Faraday Discussions

Cite this: *Faraday Discuss.*, 2018, **211**, 643

PAPER

View Article Online
View Journal | View Issue

Crystal structure prediction: reflections on present status and challenges

Artem R. Oganov [abc]

Received 30th August 2018, Accepted 30th August 2018
DOI: 10.1039/c8fd90033g

Long thought to be impossible, crystal structure prediction (CSP) is a thriving field today, with many important discoveries in fields as diverse as computational materials discovery, drug design, high-pressure chemistry and mineralogy of the Earth's and planetary interiors. However, major challenges remain, warranting more research. In these Concluding Remarks, I try to summarize my personal view of the enormous progress made in the field of CSP and the open questions and challenges that keep this field more exciting than ever.

Introduction

One can say that the first crystal structure predicted by a human was the structure of ice, first drawn by Johannes Kepler in his treatise "On the six-cornered snowflake" in 1611 (Fig. 1) – this was an intuitive model, which was meant to explain the hexagonal symmetry of snowflakes. Kepler's model turned out not to be the correct structure of ice, and corresponds to what we call today the hexagonal close packing – such a structure is adopted by Be, Mg and Cd under normal conditions, and Fe at high pressures. The first correct structure prediction, albeit for a molecule rather than a crystal, was made in 1857 by August Kekulé: in his famous dream by the fireplace he saw a swirling snake biting its tail, and this gave him a model of the cyclic structure of benzene. Then, in 1897 William Barlow created a model of the rocksalt structure – again, this model was just a fruit of intuition, but much later turned out to be correct. I could rate this as the first successful example of crystal structure prediction (CSP), but I won't, because the whole point of CSP is to find the structure based on the laws of physics, rather than human imagination or intuition. By the way, Barlow's result played an important role: W. L. Bragg knew this model, and his determination of the crystal structure of rocksalt can be described as a confirmation of Barlow's model.

[a]*Skolkovo Institute of Science and Technology, Skolkovo Innovation Center, 3 Nobel St., Moscow 143026, Russia. E-mail: a.oganov@skoltech.ru*

[b]*Moscow Institute of Physics and Technology, 9 Institutskiy Lane, Dolgoprudny City, Moscow Region, 141700, Russia*

[c]*International Center for Materials Discovery, Northwestern Polytechnical University, Xi'an, 710072, China*

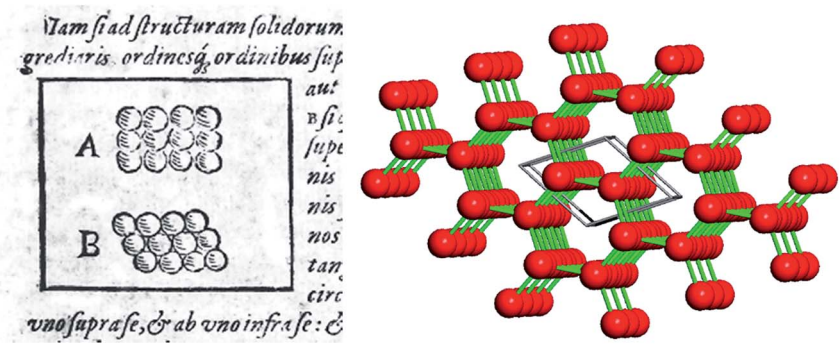

Fig. 1 Crystal structure of ice: (a) structure proposed by Kepler, (b) correct structure. In both drawings, each molecule of H_2O is represented as a sphere.

Since the pioneering works of W. L. Bragg and his father W. H. Bragg,[1,2] where the first crystal structures were experimentally determined using X-ray diffraction, an enormous number of crystal structures were solved, and this knowledge formed much of the foundation of modern solid-state physics, chemistry, biochemistry, and materials science. It all started with the simplest structures (historically, the first structures solved by Braggs were diamond, zincblende and rocksalt), and as time went on, experimental techniques improved, and nowadays it is possible to solve structures containing thousands of atoms in the unit cell. Still, challenges remain – for example, and while structure solution by single-crystal X-ray diffraction is routine, solving structures from powder data is still largely an art.

CSP, on the other hand, has only recently become possible. Compared to the contemporaneous triumphant successes of experimentalists, theorists looked hopeless for a long time. This situation can be understood: predicting a stable crystal structure (*i.e.* the arrangement of atoms with the lowest possible energy) encounters two key problems:

(1) The "Ranking problem", *i.e.* the reliable calculation of relative structural energies. This turned out to be highly non-trivial, because the energy differences between different polymorphs are often very small and correctly ranking the structures by energy has long been (and to some extent still is) a challenge.

(2) The "Search problem": the number of possible arrangements of atoms in space is astronomically large: for a unit cell with N atoms, the number of possible structures is $C \sim \exp(ad)$, where d is the number of degrees of freedom and a is some system-specific constant. If the positions of all N atoms are uncorrelated, then $d = 3N + 3$, resulting in a very high-dimensional problem; roughly, $C \sim 10^N$ (ref. 3). The problem can be greatly simplified if every generated structure is relaxed, *i.e.* brought to a local energy minimum, as this dramatically lowers the number of configurations, but the scaling of the problem is still exponential, *i.e.* it is NP-hard and very challenging.

Both problems have been largely solved, though as I will detail below, the story is far from being over. These successes are making industry more interested, especially in the pharmaceutical world: Lilly, Avant-garde, and XtalPi are examples of companies specializing in CSP for the pharmaceutical industry, and such

companies as Novartis actively participate in this research field. CSP is also rapidly revealing its potential for computational materials discovery,[4] and we have sold a number of (expensive) commercial licenses for our USPEX code (http://uspex-team.org) to companies.

Historically, the inorganic and organic crystal structure prediction communities evolved in parallel. In the organic community, the ranking problem always attracted more attention and the focus has been on the development of a sufficiently accurate forcefield; since forcefield calculations are rather cheap, it was thought that even an enumerative search through all local minima was doable. This focus on the ranking problem is partly justified by the fact that organic crystals often have $\leq$1 molecule in the asymmetric unit and the vast majority crystallize in just a dozen space groups. In the inorganic community, the search problem attracted more attention: first, the number of degrees of freedom is typically very large (and equal to $3N + 3$, *i.e.* for a medium-complexity crystal with 30 atoms per cell one has 93 degrees of freedom), making the search problem very acute. Second, it is quite clear that forcefields cannot describe with sufficient accuracy all the types of chemical bonding present in inorganic substances and one must resort to *ab initio* calculations. Third, given the high computational cost of *ab initio* calculations, one is forced to reduce their number, *i.e.* solve the search problem as efficiently (and reliably) as possible.

This historical separation of the two communities is reflected by the fact that traditional blind tests organized by the Cambridge Crystallographic Data Centre (CCDC) were and still are entirely focused on organic crystals and until the 2010s only researchers from the organic CSP community participated in it. One can notice convergence of the two communities: in the 6th blind test,[5] 3 out of 25 participating teams were "inorganic"; a sad note is that they did not score much success (in fact, ours was the only "inorganic" team that scored any successes – getting the right structure in 2 out of 5 cases, which is better than most "organic" teams). It appears that when the conformation of the molecule is known or can be correctly guessed before CSP, the crystal structure can usually be predicted easily. For example, the structures of the newly synthesized polymorphs of resorcinol[6] and coumarin[7] were easily predicted by us and gave a perfect match to experimental X-ray diffraction patterns. The presence of many flexible torsion angles within the molecule, or the possibility of the formation of zwitterions make the problem much harder, both in search and in ranking. In contrast to six organic structure prediction tests, there had been only one such test for inorganic structures,[8] and I think this has to be continued and done regularly too, and "organic" teams will need to be invited (I joking say that I am vengefully looking forward to seeing how they fare on our territory). The gradual convergence of the "organic" and "inorganic" CSP groups is very encouraging.

To sum up, CSP consists of two problems: search and ranking. Ranking is usually done at zero Kelvin, with either forcefields or quantum-mechanical approaches. Forcefields come in different flavors – pair potentials *vs.* many-body potentials, rigid-ion *vs.* shell model potentials, rigid-molecule *vs.* flexible-molecule, point charge *vs.* multipole, parametric *vs.* machine learning forcefields. Quantum-mechanical methods are usually done at the level of density functional theory (DFT) – for molecular crystals, with some dispersion correction needed to correctly describe van der Waals bonding, the most advanced approaches to which seem to be the SCAN meta-GGA functional[9] with the rVV10

van der Waals functional,[10] and various exchange-correlation functionals (*e.g.*, PBE[11] or PBE0 (ref. 12)) with a many-body dispersion potential.[13] It was a big surprise for me when Tkatchenko *et al.*[13] showed that many-body van der Waals effects are essential for molecular crystals, and proposed an elegant way to account for them. Between forcefields and DFT-based calculations are semi-empirical methods – these crude quantum-mechanical methods are much cheaper than DFT, but their accuracy has not convinced the CSP community so far. The new study by Sally Price's group (Iuzzolino *et al.*, DOI: 10.1039/c8fd00010g) utilizes a semiempirical tight-binding method for the relatively cheap yet reliable pre-relaxation of crystal structures made of flexible molecules, and perhaps this is a right niche for such methods.

To go beyond zero-Kelvin CSP, it is necessary to compute the free energy, the most difficult part of which is the entropy. The vibrational entropy can be computed quasiharmonically, or at various levels of accounting for anharmonicity. A full account of anharmonicity is possible with molecular dynamics and Monte Carlo methods, but the calculation of the free energy in these methods in not trivial,[14] and for CSP one needs a fast and automatic way. Even that will only take care of the vibrational part of the free energy, and the treatment of other contributions to the entropy (especially configurational entropy, and also magnetic entropy) is a challenge. While the calculation of the free energy at finite temperature (*i.e.* the ranking problem) is much heavier than at zero Kelvin, the search problem should become (much?) simpler as the temperature increases, because shallow energy minima merge into one broader free energy minimum. Indeed, while at 0 K we have an astronomically large number of local minima; near the melting point there are only on the order of one local minima (one or a few crystalline states, and a few or usually one liquid state). I expect the number C of local free energy minima to decrease exponentially with temperature T:

$$C = C_0 \exp\left(\beta \frac{T^* - T}{T^*}\right), \quad (1)$$

where C_0 is the number of local minima at zero Kelvin, β is a constant, and T^* is a characteristic temperature (higher than the melting temperature), at which only one free energy minimum exists.

The search problem, being NP-hard, is bound to have an upper limit of tractable complexity, which today stands at $d < 400$–500, *i.e.* we can deal with non-molecular crystals containing not more than ~150 atoms in the primitive cell. However, for molecular crystals the number of degrees of freedom is much lower than $3N + 3$, *i.e.* if we impose a molecular geometry and search for a minimum-energy packing of prespecified molecules, the problem becomes much simpler than if all atoms were treated independently. Partial or complete freezing of the molecular geometry is needed not only for computational convenience, but also because most organic compounds are metastable, *i.e.* the global minimum will not have the organic molecules we are interested in, but will contain a mixture of such small molecules as H_2O, CH_4, CO_2, NH_3, *etc.* Here, only constrained, as opposed to unconstrained, global optimization is meaningful – *i.e.* searching for the most favorable packing of the molecules of interest.

An interesting approach, nested sampling, was used by Livia Partay and Gabor Csanyi,[15,16] allowing one to compute the partition function Z in a realistic, though long, time:

$$Z = \sum_i e^{-E_i/k_B T}, \quad (2)$$

where E_i are the energies of all the quantum states and k_B is the Boltzmann constant. Taking advantage of the well-known formula for the Helmholtz free energy

$$F = -k_B \ln Z, \quad (3)$$

it is possible to predict phase transitions (they are indicated by a peak of the heat capacity), the stable phases (including even the liquid state) and the free energy as a function of temperature at a given pressure or volume.

Today we develop better algorithms to deal with the NP-hard search problem and make more complex systems tractable – but quantum computers might be able to deal with the search problem efficiently in the future, bringing it to polynomial or even linear complexity. In what follows, I will focus more on different aspects of the search problem, though from time to time I will detour into ranking-related issues and issues beyond both these problems. This is by no means a review, but a personal reflection – hence, there are many references to the works of my laboratory, and my personal thoughts.

Probing the limits of complexity

The "exponential wall", which one encounters in the search problem, is one of the faces of the phenomenon known in mathematics as the "curse of dimensionality". A simple strategy to reduce the curse of dimensionality is to reduce, whenever possible, the dimensionality of the problem.

In the FUSE method (Collins *et al.*, DOI: 10.1039/c8fd00045j) one assembles structures not from atoms, but from atomic pairs or triples – in ionic crystals (for which FUSE is designed) one expects cations to be surrounded by anions, rather than by other cations – hence, the pairs can be cation–anion. This simple trick allows rather complex systems to be studied, and Collins *et al.* (DOI: 10.1039/c8fd00045j) have reported a successful exploration of the challenging quasiternary SrO–TiO_2–Y_2O_3 system (Fig. 2). Another method for generating random structures with alternating cations and anions was proposed by Stevanovic.[17] Such random structures can then, for example, be used as an initial population in an evolutionary search.

For ionic systems, the permissible stoichiometries must satisfy charge balance – this restriction on allowed compositions is also an effective way to reduce the dimensionality of the problem. To use it, one has to know the oxidation states of all the atoms – and Davies *et al.* (DOI: 10.1039/c8fd00032h) have proposed a data mining approach to predict the likely oxidation states of all elements when present together with other elements. Note, however, that the charge balance rule is often violated under pressure: *e.g.*, in the Na–Cl system, such compounds as Na_3Cl, Na_2Cl, $NaCl_3$ and $NaCl_7$ become stable.[18] Even at ambient conditions, in the Ca–Al–O system, a very interesting material with metallic conductivity – the electride $Ca_{12}Al_{14}O_{32}$ – exists,[19] and violates charge balance. Nature, as usual, is smarter than our rules.

There are ways to avoid the curse of dimensionality – for example, the metadynamics method for CSP[20,21] (or its other variant, evolutionary metadynamics[22,23])

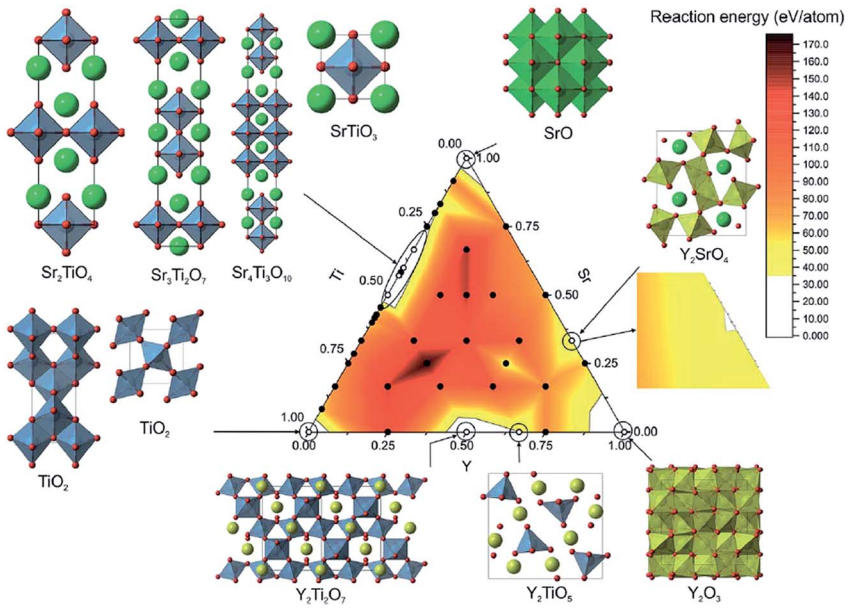

Fig. 2 Results of the FUSE method for the SrO–TiO_2–Y_2O_3 system (Collins *et al.*, DOI: 10.1039/c8fd00045j), showing numerous correctly predicted ground states. Reproduced from Collins *et al.*, DOI: 10.1039/c8fd00045j with permission from the Royal Society of Chemistry.

reduces the number of degrees of freedom from $3N + 3$ to a mere 6 – searching for the stable crystal structure in the space of 6 lattice vectors matrix components. In this case, the exponential scaling disappears, and large systems can be dealt with efficiently. For example, a complex structure of the high-pressure phase of $Li_{15}Si_4$ (the starting low-pressure structure has 152 atoms in the unit cell) was predicted at a very modest computational cost using evolutionary metadynamics, and confirmed by powder X-ray diffraction.[24] One of the bonuses of metadynamics is that it can be naturally used at finite temperatures. Piaggi and Parrinello have developed a new version of metadynamics, which was used for simulating *e.g.* the crystallization and solid–solid phase transitions of urea.[25] In their approach, they used a vibrational entropy-like descriptor as an order parameter. Metadynamics is unique in that it has no exponential scaling and allows large systems to be treated. However, there is a price to pay: the success of CSP depends on the initial crystal structure and on the appropriateness of the reduced-dimensionality description (*e.g.*, the 6-dimensional order parameter composed of lattice vectors matrix components).

Even when we have a good algorithm to solve the search problem, structure relaxation and energy evaluations may be too expensive at the *ab initio* level. Here, the greatest promise is given by machine learning forcefields (MLF). Trained on *ab initio* energies, stresses and forces on atoms for a number of configurations, MLF can predict energies, stresses and forces in a new, hitherto unseen, configuration. This works only as an interpolation, *i.e.* the new configurations should be sufficiently similar to the ones used for training the forcefield, and if a very different configuration is found, the errors can be very large, and such

a configuration should either be rejected, or calculated at the *ab initio* level and then used for re-training the MLF. In this issue, Deringer *et al.* (DOI: 10.1039/c8fd00034d) showed results on phosphorus, obtained with a combination of random sampling and MLF. Their work shows both the current level of accuracy of such methods (*e.g.*, training a MLF accurate on average to less than 100 meV per atom is cheap) and problems involved (*e.g.*, Hittorf's phosphorus with 84 atoms per cell could not be found in their random sampling search, even after imposing constraints, and errors of the MLF for this metastable phase are unusually large, above 100 meV per atom). Such studies, openly showing the difficulties, are very important for further progress. Recently, we[26] combined a MLF with our evolutionary method USPEX to search for stable and low-energy metastable structures of boron. With a MLF showing a mean average error of 37 meV per atom, we have successfully found α-boron (12 atoms per cell), γ-boron (28 atoms per cell), several versions of tetragonal T52-boron (52 atoms per cell), and even the extremely complex and disordered β-boron (for which we found ~50 energetically nearly degenerate ordered approximants with 105–108 atoms per cell, see Fig. 3a). In addition to these known phases, we predicted a new low-energy cubic *Im*3 phase (with 54 atoms per cell, see Fig. 3b), which is only 29 meV per atom higher in energy than α-boron and energetically degenerate with the experimentally known T52-boron. The use of a MLF has allowed a speedup of 100–1000 times.

Machine learning approaches, such as neural networks, are traditionally viewed as black boxes that can give the right numbers, but cannot give insight. However, there are ways to obtain insight, too. For example, describing the pair potential by a very flexible inverse-power series with fitted parameters and the many-body potential by a neural network, we[27] obtained an accurate representation of the energies in solid He, Xe and Al, and recovered a Lennard-Jones-like He–He pair potential, while for the Xe–Xe potential there were visible deviations from such shape (probably indicative of significant many-body effects renormalizing the effective pair potential), and for the Al–Al pair potential we obtained a density-dependent oscillating pair potential – which is exactly what one would expect for a metal. In all these pair potentials the minimum corresponded precisely to the equilibrium bond distance in the crystal (or a sum of van der Waals or metallic radii). I think it is a matter of time until more complex insight, currently accessible only to humans – *e.g.*, the derivation of Pauling's rules – becomes derivable from machine learning.

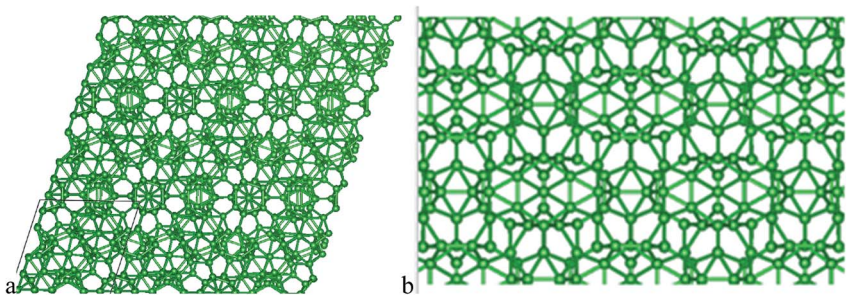

Fig. 3 (a) β-boron approximant with 106 atoms per cell, found in ref. 26. (b) *Im*3 structure predicted in ref. 26.

Though the current progress is impressive, many challenges remain. I don't think anyone can handle cases as complex as the Samson phase, β-Mg_2Al_3: its cubic structure has 1168 atoms in the conventional unit cell[28] and is, I think, too large and complex. This is just one example of a tricky complex metallic alloy, and the worst example would be quasicrystals. For quasicrystals, lacking periodicity, the energy cannot be computed by existing approaches, and one has to use periodic approximants – and, in view of their structural complexity, this should be challenging (though no one has really tried). Another limitation is magnetism: so far, CSP studies have largely avoided magnetic materials, and even when dealing with them, were confined to collinear magnetism. Can any of the existing methods predict the structure of α-Mn with non-collinear magnetism and containing 54 atoms in the unit cell (belonging to four types, one of which is non-magnetic[29])? I doubt it – though with a little bit of method development and at a high computational cost (to deal with a difficult search problem and facing technical issue of having to resolve small energy differences between different magnetic states) this should become possible. Finally, the prediction of protein structures remains an open challenge – in addition to their extreme structural complexity, we are not even sure whether the structure they adopt is thermodynamically controlled (see below).

When there is more than one solution, and searching for useful materials

Finding the stable crystal structure for a given chemical composition, mathematically formulated as global (free) energy minimization, is finally, though with certain important caveats, a solved problem. The solution of this problem is one crystal structure – the global minimum.

There are at least three important problems where the solution is not a single structure, but a set of structures or a set of materials. These problems, which can also be considered as solved, letting us do a lot more, are described below:

(I) Variable-composition structure prediction, searching for all stable compounds (and their crystal structures) formed by given elements. Let me take three simple examples. The Fe–S system has two stable iron sulfides, FeS and FeS_2. The H–O system has two well-known compounds, H_2O and H_2O_2, but the former is stable and the latter is metastable. In the Na–Cl system, only one compound is known, NaCl. Stability or instability of different compounds can be conveniently determined by the convex hull (or Maxwell) construction, an example of which is shown in Fig. 4. This construction is a convenient representation of the free energies of all possible reactions in the system. All stable compounds (*i.e.* those that have a lower free energy than any isochemical assemblage) form a convex figure on the graph, the *y*-axis of which is the free energy of formation (normalized per atom) and *x*-axis of which is the composition. The example of the Mn–B system (Fig. 4) shows that all known stable manganese borides were predicted correctly (and for MnB_4 the theoretical crystal structure was subsequently confirmed by experiment), and a new compound MnB_3 was predicted and then synthesized.[30] Our latest calculations show that this new compound, MnB_3, has very interesting mechanical properties.[31] It is interesting that in such well-studied systems as Mn–B, even at ambient conditions new stable compounds keep being discovered. Three-, four- and more-component

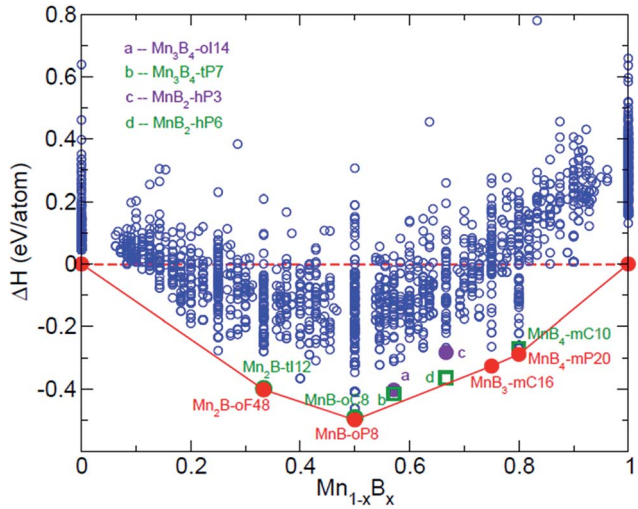

Fig. 4 Convex hull of the Mn–B system.[30] Solid red circles indicate stable phases, which are joined by lines forming the convex hull. Open green squares are experimentally known compounds Mn_3B_4 and MnB_2, which are calculated to be metastable. Reproduced from ref. 30 with permission from the PCCP Owner Societies.

systems can be studied the same way. This formalism can be applied to molecular co-crystals; in this case, one has to consider the free energy of formation normalized per molecule (rather than per atom). As there are many possible stoichiometries, doing CSP for all of them is probably unaffordable. One can limit consideration to the most common cases (*e.g.* AB, AB_2, AB_3, AB_4, and A_2B_3 stoichiometries), but this is risky as many stable compounds can be overlooked. Evolutionary codes USPEX (http://uspex-team.org) or GASP (http://gasp.mse.ufl.edu/) can efficiently sample the entire compositional space in an automatic fashion and without any assumptions about stoichiometries, and locate all stable compounds. The advantage of evolutionary algorithms is that different sampled stoichiometries compete, exchange structural information (greatly speeding up the search), and evolve, producing new stoichiometries. This is much more efficient than sampling all possible compositions independently.

(II) Prediction of materials optimal in two or more properties – *i.e.* multi-objective optimization, or Pareto optimization. Without realizing it, we use Pareto optimization in our own decision making every day. Which car to buy – a perfect and expensive one, or the cheapest but not so good? – here we optimize price and performance. Which school to choose for our children? – here, we optimize the quality of education, cost, and duration of commute. The set of optimal solutions, known as the Pareto front (or first Pareto front), is made of all non-dominated solutions, *i.e.* those solutions that cannot be beaten by any other solution on all target properties at the same time.

Materials scientists also usually want to find materials satisfying several criteria: for example, for many applications one needs materials with the highest possible hardness and fracture toughness. It is generally wise to have stability as one of the optimized properties: nobody wants to predict materials that are so unstable that they cannot be synthesized. For example, we searched[32] for

polymorphs of Bi_2Te_3 with the highest thermoelectric figure of merit *ZT*, and found that maximizing *ZT* alone leads to crazy structures, which are extremely unstable. Pareto optimization, searching simultaneously for high *ZT* and low energy, on the other hand, leads to reasonable results: it correctly finds the stable $R\bar{3}m$ polymorph of Bi_2Te_3 (a known good thermoelectric) and locates a number of low-energy polymorphs with very high *ZT*. This result shows not only the power of Pareto optimization, but also the promise of searching for new thermoelectrics: a 2–3-fold increase of *ZT*, compared to the current best materials, seems absolutely possible (Fig. 5).

(III) Searching for the highest-performance material(s) among all possible compounds of all elements. With 118 elements in Mendeleev's Periodic Table, we have 7021 binary, 273 937 ternary, and many more quaternary, quinary *etc.* systems, in each of which many stable and metastable compounds are possible. Doing so many CSP runs is impractical. Yet, it is possible to optimize the desired properties and predict the best-performing material(s) among all possible compounds. A coevolutionary method for doing this, called Mendelevian Search, was recently developed.[31] Here, the key is to arrange the chemical space in such a way that neighboring points are chemically similar (*e.g.* Na–Cl and K–Cl systems are similar) and have similar properties: then, the target property has a benign landscape that can be globally optimized. We[31] used this technique to simultaneously optimize, at 0 K, the hardness and stability, and magnetization and stability. In the former case we found diamond and lonsdaleite to be the hardest possible materials, but as these are metastable, a number of other materials also appeared on the Pareto front. In the latter case pure iron in its stable bcc structure was the only material on the first Pareto front, but a number of interesting magnetic materials appeared on higher-order Pareto fronts. Regarding the search

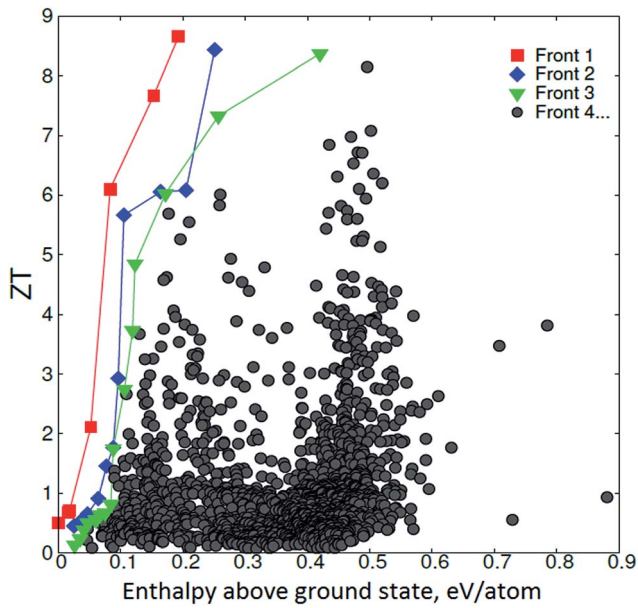

Fig. 5 Pareto optimization of thermoelectric figure of merit *ZT* and stability. Different Pareto fronts are shown. Reproduced from ref. 32 with permission from Elsevier.

for novel hard materials, two important stories spun off this calculation. In one of these studies,[33] we explored Cr–B, Cr–C, and Cr–N systems, found that the only superhard compound in these systems is CrB_4, and predicted a new stable chromium carbide, Cr_2C, which experimentalists should be able to synthesize. In the other study,[34] we explored the W–B system and predicted a new outstanding and hitherto unknown superhard compound WB_5 to be stable at normal conditions (Fig. 6); subsequently it was synthesized by V. Filonenko and V. Brazhkin and we have filed a patent on this material. This material has extraordinary hardness and fracture toughness, making it likely to replace WC in drilling equipment, machining tools and other applications.

Predicting an array of materials by just one calculation sounds like a dream, and although it is a reality, there are significant limitations. First, there are limitations related to the dimensionality of the problem. Variable-composition CSP can easily be done only for (pseudo)binary systems, already for ternary systems it is a challenge (but still doable, see ref. 35), while for quaternary systems it is probably already out of reach. Likewise, Pareto optimization works well for up to 3–4 simultaneously optimized properties.

Then, whilst some properties can easily be computed, others cannot. In general, physical properties (which characterize a particular energy minimum, and usually are response functions) are easier than chemical properties (which characterize a process of changing state). Biochemical properties, characterizing the complex interaction of a given molecule with receptors, enzymes, ion channels, *etc.*, are the hardest: for example, we do not have a general theory of toxicity, and are unlikely to have one any time soon (because there are so many mechanisms of toxicity). Today, we routinely optimize many physical properties, but I have not seen any direct optimizations of chemical or biochemical properties.

Yet, even for physical properties, we still have much to desire. Some properties are calculable, but at a great computational cost: *e.g.*, the critical temperature T_c of conventional (phonon-mediated) superconductors, or thermal conductivity. In principle, their optimization could be done, but it is just too expensive computationally. There are properties (*e.g.*, the viscosity of solids, or T_c of non-conventional superconductors) which cannot be computed at all at the

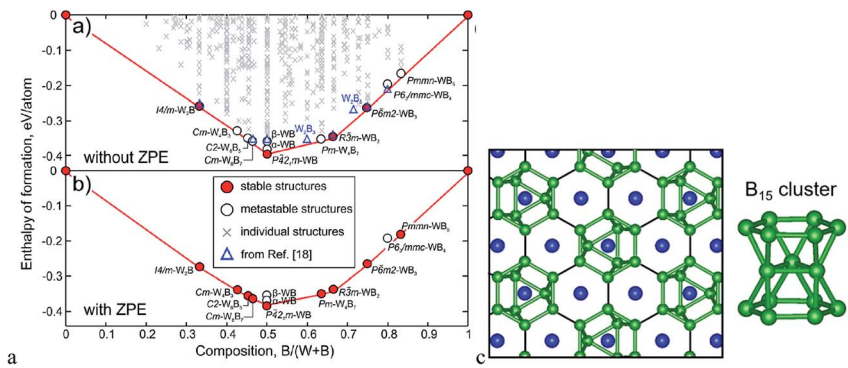

Fig. 6 (a and b) Convex hull of the W–B system with and without account for zero-point energy (ZPE), and (c) crystal structure of superhard WB_5. Interestingly, WB_5 is stabilized by ZPE. Adapted with permission from ref. 34. Copyright (2018) American Chemical Society.

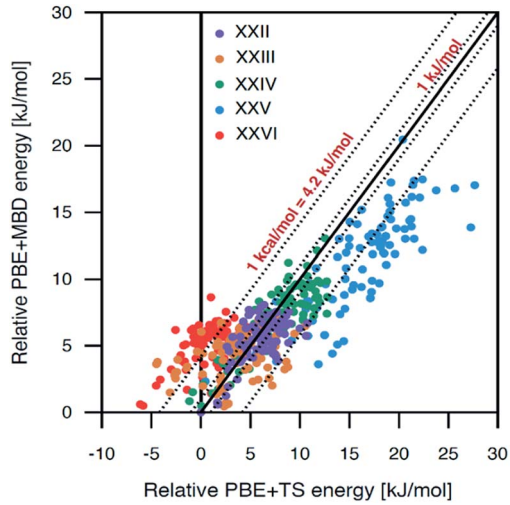

Fig. 7 Comparison of the energies obtained using two approximations for different structures of several blind test cases (PBE + TS and PBE + MBD). Reproduced from Hoja and Tkatchenko, DOI: 10.1039/c8fd00066b with permission from the Royal Society of Chemistry.

moment. In such cases, machine learning models of "difficult" properties may be the best way forward.

Even for the most basic property – the energy – for some systems the accuracy of existing approximations is sorely insufficient: this has been most clearly demonstrated by Tan *et al.* (DOI: 10.1039/c8fd00039e) for ROY (which stands for "red, orange, yellow", the colors of some of its numerous polymorphs – this is the compound with the largest number of coexisting polymorphs, ten). Comparing energies obtained by different modern approximations for different blind test molecules, Hoja and Tkatchenko (DOI: 10.1039/c8fd00066b) found a rather poor correlation, indicating that even for the energy the development of better approximations must continue (Fig. 7).

Discovering new chemical phenomena with CSP

A lot of new phenomena and surprising new compounds have been predicted with CSP (and many have already been experimentally verified). I will mention just a few such discoveries in the field of high-pressure chemistry to illustrate the ability of CSP to discover new knowledge.

In 2009, we published the prediction and experimental synthesis of a transparent insulating high-pressure allotrope of sodium.[36] This was at first surprising: first, because sodium is known as an "alkali metal", but at pressures above ~190 GPa, as it turns out, it is not a metal at all. Second, this was surprising because under pressure one expects closing of the band gap, rather than its opening. However, not long before this discovery, Ashcroft published[37] a model explaining this phenomenon (yes, the explanation appeared before the discovery!), based on the (unexpected to the classical chemist!) importance of core

electrons in strongly compressed matter. This example clearly shows a dramatic change of the physics and chemistry of the elements under pressure.

Another "simple" element, helium, long believed to be chemically inert – and for seemingly good reasons of having a closed-shell structure, record-high ionization potential and zero electron affinity – was predicted and experimentally confirmed to form a new stable compound Na_2He (ref. 38). This compound, stable from ~110 GPa to at least 1000 GPa, unlike other pressure-stabilized compounds of helium, is not an inclusion compound, because helium fundamentally changes its properties and makes it insulating. Both transparent sodium and Na_2He are electrides, *i.e.* ionic compounds where Na^+ cores are cations and interstitially localized electron pairs $(2e)^{2-}$ are anions. Another helium compound, Na_2HeO, was predicted to be stable at pressures from just 14 GPa (ref. 38), and then a number of other helium compounds were predicted as well.[39] Thanks to CSP, helium chemistry has suddenly become an active field.

It was also shown that completely new and unexpected stable stoichiometries become stable under pressure. For example, while only NaCl is stable at normal conditions (and we all thought that other stoichiometries are forbidden by the "charge neutrality rule"), nature outsmarted our rules, and at pressures amenable to experiment (starting from 22 GPa) new compounds emerge as stable – Na_3Cl, Na_2Cl, Na_3Cl_2, Na_4Cl_4, $NaCl_3$, and $NaCl_7$ (ref. 18 and 40), see Fig. 8. For two compounds, Na_3Cl and $NaCl_3$, experimental synthesis was attempted and led to their successful synthesis.[18] Most of these newly predicted sodium chlorides are metallic, despite the large electronegativity difference that we would expect to dictate ionic bonding and make Na^+Cl^- the only allowed compound. It remains to

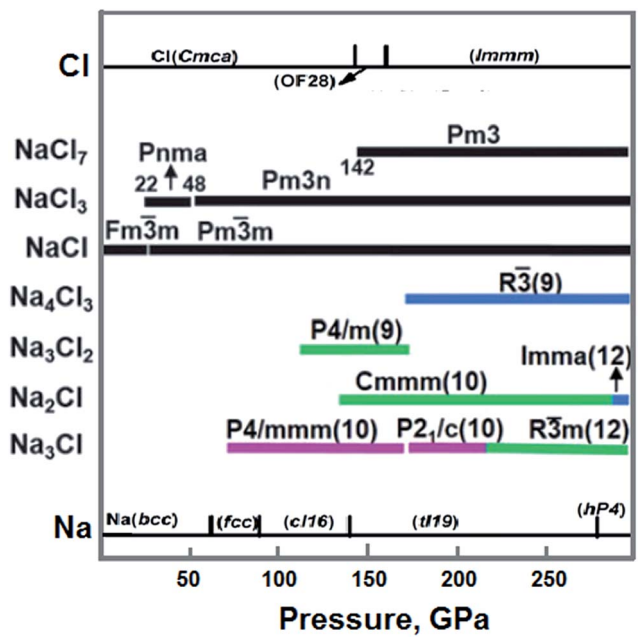

Fig. 8 Pressure-composition phase diagram of the Na–Cl system. Reproduced from ref. 40 with permission from the PCCP Owner Societies.

be explained why such compounds are formed, and whether any rules capable of explaining and predicting their stoichiometries can be formulated.

Such unexpected compounds become stable not only in the Na–Cl system, but probably in all chemical systems under sufficient pressure, and one wonders what role they may play in the interiors of the Earth and other planets, with their typically high pressures. For example, in the Si–O system, at planetary pressures two new stable oxides appear, SiO and SiO_3 (ref. 35), and in the Mg–O system, MgO_2, MgO_3 and Mg_3O_2 become stable.[35,41] Then, while we traditionally thought that ~10% of our planet is comprised by magnesium oxide (Mg,Fe)O – are we really sure that it's not (Mg,Fe)O_2? A new compound FeO_2 was recently predicted by CSP and then synthesized by Dave Mao's group,[42] and proposed to play an important role in the structure and evolution of our planet.

Such exotic compounds are bound to have exotic properties. It was predicted[43] that H_2S becomes unstable under pressure and breaks down, forming H_3S, and that compound was predicted to possess extremely high-T_c superconductivity with $T_c \sim 200$ K. One year later, an independent experimental work, confirming this startling prediction and $T_c = 203$ K, appeared.[44] Now there are several predictions of even higher-T_c superconductors, including even room-temperature superconductivity in YH_{10} and LaH_{10} (ref. 45 and 46). LaH_{10} has already been synthesized,[47] and it remains to be seen whether it is indeed a room-temperature superconductor.

Which phases are synthesizable?

So, here we are, a whole community (and a quickly growing one!), predicting numerous new phases, some stable or low-energy metastable (some already known and some not yet known from experiment), and some expected to have very interesting properties. How to synthesize these new phases? And generally, what makes a phase synthesizable? Is the number of phases that can, in principle, be synthesized small or large? Is it even a countable number?

Clearly, some phases are easily obtained, while others (sometimes lower in energy) are very difficult or seemingly even impossible to make. Why? Why is the extremely complex and pretty high in energy Hittorf's phosphorus (see, *e.g.*, Deringer *et al.*, DOI: 10.1039/c8fd00034d in this volume) synthesizable and other, lower-energy structures are seemingly not? Are there any selection rules (ideally based on crystal structure) that would tell us whether the phase is synthesizable at all? These are extremely important questions that have not been answered yet.

Directly related to these questions is the paradox of "disappearing polymorphs",[48] where instead of the easily obtained metastable polymorph (*e.g.* of a pharmaceutical compound), produced for a long time, the stable polymorph suddenly appeared and made synthesis of the metastable polymorph essentially impossible (thus, the old polymorph "disappeared"). This paradox occurs in materials where the ground state is difficult to reach, and was addressed by Neumann and van de Streek (DOI: 10.1039/c8fd00069g). Making the assumption that for rigid molecules the ground state is easily formed in ~100% cases, they estimated that for molecules with the complexity (primarily meaning conformational flexibility) typical of drugs, the ground state has been reached in 55–85% cases (probably because in the solution, from which crystals grew, one has molecules in the "wrong" conformation). Perhaps the percentage of cases where

the ground state is reached decreases exponentially with increasing the number of flexible angles. If so, in proteins, the iconic extreme case of conformational flexibility, the ground state is reached in about 0% cases. I do not know (and I think no one really does) whether the conformation of proteins in the living cell, in the solution or in the crystal is really the lowest-energy one. Anfinsen[49] hypothesized that structures of proteins are thermodynamically controlled, but the above considerations make us doubt it.

Synthesizability is related to crystal growth, and it seems to me that we poorly understand the process. It seems natural for crystals to grow by annexing not single atoms or molecules, but entire clusters of atoms or molecules. Using this idea, Anderson *et al.*[50] simulated the growth of zeolite crystals from their secondary building units in excellent agreement with experiment. But even this is a simplified picture of crystal growth. This is illustrated by the mystery of quasicrystal growth: upon hearing that quasicrystals, with structures described by his Penrose tilings, have been discovered, Sir Roger Penrose reportedly exclaimed that he thought this to be impossible because growing a perfect Penrose tiling by just following local rules of assembly is impossible, yet essentially perfect quasicrystals have been grown in experiment (see ref. 51). If we do not understand how quasicrystals grow, we also do not understand the growth of crystals: the mechanism should be similar, as atoms do not know whether they participate in a crystal or quasicrystal structure. The ease of assembling the structure should be an important criterion of its synthesizability. This means that there should exist structure-based selection rules determining the synthesizability of a structure.

Conclusions

In these Concluding Remarks I have covered a number of issues related to CSP and connecting it to other important fields – materials design, drug design, theoretical chemistry and chemical bonding, crystal growth, and planetary sciences. Essentially all fields of science related to the atomic structure of matter benefit from advances in CSP, and CSP can help in resolving many puzzles in these fields. CSP is a new powerful research tool that is already now making great contributions in all these fields, and more is to come.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

I thank the Russian Science Foundation (grant 16-13-10459) for generously supporting my research.

References

1 W. H. Bragg and W. L. Bragg, The structure of the diamond, *Nature*, 1913, **91**, 557.
2 W. L. Bragg, The structure of some crystals as indicated by their diffraction of X-rays, *Proc. R. Soc. London, Ser. A*, 1913, **89**, 248–277.

3 A. R. Oganov and C. W. Glass, Crystal structure prediction using *ab initio* evolutionary techniques: principles and applications, *J. Chem. Phys.*, 2006, **124**, 244704.
4 *Computational Materials Discovery*, ed. A. R. Oganov, G. Saleh and A. G. Kvashnin, Royal Society of Chemistry, 2018, ISBN: 978-1-78262-961-0.
5 A. M. Reilly, *et al.*, Report on the sixth blind test of organic crystal-structure prediction methods, *Acta Crystallogr., Sect. B: Struct. Sci., Cryst. Eng. Mater.*, 2016, **72**(4), 439–459.
6 Q. Zhu, *et al.*, Resorcinol Crystallization from the Melt: A New Ambient Phase and New "Riddles", *J. Am. Chem. Soc.*, 2016, **138**, 4881–4889.
7 A. G. Shtukenberg, *et al.*, Powder diffraction and crystal structure prediction identify four new coumarin polymorphs, *Chem. Sci.*, 2017, **8**, 4926–4940.
8 A. R. Oganov, J. C. Schön, M. Jansen, S. M. Woodley, W. W. Tipton and R. G. Hennig. First blind test of inorganic crystal structure prediction, in *Modern Methods of Crystal Structure Prediction*, ed. A. R. Oganov, Wiley-VCH, Berlin, 2010, pp. 223–231.
9 J. W. Sun, A. Ruzsinsszky and J. P. Perdew, Strongly constrained and appropriately normed semilocal density functional, *Phys. Rev. Lett.*, 2015, **115**, 036402.
10 H. W. Peng, Z. H. Yang, J. P. Perdew and J. W. Sun, Versatile van der Waals density functional based on a meta-generalized gradient approximation, *Phys. Rev. X*, 2016, **6**, 041005.
11 J. P. Perdew, K. Burke and M. Ernzerhof, Generalized gradient approximation made simple, *Phys. Rev. Lett.*, 1996, **77**, 3865–3868.
12 C. Adamo and V. Barone, Toward reliable density functional methods without adjustable parameters: The PBE0 model, *J. Chem. Phys.*, 1999, **110**, 6158–6170.
13 A. Tkatchenko, R. A. DiStasio, R. Car and M. Scheffler, Accurate and efficient method for many-body van der Waals interactions, *Phys. Rev. Lett.*, 2012, **108**, 236402.
14 M. P. Allen and D. J. Tildesley, *Computer Simulation of Liquids*, Oxford University Press, 2017.
15 L. B. Partay, A. P. Bartok and G. Csanyi, Efficient sampling of atomic configurational space, *J. Phys. Chem. B*, 2010, **114**, 10502–10512.
16 L. B. Partay, On the performance of interatomic potential models of iron: Comparison of the phase diagrams, *Comput. Mater. Sci.*, 2018, **149**, 153–157.
17 V. Stevanovic, Sampling polymorphs of ionic solids using random superlattices, *Phys. Rev. Lett.*, 2016, **116**, 075503.
18 W. W. Zhang, A. R. Oganov, A. F. Goncharov, Q. Zhu, S. E. Boulfelfel, A. O. Lyakhov, M. Somayazulu, V. B. Prakapenka and Z. Konopkova, Unexpected stoichiometries of stable sodium chlorides, *Science*, 2013, **342**, 1502–1505.
19 S. Matsuishi, Y. Toda, M. Miyakawa, K. Hayashi, T. Kamiya, M. Hirano, I. Tanaka and H. Hosono, High-density electron anions in a nanoporous single crystal: $[Ca_{24}Al_{28}O_{64}]^{(4+)}(4e^-)$, *Science*, 2003, **301**, 626–629.
20 R. Martonak, A. Laio and M. Parrinello, Predicting crystal structures: The Parrinello-Rahman method revisited, *Phys. Rev. Lett.*, 2003, **90**, 075503.
21 R. Martoňák, D. Donadio, A. R. Oganov and M. Parrinello, Crystal structure transformations in SiO_2 from classical and *ab initio* metadynamics, *Nat. Mater.*, 2006, **5**, 623–626.

22 Q. Zhu, A. R. Oganov and A. O. Lyakhov, Evolutionary metadynamics: a novel method to predict crystal structures, *CrystEngComm*, 2012, **14**, 3596–3601.
23 Q. Zhu, A. R. Oganov, A. O. Lyakhov and X. X. Yu, Generalized evolutionary metadynamics for sampling energy landscapes and its applications, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 2015, **92**, 024106.
24 Z. D. Zeng, Q. F. Zeng, N. Liu, A. R. Oganov, Q. S. Zeng, Y. Cui and W. L. Mao, A new phase of $Li_{15}Si_4$ synthesized under pressure, *Adv. Energy Mater.*, 2015, **5**, 1500214.
25 P. Piaggi and M. Parrinello, Predicting polymorphism in molecular crystals using orientational entropy, 2018, arXiv:1806.06006.
26 E. V. Podryabinkin, E. V. Tikhonov, A. V. Shapeev and A. R. Oganov, Accelerating crystal structure prediction by machine-learning interatomic potentials with active learning, 2018, arXiv:1802.07605.
27 P. E. Dolgirev, I. A. Kruglov and A. R. Oganov, Machine learning scheme for fast extraction of interatomic potentials and chemistry, *AIP Adv.*, 2016, **6**, 085318.
28 M. Feuerbacher, *et al.*, The Samson phase, beta-Mg_2Al_3, revisited, *Z. Kristallogr.*, 2007, **222**, 259–288.
29 D. Hobbs, J. Hafner and D. Spisak, Understanding the complex metallic element Mn. I. Crystalline and noncollinear magnetic structure of alpha-Mn, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 2003, **68**, 014407.
30 H. Niu, X. Q. Chen, W. Ren, Q. Zhu, A. R. Oganov, D. Li and Y. Li, Variable-composition structure prediction and experimental verification of MnB_3 and MnB_4, *Phys. Chem. Chem. Phys.*, 2014, **16**, 15866–15873.
31 Z. Allahyari and A. R. Oganov, Coevolutionary search for optimal materials in the space of all possible compounds, 2018, arXiv:1807.00854.
32 M. Nunez-Valdez, Z. Allahyari and A. R. Oganov, Efficient technique for computational design of thermoelectric materials, *Comput. Phys. Commun.*, 2018, **222**, 152–157.
33 A. G. Kvashnin, A. R. Oganov and Z. Allahyari, Computational search for novel hard chromium-based materials, *J. Phys. Chem. Lett.*, 2017, **8**, 755–764.
34 A. G. Kvashnin, H. Zakaryan, C. M. Zhao, Y. F. Duan, Y. A. Kvashnina, C. W. Xie, H. F. Dong and A. R. Oganov, New tungsten borides, their stability and outstanding mechanical properties, *J. Phys. Chem. Lett.*, 2018, **9**, 3470–3477.
35 H. Y. Niu, A. R. Oganov, X. Q. Chen and D. Z. Li, Novel stable compounds in the Mg–Si–O system under exoplanet pressures and their implications in planetary science, *Sci. Rep.*, 2015, **5**, 18347.
36 Y. Ma, M. I. Eremets, A. R. Oganov, Y. Xie, I. Trojan, S. Medvedev, A. O. Lyakhov, M. Valle and V. Prakapenka, Transparent dense sodium, *Nature*, 2009, **458**, 182–185.
37 B. Rousseau and N. W. Ashcroft, Interstitial electronic localization, *Phys. Rev. Lett.*, 2008, **101**, 046407.
38 X. Dong, A. R. Oganov, A. F. Goncharov, E. Stavrou, S. Lobanov, G. Saleh, G. R. Qian, C. Zh Gatti, V. Deringer, R. Dronskowski, X.-F. Zhou, V. Prakapenka, Z. Konopkova, A. I. Popo Boldyrev and H. T. Wang, A stable compound of helium and sodium at high pressure, *Nat. Chem.*, 2017, **9**, 440–445.

39 Z. Liu, J. Botana, A. Hermann, S. Valdez, E. Zurek, D. D. Yan, H. Q. Lin and M. S. Miao, Reactivity of He with ionic compounds under high pressure, *Nat. Commun.*, 2018, **9**, 951.

40 G. Saleh and A. R. Oganov, Alkali subhalides: High-pressure stability and interplay between metallic and ionic bonds, *Phys. Chem. Chem. Phys.*, 2016, **18**, 2840–2849.

41 Q. Zhu, A. R. Oganov and A. O. Lyakhov, Novel stable compounds in the Mg–O system under high pressure, *Phys. Chem. Chem. Phys.*, 2013, **15**, 7696–7700.

42 Q. Hu, D. Y. Kim, W. Yang, L. Yang, Y. Meng, L. Zhang and H. K. Mao, FeO_2 and FeOOH under deep lower-mantle conditions and Earth's oxygen-hydrogen cycles, *Nature*, 2016, **534**, 241–244.

43 D. F. Duan, Y. X. Liu, F. B. Tian, D. Li, X. L. Huang, Z. L. Zhao, H. Y. Yu, B. B. Liu and T. Cui, Pressure-induced metallization of dense $(H_2S)_2H_2$ with high-T_c superconductivity, *Sci. Rep.*, 2014, **4**, 6968.

44 A. P. Drozdov, M. I. Eremets, I. A. Troyan, V. Ksenofontov and S. I. Shylin, Conventional superconductivity at 203 kelvin at high pressures in the sulfur hydride system, *Nature*, 2015, **525**, 73–76.

45 H. Y. Liu, I. I. Naumov, R. Hoffmann, N. W. Ashcroft and R. J. Hemley, Potential high-T_c superconducting lanthanum and yttrium hydrides at high pressure, *Proc. Natl. Acad. Sci. U. S. A.*, 2017, **114**, 6990–6995.

46 F. Peng, Y. Sun, C. J. Pickard, R. J. Needs, Q. Wu and Y. M. Ma, Hydrogen clathrate structures in rare earth hydrides at high pressures: possible route to room-temperature superconductivity, *Phys. Rev. Lett.*, 2017, **119**, 107001.

47 Z. M. Geballe, H. Y. Liu, A. K. Mishra, M. Ahart, M. Somayazulu, Y. Meng, M. Baldini and R. J. Hemley, Synthesis and stability of lanthanum superhydrides, *Angew. Chem.*, 2018, **57**, 688–692.

48 J. D. Dunitz and J. Bernstein, Disappearing polymorphs, *Acc. Chem. Res.*, 1995, **28**, 193–200.

49 C. B. Anfinsen, Principles that govern the folding of protein chains, *Science*, 1973, **181**, 223–230.

50 M. W. Anderson, J. T. Gebbie-Rayet, A. R. Hill, N. Farida, M. P. Attfield, P. Cubillas, V. A. Blatov, D. M. Proserpio, D. Akporiaye, B. Arstad and J. D. Gale, Predicting crystal growth *via* a unified kinetic three-dimensional partition model, *Nature*, 2017, **544**, 456–459.

51 C. Janot, *Quasicrystals: A Primer*, Oxford University Press, 1994.