Faraday Discussions

Cite this: *Faraday Discuss.*, 2025, **256**, 61

PAPER

Beyond theory-driven discovery: introducing hot random search and datum-derived structures

Chris J. Pickard [ab]

Received 17th June 2024, Accepted 31st July 2024
DOI: 10.1039/d4fd00134f

Data-driven methods have transformed the prospects of the computational chemical sciences, with machine-learned interatomic potentials (MLIPs) speeding up calculations by several orders of magnitude. I reflect on theory-driven, as opposed to data-driven, discovery based on *ab initio* random structure searching (AIRSS), and then introduce two new methods that exploit machine-learning acceleration. I show how long high-throughput anneals, between direct structural relaxation, enabled by ephemeral data-derived potentials (EDDPs), can be incorporated into AIRSS to bias the sampling of challenging systems towards low-energy configurations. Hot AIRSS (hot-AIRSS) preserves the parallel advantage of random search, while allowing much more complex systems to be tackled. This is demonstrated through searches for complex boron structures in large unit cells. I then show how low-energy carbon structures can be directly generated from a single, experimentally determined, diamond structure. An extension to the generation of random sensible structures, candidates are stochastically generated and then optimised to minimise the difference between the EDDP environment vector and that of the reference diamond structure. The distance-based cost function is captured in an actively learned EDDP. Graphite, small nanotubes and caged, fullerene-like, structures emerge from searches using this potential, along with a rich variety of tetrahedral framework structures. Using the same approach, the pyrope, $Mg_3Al_2(SiO_4)_3$, garnet structure is recovered from a low-energy AIRSS structure generated in a smaller unit cell with a different chemical composition. The relationship of this approach to modern diffusion-model-based generative methods is discussed.

I. Introduction

The introduction of unbiased, first principles, structure prediction in the mid-2000s revolutionised materials discovery.[1] It was no longer necessary to trawl through databases of the "usual suspects", or to concoct novel structures by hand.

[a]*Department of Materials Science & Metallurgy, University of Cambridge, 27 Charles Babbage Road, Cambridge CB3 0FS, UK. E-mail: cjp20@cam.ac.uk*
[b]*Advanced Institute for Materials Research, Tohoku University, 2-1-1 Katahira, Aoba, Sendai, 980-8577, Japan*

Unknown structure types, and surprising phenomena, emerged from explorations of the density functional theory (DFT) energy landscape, where previous approaches to structure prediction depended on the fast evaluation of empirical forcefields.[2–5] DFT provides an approximation to the underlying quantum mechanical interactions governing the stability of different phases, balancing computational efficiency with a robustness[6] that permits genuine predictions. In Section II, I review several examples of theory-driven discovery enabled by my approach to structure prediction, *ab initio* random structure search (AIRSS),[7,8] while in Section III, the generation of the random sensible structures on which AIRSS depends are discussed.

There is a further revolution underway, sparked by the discovery that machine-learning techniques can routinely be exploited to accelerate the exploration of energy landscapes, either through molecular dynamics (MD) or structure prediction. From early attempts in the 1990s,[9] the groundbreaking contributions of Behler[10] and Csányi[11] have stimulated the development of a wide array of machine-learned interatomic potentials (MLIPs).[12] Among these are the ephemeral data-derived potentials (EDDPs)[13,14] – briefly reviewed in Section IV – which were introduced with the explicit aim of accelerating AIRSS.

In Section V, I will show how the multiple orders of magnitude acceleration offered by EDDPs over DFT allow for a style of calculation that would have simply been too computationally expensive previously – a novel extension to AIRSS, hot-AIRSS. hot-AIRSS exploits the integration of long MD-driven anneals as part of the high-throughput optimisation of stochastically generated structures.

In Section VII, I introduce a new approach to the generation of random sensible structures, building on the concept of constructing structures that respect measured inter-species distances, and are likely low in energy even before structural optimisation – see Section VI. The new method is based on the optimisation of a cost based on the distance to a reference structure (or potentially multiple structures) evaluated in the space of EDDP environment/feature vectors, and requires few modifications to the existing AIRSS/EDDP workflow. In Section VIII, this new approach is applied to two challenging systems – carbon, and $Mg_3Al_2(SiO_4)_3$.

Finally, in Section IX, it is recognised that the method introduced in Section VII is very closely related to modern diffusion-model-based generative approaches, providing a point of connection with traditional structure prediction methods, and AIRSS in particular.

II. Theory-driven discovery

AIRSS, introduced in ref. 7 and described in depth in ref. 8, is built on the high-throughput first-principles relaxation of diverse stochastically generated structures (from crystals, to clusters, molecules, surfaces, interfaces, and grain boundaries). The emphasis is on exploration, and the hunting for outliers, or surprises, through an attempt to uniformly sample configuration space, within a defined distribution of candidate structures.

Throughout my work, there is a focus on the discovery of unexpected phenomena, as opposed to the detail of a particular crystal structure – not forgetting that it is essential that the structural details are correctly identified in order to meaningfully predict the discovered material's properties. When

a surprising result is encountered, considerable effort is expended in attempting to identify the competing phases that might render the prediction unsound. In many cases, this is indeed the outcome. Persisting in this approach leads to a high success rate, with few false positives, and high-quality predictions.

The first applications of AIRSS were to the high-pressure sciences, beginning with an exploration of superconductivity and metallicity in the dense hydrides.[7,15] This has grown to be a very active area with many well-known successes[16] – see Section II.D. With other first-principles structure-prediction techniques,[1] USPEX,[17] CALYPSO,[18] and XtalOpt,[19] AIRSS is now a key tool for materials discovery with applications ranging from battery materials[20,21] to molecular polymorphism,[22] and nanoconfined water.[23]

The emphasis of first-principles random structure search on highly parallelisable and broad sampling ensures it is particularly well-adapted to modern computational trends, statistical physics and machine learning in particular, where it has become an indispensable source of training data.[24–26]

A. Mixed phases in hydrogen

An early application of AIRSS was an attempt to understand phase III of dense hydrogen, and in particular to identify model structures that exhibited the key vibrational spectroscopic signatures measured in diamond anvil cell experiments.[27] Our prediction of the *C*2/*c*-24 structure as the best model for phase III is standing the test of time.[28,29]

Analysing the large number of AIRSS-generated structures, I was confronted by a striking family of metastable structures, of a type that had not previously been suggested for an element. They consisted of layers, alternating between graphene-like and molecular; see Fig. 1(a). I felt these structures must be important and potentially dynamically stabilised phases (either through zero-point motion, or temperature), but the techniques were not then ready to allow a full phase diagram to be computed. Nevertheless, we published the mixed phase structures in ref. 27 and emphasised them in presentations to experimentalists.

Initially, the mixed phases did not address any open experimental questions and were largely ignored. This changed when Goncharov and Gregoryanz approached me with a puzzle – they were seeing a surprising softening in a high-frequency Raman peak in warm (room temperature) hydrogen at megabar pressures. I suggested that they were observing a mixed phase, and on investigation this proved to be the case.[30] The mixed phases are now an established feature of the hydrogen phase diagram. It is fair to say that, given the experimental challenges in determining the positions of protons, our current understanding of dense hydrogen is largely due to first-principles structure searches, with much having been mapped out in ref. 27.

Why was first-principles structure search so successful in tackling this well-explored problem? Of course, the high-throughput nature of the searches made a big difference, increasing the sheer number of structures considered. But the most important structures could probably have been found using contemporary MD methods. The fact that they were not is likely because MD was frequently conducted in cubic, or orthorhombic, unit cells, and with fixed numbers of atoms, typically multiples of 8. But my candidates for dense hydrogen, *C*2/*c*-24, *Cmca*-12 and the mixed phases, all contained multiples of 12 atoms. I had been in the habit

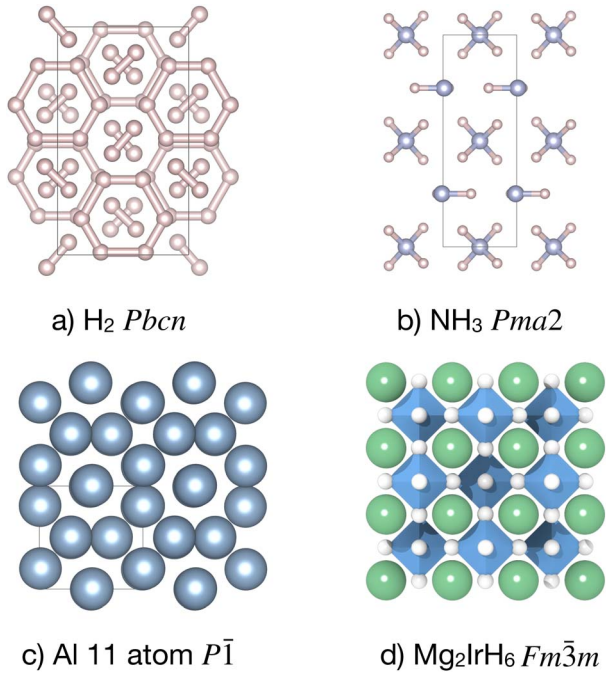

Fig. 1 (a) *Pbcn* mixed phase of hydrogen at 300 GPa, (b) *Pma*2 NH_2–NH_4 phase of ammonia at 100 GPa, (c) 11-atom host–guest phase of aluminium at 5 TPa, and (d) dynamically stable 0 GPa cubic phase of Mg_2IrH_6 with a predicted superconducting T_c of 160 K.

of not assuming the number of atoms in the unit cell and choosing them randomly as part of the structure generation. This was also to be very important for aluminium, described below in Section II.C, and highlights the importance of minimally biased stochastic searches.

B. Ionic ammonia

When searching for molecular crystal structures, a well-established protocol is to stochastically pack connected molecular units.[8,31] This shrinks the search space, as compared to a less restricted search starting from unconnected atoms, and dramatically increases the odds of finding low-energy configurations. But it is at the cost of potentially missing the most stable one, if it does not adhere to the chosen molecular unit. In the spirit of assuming as little as it is computationally feasible to, I had been searching for dense phases of NH_3 by randomly placing the N and H atoms into randomly shaped unit cells individually. It was a routine project, but I was jolted awake one early morning while checking the results of the overnight runs. The most stable units under pressure were, by some margin, NH_2^- and NH_4^+ – see Fig. 1(b) – not the expected NH_3.[32] I assumed that something was wrong with the calculations. This possibility had not been discussed for pure ammonia previously, and it was not something we were looking for. After careful testing, the result held, and the spontaneous ionisation of NH_3 has been experimentally established.[33] Spontaneous self-ionisation more generally is now considered as a possibility where it might not have been previously.

C. Complex phases of aluminium at terapascal pressures

We (and others, particularly Yanming Ma and co-workers) had been starting to find a great number of electride-type structures in the dense elements.[34,35] One striking feature of these was the localisation of states under increasing pressure, and band narrowing. I wondered whether I could find a non-magnetic element that, under the right conditions, would exhibit magnetism. I began the hunt, systematically working my way through the periodic table. Importantly, it turned out, I was randomly choosing the number of atoms in the unit cell. When it came to aluminium, I was surprised to find that the most stable structure at 3 TPa contained 11 atoms in the unit cell. At that time, few groups would even consider odd numbers of atoms as a possibility, based on the heuristic that they would unlikely be the most stable. The 11-atom cell was, however, significantly more stable than the other candidates, and initially, when I visualised it, it made no sense. It appeared to be amorphous, or still random somehow. This was unusual, as the most stable structures usually exhibit some symmetry. But I continued building supercells and spinning the structure around in the visualiser, and eventually all became clear.

The structure consisted of tubes and chains of atoms – see Fig. 1(c). I was aware of the work of Nelmes and McMahon[36] on incommensurate host–guest phases in the alkali metals,[37] as Volker Heine had publicised it in the Theory of Condensed Matter Group, Cambridge. This turned out to be exactly what I was seeing in the 11-atom structure – an approximant of a kind of 1D quasicrystal. Once I had recognised that, it was straightforward to manually construct other, larger, approximants, and estimate the ideal lattice parameters for the host and guest phases. I was also able to determine that the structure was of the electride type and construct a simple model for it,[38] based on a generalised Lennard-Jones model, which later became the basis of the EDDPs – see Section IV.

This result has not been confirmed experimentally – yet. But it has had an impact on the field – it showed that materials under extreme compression might be complex, and not just simply close packed. This has inspired the high-pressure community, particularly the shock physicists, for example being used as part of the justification for using the National Ignition Facility (NIF) to perform exploratory science.[39] Continuing my sweep through the periodic table, I did eventually manage to find magnetism in an electride phase, in potassium.[40]

D. High-throughput hunt for conventional superconductivity

Bringing the applications of AIRSS up-to-date, recent work has refocussed on the search for high-temperature superconductors, specifically the hydrides, which may be (meta-)stable at ambient pressures, and superconduct at temperatures exceeding the critical temperature (T_c) of magnesium diboride. The field of hydride superconductivity has not been without controversy,[41] and it is essential to be able to identify candidate superconductors that might maybe be synthesised at low pressures, opening the field to broad and intense experimental scrutiny.

With the growth of computational resources since the debut of AIRSS, as well as refinements in the methods and optimisations of the key DFT code used for structural optimisation (CASTEP[42]), it is now possible to add an additional layer of sampling to the searches. While early studies would concentrate on elements or compounds with a fixed composition, it later became possible to study the

composition space of a given binary, or ternary, system.[43,44] The next step has been to search over a wide range of composition spaces simultaneously, in a high-throughput manner.

In an initial study, we explored the binary hydrides over a range of pressures from 100 GPa to 500 GPa.[45] Several novel superconducting hydrides were discovered, and known ones rediscovered. The maximum superconducting transition temperatures, T_c, varied from 380 K at 500 GPa, to above 250 K at 100 GPa. A striking feature of our result was that the T_c did not drop precipitously as the pressure was reduced, and through extrapolation one might expect hydride T_cs to be as high as 200 K at ambient pressures. This stimulated an extension of this approach to the ternary hydrides at low, and ambient, pressure.[46]

The searches across composition space were performed entirely using first-principles methods – and so theory-driven at this stage – and resulted in the discovery of Mg_2IrH_6 (see Fig. 1(d)) as a dynamically stable, moderately metastable, candidate conventional superconductor with a predicted T_c of 160 K. Once Mg_2IrH_6 had been identified, detailed structure searches over the Mg–Ir–H composition space, accelerated with the EDDP machine-learned interatomic potentials (see Section IV), provided a thorough picture of the competing phases, as well as a feasible synthesis route. Having highlighted the power of theory-driven search for discovery, this most recent work touches its limits, and demonstrates the power of data-driven approaches, which will be the focus of the rest of this contribution.

III. Generating random sensible structures

Key to the success of AIRSS is the initial step of generating an ensemble of chemically sensible random structures for subsequent high-throughput structure relaxation. This step is performed by the buildcell code of the GPL2 open-source AIRSS package.[47] The random structures are constructed once an appropriate distribution of parameters has been selected – based on either chemical insights or previous calculations (see Section VI). When building a random unit cell, its volume and shape should be chosen. These must be selected from a range, and it makes sense to choose this range to adhere to experimentally reasonable values – even if only very approximately so. There is little point in searching in excessively small, or large, unit cells. Similar choices must be made for other parameters – how closely should atoms be permitted to approach each other in the initial structures? Structures might be generated to have randomly generated space (or point) group symmetries. The structural units might be molecules or fragments, rather than individual atoms. Composition can be stochastically chosen, but the ranges of compositions to be considered must be specified. Some thought should be given to load balancing the searches – each of the stochastically generated structures should have roughly the same computational cost.

The initial random structures look sensible and certainly some of them might be expected to have reasonably low energies, even before structural optimisation. Put together, these choices define a generative model, in machine-learning terminology. This will be explored further in Sections VI and VII, and the relation to modern generative approaches to structure prediction will be discussed in Section IX.

IV. Ephemeral data-derived potentials

The prospect for data-derived potentials to accelerate structure search had long been apparent, and in ref. 24 it was shown that the random structure search and Gaussian approximation potentials (GAPs)[11] could be combined to iteratively generate a robust boron potential. At that time, the development of GAPs was relatively intricate and time consuming, and the resulting potentials slow. To ensure the AIRSS could routinely benefit from the promised acceleration, with minimal interruption to the successful high-throughput workflow, ephemeral data-derived potentials (EDDPs) were introduced.[13] The emphasis on their ephemeral nature was intended to draw the attention away from the difficult task of developing high-quality benchmarked potentials, towards the generation of disposable potentials that could be trained and used rapidly.

EDDPs are based on a simple model for the interatomic interaction, inspired by Lennard-Jones style potentials, with a minimal extension to handle many-body interactions.[13,14] The resulting feature, or environment, vectors are the input for small neural networks (in many cases, a single hidden layer with just five nodes). Multiple neural networks are fitted, in parallel with random initialisations, just as in AIRSS. Early stopping, based on a validation portion of the 80 : 10 : 10 training : validation : testing data split,[48] is used to discourage overfitting. The Levenberg–Marquadt (LM) optimiser is found to be fast and produce excellent training and testing losses. Combining the many neural networks together, minimising the non-negative least squares (NNLS) error, again to the validation split, results in a sparse ensemble, with only a fraction of the neural networks being selected for the final model. The ensemble enables the variance of the predicted energies among the many fits to be evaluated, and this can be used to detect pathological structures, as well as to drive an active learning to less certain configurations.[49,50]

A key feature of EDDPs is that they are trained on the DFT energies of large numbers of small, and so rapid to compute, structures. To date, forces are not used in the training, which might be a limitation compared to other methods. However, there are advantages to this approach, and using AIRSS to generate many highly diverse structures, the resulting potentials have proven to be more than adequate for the purposes of accelerating structure prediction. In ref. 14, it is shown that EDDPs can also be used as the basis for reliable and quantitative molecular and lattice dynamics simulations. The structures encountered in a random search are extremely varied as compared to those sampled by molecular dynamics, and this diversity of the structures on which the EDDPs are trained appears to largely eliminate the problems of stability of molecular dynamics simulations.

EDDPs have been extended to be able to handle large numbers of chemical species using the alchemical ideas of Ceriotti.[51] The GPL2 open-source EDDP package is available.[52]

V. Introducing hot random structure search

For many problems, AIRSS is an extremely effective approach to discovering low-energy structures. The first-principles potential-energy surface is relatively smooth, and for moderate system sizes, the probability of encountering low-energy configurations is sufficiently high that when coupled with high-

throughput computation, AIRSS is a competitive structure-prediction technique.[8] However, as more complex problems are attempted, the exponential growth in local minima begins to dominate, and without extensive use of constraints to prepare sensible initial starting points, the likelihood of generating low-energy configurations becomes too low to justify the computational effort in searching for them. For example, in ref. 13, an EDDP was generated for boron, and a free search for γ-boron[53] was attempted. No symmetry was exploited, nor was the knowledge that boron tends to favour icosahedra, and unit cells containing 28 boron atoms at approximately the correct density were generated. A slightly distorted version of the orthorhombic *Pnnm* γ-boron structure was successfully located, but only twice out of 362 754 putative structures. In tests, the 12-atom α-boron structure can typically be found in free AIRSS searches once every 3000 attempts. Making an assumption of an exponential increase in difficulty, we might estimate that identifying the γ-boron structure in a doubled cell of 56 atoms would take something like 3×10^6 structure optimisations, unfeasible from first-principles and challenging even using EDDPs.

A difficulty that the use of fast potentials for structure search has created is the management and storage of the vast number of structures that can be generated on even modest computer hardware. The writing of the data to disk can become a bottleneck on some high-performance computing (HPC) systems. One option is to only store the most stable structures encountered, for example by rejecting any new structures that are outside a given energy threshold of any previously encountered for that composition. An alternative is to embrace the acceleration and perform more intense computation for each generated and stored structure.

Probably the greatest impact of the MLIP revolution has been the opening up of the possibility of performing long time-scale, large length-scale, MD simulations at approaching first-principles quality.[23,54–56] We exploit this here to perform random structure search integrating an extended annealing period, between local optimisations. AIRSS, and what we introduce here as hot-AIRSS, are contrasted in Fig. 2. An initial random structure is generated, just as in traditional AIRSS, potentially using the several strategies to prepare the structures described in Section III, and relaxed to its nearest local minimum using the repose code. Rather than stopping there, the ramble molecular dynamics code supplied in the EDDP package is used to perform an anneal at a fixed temperature for a given time. The resulting structure is then again relaxed to the now nearest local minimum, which, if the temperature chosen is sufficiently high, is not likely to be the same as the initial one.

The two parameters introduced are the temperature for the anneal (typically chosen to be approaching but below the melting temperature of the system), and the time for the anneal. The time is typically selected to exceed 10 picoseconds, and potentially as long as nanoseconds. There is no quenching of the system during the molecular dynamics run, and the overall process, given the final local optimisation, can be thought of as an elaborate optimisation scheme, and from the point of view of AIRSS is a direct replacement of the usual local optimiser. From this perspective, it is reasonable to permit the exploitation of symmetry during the anneal. The ramble code implements symmetrised MD, a functionality that is not generally available in more widely used codes. While not currently implemented, the ability to optimise and run dynamics on defined structural units is likely to prove useful.

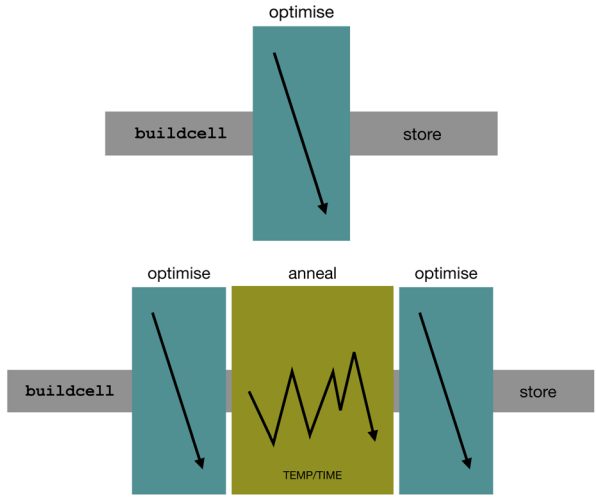

Fig. 2 Top: A representation of AIRSS: a random sensible structure is generated using the buildcell code, and is then structurally optimised to the nearest local minimum of the energy landscape, which is either described by DFT, or a fast equivalent, such as an EDDP. The resulting structure is stored. This is repeated, in parallel, a large number of times. Bottom: hot-AIRSS proceeds in a similar manner, but after the first optimisation with an EDDP, a long anneal is performed at a chosen temperature, close to but below the melting temperature, for a given time. The resulting structure is finally structurally optimised and stored.

To explore the capability of hot-AIRSS, we revisit the high-pressure phases of boron and attempt to locate the *Pnnm* γ-boron phase at 10 GPa. An EDDP is prepared so that the required high-throughput MD-driven anneals are feasible. It is generated using the chain script, with seven iterations of active learning. In the first step, 10 000 random structures containing 12, 24 and 28 boron atoms are constructed, and their PBE GGA[57] single-point energies are computed using CASTEP[42] with the default QC5 OTFG pseudopotential for boron, a *k*-point spacing of 0.07 2π Å^{-1}, a plane-wave cutoff of 340 eV, and default grid scales. Marker structures consisting of 11 known and putative phases of boron are added to the dataset, each one shaken 1000 times with an amplitude of 0.1. For each iteration of active learning, AIRSS is used to generate 10 000 structures at a randomly chosen pressure between 5 GPa and 15 GPa, which are each shaken once with an amplitude of 0.1. 30 individual potentials are trained, with NNLS selecting 12. The resulting training and testing MAEs are 13.33 and 13.67 meV per atom, respectively.

The results of three searches for 56 atoms of boron at 10 GPa are presented in Fig. 3. The structures generated using traditional AIRSS are highly disordered. The most stable are around 0.3 eV per atom less stable than the known ground-state γ-boron structure. The probability of generating low-energy structures is low, and consistent with the above estimate of the difficulty of this task. Even given the very rapid structural optimisation, this is not a viable approach to finding the ground-state structure in such a large unit cell.

In the second search, hot-AIRSS is performed. After an initial relaxation, a 10 ps anneal at 1800 K is performed. This temperature is selected after conducting

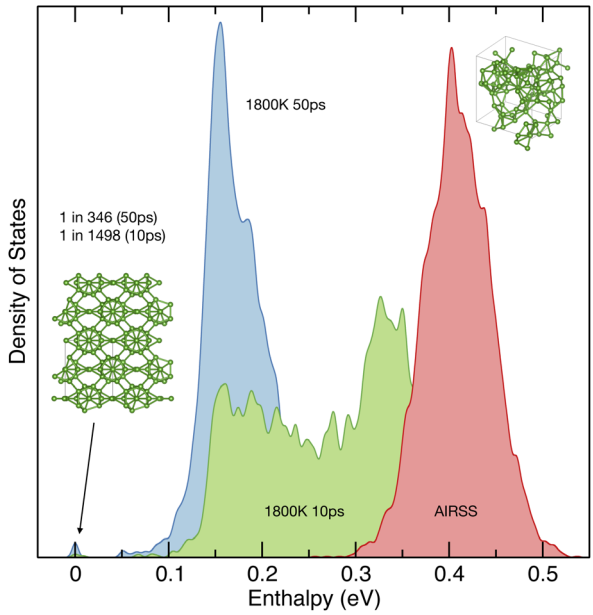

Fig. 3 Unconstrained search for 56 boron atoms at 10 GPa. Structural densities of states for (red) an AIRSS search, (green) a hot-AIRSS search at 1800 K for 10 ps, and (blue) a hot-AIRSS search at 1800 K for 50 ps. The enthalpy per boron atom relative to the ground-state *Pnnm* γ-boron phase (shown) is plotted.

a few short runs and assessing the average mobility of atoms in the unit cell. The temperature should be below the melting temperature, as fully molten configurations relax to approximately the same distribution as AIRSS. However, the atoms should be sufficiently energetic so as to be mobile enough to explore a wide range of configurations. Should a low-energy configuration be encountered, since the system is at below the melting temperature, it is liable to freezing. This is acceptable, since on further relaxation the low-energy configuration will be maintained. In principle, it should be possible to set the anneal temperature automatically, and on a per-sample basis, but this is not explored further here.

The resulting structural density of states exhibits a much broader distribution, with an increased diversity of structures. Out of 2996 samples, two of the structures located are found to be identical to the known γ-boron structure. One of them was the 56-atom *Pbcn* modification of γ-boron discussed in ref. 58. On increasing the time of the anneal to 50 ps, the distribution shifts to lower energies still, and the γ-boron phase is found 11 times out of 3806 samples. It should be noted that while the probability of encounter has increased by 4.3, each anneal was five times longer – so the length of anneal is a parameter that should be adjusted to maximise computational efficiency.

It is currently thought that rhombohedral β-boron is the most stable phase at low temperatures and pressures. The structure is complex, and likely high in defects, leading to entropic stabilisation.[59] In ref. 24, we used an actively learned GAP potential to explore the relative energy of the defects and interstitials. In ref. 60, it was shown that moment tensor potential[61] accelerated evolutionary algorithms could generate low-energy approximants of rhombohedral β-boron

without recourse to experimental information. Tetrahedral β-boron is thought to have a region of stability at elevated temperatures and pressures. Similarly to the rhombohedral phase, the tetrahedral phase is complex, with the best models containing 192 atoms in the primitive unit cell, and is also stabilised by a propensity to defect and interstitial formation. The stabilisation of these, and other, phases of boron has recently been studied in detail by Hayami *et al.*[62]

In Fig. 4, the results of AIRSS and hot-AIRSS searches for 105 to 111 boron atoms in a single rhombohedral unit cell, fixed to experimental lattice parameters, are shown.[63] The density of structural states for the AIRSS search is narrowly peaked around 0.4 eV above the most stable structure found. The distribution of states from hot-AIRSS calculations at 1800 K for 25 ps is much broader, extending to lower energy. There is a peak at low energy, consisting of many structures visually similar to known β-boron models, but exhibiting a wide range of defects and interstitials, which can be expected to contribute to entropic stabilisation. The situation for tetragonal β-boron is very similar – see Fig. 5 – although the low-energy peak of defective structures is significantly narrower in energy. Apart from the work of Podryabinkin *et al.*,[60] theoretical studies of the β-borons have proceeded by analysing defect and interstitial populations of the experimental structures. Here we see that hot-AIRSS can discover the underlying structural motifs of these complex phases.

hot-AIRSS is an elegant modification to AIRSS that maintains the trivial parallelisability of random structure search, and requires minimal changes to the computational workflow, or the provided airss.pl script in which the workflow is embodied. Temperature has been long recognised as a key parameter in structure

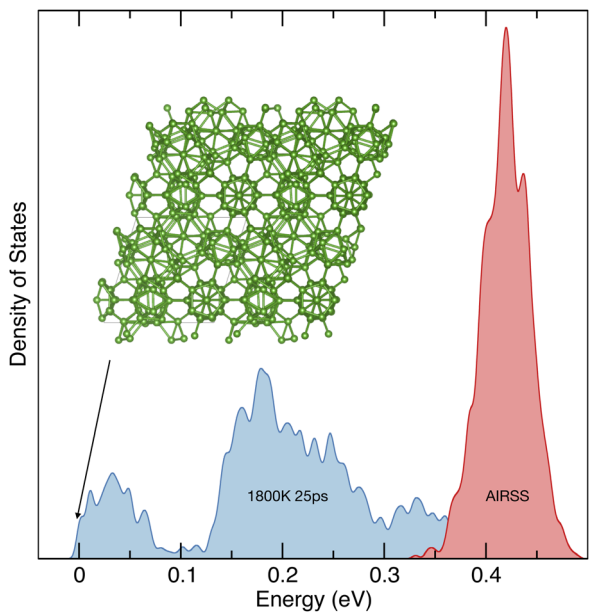

Fig. 4 Fixed cell search for 105 to 111 boron atoms. Structural densities of states for (red) an AIRSS search, and (blue) a hot-AIRSS search at 1800 K for 25 ps. The energy per boron atom relative to the most stable structure (shown) is plotted. The lattice parameters for rhombohedral β-boron were fixed and taken from ref. 63.

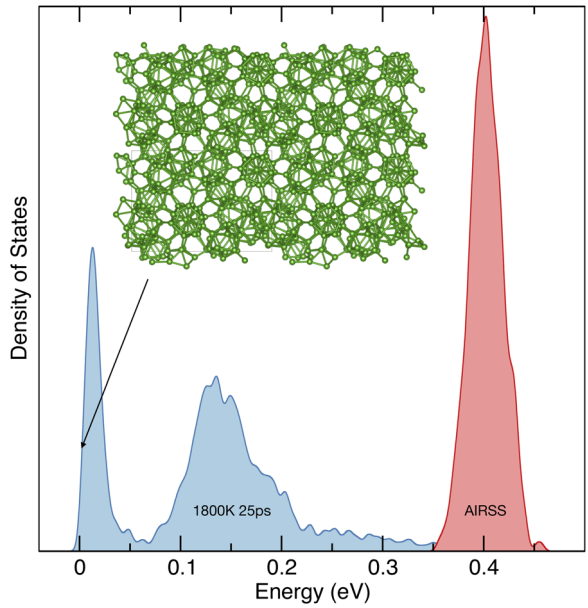

Fig. 5 Fixed cell search for 192 boron atoms. Structural densities of states for (red) an AIRSS search, and (blue) a hot-AIRSS search at 1800 K for 25 ps. The energy per boron atom relative to the most stable structure (shown) is plotted. The lattice parameters for tetrahedral β-boron were fixed and taken from ref. 64.

search, most notably in simulated annealing,[65] basin hopping,[66] and more explicitly through short molecular dynamics explorations in minima hopping.[67] The computationally efficient EDDPs now allow temperature to play a role in random structure search, and it is shown to be a powerful approach to tame complex and challenging systems.

VI. Generating structures from measured minimum separations

The computational creation of random, yet chemically sensible, structures is central to the success of AIRSS; see Section III. One of the most powerful approaches is the building of structures satisfying a defined (but potentially stochastically generated) species-wise matrix of minimum separations – the MINSEP method of the AIRSS buildcell code. With the method additionally tagged with AUTO, the minimum separations are measured from the most stable structure with the desired composition, if available, along with a target density. If there are no structures available, the specified minimum separation parameters are used.

For well-packed inorganic materials, the random structures generated in this way are likely to be chemically sensible and hence of relatively low energy when computed using DFT. The measured structures are typically the result of earlier, less constrained, searches. However, should experimentally known crystal structures be available for a given composition, the separations and density can be measured from those.

VII. A new approach to generating structures from measured EDDP feature vectors

The development of many body descriptors or feature/environment vectors as the basis for MLIPs, such as the EDDPs described in Section IV, opens the way for much more sophisticated measurements to be made of atomistic structures. Related to the measurement of the minimum atomic separations, these descriptors provide a detailed measurement and description of the environment around a chosen atomic site. If structures can be generated that have similar environment vectors to a known, stable, structure then those structures are likely to be chemically similar to the target, and similarly low lying in the potential-energy landscape.

If the so-generated structures exhibit some diversity, and are not identical to the target, this provides an alternative approach to building structures for AIRSS, and one might expect them to be not only sensible, but close to their nearby local minimum, and hence require little or no structural optimisation using DFT. Computing the single-point total energies should be sufficient to rank the candidates.

We now present such a scheme to generate structures that are closely related to a target structure. First, the feature vectors for the atomic environments in the target structure are computed. We will use the EDDP feature vectors, and these are obtained using the frank code. One might then perform an AIRSS search where the structural optimiser (for example, CASTEP in first-principles searches and repose when EDDPs are used to accelerate the search) is replaced with a code that computes the gradient with respect to atomic displacements and changes in unit cell shape, of some cost function that monotonically depends on the distance of the new structure's feature vectors from the target vectors; see Fig. 6. Here we instead actively train an EDDP on this cost function, using a modified version of the chain script, manifest. While a less direct approach, it has advantages.

Firstly, it permits the use of the AIRSS/EDDP tools with no modification – once the cost-based EDDP has been trained, it can be used as any other EDDP, permitting structure searches using repose, molecular dynamics using ramble, and lattice dynamics through wobble. Secondly, while the cost function may (or may not) be a strictly smooth function, the learned EDDP will be, by construction.

As the manifest script progresses, structures are generated randomly, as in the first step of the iterative training of an EDDP, as shakes of the target structure (a marker structure), and from shaken AIRSS structures with intermediate generations of the cost-based potential. Instead of computing the DFT single-point total energies for these configurations, the cost for each one is computed from the sum of a function of the distances from the configuration environments to the target environments. The training of the cost-based EDDP then progresses iteratively, and rapidly, as no DFT computations are required.

The cost contribution of a single environment in a structure is defined as a function of the soft minimum Euclidean distance to the potentially many environments of the target structure. This choice avoids the need to assign and pair the environments between the structure and the target structure and means that a minimum cost can be achieved if the environments of the new structure match any combination of the environments in the target structure.

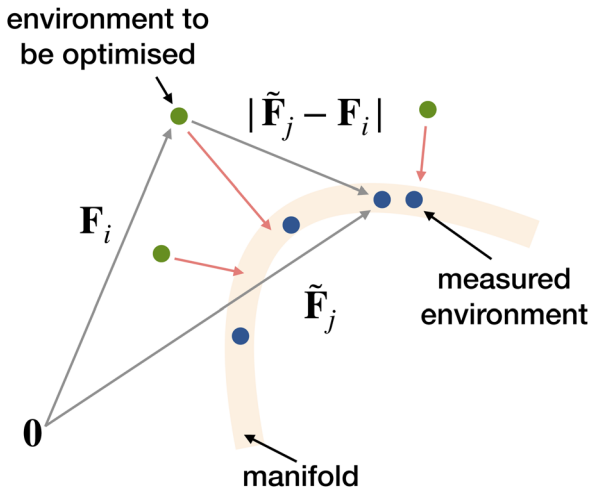

Fig. 6 Optimisation to the manifold of measured environments. The blue circles are the environments, *j*, in the chosen feature space, measured from the target structure. They are assumed to lie on a low-dimensional manifold embedded in the feature space, sketched by the light red band. The green circles represent the distinct environments, *i*, of the structure to be generated by optimisation towards the manifold, in the direction of the red arrows.

A choice of the function of the Euclidean distance might be the commonly used squared distance. However, this function becomes very large for dissimilar environments, and the optimisation scheme may lose discrimination between environments similar to the target once the EDDP has been learned from the cost data. To maintain resolution close to the target environments, the partial costs are evaluated as:

$$c_{ij} = \ln\left(\frac{\beta}{N^2}\left|\mathbf{F}_i - \tilde{\mathbf{F}}_j\right|^2 + 1\right). \tag{1}$$

For small distances between the feature vectors $\mathbf{F}_i$ and $\tilde{\mathbf{F}}_j$, of length N, the squared Euclidean distance is recovered, but for large distances, the cost is moderated, and does not grow to be too large. The parameter β controls the degree to which small distances increase the cost, and so for large β, the cost is minimised by more strictly enforcing similarity with the target environments.

To evaluate the cost for each configuration, with respect to the target environments, the most straightforward approach is to identify the minimum partial cost for each atom in the configuration:

$$E_{\text{cost}} = \sum_i \min_j \{c_{ij}\}. \tag{2}$$

This approach has the disadvantage that the resulting cost landscape is not smooth. To some extent, this could be managed through learning the EDDP representation of the cost landscape. However, it is preferable to instead construct a softened approximation to the minimum:

$$E_{\text{cost}} = -\sum_i \frac{1}{\alpha} \ln\left(\frac{1}{M}\sum_j^M e^{-\alpha c_{ij}}\right), \quad (3)$$

where M is the total number of target environments. The parameter α controls the degree of softness of the approximation. For large values of α, the strict minimum is recovered. It is worth noting that for typical values of α, the cost for the target structures computed against themselves does not evaluate to zero. However, if α is appropriately set, the cost should increase for all distortions of the target.

VIII. Datum-driven discovery

We have discussed the power of theory-driven discovery in Section II. Data-driven approaches are emerging as powerful methods to accelerate search and discovery, but it is instructive to consider what can be learned from a single data point, or datum. Using the scheme described above, we first investigate the discovery potential of a using a single, experimentally known, structure as a generative source of hypothetical structures. We then explore how the approach might be integrated within a first-principles searching strategy.

A. Carbon

Carbon is a fascinating element, with a great number of theoretically proposed allotropes,[68] and fewer iconic experimentally known structures. Graphite is the thermodynamically favoured structure under ambient conditions, with diamond becoming stable at high pressures, and being an important metastable material. At higher pressures still, several phase transitions have been predicted, from bc8,[69] to sc[70] at terapascal pressures, and sh, fcc, dhcp and bcc up to petapascal pressures.[71] Carbon structures that are metastable under all conditions include graphene, nanotubes, and fullerenes.[72]

We will now explore what can be learned about carbon from a single known carbon phase – the diamond structure. This high-symmetry $Fd\bar{3}m$ cubic structure has a single environment, so the generated structures will be optimised to have environments as close to this environment as possible.

A cost-based EDDP potential was generated using the manifest script, which performs the active learning process. A three-body neural network potential, with 16 polynomials for the two-body terms of the environment features, and 4 for the three-body, was trained, with two hidden layers of 20 nodes each. 31 individual networks were trained, with 18 selected by the NNLS ensembling procedure. 1000 structures with 1 to 12 atoms were randomly generated in the first step, along with 1000 shakes of the target diamond structure with a position and cell amplitude of 0.1. The cutoff radius was set to 3.75 Å. During the active learning phase, there were 10 cycles of adding 1000 AIRSS-generated structures, added with a 0.1 position and cell amplitude shake. The parameters for the cost function were $\alpha = 10$ and $\beta = 100$.

A search for low-energy carbon structures was performed in the following way. Using the cost-based EDDP, an AIRSS search is conducted for 8 to 48 atoms, generating initial structures with a volume per atom between 5 and 10 Å^3 and 12 to 24 randomly selected symmetry operations. The application of high symmetry ensures a diversity of generated structures, and at the same time reduces the

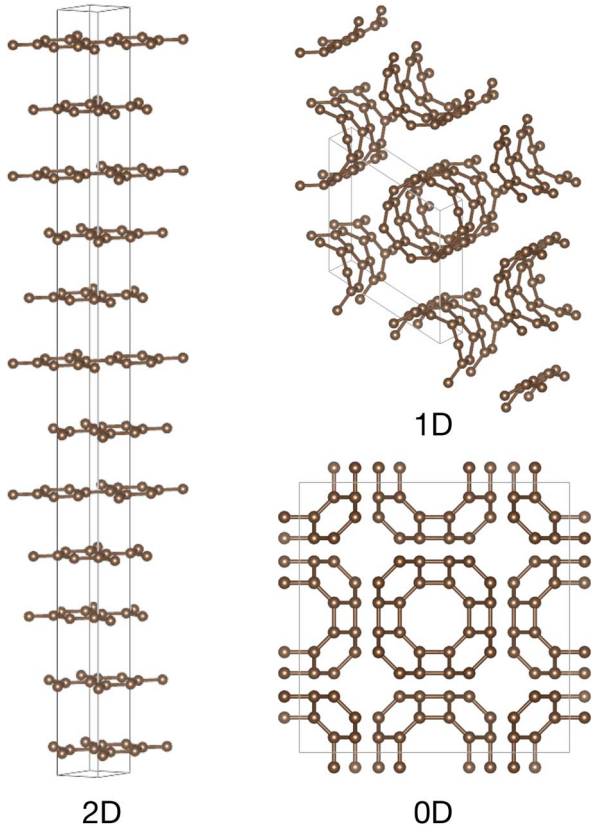

Fig. 7 Selected low-dimensional carbon structures. The zero-dimensional structure consists of a face-centered lattice of C_{48} clusters, but is relatively unstable compared to the fullerenes due to the presence of four-membered rings. The one-dimensional structure is an array of small nanotubes, and the two-dimensional structure is a complex stacking of graphite.

number of low-energy structures that are simply defect-containing versions of diamond or graphite. The ranking of the structures is performed in three stages, using PBE-DFT,[57] computed by CASTEP.[42] First, single-point DFT energies are computed for all the generated structures using the following settings: the default QC5 OTFG pseudopotential for carbon, a *k*-point spacing of 0.07 2π Å^{-1}, a plane-wave cutoff of 340 eV, and default grid scales. Next, all structures within 1 eV of the most stable structure are DFT geometry-optimised with the same settings. Finally, the structures within 0.5 eV of the ground state are re-optimised with more stringent settings: the default C9 OTFG pseudopotential for carbon, a *k*-point spacing of 0.03 2π Å^{-1}, plane-wave cutoff of 700 eV, and standard and fine grid scales of 2 and 2.3, respectively.

Analysing the structures up to 1 eV reveals a wide variety of bonding beyond that of the tetrahedral diamond from which the structures are generated, including sp, sp^2 and sp^3 bonding and mixtures. In Fig. 7, the most stable zero-, one- and two-dimensional structures are highlighted. The observation that, starting from the experimental diamond structure, isolated clusters

(foreshadowing the fullerenes), nanotubes and graphitic structures are generated is astonishing, and suggests the discovery potential of single pieces of data. Even without the DFT energetic data, which points to a given structure's stability and likely synthesisability, the existence of the low-dimensional, threefold coordinated, carbon structures among the generated structures would likely encourage speculation, had they not been previously known. It should be noted that the application of symmetry enforces the large diversity of structures. However, even without applying symmetry in a search of 8 carbon atoms, layered graphitic-like structures are generated, albeit somewhat distorted, and highly compressed.

The data relaxed to a higher level of accuracy, up to 0.5 eV above the most stable structures, are filtered so as to highlight only the three-dimensional carbon

Table 1 Three-dimensional carbon framework structures. Space groups are reported in the Hermann–Mauguin notation, along with the number of atoms in the primitive unit cell. The total energies, with respect to the graphitic two-dimensional structure shown in Fig. 7, and volumes are reported per atom. The SACADA serial number is reported where identified. A dash indicates no SACADA entry has been identified

Space group	Number	Energy (eV)	Volume (Å^3)	SACADA #
$Fd\bar{3}m$	34	0.205	6.583	158
$Pm\bar{3}n$	46	0.238	6.526	159
$P4_2/ncm$	12	0.239	5.954	107
$Pn\bar{3}m$	24	0.241	9.411	46
$P6_522$	6	0.242	6.213	29
$P6_3/mcm$	48	0.260	5.855	917
$P6_3/mmc$	36	0.292	5.908	549
$P6_3/m$	42	0.316	5.948	—
$P6_122$	36	0.323	5.907	569
$I4/mmm$	4	0.328	6.011	60
$Fd\bar{3}m$	44	0.341	7.029	—
$I\bar{4}3m$	31	0.342	5.894	—
$I\bar{4}3m$	23	0.346	6.293	204
$P\bar{6}2m$	32	0.352	5.988	—
$P6_122$	48	0.359	5.861	—
$F\bar{4}3m$	17	0.365	7.223	—
$P6_122$	48	0.373	5.868	—
$P\bar{6}2m$	15	0.378	6.043	—
$P6_3/m$	48	0.380	6.427	—
$I4/mmm$	16	0.383	5.848	916
$P6_322$	48	0.386	6.048	—
$P6/m$	48	0.388	6.048	—
$P6/mmm$	12	0.427	6.049	—
$I4_1/acd$	32	0.428	6.114	—
$P\bar{3}c1$	48	0.431	6.131	—
$P6_522$	36	0.433	5.920	—
$I4/mcm$	8	0.435	6.392	76
$F\bar{4}3m$	29	0.436	7.904	—
$P6/m$	16	0.436	7.611	—
$P6/mmm$	36	0.455	6.193	1037
$Im\bar{3}m$	24	0.462	10.046	54
$P6/m$	34	0.475	6.328	—
$Im\bar{3}m$	30	0.485	6.180	121
$P4/mnc$	40	0.494	6.103	—

framework structures. The resulting structures are listed in Table 1 and a selection highlighted in Fig. 8. The SACADA[68] online database aims to collect the many, often repeated, predictions of carbon structures from the literature. This is a challenging task, and absence in the database does not necessarily indicate the novelty of a given structure. Further, many topologies may have been reported for related systems such as silicon, and the silicates. However, it is notable that a significant fraction of the structures reported in Table 1 are not currently listed in the SACADA database, again pointing to the discovery potential of generating structures related to a single known experimental structure.

B. Pyrope garnet $Mg_3Al_2(SiO_4)_3$

To extend the investigation to a more complex example, we consider the pyrope garnet composition, $Mg_3Al_2(SiO_4)_3$. The garnet structure is rather elaborate, $Ia\bar{3}d$ cubic with 160 atoms in the conventional unit cell. With four chemical species, in contrast to the diamond structure there are multiple local environments.

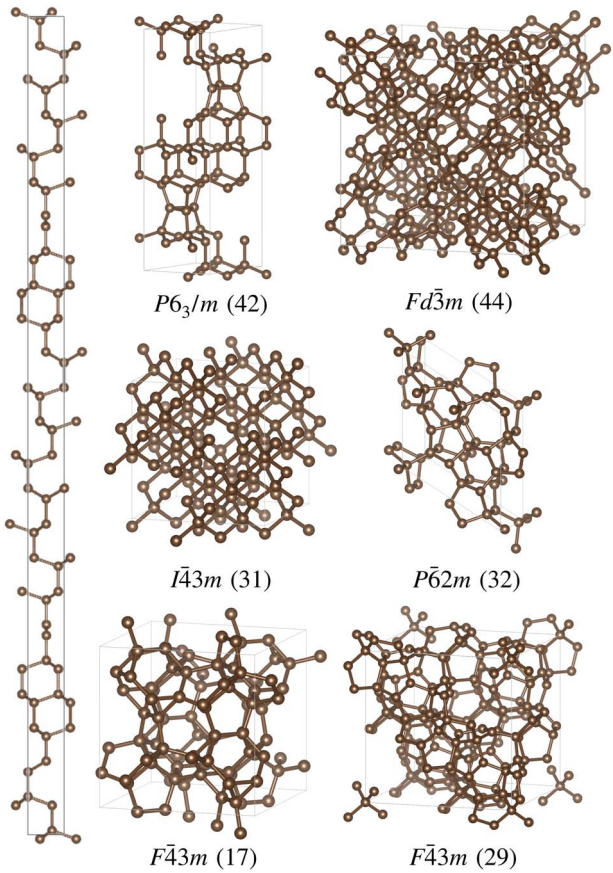

Fig. 8 Selected three-dimensional carbon framework structures. The space group and number of atoms in the primitive unit cell are indicated. The left hand, high aspect ratio, structure has space group $P6_122$ and 48 atoms. It is characterised by regions of diamond-like material, connected by graphitic regions, reminiscent of diaphite.[73]

To explore the transferability of the approach, and to test its integration into a measurement-based structure-searching strategy, rather than starting from the pyrope composition, or an experimental crystal structure, a DFT-driven AIRSS search with a single formula unit of a 1 : 1 : 1 composition of MgO, Al_2O_3 and SiO_2 was first performed. The initial random structures were generated to have a range of volumes and a random MINSEP matrix of between 2 and 3 Å. Symmetry was applied to the structures, randomly choosing 2 to 4 symmetry operations. CASTEP, QC5 OTFG pseudopotentials, a 340 eV plane-wave cutoff, and 0.07 2π $Å^{-1}$ *k*-point spacing and the PBE density functional were used to structurally optimise 29 random structures

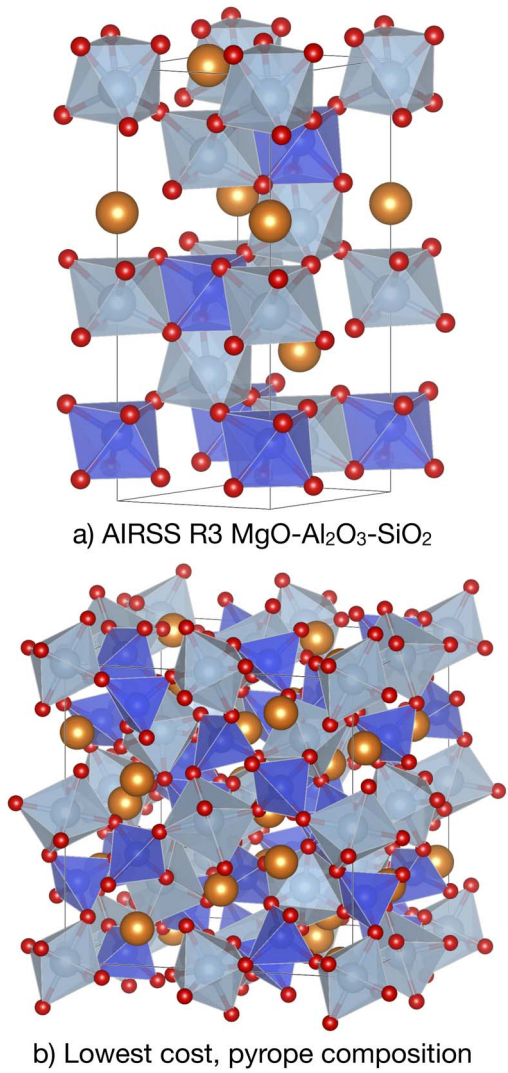

Fig. 9 Generation of pyrope garnet structure. (a) The conventional cell of the *R*3 symmetry AIRSS-generated structure for a single formula unit of $MgO–Al_2O_3–SiO_2$ at 10 GPa. (b) The lowest predicted cost structure in the pyrope $Mg_3Al_2(SiO_4)_3$ composition, which is identical to the experimentally known 180-atom conventional cell garnet structure.

under 10 GPa of applied external pressure. A structure with the space group $R3$, see Fig. 9(a), was encountered multiple (6) times, and taken as the target structure for the generation of a cost-based EDDP potential.

A two-body EDDP was trained on the cost data using manifest, with 16 polynomials for the environment features, and two hidden layers of 20 nodes each. 30 individual networks were trained, with 9 selected by the NNLS ensembling procedure. 1000 structures with a single formula unit of $MgO–Al_2O_3–SiO_2$ were randomly generated in the first step, applying 2 to 4 symmetry operations and a random MINSEP matrix of 2 to 3 Å, along with 1000 shakes of the target lowest energy $MgO–Al_2O_3–SiO_2$ structure with a position and cell amplitude of 0.1. The cutoff radius was set to 5 Å. During the active learning phase, there were 5 cycles of 1000 AIRSS-generated structures, added with a 0.1 position and cell amplitude shake. The parameters for the cost function were $\alpha = 100$ and $\beta = 10$.

Using the cost-based EDDP, a random search is performed in the pyrope, $Mg_3Al_2(SiO_4)_3$, composition, with a unit cell containing 4 formula units, 24 and 48 randomly chosen symmetry operations, and a random MINSEP matrix of 2 to 3 Å. Of the 814 structures generated, the one with the lowest EDDP predicted cost had a space group of $Ia\bar{3}d$ and was encountered three times. Already visually appearing very similar, geometry optimising the generated structure, using CASTEP, QC5 OTFG pseudopotentials, a 340 eV plane-wave cutoff and a gamma point sampling of the Brillouin zone, leads to an identical structure to the experimentally known pyrope garnet; see Fig. 9(b). The next lowest predicted cost structure, with space group $I4_132$, is 223 meV per atom less stable when optimised at 10 GPa. The rediscovery of the garnet structure demonstrates both the transferability of the approach to novel compositions, and a practical and highly computationally efficient method to uncover complex crystal structures.

IX. Relation to diffusion-based generative approaches

It can be a challenge to navigate the differences in terminology when research fields collide. Generative machine-learning methods have excited the research community. The field of structure prediction is no exception, with a wide array of generative approaches to structure prediction being explored.[26,74–79] In the above, I have tried to make the case that the building of "random sensible structures" is a generative process. But the similarities to machine-learning-based approaches go beyond that.

The scheme outlined in Section VII is in essence identical to a generative diffusion process. In a diffusion model, target images, or structures, are "noised" – or in the language of random structure searching, "shaken". The noise is increased until no remnants of the original target remain. Given the target, and the noised intermediates, a machine-learning model is trained to "find its way" from a noised to a less-noised configuration. As described illuminatingly in ref. 80, the denoising can be achieved by starting from a random configuration and minimising some cost function of the distance to the manifold of the target examples. It is clear that this is exactly the procedure described in Section VII, where the machine-learning model is an EDDP, trained on distance (in feature, or environment vector, space) derived data. Indeed, it is clear that such a diffusion-

style model is also very similar to random structure search based on an EDDP (or other MLIP) trained on DFT energetic data of marker structures – and going downhill in energy takes you back to the marker structures, or new similar ones, with similarly low energy. From this perspective, it is instructive to note the fundamental similarity of generative models (such as MatterGen[26]), and universal potentials (such as MACE0 (ref. 81)) coupled with AIRSS.[7,8]

When creating diffusion models, a lot of care is taken in designing the noising process. From the perspective of structure prediction, this is equivalent to designing appropriate shakes in AIRSS, or moves in basin-hopping-style algorithms. This suggests that there is expected to be considerable benefit from exploring the respective field's insights – for the generative models to learn the denoising process, and for MLIPs to design optimal sampling of energy landscapes for the construction of training datasets.

X. Conclusion

We have seen that first-principles, theory-driven, random structure searching, as implemented in AIRSS, is an engine for the discovery of novel arrangements of matter, exposing new science, which is frequently experimentally confirmed – almost to the point of it being routine. These searches must be thoroughly carried out, to identify all competing, and potentially less interesting, phases, and to avoid over-prediction. Purely first-principles searches are computationally demanding, and this thoroughness can be difficult to achieve. With the rise of data-driven methods – especially MLIPs, which massively accelerate traditional structure searches, but also the closely related generative approaches – AIRSS is emerging as a key source of training data. The broad sampling of structure space that AIRSS naturally offers is essential to the development of robust MLIPs, something that is challenging when restricted to the datasets derived from highly biased materials structure databases. Innovations enabled by machine-learning acceleration, such as hot-AIRSS, introduced here, broaden the applicability of AIRSS to a greater variety of ever more complex structures, combined with more sophisticated schemes for generating candidate structures, such as our new EDDP distance-based approach, emphasising data-driven discovery as an emerging and powerful force in the atomistic sciences.

Data availability

Data for this article, including the EDDPs and results of structure searches, are available at Zenodo at **https://doi.org/10.5281/zenodo.11966891**.

Conflicts of interest

There are no conflicts to declare.

References

1 A. R. Oganov, C. J. Pickard, Q. Zhu and R. J. Needs, *Nat. Rev. Mater.*, 2019, **4**, 331.

2 F. H. Stillinger and T. A. Weber, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 1985, **31**, 5262.
3 R. Biswas and D. Hamann, *Phys. Rev. Lett.*, 1985, **55**, 2001.
4 J. Tersoff, *Phys. Rev. Lett.*, 1988, **61**, 2879.
5 S. Woodley, P. Battle, J. Gale and C. A. Catlow, *Phys. Chem. Chem. Phys.*, 1999, **1**, 2535.
6 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*, 2016, **351**, aad3000.
7 C. J. Pickard and R. J. Needs, *Phys. Rev. Lett.*, 2006, **97**, 045504.
8 C. J. Pickard and R. J. Needs, *J. Phys.: Condens. Matter*, 2011, **23**, 053201.
9 D. F. R. Brown, M. N. Gibbs and D. C. Clary, *J. Chem. Phys.*, 1996, **105**, 7597.
10 J. Behler and M. Parrinello, *Phys. Rev. Lett.*, 2007, **98**, 146401.
11 A. P. Bartók, M. C. Payne, R. Kondor and G. Csányi, *Phys. Rev. Lett.*, 2010, **104**, 136403.
12 A. P. Bartók, S. De, C. Poelking, N. Bernstein, J. R. Kermode, G. Csányi and M. Ceriotti, *Sci. Adv.*, 2017, **3**, e1701816.
13 C. J. Pickard, *Phys. Rev. B*, 2022, **106**, 014102.
14 P. T. Salzbrenner, S. H. Joo, L. J. Conway, P. I. Cooke, B. Zhu, M. P. Matraszek, W. C. Witt and C. J. Pickard, *J. Chem. Phys.*, 2023, **159**, 144801.
15 C. J. Pickard and R. Needs, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 2007, **76**, 144114.
16 C. J. Pickard, I. Errea and M. I. Eremets, *Annu. Rev. Condens. Matter Phys.*, 2020, **11**, 57.
17 A. R. Oganov and C. W. Glass, *J. Chem. Phys.*, 2006, **124**, 244704.
18 Y. Wang, J. Lv, L. Zhu and Y. Ma, *Comput. Phys. Commun.*, 2012, **183**, 2063.
19 D. C. Lonie and E. Zurek, *Comput. Phys. Commun.*, 2011, **182**, 372.
20 Z. Lu, B. Zhu, B. W. Shires, D. O. Scanlon and C. J. Pickard, *J. Chem. Phys.*, 2021, **154**, 174111.
21 B. Zhu, Z. Lu, C. J. Pickard and D. O. Scanlon, *APL Mater.*, 2021, **9**, 121111.
22 C. J. Smalley, H. E. Hoskyns, C. E. Hughes, D. N. Johnstone, T. Willhammar, M. T. Young, C. J. Pickard, A. J. Logsdail, P. A. Midgley and K. D. Harris, *Chem. Sci.*, 2022, **13**, 5277.
23 V. Kapil, C. Schran, A. Zen, J. Chen, C. J. Pickard and A. Michaelides, *Nature*, 2022, **609**, 512.
24 V. L. Deringer, C. J. Pickard and G. Csányi, *Phys. Rev. Lett.*, 2018, **120**, 156001.
25 A. Merchant, S. Batzner, S. S. Schoenholz, M. Aykol, G. Cheon and E. D. Cubuk, *Nature*, 2023, **624**, 80.
26 C. Zeni, R. Pinsler, D. Zügner, A. Fowler, M. Horton, X. Fu, S. Shysheya, J. Crabbé, L. Sun, J. Smith, *et al.*, *arXiv*, 2023, preprint, arXiv:2312.03687DOI: **10.48550/arXiv.2312.03687**.
27 C. J. Pickard and R. J. Needs, *Nat. Phys.*, 2007, **3**, 473.
28 P. Loubeyre, F. Occelli and P. Dumas, *Nature*, 2020, **577**, 631.
29 L. Monacelli, M. Casula, K. Nakano, S. Sorella and F. Mauri, *Nat. Phys.*, 2023, **19**, 845.
30 R. T. Howie, C. L. Guillaume, T. Scheler, A. F. Goncharov and E. Gregoryanz, *Phys. Rev. Lett.*, 2012, **108**, 125501.
31 S. L. Price, *Chem. Soc. Rev.*, 2014, **43**, 2098.
32 C. J. Pickard and R. Needs, *Nat. Mater.*, 2008, **7**, 775.

33 S. Ninet, F. Datchi, P. Dumas, M. Mezouar, G. Garbarino, A. Mafety, C. Pickard, R. Needs and A. Saitta, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 2014, **89**, 174103.
34 C. J. Pickard and R. Needs, *Phys. Rev. Lett.*, 2009, **102**, 146401.
35 Y. Ma, M. Eremets, A. R. Oganov, Y. Xie, I. Trojan, S. Medvedev, A. O. Lyakhov, M. Valle and V. Prakapenka, *Nature*, 2009, **458**, 182.
36 M. I. McMahon and R. J. Nelmes, *Chem. Soc. Rev.*, 2006, **35**, 943.
37 M. McMahon and R. Nelmes, *Z. Kristallogr. – Cryst. Mater.*, 2004, **219**, 742.
38 C. J. Pickard and R. Needs, *Nat. Mater.*, 2010, **9**, 624.
39 M. G. Gorman, S. Elatresh, A. Lazicki, M. M. Cormier, S. Bonev, D. McGonegle, R. Briggs, A. Coleman, S. Rothman, L. Peacock, et al., *Nat. Phys.*, 2022, **18**, 1307.
40 C. J. Pickard and R. Needs, *Phys. Rev. Lett.*, 2011, **107**, 087201.
41 D. Garisto, *Nature*, 2024, **628**, 481.
42 S. J. Clark, M. D. Segall, C. J. Pickard, P. J. Hasnip, M. I. Probert, K. Refson and M. C. Payne, *Z. Kristallogr. – Cryst. Mater.*, 2005, **220**, 567.
43 L. J. Conway, C. J. Pickard and A. Hermann, *Proc. Natl. Acad. Sci. U. S. A.*, 2021, **118**, e2026360118.
44 J. R. Nelson, R. J. Needs and C. J. Pickard, *Phys. Rev. Mater.*, 2021, **5**, 123801.
45 A. M. Shipley, M. J. Hutcheon, R. J. Needs and C. J. Pickard, *Phys. Rev. B*, 2021, **104**, 054501.
46 K. Dolui, L. J. Conway, C. Heil, T. A. Strobel, R. P. Prasankumar and C. J. Pickard, *Phys. Rev. Lett.*, 2024, **132**, 166001.
47 **https://www.mtg.msm.cam.ac.uk/Codes/AIRSS**.
48 L. Prechelt, in *Neural Networks:Tricks of the Trade*, Springer, 1998, pp. 55–69.
49 L. K. Hansen and P. Salamon, *IEEE Trans. Pattern Anal. Mach. Intell.*, 1990, **12**, 993.
50 C. Schran, K. Brezina and O. Marsalek, *J. Chem. Phys.*, 2020, **153**, 104105.
51 N. Lopanitsyna, G. Fraux, M. A. Springer, S. De and M. Ceriotti, *Phys. Rev. Mater.*, 2023, **7**, 045802.
52 **https://www.mtg.msm.cam.ac.uk/Codes/EDDP**.
53 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*, 2009, **457**, 863.
54 B. Cheng, G. Mazzola, C. J. Pickard and M. Ceriotti, *Nature*, 2020, **585**, 217.
55 C. Schran, F. L. Thiemann, P. Rowe, E. A. Müller, O. Marsalek and A. Michaelides, *Proc. Natl. Acad. Sci. U. S. A.*, 2021, **118**, e2110077118.
56 V. L. Deringer, N. Bernstein, G. Csányi, C. Ben Mahmoud, M. Ceriotti, M. Wilson, D. A. Drabold and S. R. Elliott, *Nature*, 2021, **589**, 59.
57 J. P. Perdew, K. Burke and M. Ernzerhof, *Phys. Rev. Lett.*, 1996, **77**, 3865.
58 S. E. Ahnert, W. P. Grant and C. J. Pickard, *npj Comput. Mater.*, 2017, **3**, 35.
59 M. J. Van Setten, M. A. Uijttewaal, G. A. de Wijs and R. A. de Groot, *J. Am. Chem. Soc.*, 2007, **129**, 2458.
60 E. V. Podryabinkin, E. V. Tikhonov, A. V. Shapeev and A. R. Oganov, *Phys. Rev. B*, 2019, **99**, 064114.
61 A. V. Shapeev, *Multiscale Model. Simul.*, 2016, **14**, 1153.
62 W. Hayami, T. Hiroto, K. Soga, T. Ogitsu and K. Kimura, *J. Solid State Chem.*, 2024, **329**, 124407.
63 B. Callmer, *Acta Crystallogr., Sect. B: Struct. Crystallogr. Cryst. Chem.*, 1977, **33**, 1951.
64 W. Hayami, *J. Solid State Chem.*, 2015, **221**, 378.

65 K. Doll, J. Schön and M. Jansen, *J. Phys.: Conf. Ser.*, 2008, **117**, 012014.
66 D. J. Wales and J. P. Doye, *J. Phys. Chem. A*, 1997, **101**, 5111.
67 S. Goedecker, *J. Chem. Phys.*, 2004, **120**, 9911.
68 R. Hoffmann, A. A. Kabanov, A. A. Golov and D. M. Proserpio, *Angew. Chem., Int. Ed.*, 2016, **55**, 10962.
69 M. Yin, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 1984, **30**, 1773.
70 M. P. Grumbach and R. M. Martin, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 1996, **54**, 15730.
71 M. Martinez-Canales, C. J. Pickard and R. J. Needs, *Phys. Rev. Lett.*, 2012, **108**, 045704.
72 X. Lu, Connecting fullerenes with carbon nanotubes and graphene, in *Handbook of Fullerene Science and Technology*, ed. X. Lu, T. Akasaka and Z. Slanina, Springer Nature Singapore, Singapore, 2022, pp. 265–270.
73 P. Németh, K. McColl, L. A. Garvie, C. G. Salzmann, C. J. Pickard, F. Cora, R. L. Smith, M. Mezouar, C. A. Howard and P. F. McMillan, *DiamondRelat. Mater.*, 2021, **119**, 108573.
74 S. Kim, J. Noh, G. H. Gu, A. Aspuru-Guzik and Y. Jung, *ACS Cent. Sci.*, 2020, **6**, 1412.
75 C. J. Court, B. Yildirim, A. Jain and J. M. Cole, *J. Chem. Inf. Model.*, 2020, **60**, 4518.
76 T. Pakornchote, N. Choomphon-Anomakhun, S. Arrerut, C. Atthapak, S. Khamkaeo, T. Chotibut and T. Bovornratanaraks, *Sci. Rep.*, 2024, **14**, 1275.
77 X. Luo, Z. Wang, P. Gao, J. Lv, Y. Wang, C. Chen and Y. Ma, *arXiv*, 2024, preprint, arXiv:2403.10846DOI: **10.48550/arXiv.2403.10846**.
78 B. Cheng, *arXiv*, 2024, preprint, arXiv:2405.09057DOI: **10.48550/arXiv.2405.09057**.
79 N. Rønne, A. Aspuru-Guzik and B. Hammer, *arXiv*, 2024, preprint, arXiv:2402.17404DOI: **10.48550/arXiv.2402.17404**.
80 F. Permenter and C. Yuan, *arXiv*, 2023, preprint, arXiv:2306.04848DOI: **10.48550/arXiv.2306.04848**.
81 I. Batatia, P. Benner, Y. Chiang, A. M. Elena, D. P. Kovács, J. Riebesell, X. R. Advincula, M. Asta, W. J. Baldwin, N. Bernstein, *et al.*, *arXiv*, 2023, preprint, arXiv:2401.00096DOI: **10.48550/arXiv.2401.00096**.

