ARTICLE   OPEN

Revealing and exploiting hierarchical material structure through complex atomic networks

Sebastian E. Ahnert[1,2], William P. Grant[1] and Chris J. Pickard[3,4]

One of the great challenges of modern science is to faithfully model, and understand, matter at a wide range of scales. Starting with atoms, the vastness of the space of possible configurations poses a formidable challenge to any simulation of complex atomic and molecular systems. We introduce a computational method to reduce the complexity of atomic configuration space by systematically recognising hierarchical levels of atomic structure, and identifying the individual components. Given a list of atomic coordinates, a network is generated based on the distances between the atoms. Using the technique of modularity optimisation, the network is decomposed into modules. This procedure can be performed at different resolution levels, leading to a decomposition of the system at different scales, from which hierarchical structure can be identified. By considering the amount of information required to represent a given modular decomposition we can furthermore find the most succinct descriptions of a given atomic ensemble. Our straightforward, automatic and general approach is applied to complex crystal structures. We show that modular decomposition of these structures considerably simplifies configuration space, which in turn can be used in discovery of novel crystal structures, and opens up a pathway towards accelerated molecular dynamics of complex atomic ensembles. The power of this approach is demonstrated by the identification of a possible allotrope of boron containing 56 atoms in the primitive unit cell, which we uncover using an accelerated structure search, based on a modular decomposition of a known dense phase of boron, γ-B$_{28}$.

INTRODUCTION

Considerable growth in computational power and its ubiquity has been coupled with the development of efficient algorithms[1, 2] and their implementation in robust[3–5] and reliable[6] computer codes. This has permitted the first principles, quantum mechanical (through density functional theory (DFT)[7–9]), treatment of material,[10] chemical[11] and biological systems[12] of increasing complexity. High throughput computations can be performed directly on pre-existing data of more modest complexity, or some modification of it,[13] or as a way to sample configuration space to discover previously unknown structures.[14–17] Altogether, these approaches offer a route to computational materials discovery.[18]

In parallel, increasing computational power and an abundance of data has given rise to another rapidly expanding field—the study of complex networks.[19–22] A key reason for its success is the fact that related mathematical approaches can be applied to a wide range of real-world network data across many academic disciplines. The structure of networks can be studied at a variety of resolutions. Local measures of connectivity can quantify the topological properties of individual nodes or edges. Global measures, calculated across the entire network, such as the average shortest path length,[19] can help us to compare networks as a whole. Between these two extremes however lies an entire field of research that searches for meaningful descriptions of intermediate structures, such as 'cliques',[23] 'communities'[24] and 'rich clubs',[25] among others. These are sets of nodes or edges which are particularly densely connected, or which share some other defining topological feature. Many definitions of such structures have been put forward.[23–28] Here we select an approach that is particularly good at detecting hierarchical modularity[27] and apply it to atomic networks, which until now have received scant attention from the networks research community, beyond the study of proteins.[29–31] Our aim is to provide an automated coarse-graining of the atomic structures of crystal structures. The simplification of the space of possible configurations of complex atomic systems has the potential to vastly accelerate the process of computational materials discovery, among other tasks that can benefit from automatic coarse-graining based on hierarchical decomposition. We illustrate this through the identification of a possible new allotrope of boron.

RESULTS

Determining the modularity of atomic networks

For atomic structures of a single atomic species we can generate an unweighted network of atoms by simply imposing a threshold d^* on the interatomic distance and connecting atoms that are closer to each other than this threshold distance. The communities in this network can then be extracted by using the algorithm of Arenas et al.,[27] as discussed below.

The extent to which a network has well-defined community structure can be quantified with a metric known as the

[1]Theory of Condensed Matter Group, Cavendish Laboratory, University of Cambridge, JJ Thomson Avenue, Cambridge CB3 0HE, UK; [2]Sainsbury Laboratory, University of Cambridge, Bateman Street, Cambridge CB2 1LR, UK; [3]Department of Materials Science & Metallurgy, University of Cambridge, 27 Charles Babbage Road, Cambridge CB3 0FS, UK and [4]Advanced Institute for Materials Research, Tohoku University 2-1-1 Katahira, Aoba, Sendai 980-8577, Japan
Correspondence: Chris J. Pickard (cjp20@cam.ac.uk)

npj

Revealing and exploiting hierarchical material structure
SE Ahnert et al.

2

modularity[24] (Fig. 1). This is defined as the fraction of edges that run between nodes of the same community, minus the expected fraction if the edges of the network were positioned randomly:

$$Q = \frac{1}{2A} \sum_{i,j} (A_{ij} - P_{ij}) \delta(C_i, C_j) \tag{1}$$

Here $i, j \in [1, n]$, where n is the number of nodes. $\delta(C_i, C_j) = 1$ if nodes i and j belong to the same community, and 0 otherwise, and A_{ij} and P_{ij} are the adjacency matrices of the network and of the null model (the randomised network), respectively. $A = \frac{1}{2} \sum_{i,j} A_{ij}$ gives the total number of edges in the network.

This metric relies on the concept that a random graph is not expected to exhibit community structure. As such, the quality of the proposed community structure can be quantified by the difference between the network and the null model. The degree of a node i is defined as the number of edges of the node, $A_i = \sum_j A_{ij}$. Choosing the null model to have the same degree distribution as the target network gives:

$$Q = \frac{1}{2A} \sum_{i,j} \left(A_{ij} - \frac{A_i A_j}{2A} \right) \delta(C_i, C_j) \tag{2}$$

This can be rewritten as a sum over the M communities of the network:

$$Q = \sum_{s=1}^{M} \left(\frac{A_{ss}}{A} - \left(\frac{A_s}{2A} \right)^2 \right) \tag{3}$$

Where for unweighted networks A_{ss} is the number of edges within community s, and A_s is the sum of the degrees of nodes within community s. (Equivalently, A_s is the number of edges exiting the community $+2A_{ss}$.) This metric allows community detection to be recast as an optimisation problem; maximising Q minimises the number of edges between communities. However, naive optimisation of the modularity has been shown to have a fundamental

Fig. 1 A network with three communities of five nodes each. Let the white, grey and black communities be C_1, C_2, C_3. Then A_{ss}, the number of internal edges, is {8, 9, 9}. The sum of degrees of each community, A_s is {18, 21, 21}, and the total number of edges is 30, giving a final value for Q of 0.532. An alternative partition, indicated by the dotted lines, has $A_{ss} = \{18, 9\}$ and $A_s = \{39, 21\}$, giving $Q = 0.355$. As expected, the less intuitive partitioning gives a lower Q value

resolution limit.[32] Communities smaller than a certain size A_{ss}^{min} will not be detected. This threshold depends on the total number of edges in the network:

$$A_{ss}^{min} = \sqrt{\frac{A}{2}} - 1 \tag{4}$$

This resolution limit arises due to the explicit dependence on the number of edges within the null model. This introduces a preferred size for the communities in the network.

The algorithm of Arenas et al.[27] utilises this bound on the size of detectable communities. By adding a self-loop of strength w to each node in network, $\mathbf{A} \rightarrow \mathbf{A} + w\mathbf{I}$, the resolution limit inequality becomes:

$$A_{ss} < \frac{1}{2} \left(\sqrt{2A + Nw} - N_s w - 2 \right) \tag{5}$$

where N_s is the number of nodes in community s, and N the total number of nodes. By varying the effective resolution limit, the scale of the communities extracted can be varied. As w is increased, modularity optimisation will result in an increasingly fragmented decomposition. By optimising modularity for a range of this parameter w, and across a range of threshold values d^*, one obtains a variety of hierarchical decompositions into modules (see the section 'Relax and Shake Algorithm' for details of the optimisation algorithm). The simplest quantity one can establish across this two-dimensional (2D) space is the number of modules. Regions across which this value is stable represent more meaningful modules that may have real physical or biological meaning as rigid clusters or units of protein architecture.

Pauling's rule of parsimony suggests that the number of different kind of constituents in a crystal is small.[33] This suggests that in complex crystal structures, in which modules are likely to exhibit a degree of symmetry, it makes sense to minimise a more sophisticated quantity, namely the information content of a structure. The identification of modules corresponds to a compression if these modules contain symmetries or if the same module appears multiple times. We can calculate the amount of information I required to describe a given module structure in terms of the global degrees of freedom of that structure, and minimise this quantity over the space of d^* and w.

In order to calculate I, consider M modules of M' distinct types. The number of modules with one atom only is $0 \le M^* \le M$, and the number of modules with two atoms is $0 \le M^{**} \le M$. To position and rotate the M modules relative to one another we need $6M - 6$ degrees of freedom in general, with one degree less for every two-atom module, and three degrees less for every one-atom module. Now consider each of the M' distinct modules: If we have $n_i > 2$ inequivalent atoms in module i we need $3n_i - 6$ degrees of freedom to describe them. If we have $n_i = 2$ inequivalent atoms in module i we need $3n_i - 5 = 1$ degree of freedom to describe them, which corresponds to the distance between the atoms. If we have $n_i = 1$ inequivalent atom in module i we need 0 degrees of freedom to describe it internally. The global number of degrees of freedom is then:

$$I = 6M - M^{**} - 3M^* - 6 + \sum_{i=1}^{M'} 3n_i - 6 + 3\delta_{1n_i} + \delta_{2n_i} \tag{6}$$

Note that the number of inequivalent atoms n_i depends on the number of atoms in module i, as well as its point group symmetry. If all N atoms are in one module, repeated once, we have $M = 1$, $M' = 1$, $M^* = 0$, $M^{**} = 0$, and $n_i = N$. Hence $I = 3N - 6$, as required. If all N atoms are in N modules of one atom we have $M = N$, $M' = 1$, $M^* = N$, $M^{**} = 0$, $n_i = 1$. Hence in this case also $I = 3N - 6$, as required. The information I can be normalised by the maximum possible value of $3N - 6$. This normalised value is used in all the heat-map figures in this manuscript.
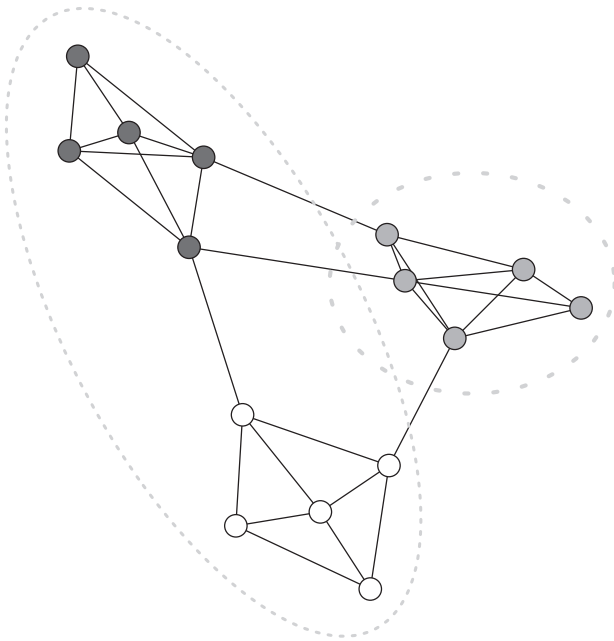
Revealing and exploiting hierarchical material structure
SE Ahnert et al.

npj

3

The decomposition which has the minimum I gives us the most concise description of a structure. This minimisation of the description length is conceptually related to the idea of algorithmic information theory,[34–36] as the symmetry operations and inequivalent atomic positions that form part of the compression can be thought of algorithms which allow us to reconstruct the original atomic structure. The length of the shortest such description is a quantitative measure of the structure's complexity. Because of the presence of crystal symmetries, we need to establish the modularity of the atomic network with high accuracy. In addition, the modularity is highly degenerate; there is a greater than exponential number of distinct possible community structures, and many will have modularity values close to that of the global maximum.[37] Moreover, these structures may have very different topologies to that of the true partition, resulting in a large change in the compression achieved. We therefore employ an algorithm similar to that of the 'relax and shake' algorithm[15] or zero temperature basin hopping.[38]

Relax and Shake Algorithm
The relax and shake (RASH) algorithm uses a repeated series of local modularity optimisations (relax) followed by the assignment of a small number of nodes into random communities (shake), in order to escape local maxima. The local optimisation follows existing work on community detection.[39, 40] The modularity is optimised by moving each node in the network to the community of the neighbouring node resulting in the highest increase in the modularity (if > 0). This is then repeated until no further local optimisations increase the modularity. Following the local optimisation, a subset of the nodes (10%) are shaken into other communities within the network, and the local optimisation repeated. This continues until 200 consecutive relax-and-shake iterations have failed to improve the modularity. As an additional check on the solution, the modularity change resulting in merging any two communities is calculated; if this results in a modularity increase, the merge is performed, and the relax-and-shake iteration process is begun again using the new partition. The above can be considered a single optimisation step; following this, a larger subset of the nodes (20%) are shaken into either pre-existing, or new communities. The optimisation is then performed again. This shake-and-relax is performed until 200 iterations fail to improve the modularity. The whole process is repeated until three consecutive runs have failed to produce a community structure with a higher modularity. The degree of repetition is parameterisable, and allows us to have confidence in the community structure obtained (at the cost of speed).

If we partition a structure of N atoms into M multi-atom modules, so that typically $M \ll N$, and assume that the modules correspond to rigid clusters, then we reduce the dimensionality of configuration space from $3N - 6$ (atomic positions minus global translation and rotation) to $6M - 6$ (as we have to specify a relative translation and rotation for each module) or less (if any of the modules have less than three atoms). We will show that this can be exploited in the first principles prediction of crystal structure.

Application to crystal structures
Boron. Boron is known to form several allotropes,[41] including α-B,[42] β-B[43, 44] and γ-B.[45–47] The structure of rhombohedral α-B is widely recognised as being made up of interconnected B_{12} icosahedra (Fig. 2). However, because the bonds between different icosahedra are shorter than the bonds within, simple thresholds on bond length will not yield the underlying modular structure of this crystal phase. This observation motivated the development of our current scheme, which, based on network modularity, does yield the scientifically agreed icosahedral modules.
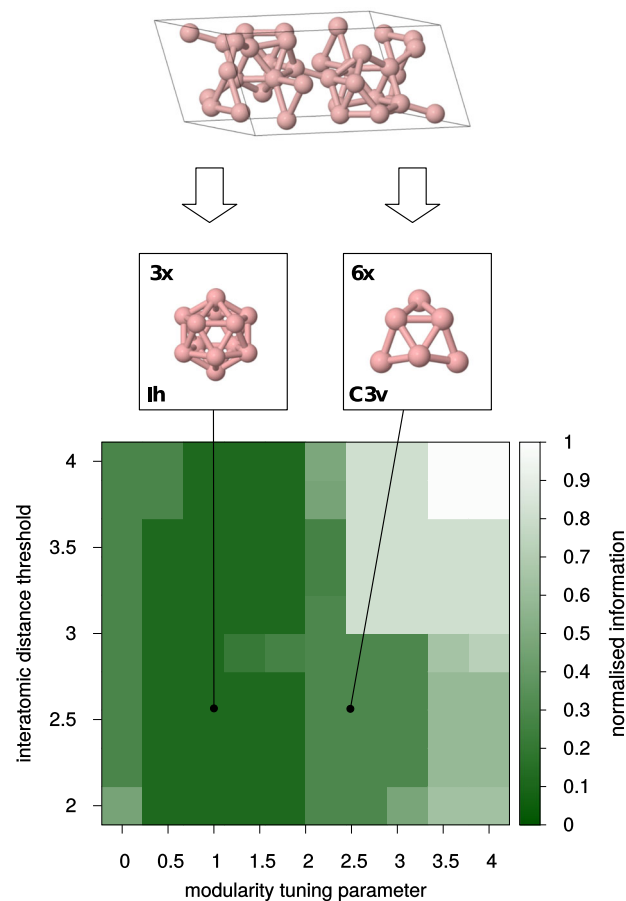

Fig. 2 The structure of α-B_{12} (in the 36 atom hexagonal unit cell) and its modular decompositions. As established by the modularity decomposition the simplest way of describing this structure is as three interconnected icosahedra. For the purpose of comparison we also show an alternative, less efficient decomposition into six units of six atoms with C_{3v} symmetry. These units correspond to halves of the icosahedra. The decomposition that minimises the amount of information I (normalised here by the maximum value of $3N - 6$) required to describe the crystal structure, occurs for values around $w_s = 1$ (modularity tuning parameter) and $d^* = 2.5$ Å (interatomic distance threshold). This decomposition occupies a broad plateau (darkest shade of *green*) in the 2D space of the modularity tuning parameter w_s and the interatomic distance threshold d^*

The structure of β-B is much more complicated, and various models have been proposed, typically with 105 or 106 atoms per unit cell. We choose the 105-atom model of Geist et al.[44] for further investigation (Fig. 3). Our modularity detection scheme identifies four icosahedra, two larger 25-atom modules with threefold cyclic point group symmetry C_{3v}, and one module with threefold dihedral symmetry D_{3d} of seven atoms. Two of the four icosahedra are slightly distorted, resulting in C_{2v} symmetry, rather than icosahedral I_h symmetry. Of course, the decomposition of complex boron structures into compact and symmetric sub units is not unprecedented, see for example Fig. 2 in ref. 41. We emphasise, however, that our scheme performs the decomposition automatically and is suitable for integration into complex computational workflows.

Recently the structure of a high-pressure phase of boron, γ-B_{28}, has been described in the literature[41, 46, 47] (Fig. 4). A w_s vs. d^* heat-map of I for the unweighted network reveals a global minimum at $w_s = 1.0$ and $d^* = 2.0$ Å. This corresponds to two 14-atom modules with D_{2h} symmetry, which are icosahedra plus two

npj
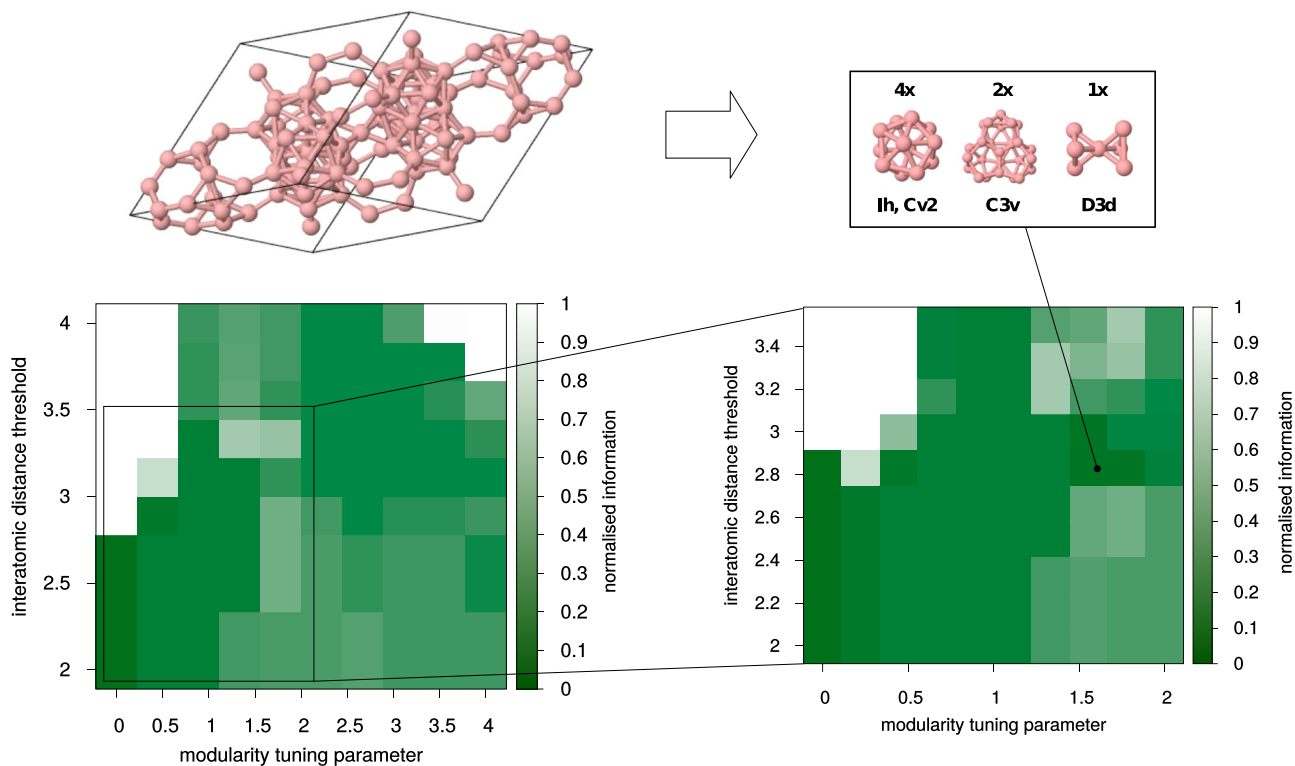Revealing and exploiting hierarchical material structure
SE Ahnert et al.

4

Fig. 3 The highly complex structure of β-B_{105} (in the 105-atom rhombohedral unit cell) and its modular decompositions. The complexity of the structure means a more complex landscape of possible decompositions. The module structure that minimises the amount of information I (normalised here by the maximum value of $3N - 6$) required to describe the crystal structure, occurs for values around $w_s = 1.66$ and $d^* = 2.82$Å. This decomposition consists of four icosahedra, two large 25-atom structures with threefold symmetry (C_{3v}), and a further module with dihedral symmetry (D_{3d}). Two of the four icosahedra have exact icosahedral symmetry, whereas the other two are slightly distorted, and exhibit C_{2v} symmetry

atoms on either side (Fig. 4). Note that this contrasts with the conventional decomposition into two icosahedra and two dimers found in the literature,[46] which is less favoured in our scheme as it corresponds to a higher value of I. Our approach offers a meaningful partition of this structure into modules, providing insight into the organisation and visualisation of this structure and opening the door to the systematic exploration of the structure space that neighbours this γ-B_{28} allotrope.

Phosphorus. Like boron, phosphorous exhibits rich allotropism, from the highly metastable white phosphorous, to layered black phosphorous, and extremely complex fibrous, or layered, structures.[48] There is considerable current interest in 2D black phosphorous, or phosphorene[49] and other layered forms.[50] Here we investigate the crystal structure of red phosphorus,[51] and attempt to identify a simple decomposition into modules using our current scheme. In the 42-atom primitive unit cell (space group P$\bar{1}$) the modularity decomposition finds two modules that each occur twice (Fig. 5). The bigger module has symmetry (C_s) and contains 13 atoms, while the smaller has C_{2v} symmetry and contains eight atoms. The relatively low degree of symmetry of red phosphorous means that the landscape of I with varying w_s and d^* is flatter, but our approach nevertheless finds a parsimonious decomposition of the crystal structure.

Metal-organic frameworks. We extend our framework to multi-species structures requiring only a definition of the relationship between d^* and the interatomic distances used to determine the network connectivity. In principle, a separate d^* could be defined for each pair of atomic species. However, this introduces the cost of exploring a higher dimensional space in the search for an optimal I. Instead, we define a single dimensionless parameter d^*_{eff}

that specifies the distance threshold as the d^*_{eff}-fold multiple of the sum of the fixed atomic radii for a given pair of atoms. We apply this multi-species version of our approach to the metal-organic framework MOF-5, or $Zn_4O(BDC)_3$, where BDC^{2-} is 1,4-benzenedicarboxylate.[52] Metal-organic frameworks exhibit a vast range of structures and are of great interest because their porosity allows them to be used for the storage of gases, such as hydrogen, or carbon dioxide.[53] As can be seen in Fig. 6 our algorithm finds two similar decompositions with almost equally low I-values. The lowest minimum corresponds to a decomposition of the structure into six 16-atom modules with D_{2h} symmetry, and two six-atom modules with tetrahedral symmetry (T_d). The 16-atom modules correspond to the BDC^{2-} molecules that are sometimes referred to as the 'struts' of metal-organic frameworks. This suggests that the decomposition derived automatically through our procedure is chemically meaningful.

Structure prediction

Ab initio random structure searching (AIRSS)[14, 15] is a simple, and demonstrably effective, approach to first principles structure prediction. It has been applied to a wide range of systems, from the crystal structures of dense hydrogen[54] and hydrogen rich compounds,[14] to matter under extreme compression,[55] and interfaces.[56] The approach involves selecting initially random structures from distributions defined by physically motivated constraints (for example, density, composition, atomic distances, symmetry, molecular units or fragments). These random 'sensible'[15] structures are fully relaxed (moved to the nearest local minimum in the energy landscape) under forces derived from DFT. Once a large number of computations have been
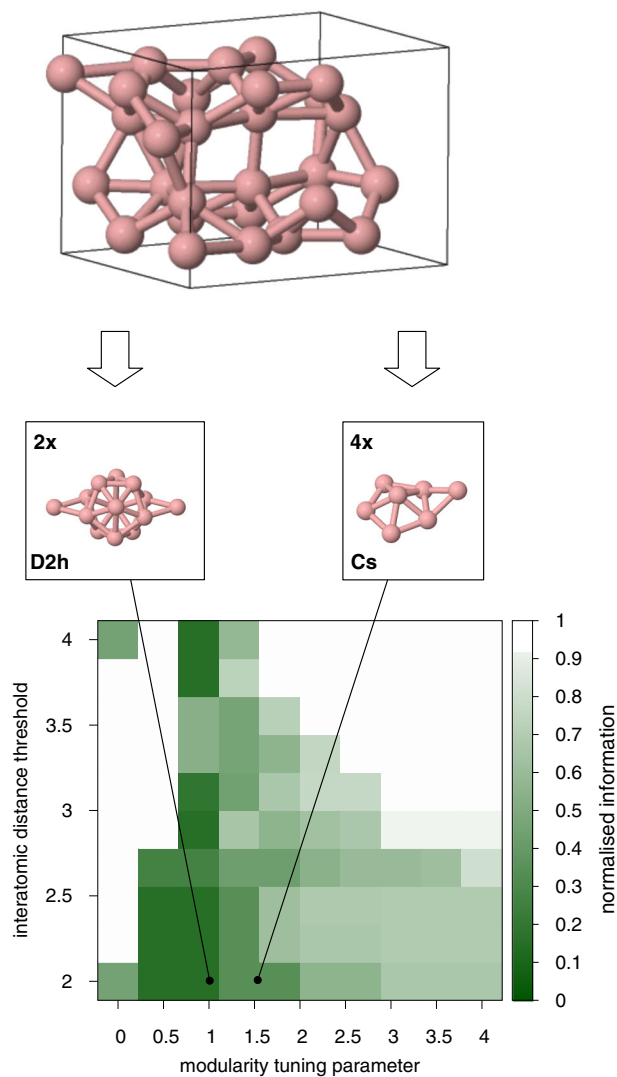
Revealing and exploiting hierarchical material structure
SE Ahnert et al.

npj

5

Fig. 4 The structure of γ-B_{28} (Pnnm unit cell) and its modular decompositions. The most parsimonious decomposition, which minimises the amount of information I (normalised here by the maximum value of $3N - 6$) required to describe the crystal structure, occurs for values around $w_s = 1.0$ and $d^* = 2.0$ Å. This decomposition consists of two identical modules of fourteen atoms with dihedral symmetry (D_{2h}), which in turn are each composed of an icosahedron with two adjoining atoms, one on either side. For the purposes of comparison we also show a less efficient decomposition, with C_s symmetry
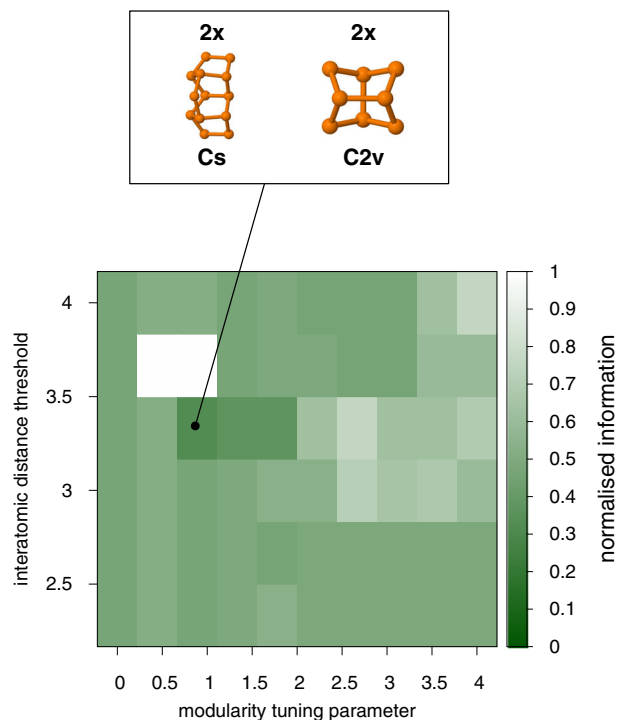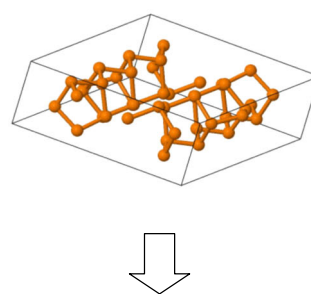

Fig. 5 The structure of red phosphorous (42 atom P$\bar{1}$ unit cell) and its modular decomposition. The relatively low degree of symmetry in the structure as a whole results in a relatively flat landscape. The information I (normalised here by the maximum value of $3N - 6$) required to describe the structure is minimised for values around $w_s = 1$ and $d^* = 3.33$ Å. This decomposition consists of two modules, which each appear twice. One module consists of 13 atoms with C_s point group symmetry. The other consists of eight atoms with C_{2v} symmetry

performed the resulting structures can be ranked according to energy (free enthalpy) or any computable property of interest.

It has been a surprise to many that such a naive approach performs well, but the method's success is linked to intrinsic features of the first principles energy landscape, such as its smoothness (a result of the quantum mechanical interactions between the atoms and electrons) and the relatively large number of large energy basins. In a smooth energy landscape the size of the basins correlates with their depth (deep basins occupying a large volume of configuration space), there is a natural bias in random sampling, plus relaxation, to the stable, low energy and relevant structures. In what follows we exploit our new approach for the automatic decomposition of known structures into minimum I fragments to accelerate the search for complex structures by restricting the regions of configuration space that must be explored.

Application to dense boron. We generate 3303 initial random structures based on packing four 14 atom D_{2h} modules derived from γ-B_{28} into unit cells with a randomly chosen shape, and the same density as γ-B_{28}. The units are not permitted to overlap each other, or be closer than 1.63 Å (the measured inter-icosahedral distance in a computed α-B structure), and are related to each other by symmetry. The symmetry is chosen at random from those space groups with four symmetry operators in the primitive cell. The random initial structures are then relaxed to nearby local energy minima (see 'Methods' for computational details). Four of the initial structures relaxed to supercells of the Pnnm γ-B_{28} structure, and three of them relaxed to a previously unreported structure with space group Pbcn and 56 atoms in a unit cell. This structure has a density very close to that of γ-B_{28}, and is just 3 meV/atom less stable. In Fig. 7 it is shown that this near degeneracy persists over a wide range of pressures.

Our new 56 atom structure corresponds to a distorted hexagonal packing of boron icosahedra, whereas γ-B_{28}
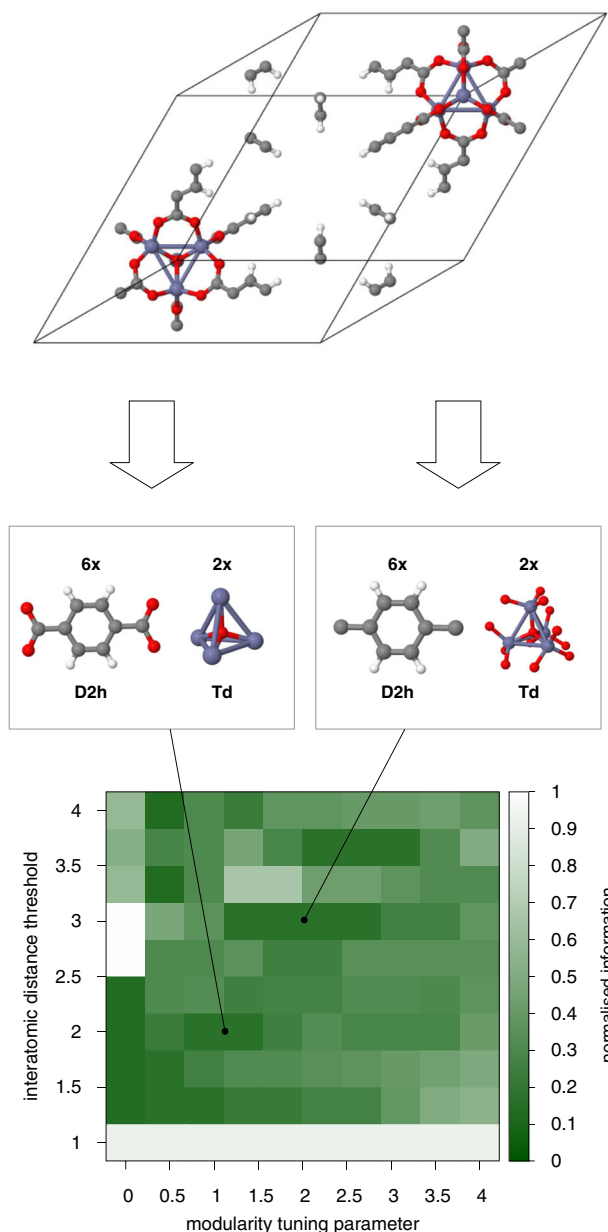
npj

Revealing and exploiting hierarchical material structure
SE Ahnert et al.

6

Fig. 6 The structure of the metal-organic framework MOF-5 (Zn_4O $(BDC)_3$, where BDC^{2-} is 1,4-benzenedicarboxylate) and its modular decompositions. This structure has 106 atoms per unit cell. The two lowest minima have very similar values for the normalised information, with $I = 0.163$ at $w_s = 2$ and $d^*_{eff} = 3$ and $I = 0.166$ at $w_s = 1$ and $d^*_{eff} = 2$. Unlike d^*, which is a physical distance threshold, the distance threshold parameter d^*_{eff} is dimensionless and expressed as the multiple of the summed atomic radii for a given pair of atoms (see text). Both decompositions consist of two modules, one of which appears six times and has D_{2h} symmetry, and the other one of which appears twice and has T_d symmetry. The modules for $I = 0.163$ are one with 16 atoms (8 C, 4 O, 4 H) and one with five atoms (4 Zn, 1 O)

corresponds closely to cubic packing. The small energy difference indicates that γ-B may be susceptible to polytypism or stacking disorder. The 3 meV/atom energy difference between the hexagonal and cubic polytypes is small compared to the 25 meV difference between cubic and hexagonal (carbon) diamond computed at the same level. The situation is very similar to that for $\alpha - B_{12}$, for which an alternatively packed structure of 24 atom

and space group Cmcm has been identified,[15] and further discussed.[57, 58]

DISCUSSION

In contrast to other applications of community-detection algorithms, in which a degree of ambiguity in the definition of communities is often tolerated, our application relies on the detection of robust and compressible modules of atoms with a maximum amount of symmetry. Small differences in the assignments of atoms to modules can have large effects on the information measure I if modular symmetries are established or broken as a result. For this reason the implementation of the modularity optimisation requires particular care in order to find maximally robust modular decompositions.

In the context of multi-species atomic structures, and particularly biological molecules, it can be valuable to consider weighted atomic networks, in which the edge weight scales with the interatomic distance. The simplest choice is a linear one:

$$w_{ij} = K\left(1 - \frac{d_{ij}}{d^*}\right)$$

for $d_{ij} < d^*$ and $w_{ij} = 0$ otherwise, where d_{ij} is the distance between atoms i and j, and K is an arbitrary constant. Increasing all edge weights by a constant factor leaves the modularity unchanged, so in practice this is chosen to ensure numerical precision. This choice of edge weighting reflects the fact that equilibrium bond lengths scale with the equilibrium bond energies.

Our approach can be applied in the context of proteins, similar to,[29–31] where the vastness of configuration space has traditionally also been a difficult barrier to overcome. The lack of symmetry in biological molecules mean that the information required to describe the structure is less useful than in crystal structures. Other measurements, such as the stability of the module number, can replace the information measurement as a criterion for assessing the quality of a modular decomposition in biomolecules. While methods for rigidity analysis in proteins already exist, such as the FIRST algorithm,[59] the tuning parameter in our method allows for the detection of rigid clusters on a variety of length scales, making it a complementary approach. Like FIRST, our method could inform coarse-grained multi-scale molecular dynamics methods such as FRODA,[60] supplying the rigid units described as 'ghost templates' in FRODA, and lead to more efficient computational models of protein dynamics.

In the context of complex crystal structures this approach has numerous potential applications. It first of all suggests an automatic coarse-graining and thereby provides an intuitive simplification and visual aid. The discovery of modules also facilitates the accelerated exploration of configuration space, particularly in the context of random structure search. This has been demonstrated by the new structure of boron closely related to γ-B_{28} that we find using a module-based search. We believe that these results show the potential of atomic network analysis as a tool for materials discovery. Our algorithm is evidently fast enough for these purposes—the run time for calculating the metal-organic framework MOF-5 decomposition (106 atoms per unit cell) was 2.3 s on a 3.1 GHz Intel Core i7 processor.

METHODS

Revealing and exploiting hierarchical material structure
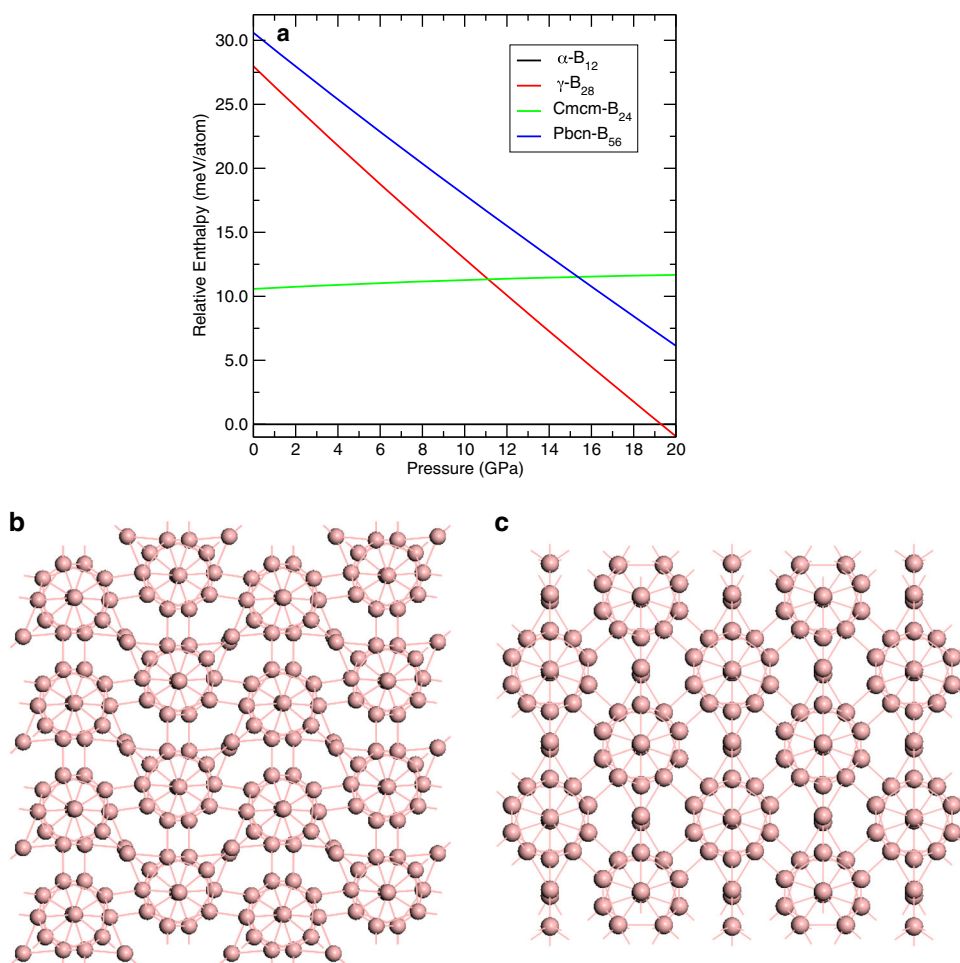SE Ahnert et al.

npj

7

Fig. 7 **a** Relative enthalpy of boron structures up to 20 GPa. **b** The new Pnnm-B_{56} structure is practically degenerate in energy with **c** the known γ-B_{28} structure

Data availability

The data associated with this manuscript is made available in ref. 63.

AUTHOR CONTRIBUTIONS

SEA and CJP designed the study and methodology. SEA, CJP and WPG performed computations, generated data and wrote the manuscript.

ADDITIONAL INFORMATION

REFERENCES

1. Car, R. & Parrinello, M. Unified approach for molecular dynamics and density-functional theory. *Phys. Rev. Lett.* **55**, 2471 (1985).
2. Payne, M. C., Teter, M. P., Allan, D. C., Arias, T. & Joannopoulos, J. Iterative minimization techniques for ab initio total-energy calculations: molecular dynamics and conjugate gradients. *Rev. Mod. Phys.* **64**, 1045 (1992).
3. Clark, S. J. et al. First principles methods using castep. Zeitschrift für Kristallographie-Crystalline. *Materials* **220**, 567–570 (2005).
4. Giannozzi, P. et al. Quantum espresso: a modular and open-source software project for quantum simulations of materials. *J. Phys.: Condens. Matter* **21**, 395502 (2009).
5. Kresse, G. & Furthmüller, J. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Phys. Rev. B* **54**, 11169 (1996).
6. Lejaeghere, K. et al. Reproducibility in density functional theory calculations of solids. *Science* **351**, aad3000 (2016).
7. Hohenberg, P. & Kohn, W. Inhomogeneous electron gas. *Phys. Rev.* **136**, B864 (1964).
8. Kohn, W. & Sham, L. J. Self-consistent equations including exchange and correlation effects. *Phys. Rev.* **140**, A1133 (1965).
9. Parr, R. G. & Yang, W. Density-functional theory of the electronic structure of molecules. *Annu. Rev. Phys. Chem.* **46**, 701–728 (1995).
10. Hasnip, P. J. et al. Density functional theory in the solid state. *Philos. Trans. R. Soc. A* **372**, 20130270 (2014).
11. Zhao, Y. & Truhlar, D. G. Density functionals with broad applicability in chemistry. *Acc. Chem. Res.* **41**, 157–167 (2008).
12. Cole, D., Skylaris, C.-K., Rajendra, E., Venkitaraman, A. & Payne, M. Protein-protein interactions from linear-scaling first-principles quantum-mechanical calculations. *Europhys. Lett.* **91**, 37004 (2010).
13. Jain, A., Shin, Y. & Persson, K. A. Computational predictions of energy materials using density functional theory. *Nat. Rev. Mater.* **1**, 15004 (2016).

npj

Revealing and exploiting hierarchical material structure
SE Ahnert et al.

8

14. Pickard, C. J. & Needs, R. J. High-pressure phases of silane. *Phys. Rev. Lett.* **97**, 045504 (2006).
15. Pickard, C. J. & Needs, R. J. Ab initio random structure searching. *J. Phys.: Condens. Matter* **23**, 053201 (2011).
16. Oganov, A. R. & Glass, C. W. Crystal structure prediction using ab initio evolutionary techniques: Principles and applications. *J. Chem. Phys.* **124**, 244704 (2006).
17. Wang, Y., Lv, J., Zhu, L. & Ma, Y. Calypso: a method for crystal structure prediction. *Comput. Phys. Commun.* **183**, 2063–2070 (2012).
18. Needs, R. J. & Pickard, C. J. Perspective: role of structure prediction in materials discovery and design. *APL Mater.* **4**, 053210 (2016).
19. Watts, D. & Strogatz, S. Collective dynamics of 'small-world' networks. *Nature.* **393**, 440–442 (1998).
20. Barabasi, A. & Albert, R. Emergence of scaling in random networks. *Science* **286**, 509–512 (1999).
21. Albert, R. & Barabasi, A. Statistical mechanics of complex networks. *Rev. Mod. Phys.* **74**, 47–97 (2002).
22. Newman, M. *Networks: An Introduction* (Oxford University Press, 2010).
23. Palla, G., Derényi, I., Farkas, I. & Vicsek, T. Uncovering the overlapping community structure of complex networks in nature and society. *Nature.* **435**, 814–818 (2005).
24. Newman, M. & Girvan, M. Finding and evaluating community structure in networks. *Phys. Rev. E* **69**, 026113 (2004).
25. Colizza, V., Flammini, A., Serrano, M. A. & Vespignani, A. Detecting rich-club ordering in complex networks. *Nat. Phys.* **2**, 110–115 (2006).
26. Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech.-Theory Exp.* **2008**, P10008 (2008).
27. Arenas, A., Fernández, A. & Gómez, S. Analysis of the structure of complex networks at different resolution levels. *New. J. Phys.* **10**, 053039 (2008).
28. Ahn, Y.-Y., Bagrow, J. P. & Lehmann, S. Link communities reveal multiscale complexity in networks. *Nature* **466**, 761–U11 (2010).
29. Delvenne, J. C., Yaliraki, S. N., Barahona, M. & Newman, M. Stability of graph communities across time scales. *Proc. Natl Acad. Sci. USA* **107**, 12755–12760 (2010).
30. Delmotte, A., Tate, E. W., Yaliraki, S. N. & Barahona, M. Protein multi-scale organization through graph partitioning and robustness analysis: application to the myosin–myosin light chain interaction. *Phys. Biol.* **8**, 055010 (2011).
31. Amor, B., Yaliraki, S. N., Woscholski, R. & Barahona, M. Uncovering allosteric pathways in caspase-1 using Markov transient analysis and multiscale community detection. *Mol. Biosyst.* **10**, 2247 (2014).
32. Fortunato, S. & Barthelemy, M. Resolution limit in community detection. *Proc. Natl Acad. Sci. USA* **104**, 36–41 (2007).
33. Pauling, L. The principles determining the structure of complex ionic crystals. *J. Am. Chem. Soc.* **51**, 1010–1026 (1929).
34. Kolmogorov, A. N. Three approaches to the quantitative definition of information'. *Probl. Inf. Transm.* **1**, 1–7 (1965).
35. Chaitin, G. J. *Algorithmic Information Theory* (Wiley Online Library, 1982).
36. Ahnert, S., Johnston, I., Fink, T., Doye, J. & Louis, A. Self-assembly, modularity, and physical complexity. *Phys. Rev. E* **82**, 026117 (2010).
37. Good, B. H., de Montjoye, Y.-A. & Clauset, A. Performance of modularity maximization in practical contexts. *Phys. Rev. E* **81**, 046106. doi:10.1103/PhysRevE.81.046106 (2010).
38. Leary, R. H. Global optimization on funneling landscapes. *Journal of Global Optimization* **18**, 367–383 (2000).
39. Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech.-Theory. E* **10**, P10008. http://iopscience.iop.org/article/10.1088/1742-5468/2008/10/P10008 (2008).
40. Massen, C. & Doye, J. Identifying communities within energy landscapes. *Phys. Rev. E* **71**, 046101. https://journals.aps.org/pre/abstract/10.1103/PhysRevE.71.046101 (2005).
41. Albert, B. & Hillebrecht, H. Boron: elementary challenge for experimenters and theoreticians. *Angewandte Chemie International Edition* **48**, 8640–8668 (2009).
42. Decker, B. & Kasper, J. The crystal structure of a simple rhombohedral form of boron. *Acta. Crystallogr.* **12**, 503–506 (1959).
43. Talley, C. P., La Placa, S. & Post, B. A new polymorph of boron. *Acta. Crystallogr.* **13**, 271–272 (1960).
44. Geist, D., Kloss, R. & Follner, H. Verfeinerung des β-rhomboedrischen bors. *Acta. Crystallogr. B.* **26**, 1800–1802 (1970).
45. Wentorf, R. H. Boron: Another form. *Science* **147**, 49–50 (1965).
46. Oganov, A. R. et al. Ionic high-pressure form of elemental boron. *Nature.* **457**, 863–867 (2009).
47. Zarechnaya, E. Y. et al. Superhard semiconducting optically transparent high pressure phase of boron. *Phys. Rev. Lett.* **102**, 185501 (2009).
48. Bachhuber, F. et al. The extended stability range of phosphorus allotropes. *Angewandte Chemie International Edition* **53**, 11629–11633 (2014).
49. Liu, H. et al. Phosphorene: An unexplored 2d semiconductor with a high hole mobility. *ACS nano* **8**, 4033–4041 (2014).
50. Schusteritsch, G., Uhrin, M. & Pickard, C. J. Single-layered hittorfs phosphorus: a wide-bandgap high mobility 2d material. *Nano. Lett.* **16**, 2975–2980 (2016).
51. Ruck, M. et al. Faserförmiger roter phosphor. *Angew. Chem.* **117**, 7788–7792 (2005).
52. Li, H., Eddaoudi, M., O'Keeffe, M. & Yaghi, O. M. Design and synthesis of an exceptionally stable and highly porous metal-organic framework. *Nature* **402**, 276–279 (1999).
53. Kitagawa, S. et al. Metal–organic frameworks (mofs). *Chem. Soc. Rev.* **43**, 5415–5418 (2014).
54. Pickard, C. J. & Needs, R. J. Structure of phase iii of solid hydrogen. *Nat. Phys.* **3**, 473–476 (2007).
55. Pickard, C. J. & Needs, R. Aluminium at terapascal pressures. *Nat. Mater.* **9**, 624–627 (2010).
56. Schusteritsch, G. & Pickard, C. J. Predicting interface structures: from SrTiO$_3$ to graphene. *Phys. Rev. B* **90**, 035424 (2014).
57. He, C. & Zhong, J. Structures, stability, mechanical and electronic properties of α-boron and α*-boron. *AIP Adv.* **3**, 042138 (2013).
58. Zhu, Q., Oganov, A. R., Lyakhov, A. O. & Yu, X. Generalized evolutionary metadynamics for sampling the energy landscapes and its applications. *Phys. Rev. B* **92**, 024106 (2015).
59. Jacobs, D. J., Rader, A. J., Kuhn, L. A. & Thorpe, M. F. Protein flexibility predictions using graph theory. *Proteins* **44**, 150–165 (2001).
60. Wells, S., Menor, S., Hespenheide, B. & Thorpe, M. F. Constrained geometric simulation of diffusive motion in proteins. *Phys. Biol.* **2**, S127–S136 (2005).
61. Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.* **77**, 3865 (1996).
62. Vanderbilt, D. Soft self-consistent pseudopotentials in a generalized eigenvalue formalism. *Phys. Rev. B* **41**, 7892 (1990).
63. Pickard, C. Complex atomic networks, Ahnert, Grant, and Pickard, 2017 URL https://figshare.com/articles/Complex_atomic_networks_Ahnert_Grant_and_Pickard_2017/4780456 (2017).