions of the energy landscape. AIRSS depends on features of the first principles energy landscape for its effectiveness, in particular its relative smoothness.[4]

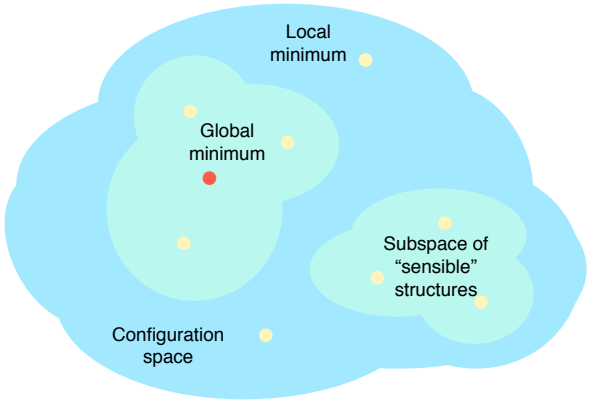

FIG. 1. A sketch of configuration space, highlighting regions which may be reached starting from structures assembled according to physically motivated biases and/or constraints. In general, the volume of configuration space accessible from these "sensible" initial structures will be very small compared to the total volume of configuration space.

The development of robust first principles codes to calculate the total energy of extended systems, through periodic boundary conditions,[17–19] along with databases of accurate pseudopotentials,[20] has enabled high throughput computational approaches. One high throughput approach is to compute properties of structures derived from experimental databases, such as the ICSD.[21,22] Structure prediction, and especially AIRSS, also depends on high throughput computations, with the structures rather generated stochastically.

Density functional theory (DFT) offers a very efficient way to compute electronic properties from first principles at the quantum mechanical level,[23] but it remains computationally expensive in the Kohn-Sham formulation, as single particle wavefunctions must for optimised for all the electrons in the system. During the 1980s, as the techniques behind modern DFT codes were being developed, there was a parallel interest in accelerating computations using empirical potentials.[24–26] Physically inspired functional forms for the interatomic potentials were constructed, and the free parameters fit to experimental data, or small datasets of first principles data.[25] With the advent of high throughput computation, which can rapidly generate large datasets, these approaches to fit potentials have been revisited, in the context of machine learning.[27]

Machine learning has a long history in the materials sciences.[28,29] In the 1990s attempts were made to use neural networks to learn electronic band-structures, to accelerate Brillouin-Zone integration for Electron Energy Loss Spectra prediction.[30] Neural networks were also used to fit complex energy landscapes of isolated systems,[31] density functionals,[32] and to predict alloy properties.[33]

Hampered by a relative lack of data, and the computational costs of training neural networks, it has taken some time for these approaches to become ubiquitous. Key to a revitalisation of the application of machine learning to interatomic potentials has been the work of Behler and Parrinello,[34] who emphasised the importance of decomposing the total energy into atomic contributions for neural network potentials, and Csanyi and coworkers,[35] who introduced the alternative gaussian approximation potentials. A wide variety of machine learning potentials are now available.[36–43] They vary depending on the strategy for assembling the training data,[44] describing the local environments,[45] and the machine learning model for regressing the energy landscape.

Structure prediction can be accelerated if the computational cost of evaluating the energy landscape can be reduced through efficient approximation.[46–54] If that approximation is robust and of sufficiently high quality, for all, or most, sampled configurations, AIRSS can be attempted. Here the development of a data derived potential, based on a many-body environment descriptor and the combination of many small neural networks, is described. Coupled with an iterative training scheme it is shown that potentials can be constructed, as needed, for a given set of search parameters. They are described as ephemeral, as there is no attempt to build a definitive potential for any given chemical system, and a new potential can be constructed from scratch at little cost.

In what follows, the scheme for generating the data derived, ephemeral or disposable, potential designed for random structure search is described. It is benchmarked first against a CH_4 dataset, then validated for boron nitride, elemental boron and urea. Finally, in an true test of the approach, it is used to uncover a complex dense phase of silane.

II. A DATA DERIVABLE POTENTIAL

An idea central to the development of potentials is that the total energy of a collection of N atoms can be decomposed into the individual contributions of each atom:

$$E = \sum_i^N E_i. \tag{1}$$

When combined with the approximation that the energy of each atom, E_i, depends on the environment of that atom within some localised region, typically a sphere with cutoff radius r_c, fast linear scaling computational schemes are possible.

The energy of each atom, E_i, can be further decomposed into terms that depend on the interactions between increasing numbers of surrounding atoms:

$$E_i = E_i^{(0)} + E_i^{(1)} + E_i^{(2)} + E_i^{(3)} + E_i^{(4)} + \cdots . \tag{2}$$

The zero body term, $E_i^{(0)}$, is typically dropped as it describes a chemical species independent energy offset, leading to a rigid shift of the total energy of the system regardless of composition.

The one body term, $E_i^{(1)}$, depends only on the chemical species of atom i. In an elemental system, or one of any fixed composition, it again leads to an overal rigid shift of the total energy, and can be ignored. It is vital, however, for the description of compounds with variable composition.

A. Two body interactions

The two body term, $E^{(2)}$, is the first that leads to a non-trivial energy landscape. Physically, it describes the attraction, or repulsion between pairs of atoms. The earliest potentials applied to model materials, such as the Lennard-Jones potential, were two body potentials. The Lennard-Jones potential, with its linear, homogenous, form compromises between computational efficiency and physical motivation. This might be contrasted with the inhomogeneous, and non-linear, Buckingham potential

with an exponential term describing the repulsion between closed electron shells, a $1/r^6$ term describing attractive dispersion interactions, and a Coulomb term.

Here, we follow the compromise made by Lennard-Jones, and choose a homogeneous linear potential with the form:

$$E_i^{(2)} = \sum_{j\neq i}^{N} \left(w_1^{(2)} f(r_{ij})^{p_1} + w_2^{(2)} f(r_{ij})^{p_2} \right), \tag{3}$$

or

$$E_i^{(2)} = \sum_{j\neq i}^{N} \sum_{m}^{2} w_m^{(2)} f(r_{ij})^{p_m}, \tag{4}$$

in the case of two terms (as for the Lennard-Jones potential), and with a general form:

$$E_i^{(2)} = \sum_{j\neq i}^{N} \sum_{m}^{M} w_m^{(2)} f(r_{ij})^{p_m}. \tag{5}$$

The sum is over the $N-1$ other atoms, and over M fixed exponents, or powers, p_m. The linear weights w_m are parameters to be determined, and the $f(r)$ is a fixed functional form.

For the original Lennard-Jones potential, $f(r) = 1/r$, $w_1 = 1$, $w_2 = -1$, $p_1 = 12$, and $p_2 = 6$. Extended Lennard-Jones potentials[55] resemble our general form, which can be written as a scalar product between a weight vector $\mathbf{w}_{(2)}$, and a vector $\mathbf{F}_i^{(2)}$, which contains information about the environment of atom i:

$$E_i^{(2)} = \sum_{m}^{M} w_m^{(2)} \sum_{j\neq i}^{N} f(r_{ij})^{p_m} = \mathbf{w}_{(2)}^{\intercal} \mathbf{F}_i^{(2)}. \tag{6}$$

B. Range cutoff

The Lennard-Jones potential is long ranged, in that there is no natural cutoff. This range is physically motivated, but it presents problems for computations of condensed systems. This has long been recognised, and managed through the imposition of range cutoffs, along with shifting and adjusting the potential so that it is zero at the cutoff radius, r_c, potentially along with the gradient and higher derivatives. This is known to have an important impact on the energy landscape, and indeed the ground state crystal structures.[56] Recently, Wang *et. al.*[57] proposed an alternative to the Lennard-Jones potential that is appropriately cutoff by construction, recognising the importance of both computationally efficient and well defined potentials. Their approach is taken here, and $f(r)$ is constructed so that it is zero at and beyond r_c. There are many functions which satisfy this condition, but we choose:

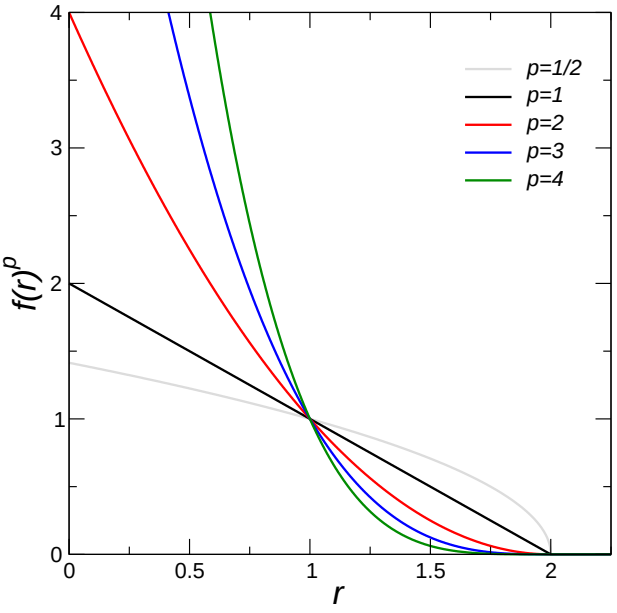

FIG. 2. The function $f(r)$, defined in Eqn. 7, raised to a range of exponents for the cutoff radius, $r_c = 2$.

$$f(r) = \begin{cases} 2(1 - r/r_c) & r \leq r_c \\ 0 & r > r_c. \end{cases} \tag{7}$$

When all the exponents, p_k, to which $f(r)$ is raised are two or greater both the resulting potential, and its gradient, at r_c are zero, by construction. Higher derivatives can also be forced to be zero by further increasing the minimum exponent. Exponents that are less than one (but greater than zero) generate step-like functions, with steep gradients approaching r_c, as shown in Fig. 2 for $p = 1/2$. In what follows all exponents are chosen to be two or greater.

C. Three body interactions

Without the careful design of unphysical two body potentials,[58] the range of structures that can be supported in the elements is extremely limited, to those that are well packed. However, the elements are known to exhibit extremely rich, and potentially open structures. For example, the diverse polymorphism in carbon and the extremely complex phosphorous and boron structures. Contributions are required to the potential that can distinguish between bond angles in triplets of atoms. A three body interaction term can achieve this, and since three distances r_{ij}, r_{ik}, and r_{jk} uniquely determine the triangle formed by the three atoms, i, j, and k, it can be written generally as:

$$E_i^{(3)} = \sum_{j\neq i}^{N} \sum_{k>j\neq i}^{N} V(r_{ij}, r_{ik}, r_{jk}). \tag{8}$$

The function $V(r_{ij}, r_{ik}, r_{jk})$ remains to be parameterised. Consistently with our treatment of the two body interactions, we write it as a linear, homogeneous, and separable approximation[25]:

$$E_i^{(3)} = \sum_{j\neq i}^{N} \sum_{k>j\neq i}^{N} \sum_{m}^{M} \sum_{o}^{O} w_{mo}^{(3)} f(r_{ij})^{p_m} f(r_{ik})^{p_m} f(r_{jk})^{q_o}. \tag{9}$$

The individual terms must be invariant to the swapping of the j and k atoms, as is the case in the above by construction. The summation can be rearranged, as for the two body terms:

$$E_i^{(3)} = \sum_{m}^{M} \sum_{o}^{O} w_{mo}^{(3)} \sum_{j\neq i}^{N} \sum_{k>j\neq i}^{N} f(r_{ij})^{p_m} f(r_{ik})^{p_m} f(r_{jk})^{q_o}, \tag{10}$$

and so

$$E_i^{(3)} = \sum_{m}^{M} \sum_{o}^{O} w_{mo}^{(3)} F_{i,mo}^{(3)} = \mathbf{w}_{(3)}^{\intercal} \mathbf{F}_i^{(3)}. \tag{11}$$

The three body terms can therefore also be written as a scalar product between the weight vector $\mathbf{w}_{(3)}$ and the vector $\mathbf{F}_i^{(3)}$, which describes the environment around atom i, taking into account three body interactions.

In principle the construction we have adopted to describe the three body interactions can be readily extended to four body interactions (see Fig. 3) and beyond. However, what follows is limited to three body potentials througout.

Our construction is related to atomic body-ordered permutation-invariant polynomials, where our basis is not complete, but carefully chosen to be computationally efficient and provide sufficient accuracy.[59]

D. Vectorisation and Multiple species

For a system containing multiple species the one body contribtion to the atomic energy, E_i, is important.

$$E_i^{(1)} = \mathbf{w}_{(1)}^{\intercal} \mathbf{F}_i^{(1)}. \tag{12}$$

The one body environment vector, $\mathbf{F}_i^{(1)}$, has the size of the total number of species, and assuming full occupancy, one (1) is added to the n^{th} element if atom i is of species n. The two body environment vector, $\mathbf{F}_i^{(2)}$, is constructed by concatenating environment vectors for each of the species pairs. For example, for two species, A and B:

$$\mathbf{F}_i^{(2)} = \mathbf{F}_{\mathrm{AA},i}^{(2)} \oplus \mathbf{F}_{\mathrm{AB},i}^{(2)} \oplus \mathbf{F}_{\mathrm{BA},i}^{(2)} \oplus \mathbf{F}_{\mathrm{BB},i}^{(2)}. \tag{13}$$

Note that in the case of full occupancy, and if atom i is of species A then the second half of the vector will be precisely zero. This leads to substantial sparsity. The three body environment vector is similarly constructed from concatenated contributions from triplets of species, where $\mathbf{F}_{\mathrm{ABA},i}^{(3)}$, for example, is equivalent to $\mathbf{F}_{\mathrm{AAB},i}^{(3)}$, and dropped. While it is not explored further here, this construction is suited to fractional and mixed occupation.

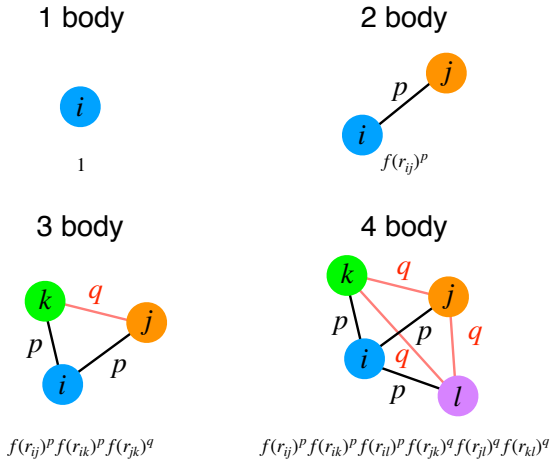

FIG. 3. Contributions to the environment vectors due to one, two, three and four bodies. The exponent p is applied to functions of the distance from the central atom, i, and the exponent q between the other atoms.

It is computationally convenient to further concatenate the one, two and three body environment vectors through the direct sum:

$$\mathbf{F}_i = \mathbf{F}_i^{(1)} \oplus \mathbf{F}_i^{(2)} \oplus \mathbf{F}_i^{(3)}. \tag{14}$$

This single vector, $\mathbf{F}_i$, describes the environment of the atom i, considering up to three bodies, and taking atomic species into account.

III. FITTING THE POTENTIAL

Once the environmental (or feature) vectors have been chosen, there are many possible choices when it comes to the functional form and fitting procedure. We now describe the scheme selected in this work. To guide the choices, a number of considerations are made. The goal is to produce a method that is robust, in that a large fraction of the structures obtained, on relaxing random sensible structures, remain sensible and physical. Further, the method should be computationally rapid. The aim is structure prediction, and the more time and computational resources spent searching for structures the better. There should also be a minimum number of parameters, and reasonable settings that apply to many systems are preferred. The overall method should demand as little intervention from the user as feasible.

A. Cost function

The design of the cost function influences the nature of the resulting fit. While it is common to fit to both the energy landscape itself, and the forces (and sometimes stresses), which are readily available within DFT, here we construct a cost function based on total energy alone:

$$C = \frac{1}{S}\sum_s \left| \sum_i^{N_s} (E(\mathbf{F}_{s,i}) - E_s) \right|^p . \qquad (15)$$

The sum is over the S structures, s, in the training data set, with energies E_s and number of atoms N_s. The concatenated vectors, $\mathbf{F}_{s,i}$, describing the environment of atom i in structure s are the input for the function $E(\mathbf{F})$ which computes the local energy for an atom with environment $\mathbf{F}$. The magnitude of the difference between the predicted and target energies is raised to the power p. For $p = 2$ the standard least squares cost function is recovered, whereas for $p = 1$, minimising the cost function reduces the mean absolute error. To deemphasise the impact on the cost function of a few very poorly predicted local energies (which will typically be encountered in highly energetic and unphysical structures far from the low energy structural minima) an intermediate value of $p = 1.25$ is chosen. In principle the individual terms in the cost function can be weighted. This is not found to be necessary in the current scheme.

B. Neural network

In Section II, a linear potential was developed from the environment vectors, $\mathbf{F}$, and weights $\mathbf{w}$: $E_i = \mathbf{w}^\intercal \mathbf{F}_i$. For $p = 2$, a closed form for the weights that minimises the cost function C can be computed. However, such a potential is limited in the form of the potential energy surface that can be modelled. Non-linear fits promise to describe more complex surfaces, but are more challenging to perform. Neural networks are recognised as a particularly powerful way to carry out general non-linear fits.[60] They have proven to be particularly adept for tasks of computational two dimensional image processing, such as classification. These breakthroughs have been built on deep (multilayer) neural networks,[61] with large number of nodes in each layer. The resulting very large number of weights are optimised through specialist computer codes running on GPUs.[62,63] In this work, in contrast, shallow narrow neural networks are found to be sufficient, and considerably easier to manage computationally. The architecture consists of an input layer of the size of the vector $\mathbf{F}$, a hidden layer with between 5 and 10 nodes, and a single output node for the predicted atomic energy. The total number of weights required is modest. Both the inputs and outputs are normalised on the training data, and a tanh activation is used between the input and hidden layer, and a linear activation on output.

C. Levenberg-Marquardt Iteratively Reweighted Least Squares

Deep neural networks are typically fit (trained) using stochastic gradient descent,[64] in which gradients are computed from random subsets (batches) of the training data. Given the small size of the neural networks employed here, direct minimisation is more appropriate. General quasi-Newton optimisers empirically did not perform particularly well for this task, converging slowly to poor solutions. Given the suitable structure of the cost function, the powerful Levenberg-Marquardt algorithm can be used.[65,66] Excellent fits are reliably obtained in modest numbers of iterations. Although implemented, geodesic acceleration[67] was not observed to significantly improve or speed up the fits in this case. As originally formulated, the Levenberg-Marquardt algorithm performs an optimisation of a least squares cost function. For $p \neq 2$, an approach based on iteratively reweighted least squares is required.[68] Overfitting is avoided through early stopping.[69] As the optimisation progresses the cost of a validation data set, C_v, is monitored. If the validation cost increases for, typically, ten steps the optimisation is halted and the weights for the minimum C_v are selected.

D. Non-negative least squares combination

In contrast to linear least square fits, fitting non-linear functions is a task of non-convex optimisation, leading to a multitude of potential solutions corresponding to the many local minima of the cost function depending on the initialisation of the weights. It is claimed that for neural networks many of these individual solutions lead to good fits.[70] An alternative is to average a number of fits to produce stabilised ensemble neural networks.[71,72] An attempt was made to linearly combine multiple fits to minimise the cost function for the validation data set (to which the neural networks had not been directly fitted). Extremely low cost functions for both the training and validation sets can be achieved, given a sufficient number of individual fits, suggesting that these fits are diverse. However, it was observed that many of the weights were large and alternating in sign, and the large costs for the held out testing set implied overfitting. In any case, such a combination is unphysical. Ideally one would hope to observe many small positive weights, resulting in an "adding" of the individual potentials or fits. To directly enforce positive weights, non-negative least squares (NNLS)[73] can be employed. NNLS has the property of producing sparse solutions, in that the weights are either positive, or precisely zero. For this application, it is found that out of, for example, 256 individual neural network fits, around 20 are selected by the NNLS. The combined NNLS potentials are found to be considerably more robust than potentials based on single fits. At the same time they are more computationally efficient than ensemble averages, automatically discarding

any relatively poor individual fits.

IV. ITERATIVE FITTING

Closely following the approach developed in Refs. 49 and 74, the fitting is carried out iteratively, in the manner of the scheme described in Fig. 4. First, random sensible structures are generated, according to the structure building parameters chosen for the specific AIRSS search for which the potential will be used. Without relaxation, the total energies are computed using DFT and stored along with the structures. These structures will span the entire region of configuration space accessible consistent with the biases implied by the AIRSS parameters (for example, unit cell volume ranges, minimum separations and space groups). Because the structures are unrelaxed, the typical total energies will be high. These samples instruct the potential about the high energy regions of the energy landscape and play an important role in the generation of robust potentials that are suitable for random search. Without these samples at high total energy it is likely that the potential will adopt low and unphysical total energies for these regions of configuration space. On structural optimisation this can lead to pathological structures with, for example, extremely close contacts.

The second step is optional, and involves taking so-called "marker" structures and applying random small amplitude displacements to their ionic positions, and lattice vectors. Again, the unrelaxed DFT total energies are computed and stored. The marker structures are typically chosen to be known structures in the system of interest. They may be derived from experiment, or earlier traditional AIRSS searches. Given that forces and stresses are not present in the cost function, the role of the shaking of the structures is to provide information about the gradients of the potential energy landscape. A related approach is the Taylor expansion method of Ref. 75.

At this point, a data set has been generated that is both broadly representative of the accessible configuration space, and, if marker structures as selected, of some of the low energy portion of the energy landscape. The environmental vectors $\mathbf{F}_{s,i}$ are computed for all the structures, which are randomly divided into training, validation and testing subsets in an approximately 80:10:10 ratio. A potential is then generated using the scheme described in Sections III B to III D.

It is quite likely that the quality of this first fit will not be particularly good, as monitored through the cost of the held out testing set, C_t. In order to expand the data set, and to ensure the final potential does not lead to a large number of unphysical low energy local minima, the following iterative procedure is followed. An AIRSS calculation is carried out, using the same structure building parameters and the most recently generated potential, to generate a number of local minima of the potential energy landscape. These structures are subjected to a number of random distortions, as for the marker structures, and the DFT total energies are computed and stored without relaxation. The combined data set is again randomly split into training, validation and testing subsets, and a new potential computed. The next iteration then begins. Either a fixed number of iterations can be performed, or the procedure halted when the quality of the fit, as measured by C_t, no longer significantly improves.

V. IMPLEMENTATION

The implementation consists of a collection of `OpenMP Fortran` codes, and `bash` scripts, assembled into three separate packages. The `nn` package is a `Fortan` implementation of multilayer neural networks, which is used by the `ddp` package to generate the EDDP potentials, and the `repose` code which performs variable cell structural optimisations using a preconditioned[76] Barzilai-Borwein[77] scheme. The `ddp` package consists of several codes. The `frank` code and `franks` script generate the environment vectors for a given input structure, singly and multiply, respectively. The `franks` script exploits the `parallel` tool[78] to parallelise the environment vector generation. The `forge` code performs individual potential neural network fits, while the `farm` script manages the high throughput multiple fits. The `flock` code combines the multiple individual fits into a single EDDP using NNLS. The `chain` script automates the iterative fitting scheme, and the `repose` code is integrated into the GPL2 `AIRSS` package.[79] The `ddp`, `repose`, and `nn` packages are also available under GPL2.[80]

The following examples were computed using a head node with 28 cores attached to 32 compute nodes, each with 32 cores and accessible by `ssh`. Each neural network was trained using 4 `OpenMP` cores, permitting 256 fits to be performed in parallel. The CASTEP plane wave total energy package[18] is used to compute the non-spin polarised DFT properties throughout.

VI. METHANE MOLECULE

As a first, and challenging, test we follow Ref. 81 and generate a data set of randomly distorted methane (CH_4) molecules. As in Ref. 81 the central carbon atom is fixed, and the four hydrogen atoms are randomly added within a sphere of radius 3Å. If any interatomic distance is less than 0.5Å, the configuration is rejected. The molecule is placed in a unit cell of side length 10Å, and the single point total energies computed using DFT as implemented in the CASTEP code[18] with the PBE exchange correlation functional.[82] The QC5 on-the-fly pseudopotentials (`1|0.9|7|7|9|10(qc=5)` for H, and `2|1.4|8|9|10|20:21(qc=5)` for C) are used, with a plane wave cutoff of 340eV. Generating 10,000 configurations, and dividing them into training, validation and testing subsets in an approximately 80:10:10 ratio, a

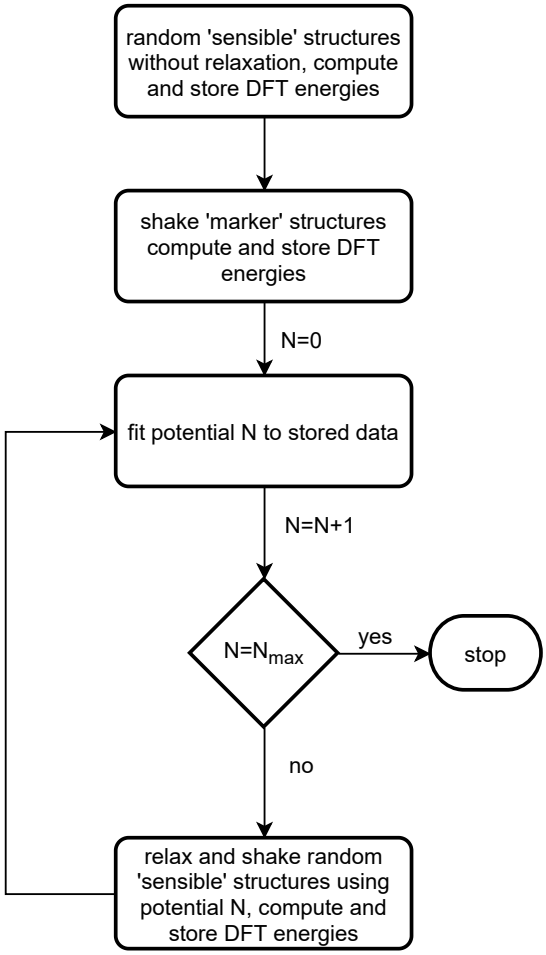

FIG. 4. A flow diagram outlining the iterative approach to fitting. The use of *marker* structures is optional. A typical value for N is 5.

three body EDDP is generated five times, with $r_c = 6$, 8 exponents ranging from 2 to 10, and 5 hidden nodes. Typically, of 256 individual fits, NNLS selects less than 10%. The best potential of the five resulted in a root mean square error (RMSE) of 0.13 eV/mol, and the worst 0.18 eV/mol. Repeating with 50,000 configurations the best and worst were 0.12 eV/mol and 0.13 eV/mol respectively. The RMSE for 10,000 configurations is somewhat lower than the best reported in Fig. 4c of Ref. 81, but the 50,000 configuration result is similar. This suggests that this EDDP, with its modest number of parameters, performs very well, but the fit does not improve rapidly with larger datasets. This is an acceptable compromise for the current application, where low energy candidate structures will ultimately be relaxed using DFT.

VII. BORON NITRIDE

As a first test of the iterative scheme described in Section IV we explore the construction of a three body EDDP for boron nitride. Boron nitride adopts a hexagonal layered polymorph as its most stable form, with the denser tetrahedral cubic polymorph being metastable. Cubic boron nitride can be synthesised at high pressures and temperatures. A hexagonal dense wurztite tetrahedral structure can also be formed at high pressure.

A. Potential Generation

The EDDP is generated from 4 formula unit (f.u.) boron nitride structures (8 atoms). The volumes of the unit cells are chosen randomly and uniformly from 4 to 8 Å^3/atom, no symmetry is applied, and minimum separations of 1 to 2 Å are randomly selected. No marker structures are used. 1000 fully random structures are generated in the first phase, and then 5 cycles of performing random searching using the current EDDP is performed, generating 100 local minima per cycle. Each of these minima are shaken 10 times, with an amplitude of 0.02 (AIRSS parameters POSAMP and CELLAMP). The total energy of each configuration is computed using CASTEP,[18] the PBE exchange correlation functional,[82] QC5 on-the-fly pseudopotential (boron definition string `2|1.4|7|7|9|20:21(qc=5)`, and nitrogen `2|1.4|13|15|17|20:21(qc=5)`), with a 440 eV plane wave cutoff and k-point sampling of $0.05\times2\pi$ Å^{-1}. Each generation of EDDP is constructed using the same parameters. The cutoff radius, r_c, is 3.75Å, and 4 exponents, ranging from 2 to 10, are used. Non-linear fits (256 in total) are performed with a neural network with 114 inputs, 5 hidden nodes in a single layer, and a single output for the predicted atomic energy, and 581 weights in total. The subsequent NNLS fit to the validation data selects 28 potentials with a non-zero weight. The final EDDP is based on 6495 structures and energies, split into training, validation sets in the ratio 5196:649:650, and has training, validation and testing RMSE of 42, 55, and 86 meV/atom, respectively. The testing RMSE is considerably larger then those of the training and validation data sets. However, as is clear in Figure 5, this is the result of deviations of the predicted energy landscape from the DFT one only at high energies, and so is benign. The data set contains structures with energies up to 11.84 eV/atom above the minimum. The Spearman rank correlation coefficient is above 0.99 for all sets, suggesting a good ordering of the predicted energies. Including iteratively building the DFT data set, the EDDP took just 23 minutes to construct.

B. Structure searches

Extensive structure searches with the final EDDP and the same structure generation parameters as used in its construction were performed for a larger unit cell of 8 f.u. None of the 55,000 fully relaxed structures contained close contacts. The lowest energy structures were either

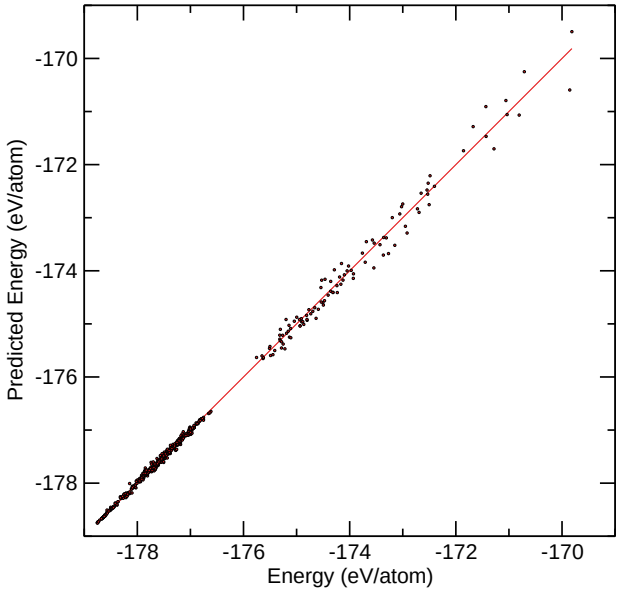

FIG. 5. The energy per atom predicted by the EDDP plotted against PBE DFT energies for the 650 boron nitride testing configurations. Note that despite the relatively large overall RMSE of 86 meV/atom, the error at low energies is small, around 18 meV/atom up to 0.5 eV above the ground state, and around 34 meV/atom up to 3 eV.

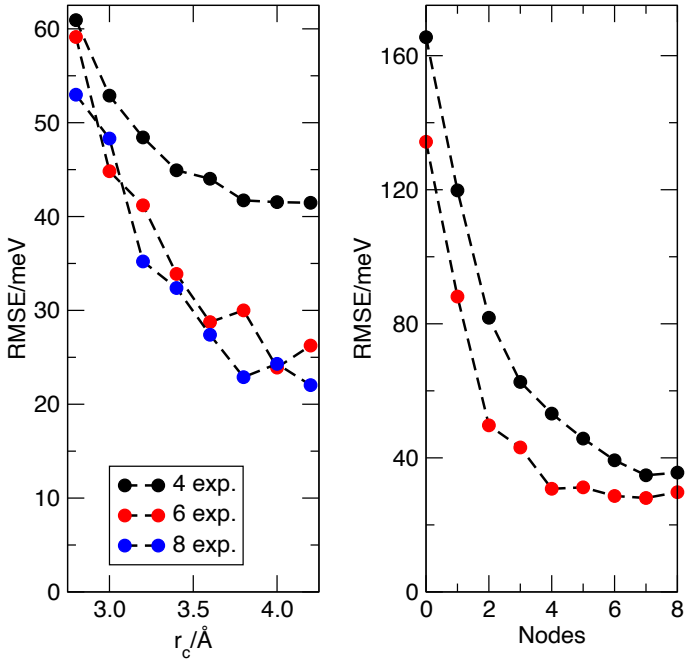

FIG. 6. The RMSE per atom for EDDPs refit to the iteratively generated boron nitride dataset. *Left:* Variation in the fit with the cutoff radius and number of exponents. *Right:* Variation in the fit with the number of hidden nodes in the neural networks, and number of exponents.

layered hexagonal or dense cubic boron nitride, or related stackings. The energy difference between relaxed hexagonal and cubic boron nitride is 77.5 meV/atom in PBE DFT, and 79.5 meV/atom using the EDDP, suggesting that the potential provides an excellent ranking at a greatly reduced computational cost. The 55,000 structures were generated in just 12 minutes using 1024 Intel Xeon Gold 6142 CPU @2.60GHz compute cores. Performing an identical structure search, using CASTEP for the first principles structural optimisations, results in 1080 structure over 11.5 hrs. This suggests that searching using an EDDP is over 250 times faster than DFT for this application. It should be noted that the EDDP optmisations are performed to machine precision, while the DFT relaxations are terminated when the forces and stresses fall below 0.05 eV/Å and 0.1 GPa, respectively, which results in far fewer DFT optimisation steps. The EDDP calculations scale linearly with number of atoms, so the acceleration for larger systems will grow rapidly. For example, the computation of the forces and stresses for a 256 atom boron nitride structure is nearly 10^5 times faster using the EDDP as compared to DFT. Should the DFT data have been computed using, for example, a denser k-point mesh, as would be required for the accurate description of a metallic system, the acceleration would be larger still.

C. Parameters

The EDDP potential for boron nitride was created without particular consideration as to the optimal parameters, such as the cutoff radius, number of exponents, or size of the neural network. The aim is to perform an accelerated structure search with as little time invested into potential generation and parameter refinement as possible. However, it is interesting to investigate how sensitive the resulting potential might be to the chosen parameters. In Figure 6 the impact of varying the number of exponents, cutoff radius, and number of hidden nodes, is explored. The previously iteratively generated data is randomly resplit into training, validation and testing sets (in the ratio 80:10:10) for each refitting of the EDDP. It is clear that the 3.75Å cutoff radius was a reasonable choice, but that increasing the number of exponents from 4 to 6 significantly improves the fit. However, increasing further to 8 exponents provides relatively little further improvement, at an increased computational cost. The fit is also seen to only improve marginally, if at all, for more than 5 hidden nodes in the neural networks. Repeating the iterative generation of a three body EDDP with 6 rather that 4 exponents leads to improved training, validation and testing RMSEs of 26, 38, and 67 meV/atom, respectively. The testing RMSE is just 20 meV/atom up to 3 eV above the ground state.

VIII. BORON

Elemental boron exhibits extremely complex crystal structures, from the purely icosahedral α-boron, to high pressure γ-boron, which consists of icosahedra and dimers which exchange charge to form an elemental ionic solid,[83,84] and the exceedingly complex β-boron[85,86], the structure of which continues to be studied,[87] but is thought to consist of icosahedra and larger defected clusters in a complex arrangement. This structural richness has ensured boron has played an important role in the development of first principles crystal structure prediction.[27,53,88] We explore boron as a case study in crystal structure prediction using EDDPs.

A. Potential Generation

To reproduce the experience of investigating the boron system without any prior knowledge, the following procedure is followed. A three body EDDP is constructed using the iterative scheme detailed above. In the absence of the knowledge that 12 atom icosahedra are an important feature of low energy boron structures, the EDDP is generated from smaller 8 atom unit cells. The volumes of the unit cells are chosen randomly and uniformly from 3 to 10 Å^3/atom, no symmetry is applied, and minimum separations of 1 to 3 Å are randomly selected. In the spirit of a naive search, initially no marker structures are used. 1000 fully random structures are generated in the first phase, and then 5 cycles of performing random searching using the current EDDP is performed, generating 100 local minima per cycle. Each of these minima are shaken 10 times, with an amplitude of 0.02 (AIRSS parameters POSAMP and CELLAMP). The total energy of each configuration is computed using CASTEP,[18] the PBE exchange correlation functional,[82] the same boron QC5 on-the-fly pseudopotential as used for boron nitride, with a 340 eV plane wave cutoff and k-point sampling of $0.05\times2\pi$ Å^{-1}. Each generation of EDDP is constructed using the same parameters. The cutoff radius, r_c, is 3.75Å, and 4 exponents, ranging from 2 to 10, are used. Non-linear fits (256 in total) are performed with a neural network with 21 inputs, 5 hidden nodes in a single layer, and a single output for the predicted atomic energy, and 116 weights in total. The subsequent NNLS fit to the validation data selects just 15 potentials with a non-zero weight. The final EDDP is based on 6499 structures and energies, split into training, validation sets in the ratio 5199:650:650, and has training, validation and testing RMSE of 52, 52, and 59 meV/atom, respectively. The data set contains structures with energies up to 11.5 eV/atom above the minimum. The Spearman rank correlation coefficient is 0.98 for all sets, suggesting a good ordering of the predicted energies.

B. "Discovery" of α-boron

As a first test of the EDDP, a random search is performed using the same structure building parameters as used during the iterative fit, but with 12 atoms rather than the original 8. Despite the fact that the training set cannot contain α-boron, it is identified as the most stable structure (once some obviously pathological results, about 1 in 6000, are removed). How is this possible, given that the training structures can contain no icosahedra? Examining the most stable 8 atom structure in the training set (see Fig. 7) it appears that there are hints of icosahedral fragments in the small cell, which the EDDP is able to learn, without overfitting, given the relatively inflexible functional form. It should be noted, unsurprisingly, that this EDDP does not perfectly reproduce the DFT energy landscape. For instance, it would be expected to find the α-boron structure about 1 in 3000 random samples in a 12 atom unit cell, but using this EDDP it is reduced to about 1 in 10000 samples. Furthermore, the volume of the relaxed alpha boron structure differs substantially from the DFT result, by about 9%.

C. Structure solution for γ-boron

A second, more ambitious test, is the solution of the 28 atom gamma boron structure, from the knowledge of the lattice parameters alone. A random search, with initial structures with minimum separations of 1.7Å and randomly selected space-groups with two to four symmetry operators, was performed. The fixed unit cell search resulted in about 1 in 3000 obviously pathological structures. The otherwise lowest energy structures had the Pnn2 space group, a subgroup of Pnnm, adopted by the γ-boron structure, see Fig.7. On inspection the structure appears closely related to the known γ-boron structure, and subsequent structural optimisation of the Pnn2 structure within DFT recovers it precisely. This result is impressive - it is difficult to conceive that the 8 atom structures contain obvious hints of the complex icosahedral/dimer interactions.

D. Free search for γ-boron

Next, the challenging task of a symmetry and lattice free search for the γ-boron structure is attempted. The EDDP is regenerated using the α-boron structure, which has already been located, as a marker, which is shaken 500 times. The shake amplitude is increased to 0.04, the r_c to 4.5Å, the number of exponents to 8 and the hidden nodes to 10. To increase the chance of encountering pathological structures during the generation procedure, and to "dig deeper" into the EDDP's energy landscape, on the N^{th} step, in order to generate a single retained structure, 2^N relaxed random structures are generated,

and the lowest energy one selected. Using this potential, 362,754 structures containing 28 atoms are randomly generated and relaxed. A dense metastable structure with space group $P2_1/c$ is encountered twice. On inspection the structure appears to be only a very slight distortion of the γ-boron structure, and indeed, on relaxation using CASTEP, it becomes precisely the Pnnm γ-boron structure.

E. Structural distortion and potential range

To test a hypothesis that the observed distortions are due to the relatively short range of the potentials, a new EDDP is generated, this time increasing the cutoff to 5.5Å. The Pnn2 and $P2_1/c$ structures relax directly to the Pnnm structure using this EDDP. There is clearly a tradeoff between the number of samples that can be generated, which depends on the computational cost of the potential used, and the quality of the generated structure. Given that all important structures in a study will ultimately be relaxed using DFT, imperfections in computationally cheaper EDDPs can be tolerated in the pursuit of a more thorough coverage of the energy landscape. However, care must be taken as a poorly described energy landscape may contain more local minima, hence be more challenging to search.

Overall, these results for boron suggest that EDDPs are a promising basis for general random structure prediction tasks.

IX. UREA

Constructing a fully reactive potential for the entire C-H-N-O chemical space is expected to present challenges, not least in the generation and manipulation of suitably large training data sets. In the spirit if this work, here we generate and apply a three body EDDP for the specific region of C-H-N-O's configuration space that contains the urea (CH_4N_2O) molecule, at around atmospheric and moderate positive pressures. Phase transitions in urea (carbamide) under pressure were first studied by Bridgeman. Polymorphism in urea remains under active investigation, both experimentally[89] and computationally.[90–92] Here we explore the application of random searching and EDDPs to identify the low energy polymorphs of urea.

A. Potential Generation

In the first phase of the iterative construction of the potential for urea, structures are generated by constructing 10000 randomly shaped unit cells with volumes from 60 to 80 Å^3/mol, and placing two urea molecules with random positions and orientations, ensuring that the molecules are no closer to each other than a randomly selected distance from 1 to 2 Å . The positions of the atoms in the molecules are then perturbed by up to 0.3 Å. The same settings as for the first potential of boron, Section VIII A, are used for the iterative phases of relaxation and shaking, as well as the final construction of the potential. The energy of each configuration is computed using CASTEP,[18] the QC5 on-the-fly pseudopotentials (`2|1.4|13|15|17|20:21(qc=5)` for N, and `2|1.5|12|13|15|20:21(qc=5)` for O, and the same definitions for C and H as for the methane example), and a high plane wave cutoff of 540 eV. A coarse k-point grid

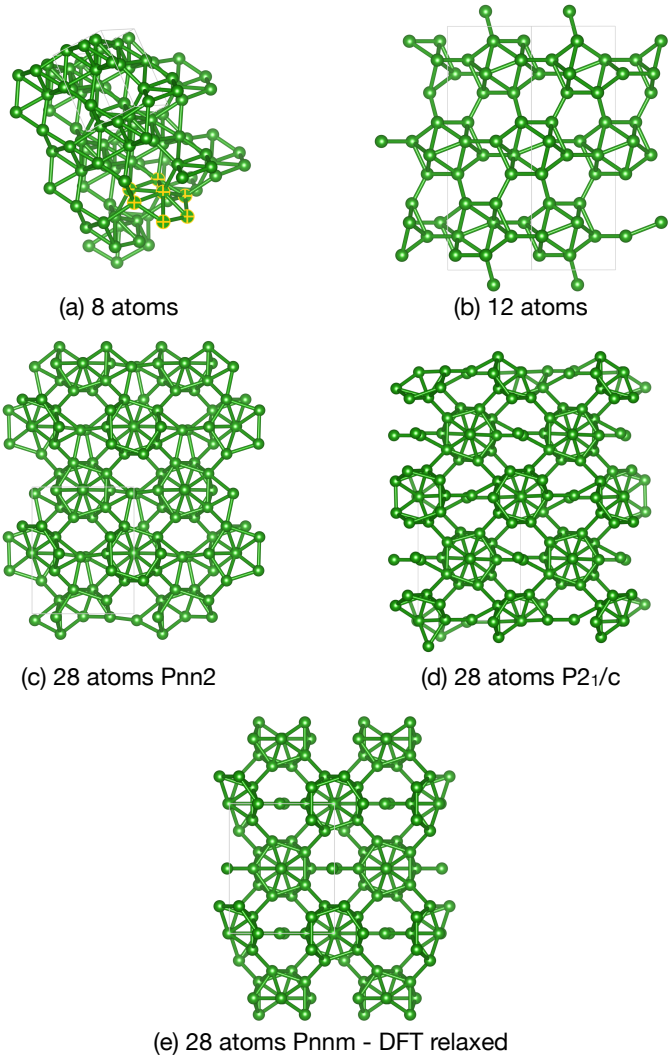

FIG. 7. Structure (a) is the lowest DFT energy configuration contained in the potential training data set for Section VIII A. A subset of the 8 atoms are highlighted as they resemble configuration encountered in icosahedral alpha boron. Structure (b) is the result of structure searches using this iteratively generated potential, and is that of alpha boron. Structure (c) is the result of structure searches in a unit cell with shape constrained to that of γ-boron. Structure (d) is the result of structure searches in variable unit cell with no imposed symmetry. On relaxation in DFT both the (c) and (d) structures become that of γ-boron, shown in (e).

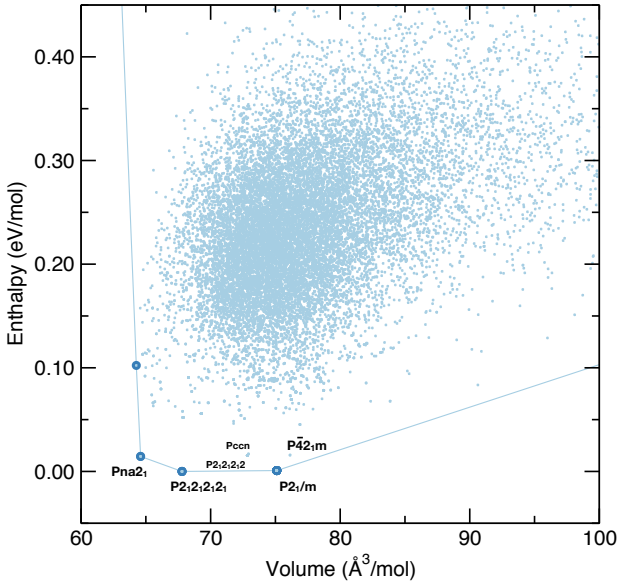

FIG. 8. Energy versus volume for the 16045 Z=4 urea structures relaxed using the EDDP. The fine blue line is the convex hull of the point, highlighting the structures that might become stable at positive and negative pressures.

spacing of $0.1\times2\pi$ Å^{-1} was used along with the PBE+TS dispersion corrected functional.[93] Of the 256 non-linear neural network fits, 38 were selected by NNLS. The final potential is based on 15500 structures and energies, split into training, validation and testing in the ratio 12400:1550:1550. The training, validation, and testing RMSE (MAE) is 20.65 (9.99), 27.50 (17.42), and 39.02 (18.76) meV/atom respectively. The data set contains structures with energies up to 5.52 eV/atom above the minimum and a Spearman rank correlation coefficient of 0.999 for all sets, demonstrating an excellent ordering of the predicted energies.

B. Structure Searches

Having generated the EDDP for urea using just two molecules per unit cell (Z=2), it is tested for Z=4. Unit cells with volumes ranging from 60 to 80 Å^3/mol are filled with four molecular units of urea. No symmetry is used to generate the structures, so in principle structures with up to Z′=4 are accessible. The molecules are placed so that they do not overlap, with a minimum separation of 2Å. The initial structures are relaxed to their nearby local minima 16045 times, generating a diverse set of structures. A scatter plot of the energy and volume of these structures is shown in Figure 8.

The lowest energy structure identified had Z=4 and space group P2$_1$2$_1$2$_1$. It was located 4 times, and is known as the high pressure Form III of urea. The ambient pressure P$\bar{4}$2$_1$m (Z=2) form I was located twice, and the high pressure P2$_1$2$_1$2 (Z=2) form IV was located 18 times. Additional structures with P2$_1$/m (Z=4), Pna2$_1$ (Z=4) (see Figure 9), and Pccn (Z=4) were identified at energies within 40 meV/mol of Form III. To assess the reliability of the ranking, the structures and energies are recomputed at both the PBE+TS level (using the same computational parameters as for the potential generation), and PBE+MBD* (the default CASTEP OTFG parameters, a plane wave cutoff of 900 eV and k-point sampling density of $0.07\times2\pi$ Å^{-1}).[94] As shown in Table I, in all cases Form III is found to be the lowest energy structure, with the maximum difference in relative enthalpy of 40 meV/mol, or 5 meV/atom. It is clear that the EDDP is capable of resolving differences in energy well below the testing RMSE.

Space	EDDP		PBE+TS		PBE+MBD*	
Group (Z)	V/Å^3	E/meV	V/Å^3	E/meV	V/Å^3	E/meV
P2$_1$2$_1$2$_1$ (4)	67.79	0	68.20	0	72.15	0
P2$_1$/m (4)	75.10	1	74.06	24	75.31	41
Pna2$_1$ (4)	64.59	14	65.37	2	67.20	18
P2$_1$2$_1$2 (2)	72.83	15	70.70	17	72.87	27
P$\bar{4}$2$_1$m (2)	76.13	16	71.35	13	71.86	23
Pccn (4)	72.93	17	70.00	53	70.84	55

TABLE I. Relative energies and volumes (per urea molecule) for the low energy structures, evaluated using the EDDP, at the PBE+TS level used to construct the potential, and PBE+MBD*.

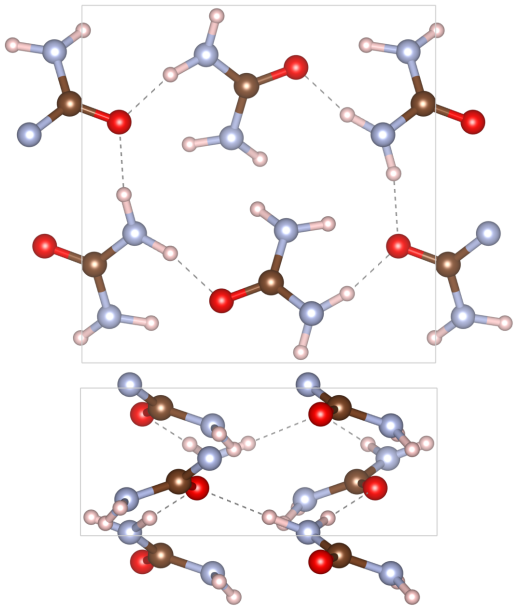

FIG. 9. The Pna2$_1$ (Z=4) urea structure is energetically competitive in all cases, and a candidate high pressure phase of urea given its high density.

X. APPLICATION TO DENSE SILANE

The earliest published application of first principles random searching (later referred to as *ab initio* random structure searching, AIRSS[4]) was to the study of high pressure polymorphism in silane.[3] Feng *et al.*[95] had proposed silane as a potential candidate for high temperature conventional superconductivity,using structures based on chemical intuition and local structural optimisation using DFT. In Ref. 3 random searches at around 100GPa using two f.u. of SiH_4 and just 40 initial configurations uncovered a more stable, semiconducting, phase of silane with space group $I4_1/a$. The presence of an electronic band-gap postponed any expectation of superconductivity to higher pressures. Shortly afterwards the $I4_1/a$ structure was encountered experimentally[96], and subsequent theoretical work, exploring larger unit cells of up to 6 f.u., identified further candidate structures at both higher and lower pressures.[97,98] Despite refinements to searching algorithms, and increased computational resources, structure predictions for binary and ternary compounds are still typically restricted to relatively small unit cells. Here we revisit silane, exploiting the computational acceleration afforded by EDDPs to search in larger unit cells (up to 16 f.u.).

A. Potential Generation

A three body EDDP was generated using the iterative scheme described in Section IV. Random unit cells were constructed with volumes ranging from 5 to 15 Å/f.u., containing just two f.u. The minimum separations between the species were randomly chosen to be between 1 and 2 Å, and no symmetry was imposed. The total energy of each configuration is computed using CASTEP,[18] the PBE exchange correlation functional,[82] QC5 on-the-fly pseudopotential (definition strings `3|1.8|4|5|5|30:31:32(qc=5)` for Si, and `1|0.9|7|7|9|10(qc=5)` for H), with a 340 eV plane wave cutoff and k-point grid spacing of $0.05\times2\pi$ Å^{-1}. The settings for the iterative scheme, and parameters for the potential, were identical to those used for boron, with one key difference. The random searches using each generation of the potential were performed by minimising the enthalpy at an elevated pressure of 500GPa. This ensures that the potential will be suitable for high pressure searches, around this pressure. Non-linear fits (256 in total) are performed with a neural network with 114 inputs, 5 hidden nodes in a single layer, and a single output for the predicted energy and 581 weights in total. The subsequent NNLS fit to the validation data set selects just 23 potentials with a non-zero weight. The final EDDP is based on 6500 structures and energies, split into training, validation sets in the ratio 5200:650:650, and has training, validation and testing RMSE of 9.98, 13.40, and 44.22 meV/atom, respectively. The MAE error for the testing set is considerably lower, at 10.77 meV/atom, which is an indication the higher RMSE is the result of a few structures with significant error. Indeed, the maximum error for the testing set is 876.02 meV/atom. The data set contains structures with energies up to 126.4 eV/atom above the minimum. The Spearman rank correlation coefficient is 0.999 for all sets, suggesting an excellent ordering of the predicted energies.

B. Structure searches

Having generated the EDDP suitable for SiH_4 at pressures around 500GPa, structures searches may be carried out. As a first test, an extensive search using the same structure generation parameters as used for the iterative construction of the EDDP was performed at 500GPa. Any structure that encounters close contacts (by default, defined at 0.5Å) during optimisation is rejected. Of the structures that survive optimisation, the most stable is the C/2c structure proposed in Ref. 3 as the very high pressure form of SiH_4. The 4 f.u. $P2_1/c$ structure reported in Ref. 98 is not accessible to a search restricted to 2 f.u.

The promise of using fast data derived potentials for structure searching is that much larger systems could be investigated if those potentials are sufficiently transferable. The challenge of larger systems is that both each individual structural optimisation is slower, with each step being more computationally expensive, and the structural optimisation requiring more of those steps, and that many more structures must be sampled to ensure the low energy regions of the energy landscape are adequately explored. Even if the same structure generation parameters are used for the potential generation and the search, exploring larger systems is necessarily an extrapolation. As such, an iteratively generated potential cannot be expected to result in the precise structures, and energy ordering, that a DFT search would. However, as we saw in the case of boron above, the EDDP does appear to offer extrapolation, and generates appropriate low energy structures. A pragmatic approach is to simply perform single point energy DFT computations at the end of each local optimisation using the EDDP. If the EDDP relaxed structures are reasonably close to what they would be within DFT the ranking obtained will be reliable, with any poor structures being pushed to the bottom of the ranking. This is the approach taken here.

We next perform searches at 500GPa with 3 and 4 f.u. of SiH_4, again using the same structure generation parameters, but this time constructing symmetric initial structures with 2 to 4 symmetry operations in the primitive cell. The $P2_1/c$ structure of Ref. 98 is rapidly recovered, along with the high pressure C2/c phase of Ref. 3.

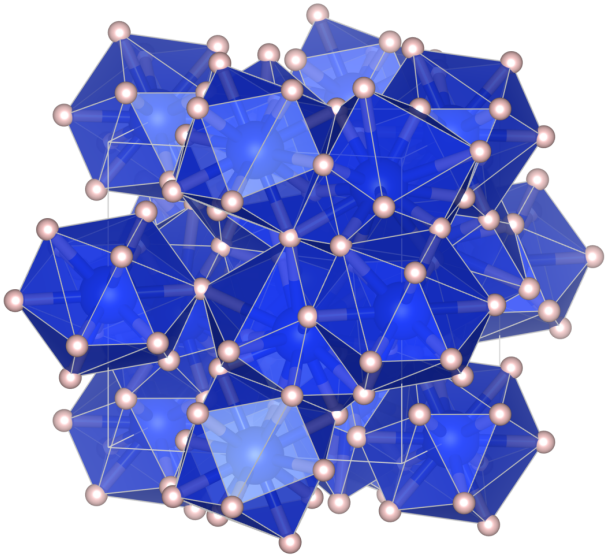

FIG. 10. A visualisation of the Pa$\bar{3}$ structure , created using the VESTA package.[99] This complex structure consists of 12 f.u. of SiH_4, or 60 atoms, in the primitive unit cell, and does not appear to be a named structure type.

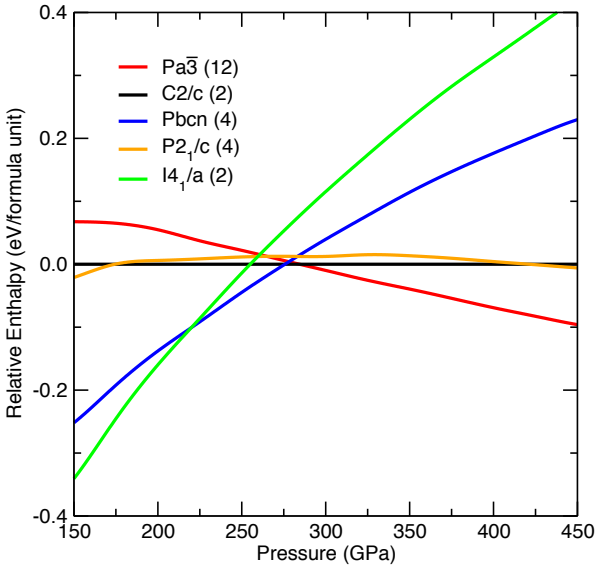

FIG. 11. Relative PBE DFT enthalpy plotted for a selection of SiH_4 polymorphs. The 60 atom Pa$\bar{3}$ structure is increasingly more stable than the P$2_1/c$ structure above 285GPa, leaving only a small window of stability for the C2/c structure from 276 to 285 GPa.

Space group	Lattice parameters (Å, °)	Atomic coordinates (fractional)			
Pa$\bar{3}$	a=b=c=4.998	Si1	0.1168	0.1168	0.1168
	α=β=γ=90.00	Si2	0.0000	0.0000	0.5000
		H1	0.1664	0.2236	0.3800
		H2	0.2246	0.4858	0.3756

TABLE II. Parameters for the Pa$\bar{3}$ structure of SiH_4 at 500GPa.

C. Identification of complex high pressure phase

Having demonstrated that the potential can recover the theoretically known high pressure structures of SiH_4, its computational efficiency can be exploited to explore much larger unit cells. A search at 500GPa is performed with up to 16 f.u. and using between 4 and 12 symmetry operators. A low enthalpy cubic structure with 12 f.u. is identified, see Figure 10 and Table II.

This structure adopts the high symmetry Pa$\bar{3}$ space group, and is characterised by two distinct silicon sites, one octahedrally coordinated by nearest neighbour silicon atoms, and the other tetrahedrally. To assess its dynamic stability, a hundred $3 \times 3 \times 3$ supercells of the cubic primitive cell, containing 1620 atoms, were constructed and "shaken" with a 0.1 amplitude. On relaxation with the EDDP all the distorted structure returned to the 60 atom Pa$\bar{3}$ space group unit cell. Computing the enthalpy of this structure, along with those previously reported, reveals that it has a wide range of stability at the static lattice level, from 285 GPa upwards using the PBE density functional. Using the rSCAN[100] functional it is stable above 305 GPa. It is significantly more dense than the competing phases, and so its relative stability grows with pressure, see Figure 11. The enthalpy curves were computed using CASTEP, a more accurate potential for hydrogen (`1|0.6|13|15|17|10(qc=8)`) and an increased plane wave cutoff of 700 eV. The electronic density of states (eDOS) for the Pa$\bar{3}$ and C2/c structures are reported in Figure 12. They were computed[101] with the same settings as for the enthalpy curves, but with a finer k-point grid spacing of $0.01 \times 2\pi$ Å^{-1}. The eDOS at the Fermi level for the Pa$\bar{3}$ structure is considerably lower than for the C2/c structure at 300 GPa, which can be attributed to its greater stability. Furthermore, without performing extremely costly density functional perturbation theory computations of T_c it is expected that this reduced eDOS would lower the prospects for high temperature superconductivity in silane at these pressures. Given that silane has been extensively studied theoretically, the emergence of such an important, and large unit cell, structure should inform our confidence in the status of our knowledge of the dense hydrides. It is very likely that more extensive searches for the dense binary hydrides, in large unit cells, will reveal a significant revision of our knowledge of these candidate high temperature superconductors.[11]

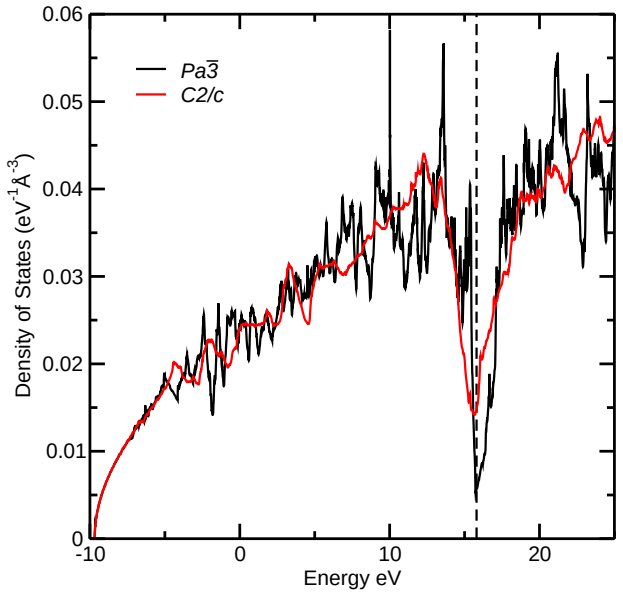

FIG. 12. The PBE DFT electronic density of states for the Pa$\bar{3}$ and C2/c structures computed at 300GPa. The density of states around the Fermi level (vertical dashed line) is considerably lower for the Pa$\bar{3}$ structure.

XI. DISCUSSION

First-principles methods owe their flexibility and applicability to databases of high-quality pseudopotentials, which allow arbitrary chemical systems to be explored. The CASTEP code[18] is unique in its on-the-fly pseudopotential methodology, where the pseudopotential is generated as needed, and consistently with the density functional chosen. This has opened the door to structure predictions at extreme densities, with small core potentials being generated as needed, and independently of the provided databases.

Here, the same flexibility is introduced to data derived potentials, which are generated specifically for the structure building parameters, and pressures, that will be used for each search. These potentials are ephemeral, in the sense that the next search performed will likely require a new, bespoke, potential. The ease and robustness of the scheme described makes this possible.

Random structure search is a challenging application of data derived potentials. It is very difficult to construct potentials that are stable across the entire space of possible inputs, or configurations. The initial random structures are extremely diverse, exploring many different regions of configuration space. Constructing the EDDPs from these diverse structures, generated from a given set of structure building parameters, is essential to ensure robustness.

For any finite training dataset, some failures are to be expected in an extended sampling of configuration space. A typical pathological behaviour is the encounter of very close contacts during structural optimisation or evolution. This could cause severe problems in a lengthy molecular dynamics simulations. However, during a random structure search such configurations may simply be rejected. A very similar situation is encountered in first-principles structure searches – for heavier elements, overlapping pseudopotentials cores can lead to problems in the calculation of the electronic structure, and common practice is to reject those configurations.

The pioneering work of Behler, and Csanyi, who introduced neural network, and Gaussian process based atomic potentials respectively, which can be fit to extensive databases of first-principles data, has led to an explosion of alternative schemes based on their key insights. It is worth reflecting on the justification of introducing yet another. In some sense, it is inevitable – there are countless valid approaches to the fitting of high dimensional functions, and while any scheme will share commonalities with the others in use, the details may differ, depending on the intended application. While the electronic structure community has coalesced around a few, very complex, computer codes, the relative simplicity of data derived potentials is likely to favour persistent diversity. In this case a scheme has been designed for random structure search.

The functional form for the EDDP has its origin in an earlier attempt to develop a few-parameter model 3-body potential that could describe the rich structure of the elements, going beyond simple close packing. Starting with the Lennard-Jones potential, this original model potential was written as follows:

$$E_i = \sum_{i \neq j} \left(\frac{A}{r_{ij}^{12}} - \frac{B}{r_{ij}^{6}} \right) + \sum_{j \neq i} \sum_{k > j \neq i} \frac{C}{r_{ij}^{n} r_{ik}^{n} r_{jk}^{m}}. \qquad (16)$$

By manually adjusting the parameters, A, B, C, n, and m, and performing random searches for each choice, it was found to be possible to navigate the space of possible elemental structures, from close packed, to the diamond lattice, and even the icosahedral α-boron structure. Exploring the properties of the simplified potential described in Eqn. 16 would be a fruitful topic of further investigation.

XII. CONCLUSION

Fitting of potentials to data generated across the whole accessible energy landscape ensures that the benign properties of the first-principles energy landscape are retained, and random search can be successfully performed. The computational simplicity of the form of the potential ensures that these searches are much accelerated compared to a purely first-principles approach. Close attention has been paid to develop a bespoke scheme that complements the computational workflow of structure search.

It has been shown that the EDDP potentials can be fit to first-principles data derived from much smaller unit

cells than are typically chosen for training. These potentials can be used to discover novel structural features in much larger unit cells. For example, a potential trained using unit cells containing just eight boron atoms was used to generate approximations to the 12-atom icosahedral alpha-boron structure, and the 28 atom gamma-boron. This extrapolation to larger unit cell sizes is essential if these potentials are to be successfully used to accelerate structure prediction.

EDDPs been used to revisit the high-pressure phase diagram of silane, uncovering a large (60-atom) unit cell structure that is considerably more stable at high pressures than those currently known. This structure had been overlooked, despite extensive investigation using both random search and evolutionary approaches. This is strong evidence that EDDPs are a powerful tool for the thorough exploration of structure space. At the same time, it suggests that many of the systems that have been explored using first-principles structure prediction should be revisited.

ACKNOWLEDGMENTS

CJP is supported by the EPSRC through grants EP/P022596/1, and EP/S021981/1, and thanks Gabor Csanyi, Chuck Witt and Lewis Conway for discussions, and further thanks Gabor Csanyi for his careful reading of the manuscript.

* cjp20@cam.ac.uk

[1] A. R. Oganov, C. J. Pickard, Q. Zhu, and R. J. Needs, Nature Reviews Materials **4**, 331 (2019).
[2] A. R. Oganov and C. W. Glass, The Journal of Chemical Physics **124**, 244704 (2006).
[3] C. J. Pickard and R. J. Needs, Physical Review Letters **97**, 045504 (2006).
[4] C. J. Pickard and R. J. Needs, Journal of Physics: Condensed Matter **23**, 053201 (2011).
[5] D. C. Lonie and E. Zurek, Computer Physics Communications **182**, 372 (2011).
[6] Y. Wang, J. Lv, L. Zhu, and Y. Ma, Computer Physics Communications **183**, 2063 (2012).
[7] Y. Ma, M. Eremets, A. R. Oganov, Y. Xie, I. Trojan, S. Medvedev, A. O. Lyakhov, M. Valle, and V. Prakapenka, Nature **458**, 182 (2009).
[8] C. J. Pickard and R. J. Needs, Nature Physics **3**, 473 (2007).
[9] C. J. Pickard and R. Needs, Nature Materials **9**, 624 (2010).
[10] N. Ashcroft, Physical Review Letters **92**, 187002 (2004).
[11] C. J. Pickard, I. Errea, and M. I. Eremets, Annual Review of Condensed Matter Physics **11**, 57 (2020).
[12] A. Drozdov, M. Eremets, I. Troyan, V. Ksenofontov, and S. I. Shylin, Nature **525**, 73 (2015).
[13] A. Drozdov, P. Kong, V. Minkov, S. Besedin, M. Kuzovnikov, S. Mozaffari, L. Balicas, F. Balakirev, D. Graf, V. Prakapenka, *et al.*, Nature **569**, 528 (2019).
[14] D. Duan, Y. Liu, F. Tian, D. Li, X. Huang, Z. Zhao, H. Yu, B. Liu, W. Tian, and T. Cui, Scientific Reports **4**, 1 (2014).
[15] F. Peng, Y. Sun, C. J. Pickard, R. J. Needs, Q. Wu, and Y. Ma, Physical Review Letters **119**, 107001 (2017).
[16] H. Liu, I. I. Naumov, R. Hoffmann, N. Ashcroft, and R. J. Hemley, Proceedings of the National Academy of Sciences **114**, 6990 (2017).
[17] G. Kresse and J. Furthmüller, Computational Materials Science **6**, 15 (1996).
[18] S. J. Clark, M. D. Segall, C. J. Pickard, P. J. Hasnip, M. I. Probert, K. Refson, and M. C. Payne, Zeitschrift für Kristallographie-Crystalline Materials **220**, 567 (2005).
[19] P. Giannozzi, S. Baroni, N. Bonini, M. Calandra, R. Car, C. Cavazzoni, D. Ceresoli, G. L. Chiarotti, M. Cococcioni, I. Dabo, *et al.*, Journal of physics: Condensed matter **21**, 395502 (2009).
[20] K. Lejaeghere, G. Bihlmayer, T. Björkman, P. Blaha, S. Blügel, V. Blum, D. Caliste, I. E. Castelli, S. J. Clark, A. Dal Corso, *et al.*, Science **351** (2016).
[21] S. Curtarolo, W. Setyawan, S. Wang, J. Xue, K. Yang, R. H. Taylor, L. J. Nelson, G. L. Hart, S. Sanvito, M. Buongiorno-Nardelli, *et al.*, Computational Materials Science **58**, 227 (2012).
[22] A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, *et al.*, APL Materials **1**, 011002 (2013).
[23] K. Burke, The Journal of Chemical Physics **136**, 150901 (2012).
[24] F. H. Stillinger and T. A. Weber, Physical Review B **31**, 5262 (1985).
[25] R. Biswas and D. Hamann, Physical Review Letters **55**, 2001 (1985).
[26] J. Tersoff, Physical Review Letters **61**, 2879 (1988).
[27] V. L. Deringer, M. A. Caro, and G. Csányi, Advanced Materials **31**, 1902765 (2019).
[28] H. K. D. H. Bhadeshia, ISIJ International **39**, 966 (1999).
[29] A. Skinner and J. Broughton, Modelling and Simulation in Materials Science and Engineering **3**, 371 (1995).
[30] C. J. Pickard, M. C. Payne, L. M. Brown, and M. N. Gibbs, in *Conference Series-Institute Of Physics*, Vol. 147 (IOP Publishing Ltd, 1995) pp. 211–214.
[31] D. F. R. Brown, M. N. Gibbs, and D. C. Clary, The Journal of Chemical Physics **105**, 7597 (1996).
[32] D. J. Tozer, V. E. Ingamells, and N. C. Handy, The Journal of Chemical Physics **105**, 9200 (1996).
[33] H. K. D. H. Bhadeshia, D. J. C. MacKay, and L.-E. Svensson, Materials Science and Technology **11**, 1046 (1995).
[34] J. Behler and M. Parrinello, Physical Review Letters **98**, 146401 (2007).
[35] A. P. Bartók, M. C. Payne, R. Kondor, and G. Csányi, Physical Review Letters **104**, 136403 (2010).
[36] A. P. Bartók, S. De, C. Poelking, N. Bernstein, J. R. Kermode, G. Csányi, and M. Ceriotti, Science Advances **3**, e1701816 (2017).
[37] J. Behler, Chemical Reviews (2021).
[38] A. Seko, A. Takahashi, and I. Tanaka, Physical Review B **92**, 054113 (2015).

[39] N. Artrith and A. Urban, Computational Materials Science **114**, 135 (2016).
[40] A. V. Shapeev, Multiscale Modeling & Simulation **14**, 1153 (2016).
[41] S. Hajinazar, J. Shao, and A. N. Kolmogorov, Physical Review B **95**, 014114 (2017).
[42] L. Zhang, J. Han, H. Wang, R. Car, and E. Weinan, Physical Review Letters **120**, 143001 (2018).
[43] M. Benoit, J. Amodeo, S. Combettes, I. Khaled, A. Roux, and J. Lam, Machine Learning: Science and Technology **2**, 025003 (2020).
[44] A. Skinner and J. Broughton, Computational Materials Science **4**, 1 (1995).
[45] F. Musil, A. Grisafi, A. P. Bartók, C. Ortner, G. Csányi, and M. Ceriotti, Chemical Reviews **121**, 9759 (2021).
[46] S. Wu, M. Ji, C.-Z. Wang, M. C. Nguyen, X. Zhao, K. Umemoto, R. Wentzcovitch, and K.-M. Ho, Journal of Physics: Condensed Matter **26**, 035402 (2013).
[47] R. Ouyang, Y. Xie, and D.-e. Jiang, Nanoscale **7**, 14817 (2015).
[48] T. K. Patra, V. Meenakshisundaram, J.-H. Hung, and D. S. Simmons, ACS Combinatorial Science **19**, 96 (2017).
[49] V. L. Deringer, C. J. Pickard, and G. Csányi, Physical Review Letters **120**, 156001 (2018).
[50] Q. Tong, L. Xue, J. Lv, Y. Wang, and Y. Ma, Faraday Discussions **211**, 31 (2018).
[51] E. L. Kolsbjerg, A. A. Peterson, and B. Hammer, Physical Review B **97**, 195424 (2018).
[52] A. Thorn, J. Rojas-Nunez, S. Hajinazar, S. E. Baltazar, and A. N. Kolmogorov, The Journal of Physical Chemistry C **123**, 30088 (2019).
[53] E. V. Podryabinkin, E. V. Tikhonov, A. V. Shapeev, and A. R. Oganov, Physical Review B **99**, 064114 (2019).
[54] S. Hajinazar, A. Thorn, E. D. Sandoval, S. Kharabadze, and A. N. Kolmogorov, Computer Physics Communications **259**, 107679 (2021).
[55] M. Born and R. D. Misra, in *Mathematical Proceedings of the Cambridge Philosophical Society*, Vol. 36 (Cambridge University Press, 1940) pp. 466–478.
[56] L. B. Pártay, C. Ortner, A. P. Bartók, C. J. Pickard, and G. Csányi, Physical Chemistry Chemical Physics **19**, 19369 (2017).
[57] X. Wang, S. Ramírez-Hinestrosa, J. Dobnikar, and D. Frenkel, Physical Chemistry Chemical Physics **22**, 10624 (2020).
[58] É. Marcotte, F. H. Stillinger, and S. Torquato, The Journal of Chemical Physics **138**, 061101 (2013).
[59] C. van der Oord, G. Dusson, G. Csányi, and C. Ortner, Machine Learning: Science and Technology **1**, 015004 (2020).
[60] C. M. Bishop, *Neural networks for pattern recognition* (Oxford University Press, 1995).
[61] Y. Bengio, *Learning deep architectures for AI* (Now Publishers Inc, 2009).
[62] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, in *Advances in Neural Information Processing Systems 32* (Curran Associates, Inc., 2019) pp. 8024–8035.
[63] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, "TensorFlow: Large-scale machine learning on heterogeneous systems," (2015), software available from tensorflow.org.
[64] S. Ruder, arXiv preprint arXiv:1609.04747 (2016).
[65] K. Levenberg, Quarterly of applied mathematics **2**, 164 (1944).
[66] J. J. Moré, in *Numerical Analysis* (Springer, 1978) pp. 105–116.
[67] M. K. Transtrum, B. B. Machta, and J. P. Sethna, Physical Review Letters **104**, 060201 (2010).
[68] R. Chartrand and W. Yin, in *2008 IEEE international conference on acoustics, speech and signal processing* (IEEE, 2008) pp. 3869–3872.
[69] L. Prechelt, in *Neural Networks: Tricks of the trade* (Springer, 1998) pp. 55–69.
[70] Y. Dauphin, R. Pascanu, C. Gulcehre, K. Cho, S. Ganguli, and Y. Bengio, arXiv preprint arXiv:1406.2572 (2014).
[71] L. K. Hansen and P. Salamon, IEEE Transactions on Pattern Analysis and Machine Intelligence **12**, 993 (1990).
[72] C. Schran, K. Brezina, and O. Marsalek, The Journal of Chemical Physics **153**, 104105 (2020).
[73] D. Chen and R. J. Plemmons, in *The birth of numerical analysis* (World Scientific, 2010) pp. 109–139.
[74] N. Bernstein, G. Csányi, and V. L. Deringer, npj Computational Materials **5**, 1 (2019).
[75] A. M. Cooper, J. Kästner, A. Urban, and N. Artrith, npj Computational Materials **6**, 1 (2020).
[76] D. Packwood, J. Kermode, L. Mones, N. Bernstein, J. Woolley, N. Gould, C. Ortner, and G. Csányi, The Journal of Chemical Physics **144**, 164109 (2016).
[77] J. Barzilai and J. M. Borwein, IMA Journal of Numerical Analysis **8**, 141 (1988).
[78] O. Tange, "Gnu parallel 20200922 ('ginsburg')," (2020), GNU Parallel is a general parallelizer to run multiple serial command line programs in parallel without changing them.
[79] https://www.mtg.msm.cam.ac.uk/Codes/AIRSS.
[80] https://www.mtg.msm.cam.ac.uk/Codes/EDDP.
[81] S. N. Pozdnyakov, M. J. Willatt, A. P. Bartók, C. Ortner, G. Csányi, and M. Ceriotti, Physical Review Letters **125**, 166001 (2020).
[82] J. P. Perdew, K. Burke, and M. Ernzerhof, Physical Review Letters **77**, 3865 (1996).
[83] A. R. Oganov, J. Chen, C. Gatti, Y. Ma, Y. Ma, C. W. Glass, Z. Liu, T. Yu, O. O. Kurakevych, and V. L. Solozhenko, Nature **457**, 863 (2009).
[84] E. Y. Zarechnaya, L. Dubrovinsky, N. Dubrovinskaia, Y. Filinchuk, D. Chernyshov, V. Dmitriev, N. Miyajima, A. El Goresy, H. F. Braun, S. Van Smaalen, *et al.*, Physical Review Letters **102**, 185501 (2009).
[85] C. P. Talley, S. La Placa, and B. Post, Acta Crystallographica **13**, 271 (1960).
[86] J. Hoard, D. Sullenger, C. Kennard, and R. Hughes, Journal of Solid State Chemistry **1**, 268 (1970).
[87] M. Widom and M. Mihalkovič, Physical Review B **77**, 064113 (2008).
[88] S. E. Ahnert, W. P. Grant, and C. J. Pickard, NPJ Computational Materials **3**, 1 (2017).

[89] K. Dziubek, M. Citroni, S. Fanetti, A. B. Cairns, and R. Bini, The Journal of Physical Chemistry C **121**, 2380 (2017).
[90] F. Giberti, M. Salvalaglio, M. Mazzotti, and M. Parrinello, Chemical Engineering Science **121**, 51 (2015).
[91] P. M. Piaggi and M. Parrinello, Proceedings of the National Academy of Sciences **115**, 10251 (2018).
[92] C. Shang, X.-J. Zhang, and Z.-P. Liu, Physical Chemistry Chemical Physics **19**, 32125 (2017).
[93] A. Tkatchenko and M. Scheffler, Physical Review Letters **102**, 073005 (2009).
[94] A. Ambrosetti, A. M. Reilly, R. A. DiStasio Jr, and A. Tkatchenko, The Journal of Chemical Physics **140**, 18A508 (2014).
[95] J. Feng, W. Grochala, T. Jaroń, R. Hoffmann, A. Bergara, and N. Ashcroft, Physical Review Letters **96**, 017006 (2006).
[96] M. Eremets, I. Trojan, S. Medvedev, J. Tse, and Y. Yao, Science **319**, 1506 (2008).
[97] M. Martinez-Canales, A. R. Oganov, Y. Ma, Y. Yan, A. O. Lyakhov, and A. Bergara, Physical Review Letters **102**, 087005 (2009).
[98] H. Zhang, X. Jin, Y. Lv, Q. Zhuang, Y. Liu, Q. Lv, K. Bao, D. Li, B. Liu, and T. Cui, Scientific Reports **5**, 1 (2015).
[99] K. Momma and F. Izumi, Journal of Applied Crystallography **44**, 1272 (2011).
[100] A. P. Bartók and J. R. Yates, The Journal of Chemical Physics **150**, 161101 (2019).
[101] A. J. Morris, R. J. Nicholls, C. J. Pickard, and J. R. Yates, Computer Physics Communications **185**, 1477 (2014).