Faraday Discussions

Cite this: *Faraday Discuss.*, 2018, **211**, 613

Applications of crystal structure prediction – inorganic and network structures: general discussion

Virginia Burger, Frederik Claeyssens, Daniel W. Davies, Graeme M. Day, Matthew S. Dyer, Alan Hare, Yi Li, Caroline Mellot-Draznieks, John B. O. Mitchell, Sharmarke Mohamed, Artem R. Oganov, Sarah L. Price, Michael Ruggiero, Matthew R. Ryder, German Sastre, J. Christian Schön, Peter Spackman, Scott M. Woodley and Qiang Zhu

DOI: 10.1039/c8fd90034e

J. Christian Schön opened discussion of the paper by Yi Li: It is not quite clear to me, to what extent you treat symmetry as a deciding criterion as far as the acceptance of a structure candidate is concerned. We had lots of discussions with crystallographers and solid state chemists in the early 1990s at the beginning of crystal structure prediction for inorganic compounds, who essentially claimed that "the higher the symmetry, the better the structure" – an argument which we have always felt to be dangerously flawed.[1] Do you have the impression from your study that for your type of systems, high symmetry is a good indicator of the quality of the structure? After all, there could be many systems with space group *P*1 which are very low in energy, and furthermore, the zeolites with space group *P*1 would be expected to greatly outnumber the ones with higher symmetry. Could you comment on this?

[1] J. C. Schön and M. Jansen, *Angew. Chem., Int. Ed.*, 1996, **35**, 1286–1304.

Yi Li replied: Yes, high symmetry is not an indicator of the quality of a zeolite structure. Among the 239 known zeolite framework types approved by the International Zeolite Association, the numbers of zeolites in triclinic, monoclinic, orthorhombic, tetragonal, trigonal/hexagonal, and cubic systems are 2, 40, 83, 36, 53, and 21, respectively. Moreover, in each crystal system, space groups with the highest multiplicities occur much more frequently than the others (such as *C*2/*m* in the monoclinic system, *Cmcm* in the orthorhombic system, *I*4/*mmm* in the tetragonal system, *P*63/*mmc* in the hexagonal system, and *Im*$\bar{3}$*m* in the cubic system). In theory, we cannot rule out the possibility of zeolites with low symmetry, but it will be a good start to predict an unknown zeolite structure from higher symmetry to lower symmetry. Another important fact is that all known zeolite structures exhibit similar lattice energies. Zeolites with low symmetry do not have an obvious advantage in lattice energy over those with high symmetry. To date, there is no report on a zeolite with *P*1 symmetry. On the other hand, our method is not only for high symmetry space groups. In fact, it will have good performance on structures with multiple Wyckoff positions, which may occur in many low symmetry space groups.

J. Christian Schön remarked: In the structures you consider/generate, do you also have T atoms that are not located on special sites?

Yi Li responded: Absolutely yes. Over 70% of the known zeolite structures possess T atoms on general positions. Our approach does not rule out general positions. We simply do not add special constraints on general positions.

J. Christian Schön continued: Conversely to the previous question, do you have symmetry elements (rotation axes, mirror planes), which do not contain any T atoms?

Yi Li answered: Yes, we do have such empty site symmetries. The number of T atoms we put in a unit cell is decided by the expected framework density. Zeolites have relatively lower framework densities than other inorganic materials. In fact, zeolites with all special site symmetries occupied by T atoms are quite rare.

German Sastre asked: Using your algorithm, is there a limit to the degree of complexity of new zeolites that you can find? By complexity I mean the maximum number of symmetry unequivalent tetrahedral sites. Similar approaches tend to stop the search at 6 T sites and so if your approach is able to go further it would be the first to include more complex zeolites in a database. If so, could you plot the number of zeolites found containing x T sites (where x goes from 1 to n, with n being the maximum complexity) for each space group? In other words, are there many complex ($x > 6$) zeolites?

Yi Li answered: We use our home-made program FraGen to generate zeolite models within a given unit cell. FraGen is based on parallel tempering Monte Carlo algorithms, which is a powerful global optimization technique. We do not need to set the maximum number of unequivalent T atoms in our approach. It is decided by the cell volumes and the upper limit in framework density. In our experience, FraGen will find structures with 10 T atoms easily. So, we did generate many structures with >6 T atoms, and the number of the generated structures grew rapidly with the number of unique T atoms. We have tried our methods in two space groups, and the most complex structures we have generated so far contain 14 T atoms.

German Sastre said: Does your algorithm require significantly more computing time for those difficult cases? Could you plot, for a significant space group containing simple (less or equal than 6 T sites) and complex (more than 6 T sites) zeolites, the average CPU time taken for each number of T sites? I mean, take all the zeolites (say n) containing 2 T sites, add all the CPU times and divide by n. And do the same for all the numbers of T sites found in the space group of your choice.

Yi Li answered: The CPU time is determined by not only the complexity of the modeled system but also the number of Monte Carlo steps we set in FraGen. Normally, we use 20 000–30 000 MC steps for each cycle of FraGen, which is already enough for the building of complex structures. Increasing the number of MC steps does not significantly increase the number of plausible structures, since FraGen locates local minima very quickly. In an orthorhombic space group, taking 30 000 MC steps in each cycle, FraGen will find a tetrahedrally coordinated structure with 6 unequivalent T atoms in 2 s, a 10-T structure in 5 s, and a 14-T structure in 7 s.

Scott M. Woodley remarked: Atomic configurations predicted for nanoclusters of a compound (examples are given in my paper) can often resemble a fragment cut from one of its observed bulk phases. Sometimes, as found for $(ZnO)_n$, where n is the number of formula units the cluster is composed of, this is not the case. The global minimum for $n = 12$ actually resembles a secondary building block of the

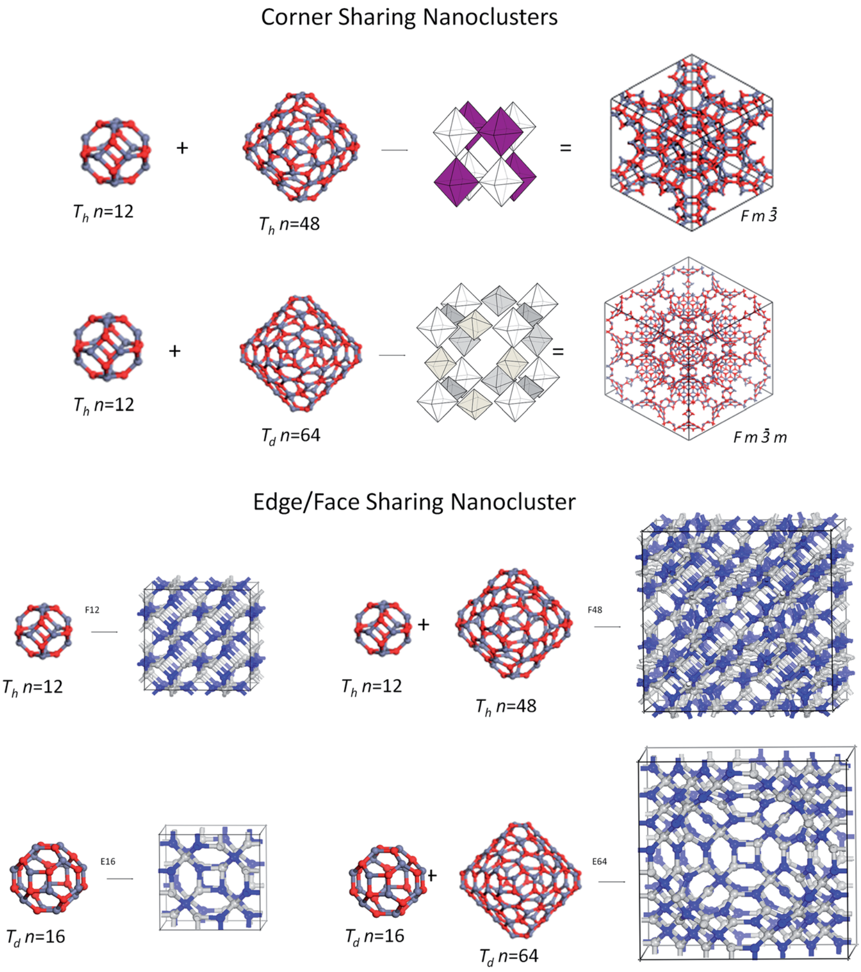


Fig. 1 Framework structures created by using predicted low energy nanoclusters for ZnO as secondary building units.[1,2]

sodalite framework. Thus, we have used the more stable sized global minima to generate new frameworks; one of these is composed of both $n = 12$ and $n = 48$ units (Fig. 1 shows this and other examples). All my atoms are four coordinated, *i.e.* sit on T sites, so switching all the zinc and oxygen atoms with a silicon and placing an oxygen atom between T sites will generate silica frameworks. Using your method, have you already predicted and published the equivalent structures that were shown during the discussions (and in the figure here)? If not, are they, particularly the structure composed of $n = 12$ and $n = 48$ units as it has a relatively low energy for ZnO (or SiC), already known in one of the zeolite databases you refer to? I can supply you with the atomic coordinates. Sometimes this generates a familiar framework, for example, the $n = 12$ example shown in the bottom left hand corner of Fig. 1 generates the sodalite framework (where atoms are located on the so called T sites of the Zeolite).

[1] S. M. Woodley, M. B. Watkins, A. A. Sokol, S. A. Shevlin and C. R. A. Catlow, *Phys. Chem. Chem. Phys.*, 2009, **11**, 3176–3185.
[2] M. B. Watkins, S. A. Shevlin, A. A. Sokol, B. Slater, C. R. A. Catlow and S. M. Woodley, *Phys. Chem. Chem. Phys.*, 2009, **11**, 3186–3200.

Yi Li responded: We have not performed structure prediction in space group $Fm\bar{3}$ and $Fm\bar{3}m$ yet. So your structure is not yet in our database. If you can provide a CIF file, I can check your structure in other databases, such as those developed by Prof. Deem and Prof. Treacy. If your structure cannot be found in any known database, I think there will be two reasons. The first is that your structure may have a big unit cell, whose dimensions are beyond current approaches. The second reason is that it may contain distortions and have a relatively high lattice energy, whereas current approaches tend to find structures with low energies. Anyway, please provide your structure file.

Caroline Mellot-Draznieks asked: There are many methods for generating thousands/millions of zeolite structures. Your method allows you to focus more efficiently on a single space group and search for structures in that space group. How do you check that the structures generated have not been predicted before, *i.e.* are new? Which is the cost function used to estimate their ranking or feasibility?

Yi Li responded: We have calculated the coordination sequences for all the zeolite structures we predict. By comparing their coordination sequences, we have removed all duplicated structures and ensure that all the structures we generate are unique and new. During model building, we use a simple cost function regarding the coordination numbers, bonding distances, and bonding angles. Then, we perform more sophisticated geometry optimizations on all the models we have built. Finally, the optimized models are evaluated by framework energy and local interatomic distance criteria. The local interatomic distance criteria are a set of structure regulations regarding the relationship between T–O, O–T–O, and T–*O*–T distances, which are obeyed by nearly all known zeolite frameworks. All structures violating any of these criteria are removed from our final result.

Scott M. Woodley asked: How does your method compare to the topological approach that employs tiling theory (as developed by Rob Bell and co-workers,[1] for example)?

[1] M. D. Foster, *et al.* Chemically feasible hypothetical crystalline networks, *Nat. Mater.*, 2004, **3**, 234–238.

Yi Li responded: As far as I understand, the tiling theory will produce all four-connected network topologies without considering much geometric information. To my knowledge, the tiling theory method has produced all uni-, bi-, and tri-nodal networks (*i.e.*, networks consisting of one, two, and three unequivalent T atoms). I think the tiling theory will generate too many networks if we consider more T atoms, and most of them are inaccessible in experiment (as shown in the mentioned reference). By considering geometric constraints/restraints, our method can generate much more complicated and chemically feasible zeolite structures (say, structures consisting of ten unequivalent T atoms), but we cannot enumerate all the possibilities because there are too many. I think the aims of these two approaches are different. Tiling theory aims to enumerate all the possible networks in a mathematical sense, whereas ours aims to find feasible synthetic targets/structure solutions as quickly/many as possible, which in most cases are complicated and inaccessible by previous approaches.

Graeme M. Day commented: You mentioned that you use the calculated energy in evaluating structures, but also geometric criteria, and that structures that have low energies are sometimes excluded because they do not fit the structural trends seen in known zeolites. Coming from the field of molecular organic crystal structure prediction, I find this interesting. For organic molecules, if we predict low energy crystal structures that have geometrical features that are different from what we have seen before, we would not exclude them from our predictions, as long as we trust the calculated energies. This is because we want to find structures that are different and might have different properties to known structures. Are those zeolite structures that are excluded based on their geometry of interest in looking for zeolites with potentially new properties?

Yi Li replied: The lattice energies of different known zeolite structures are quite similar. But there are millions of tetrahedral networks exhibiting energies similar to known ones. On the other side, many new zeolites have been synthesized with lattice energies much higher than previously reported ones. This is why we cannot use lattice energy as the main criterion to judge whether a predicted zeolite structure is feasible or not. We find that all known zeolite structures obey a set of geometric rules, the local interatomic distance criteria, which are much more reliable than the energy criterion. Recently reported high-energy zeolites obey these criteria too. The properties of zeolites, such as adsorption, separation, and catalysis, are mainly determined by their porous networks (accessible cavities, channels, and their connectivity and orientation, *etc.*). So the local geometry is not very important to these properties.

Graeme M. Day added: You mentioned that observed zeolite structures can be relatively high in energy compared to the predicted structures. Is this observation dependent on the method used to evaluate their energies? Do the high energy structures that are observed experimentally remain high in energy if they are re-evaluated using a different computational method?

Yi Li responded: Because of their framework complexity, force-field-based calculations are the most used methods for the evaluation of zeolite lattice energies. Many researchers have systematically calculated the lattice energies of known zeolites, and some of them always exhibit much higher energies than the others, no matter what program and potential functions are used. No one has tried systemic periodic DFT calculations on all known zeolites, but I would expect the same trend. One of the most important reasons why high energy zeolites exist in experiment is that there are many other species being introduced into zeolite synthesis. For instance, metal cations and organic amines are often used as key additives in zeolite synthesis. These species do not form zeolite frameworks, but reside in the extraframework voids and co-crystallize with zeolite frameworks. The host–guest interactions may compensate the high lattice energy of zeolite frameworks. Other unconventional synthetic methods, such as high-pressure approaches and ADOR approaches, can also lead to the formation of high-energy zeolites. Unfortunately, we cannot consider these practical synthetic factors among such a large number of hypothetical zeolite structures.

Graeme M. Day said: You found limits for the densities of T atoms (TAD densities) along symmetry axes and in symmetry planes, and use these limits in predicting structures. You demonstrated that the limits on TAD density are related to the distortion of the structure around the T atoms. I would expect that geometrical distortion would lead to high calculated energies. Did you look for a correlation between the TAD densities and the resulting energies of the predicted structures? I wonder if the limits that you apply to the TAD density reflect an energetic limit.

Yi Li responded: Yes. We propose defining the TAD density limit because overly crowed T atoms may lead to distortions in the bonding geometry and violate the required tetrahedral coordination in the most extreme cases, which definitely will increase the framework energy (to an unrealistic level). So, setting an upper limit for the TAD density is equivalent, more or less, to setting an upper limit for the lattice energy. We have not studied the exact correlation between the TAD densities and framework energies, because we believe the correlation is not smooth and is quite specific to different zeolite structures. Our aim is to rule out as many Wyckoff position combinations as possible, so only the highest allowed TAD density matters to us.

Qiang Zhu asked: What is the physics behind the observed empirical relation in the geometries? If you already know the relation, can you impose these geometry constraints when you generate structures for zeolite crystal structure prediction?

Yi Li replied: The local interatomic distance criteria reflect the bonding geometry of zeolite frameworks. The correlation between T–O and O–T–O distances corresponds to the constraints on the shape of a TO_4 tetrahedron. It cannot be far away from an ideal tetrahedron. The correlation between T–O and T–*O*–T reflects the specific way that TO_4 tetrahedra are linked. In general, we can impose constraints on coordination numbers, bonding distances, and bonding angles for zeolite structure prediction, but we cannot ensure all generated structures will satisfy the local interatomic distance criteria. However, for some specific zeolite structures, such as ABC-6 zeolites, which are constructed by the stacking of 6-rings, we can encode these structural regulations into the structure prediction procedure, which ensures all the structures predicted satisfy the local interatomic distance criteria.

Qiang Zhu continued: If these are purely geometry constraints, could you start the structure generation with building blocks that have these geometry relationships?

Yi Li responded: Yes, our approach is mainly based on geometric constraints/restraints. Currently, we cannot generate structures with building blocks. We notice that some methods are capable of generating structure models with building blocks, such as the AASBU method developed by Caroline Mellot Draznieks *et al.* This is a good suggestion, but it requires a lot of code writing. We may implement this idea in the future in our program, but it will require a lot of code writing.

Matthew S. Dyer opened discussion of the paper by Daniel W. Davies: This could be taken in a positive or negative sense, but one of the two new stable phases proposed in Table 2 and Fig. 9 of the paper has been previously reported. The synthesis of $YZrF_7$ is reported in ref. 1. It is isostructural with $SmZrF_7$, with the crystal structure reported in ref. 2.

This can be seen as a validation of the approach in this paper, as an existing compound outside of the original database has been identified. However, I believe that more care must be taken to perform a wide literature search before proposing a compound as a particularly promising new target.

[1] M. Poulain, M. Poulain and J. Lucas, Les fluorozirconates de terres rares $LnZrF_7$, *Mater. Res. Bull.*, 1972, **7**, 319–325.

[2] M. Poulain, M. Poulain and J. Lucas, Structure cristalline de $SmZrF_7$. Relations structurales avec le type ReO_3, *J. Solid State Chem.*, 1973, **8**, 132–141.

Daniel W. Davies replied: Thank you for highlighting this important previous work. I would argue that this should be seen as a positive result in the context of this study for the reasons you presented, and would add that for this work only minimal checks were carried out to establish the existence of a compound: the Materials Project database was queried automatically as a screening step and the ICSD was searched manually at the end of the process for any leading compounds. More careful checks should indeed be performed before taking compounds forward to experimental synthesis.

Matthew S. Dyer remarked: When allocating oxidation states on the basis of bond valence sums in a crystal structure, how does your method cope when less common anion oxidation states such as peroxide, superoxide and disulphide anions are present? For example, does it correctly ascribe the +2 oxidation state to Fe in FeS_2?

Daniel W. Davies replied: The approach we used as implemented in Pymatgen[1] uses a maximum *a posteriori* estimation method, so some prior knowledge of the distribution of oxidation states for each particular metal is needed.

For some very uncommon oxidation states, this model would lack the necessary data to assign them correctly. For example, for potassium and chlorine, it only has the required parameters for the +1 and −1 oxidation states, respectively. It would therefore be unable to reconcile the unusual compositions KCl_3, KCl_7 and K_3Cl_5, as characterised under high pressures.[2] For the particular example of FeS_2, the algorithm is able to assign these oxidation states correctly, as it has data on the distribution of the +1 to +4 oxidation state of iron. Furthermore, the benefit of using a structure-based approach can be seen when treating mixed valence compounds, for example magnetite (Fe_3O_4), which contains both Iron(II) and Iron(III).

[1] S. P. Ong *et al.*, *Comput. Mater. Sci.*, 2013, **68**, 314–319.
[2] W. Zhang *et al.*, *Sci. Rep.*, 2016, 26265.

J. Christian Schön remarked: Concerning the height of the bars in Fig. 3 in your paper, indicating the distribution of oxidation states: did you take into account that these might be actually due to spatial constraints? After all, there are only so many *e.g.* halogen atoms (with oxidation state −1) that you can pack around a cation.

Daniel W. Davies answered: This is would certainly be an explanation for some of the unexplained trends that are seen in Fig. 3–5, and one that we have not explored in the main text. For example, it may go some way to explaining why for higher oxidation states, the likelihood of finding the metal with an anion of moderate electronegativity often goes to zero before the likelihood of finding it with an anion of low electronegativity.

J. Christian Schön added: As a concrete example regarding the previous question, in spite of the fact that F has a higher electronegativity that O, you would end up preferring to use O (in oxidation state −2) when aiming for cations in a high oxidation state.

J. Christian Schön continued: In eqn 3 in your paper, you write down a probability as the product of other probabilities. This only makes sense if the product involves only independent probabilities. I am not sure whether this assumption really holds. If you look at the dataset where you get the probabilities from, this might already contain lots of correlations between certain types of cations/anions (when discovering certain elements, people often talked about “Vergesellschaftung”, indicating that in minerals certain elements liked to appear together[1]). Since the formula is very nice and suggestive, it would be great if you could show to what extent the assumption of statistical independence is justified.

[1] V. M. Goldschmidt, H. Hauptmann and Cl. Peters, *Naturwissenschaften*, 1933, **20**, 362–365.

Daniel W. Davies replied: This is an important point, as the independence of the individual probabilities is a key assumption of this approach. This is done for simplicity only and I agree that there is likely to be a range of other correlations between species and other trends whose incorporation into the model would improve its predictive power.

For the present study, we are satisfied that sufficient information is captured by the model for it to act as a pre-screening filter, as evidenced by Fig. 6 in the paper. The number of compositions generated using the SMACT code[1] that pass through such a filter decreases rapidly as the threshold is increased. Further work is planned to investigate how many of the compositions that are disregarded are false negatives. This will act as a first step towards establishing which other correlations are the most important to consider and incorporate.

[1] D. W. Davies *et al.*, *Chem*, 2016, **1**, 617–627.

Sarah L. Price remarked: This paper was a valuable update for me on inorganic chemistry. I was struck by the statement “it was recently estimated that half of all inorganic compounds are metastable”. Whilst it seems reasonable, with our current lack of knowledge about the causes of polymorphism, to assume that there is an exponential decrease in probability with thermodynamic instability, the limit of 100 meV is quite large. Your analysis is more complex than in the organic solid state where we are less often addressing the problem of the relative stability of multi-component systems such as solvates, hydrates, and cocrystals relative to their components or other stoichiometries. When you are comparing your ternary systems with the possible binary phases *etc.*, to what extent does the idea that the structures may be metastable complicate the analysis of whether the predicted structures are feasible?

Daniel W. Davies replied: The recent estimation comes from an analysis carried out by Sun *et al.*[1] in which they explore the “thermodynamic stability limit” for different chemistries. It is certainly true to say that 100 meV above the convex hull is a large limit; for certain chemistries such as chlorides, Sun *et al.* show that the limit is more like 50 meV, and for others such as iodides, it is considerably lower at around 25 meV. 100 meV is a reasonable limit for fluorides, however, so we have used it universally here in our search for mixed halide systems. The concept of metastability does complicate the prediction of new feasible structures significantly. One extension of the workflow presented here that we have successfully used previously[2] is to carry out phonon calculations as a logical next step, to confirm dynamic stability. This comes at a significantly higher computational price than carrying out DFT total energy calculations alone, but does provide useful information about the potential energy surface. Some more recent studies using, for example, energies of the amorphous states of compositions,[3] are beginning to provide a clearer definition to the upper limits of metastability for inorganic systems, but further work is needed to clarify the relationship between metastability and whether or not a compound can be synthesised.

[1] Sun *et al.*, *Sci. Adv.*, 2016, **2**, e1600225.
[2] D. W. Davies *et al.*, *Chem. Sci.*, 2018, **9**, 1022–1030.
[3] M. Aykol *et al.*, *Sci. Adv.*, 2018, **4**, eaaq0148.

German Sastre asked: Can you use your approach with mixed metallic oxides containing molybdenum, vanadium and niobium where the stoichiometry is known? Vanadium and molybdenum can have different oxidation states in the same compound. The challenge is to find, in the unit cell corresponding to a given compound, the oxidation state of each atom.

Daniel W. Davies replied: I believe this can be done using the approach that we used to assign oxidation states,[1] as this employs parameters based on the size and electronegativity of elements to establish bond valence (similar to bond order in molecules).

This is implemented in the Pymatgen code[2] using a maximum *a posteriori* estimation method, so some prior knowledge of the distribution of oxidation states for each particular metal is needed. In your particular example, I would suggest that given the wide range of oxidation states that can be adopted by each transition metal, the limits of this method would be tested. It would be an interesting test case.

[1] M. O'Keeffe and N. E. Brese, *J. Am. Chem. Soc.*, 1991, **113**, 3226–3229.
[2] S. P. Ong *et al.*, *Comput. Mater. Sci.*, 2013, **68**, 314–319.

Alan Hare enquired: Is the geometrical shape of the YF_8 polyhedron known so precisely that you can calculate the packing density of its ternary halide with ZrF_6? The ternary structure looks potentially to be a 3D space-filler: almost a cube, in fact (or possibly Menger sponge, or something like that). If so, it might be possible to suggest a further solution to Hilbert's 18th problem in mathematics, and perhaps to propose the computation of a supercell on a supercell recursively.

Daniel W. Davies responded: You have raised a very interesting series of points that merit further investigation beyond what has been reported in our work on the role of oxidation states in materials discovery.

John B. O. Mitchell added: I understood Hilbert's 18th problem to have been solved. The anisohedral tiling part was solved by Reinhardt in 1928. As for the sphere packing part, Thomas Hales (1998) produced a computer-assisted enumeration of all the possibilities, which seems to have been accepted as effective proof that cubic and hexagonal close packings can't be bettered.

Artem R. Oganov also responded: You are talking about what was known in mathematics as Kepler's hypothesis and is now known as Kepler's theorem. Let me update you: in 2015, Thomas Hales produced its formal proof, and this is why now this is a theorem.

Alan Hare answered: Yes. I understood this, too. The close packings (of just over 74%, as I recall) can't be bettered by spheres. They can of course be bettered by cubes (which fill space 100%). Octahedra, tetrahedra and icosahedra don't fill space at all (except for their own individual volume). Also, there are (100%) space-fillers beyond these Platonic solids.

An example is the Bilinski rhombic dodecahedron. This comprises two unlike rhombohedra (compare this to Penrose in 2D, with its kites and darts). There may well be other combinations of polyhedra, besides.

When I saw the ternary $YZrF_7$ pairing drawn, I thought this could be another: it looks almost cubic. But I don't know the precise geometry of the YF_8 polyhedron. So I asked the question.

Given molecular orbitals, I suppose it depends if we can ever really be satisfied with either cubes *or* polyhedra.

I recognise that years ago Reinhardt's "jigsaw" tiling solved, and latterly Hales's sphere packing addressed, an important part of the problem; and that close packings had already been seen as maximising the *sphere* packing density. But here in the crystal chemistry we are looking at two unlike *polyhedra* that have come together; and so I wondered if this might suggest a further solution to the part of the problem that concerns anisohedral tiling.

Once again we know that while the cube is the only Platonic solid that fills space, beyond these solids there are numerous space-filling polyhedra; although small unlike pairs, such as we see here perhaps, may be of particular significance physicochemically.

J. Christian Schön asked: In Fig. 8 in your paper, you state a probability of 1.0 for magnesium bromide and zinc bromide. But what about solid solution phases? Are they also captured in your scheme?

Daniel W. Davies responded: The probability of 1.0 is associated with the hypothetical ternary phase $MnZnBr_4$. This comes about due to the existence of stable $ZnBr_2$ and $MnBr_2$ phases (as depicted in Fig. 8), *i.e.*, the necessary species-anion pairs (Mn^{2+}–Br^- and Zn^{2+}–Br^-) both have probabilities of 1.0 according to eqn 2. This scheme cannot strictly be applied to solid solutions in its current form as it is stoichiometry agnostic (eqn 3), however this would form an interesting extension.

Qiang Zhu enquired: You mentioned that you used both first principle CSP and data mining approaches. There are structures from evolutionary structure prediction that belong to (un)known structure types. Could you comment on these two different approaches?

Daniel W. Davies answered: The structure prediction by substitution (data mining) approach[1] finds structures within our thermodynamic limit with very low computational cost; a database of possible parent structures can be searched on a desktop computer within a few minutes. Evolutionary structure prediction explores large areas of the potential landscape at a much higher computational cost, but is not restricted to known structure types. Hence, it was able to find lower energy structures for all four compounds in our previous study.[2]

[1] G. Hautier *et al.*, *Inorg. Chem.*, 2011, **50**, 656–663.
[2] D. W. Davies *et al.*, *Chem. Sci.*, 2018, **9**, 1022–1030.

Qiang Zhu continued: There are also some structures found by a loose fit that are analogous to known structure types. However, they were not found by the substitution method – do you know why?

Daniel W. Davies answered: Specifically, in the study to which you refer[1] the compound $Cd_5S_4Cl_2$ adopts the same structure type as Li_5BiO_5. The partial inversion in terms of the anion/cation occupancy means that the substitution algorithm does not consider this structure type.

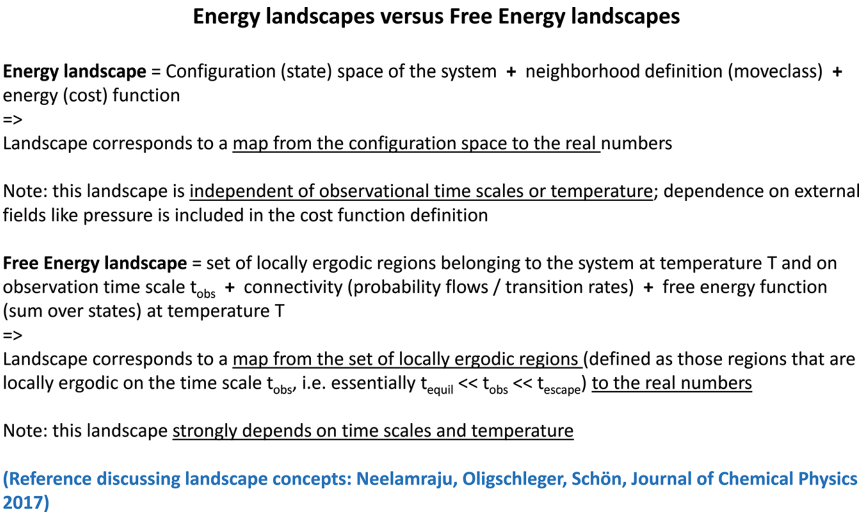
[1] D. W. Davies *et al.*, *Chem. Sci.*, 2018, **9**, 1022–1030.

Artem R. Oganov added: Among the four compounds that we have looked at[1] with Daniel W. Davies, data mining with the substitution algorithm failed in all four cases – USPEX found lower-energy structures in all four cases. In 3 cases this was because there is no known prototype. In 1 case a prototype could be found in the database, but the substitution was deemed improbable as it involved a change of charges (through a coupled substitution) and exchange of extremely different atoms.

[1] D. W. Davies, K. T. Butler, J. M. Skelton, C. Xie, A. R. Oganovcde and A. Walsh, *Chem. Sci.*, 2018, **9**, 1022–1030.

J. Christian Schön remarked: Based on the many discussions regarding "energy landscapes" and "free energy landscapes" I have had over the past day and a half, I wanted to show a slide to illustrate the central aspects and differences of these entities (Fig. 2). Mathematically, an energy landscape consists of three pieces: the configuration (or state or solution) space of the system, a neighborhood definition (called moveclass in the case of an algorithm), and an energy (or more generally cost) function. In chemistry, the configuration space is the set of all atom arrangements in the system, possibly with some constraints, *e.g.* if we already know what the unit cell of a crystalline structure is supposed to be, and in the case of a general complex optimization problem, we speak of the solution space (in which we search for the optimal solution to our task such as *e.g.* allocating money to different goods we want to invest in).

A neighborhood on the energy landscape of a chemical system is either an algorithmic one or a physical one; in the latter case, the neighborhood is associated with small displacements of individual atoms, or a combination of them



Energy landscapes versus Free Energy landscapes

Energy landscape = Configuration (state) space of the system + neighborhood definition (moveclass) + energy (cost) function
=>
Landscape corresponds to a map from the configuration space to the real numbers

Note: this landscape is independent of observational time scales or temperature; dependence on external fields like pressure is included in the cost function definition

Free Energy landscape = set of locally ergodic regions belonging to the system at temperature T and on observation time scale t_{obs} + connectivity (probability flows / transition rates) + free energy function (sum over states) at temperature T
=>
Landscape corresponds to a map from the set of locally ergodic regions (defined as those regions that are locally ergodic on the time scale t_{obs}, i.e. essentially $t_{equil} << t_{obs} << t_{escape}$) to the real numbers

Note: this landscape strongly depends on time scales and temperature

(Reference discussing landscape concepts: Neelamraju, Oligschleger, Schön, Journal of Chemical Physics 2017)

Fig. 2 Energy landscapes *vs.* free energy landscapes: Basic concepts.[9]

(which can occur on very short time scales during atom-level physical processes). The landscape for this kind of moveclass is the one we are familiar with, and the one we have in mind when we discuss the physics and chemistry of the chemical system. But this is complemented by an algorithmic neighborhood, *i.e.* the general moveclass which defines the way we explore the configuration space and thus tells us which configurations are neighbors to one another for the purpose of the algorithm. For example, in an algorithmic landscape, we can have jumpmoves or large hops such as the exchange of different cations or whole building blocks *etc.*. Typically, such moves would require a much longer time in the natural evolution of the system according to physical laws (Newtonian dynamics, *etc.*). But for the global exploration of the landscape, the moveclass of our algorithm should help us zero in on low-energy states very quickly (*i.e.* make the landscape look very "smooth" with "easy" paths towards low-energy states). Choosing the optimal moveclass for exploring the energy landscape is the "black art" behind every stochastic global optimization algorithm! But this algorithmic landscape is not connected to the stability of the predicted polymorphs – for that we need to look at the first kind of landscape that is defined by the physically relevant neighborhoods! Taken together, an energy landscape is a map from the configuration space to the real numbers, where we have some freedom to choose our neighborhoods (with all that implies for mathematical properties such as differentiability!).

As an aside, we note that *e.g.* basin hopping[1] or evolutionary algorithms (equal to a genetic algorithm operating on local minima)[2] which combine jumpmoves (hops or cross-over moves) with local optimizations actually operate on two different landscapes: for the move part, they allow large jumps from current configurations to some that are far removed (compared to what a physical move would be), but the subsequent local minimization takes place on the physical landscape. As a consequence, it is essentially impossible to derive information about the physical stability from such an exploration – we only gain information about the energy of local minima but nothing about the barriers surrounding them. An important feature is that this landscape is independent of observational time scales or temperature; the dependence on external thermodynamic parameters such as pressure p or electric fields E is introduced *via* additional terms in the potential energy (*e.g.* by adding a $+pV$ term or an $+PE$ term). Another quantity that is independent of observational times would be the density of states for any pre-defined subregion of the configuration space; however, this does not mean that this subregion is physically relevant! From this discussion, we also see that calling an energy–volume or energy–density correlation diagram an "energy landscape" is a misnomer: such correlation plots do not provide any information about the connectivity of these minima, *i.e.* which ones are neighbors on the landscape, or the energy barriers separating them!

What is a free energy landscape now? As is known from statistical mechanics, quantities like the entropy or the free energy cannot be associated with single configurations – instead we need to sum over a set of states; the choice of weight function depends on the type of ensemble we use, and the quantities kept constant, *e.g.* the microcanonical ensemble (E, N, V) with weight one for each state, the canonical ensemble (T, N, V) with the weight being the Boltzmann factor, *etc.*. In experiments, however, our measurements are always time-averages of some quantities, and the applicability of statistical mechanics, and thus our ability to define *e.g.* a free energy, depends on whether the time averages of our observables are (approximately) equal to the ensemble averages of these observables. This property is called ergodicity, and we can only compute a free energy for an ergodic system. For systems with complex (physical) energy landscapes, it often happens that we cannot explore the whole landscape on the time scale of observation, and thus we never achieve global ergodicity. However, one often finds that the system can equilibrate within a subregion R of the full landscape, and "remain" in this region for a long time. On observational time scales between the equilibration time and the escape time of this region, the system behaves as if it were ergodic, and we speak of local ergodicity.[3–5] (Note that you can even have locally ergodic regions that do not contain a local minimum – they are separated by entropic barriers from the rest of the system![6]) In this situation, we can define a local free energy by computing the sum over states restricted to the subregion, *e.g.* via the local density of states we have measured for this subregion. One should note that the equilibration and escape times clearly depend on temperature, *e.g. via* the Arrhenius law for the escape time across energetic barriers. But keep in mind that you can associate a local free energy with a local density of states only if the region of interest is locally ergodic. For a given observational time scale, we can thus determine all the locally ergodic regions and compute their local free energy.

If we now add information about the probability flows between these regions on the observational time scale (*e.g.* measured using transition path sampling[7] or the threshold algorithm[8]), then we have all the pieces together for constructing the free energy landscape of the system as function of the observational time scale: the free energy landscape corresponds to a map from the set of locally ergodic regions to the real numbers where the neighborhood is defined by the probability flows of the system. It is important to keep in mind that this landscape strongly depends on the time scales and temperature. For more details about (free) energy landscape concepts I would refer to the literature.[9]

[1] D. J. Wales and J. P. K. Doye, *J. Phys. Chem. A*, 1997, **101**, 5111–5116.
[2] D. M. Deaven and K. M. Ho, *Phys. Rev. Lett.*, 1995, **75**, 288.
[3] J. C. Schön, in *Proceedings of RIGI-workshop 1998*, ed. J. Schreuer, ETH Zürich, Zürich, 1998, pp. 75–93.
[4] J. C. Schön and M. Jansen, *Z. Kristallogr.*, 2001, **216**, 307–325.
[5] J. C. Schön and M. Jansen, *Z. Kristallogr.*, 2001, **216**, 361–383.
[6] J. C. Schön, M. A. C. Wevers and M. Jansen, *J. Phys.: Condens. Matter*, 2003, **15**, 5479–5486.
[7] C. Dellago, P. Bolhuis, F. S. Csajka and D. Chandler, *J. Chem. Phys.*, 1998, **108**, 1964.
[8] J. C. Schön, *Ber. Bunsenges. Phys. Chem.*, 1996, **100**, 1388–1391.
[9] S. Neelamraju, C. Oligschleger and J. C. Schön, *J. Chem. Phys.*, 2017, **147**, 152713.

J. Christian Schön opened discussion of the paper by Frederik Claeyssens: I am wondering about Fig. 8 in your paper. The energy *vs.* volume curve for $\xi_5 = 0.48$ shows a very unusual shape: when going from the minimum to larger volumes, the slope is much larger than when moving towards smaller volumes – exactly the opposite one finds in most bulk inorganic materials. Could you comment on this?

Frederik Claeyssens responded: You are referring to the energy *vs.* volume curve in Fig. 8 for beta-cristobalite itself. This is in very good agreement with previous theoretical work on this system and the shape of this curve is indeed unusual, as you point out. Nevertheless, it is consistent with the work of Wells and co-workers (*e.g.* ref. 1 and 2), who have developed the concept of flexibility windows in covalent frameworks. They have noted that silicates are atypical in their tendency to be maximally extended in their relaxed (minimum energy) states. The behaviour of the borate analogue (beta-cristobalite-borate-a), also shown in Fig. 8 is different and more common; the minimum energy structure lies towards the low-volume end of the flexible region. Compare also the results for other B_2O_3 polymorphs in Fig. 1 in the paper.

[1] S. A. Wells and A. Sartbaeva, *Mol. Simul.*, 2015, **41**, 1409–1421.
[2] S. A. Wells *et al.*, *R. Soc. Open Sci.*, 2017, **4**, 170757.

J. Christian Schön commented: Concerning your study, there has been similar work by the group of Catlow in the mid-1990s.[1] How does your work compare to their investigations?

[1] A. Takada, C. R. A. Catlow and G. D. Price, *J. Phys.: Condens. Matter*, 1995, **7**, 8659–8692.

Frederik Claeyssens responded: We are aware of the work of Catlow and co-workers[1,2] and refer to these in our paper. These authors reported the results of classical molecular dynamics simulations using a conventional harmonic bond-angle potential for the B–*O*–B bond angle. Thus these simulations take no account of the changes in bonding in the B–*O*–B bridges with angle that are responsible for the flexibility which plays such a crucial role in our own findings.

[1] A. Takada, C. R. A. Catlow and G. D. Price, *J. Phys.: Condens. Matter*, 1995, **7**, 8659–8692.
[2] A. Takada, C. R. A. Catlow and G. D. Price, *J. Phys.: Condens. Matter*, 1995, **7**, 8693–8722.

Sarah L. Price commented: I was struck by your comments about boron oxide B_2O_3 being a good glass former and wondered how this may link to the plethora of structures being generated as possible polymorphs. The zeolites are polymorphs of SiO_2, the definitive glass former. In the organic/pharmaceutical solid state world, the amorphous state is of great interest. Why won't a given organic molecule crystallise? Can we develop the amorphous form and be confident that it will not crystallise? There is the possibility that the formation of amorphous phases is linked with the generation of a suitably diverse set of hypothetical crystal structures with, for example, a diversity of possible hydrogen bonding motifs, that impedes the formation of a stable crystal structure.[1] The glass forming ability could be linked with having many polymorphs or no crystallisation. Could you comment on whether being able to readily form an amorphous state may make it difficult to experimentally realise your predicted new families of frameworks?

[1] M. Habgood *et al.*, *Cryst. Growth Des.*, 2013, **13**, 1771–1779.

Frederik Claeyssens responded: This remains an open question. Zeolites can of course be synthesised even though SiO_2 as you comment is the definitive glass former.

Michael Ruggiero added: This is a follow up to the previous question by Prof. Price – it is important to differentiate between flexibility and stability. It is true that molecular flexibility often increases the number of accessible polymorphs, but once crystallised (or quenched into amorphous solids), care must be taken when referring to the same process, as now the energy barriers to reorganisation must be considered. A highly flexible molecule will inherently have shallow potential curves at some finite displacement, but to completely reorganise might (or might not) require overcoming large energy barriers, limiting the process. This might not be necessarily true in the solution/liquid state, which is why numerous polymorphs might form, each individually stable. This is also why thermodynamically unstable (metastable) polymorphic forms are shown on occasion to never undergo a solid-state phase transformation, since presumably the barriers are simply too large.

Matthew R. Ryder remarked: It would be interesting to have a better idea of the mechanical and thermal stability of these ultra-flexible boron oxide frameworks. Have you done any work on the single-crystal elasticity or thermal expansion?

Frederik Claeyssens answered: As yet, we haven't done so – these remain important studies for the future.

J. Christian Schön remarked: You have rather big empty rings in your structure, reminding me of some unpublished preliminary calculations we did about twenty years ago on B_2O_3. We also found such rings, sometimes even in a three-dimensional "stacking", with B sometimes three- and sometimes fourfold coordinated by O. These rings were big enough to allow other rings to penetrate, in principle, although I am not sure whether this ever happened in our searches. Have you observed such stacking of loops on top of each other?

Frederik Claeyssens replied: We have not so far examined structures in which rings penetrate other rings.

J. Christian Schön continued: Following up on the previous question, since there appear to be rather large "holes" in some of the boron oxide structures, I am wondering whether you can knit them in a loop, or generate some interpenetrating B–O network?

Frederik Claeyssens answered: Further interesting possibilities, and we suspect quite challenging ones experimentally.

Caroline Mellot-Draznieks asked: You show that boron chemistry may lead potentially to new structures directly inspired from zeolite topologies, however with much greater flexibility. When compared to zeolites, have you tried to establish the nature of "universal" descriptors that might indicate whether such structures are attainable experimentally?

Frederik Claeyssens replied: This is an interesting question which we will address further in future work, where we will study in more detail the enthalpies of formation of the new structures derived from zeolite topologies.

Caroline Mellot-Draznieks noted: there is some analogy between the boron-based porous structures you predict and MOFs possessing very large flexibility (breathing, gate opening). Such flexibility in MOFs is highly dependent on temperature, pressure, and guest induced transitions. Are your predictions currently able to incorporate such effects (responsive materials)?

Frederik Claeyssens replied: Pressure effects are readily included in the calculations and indeed we report in our paper a number of results at high pressure (and negative pressure). Guest molecules could be incorporated straightforwardly although calculations on cases where there are many possible sites for their incorporation would be computationally expensive. Temperature effects are more challenging and remain for future work. Expensive *ab initio* molecular dynamics (MD) rather than its much cheaper classical equivalent is required because it is essential to take account of the changes in electronic structure at the boron atoms which are responsible for the crucial angular flexibility of the B–*O*–B bridges. It is worth also nothing that unfortunately it has been usual in classical molecular dynamics of B_2O_3 to use a conventional three-body bond-bending term for the B–*O*–B bond angle which stiffens the angle and so removes this crucial flexibility from such simulations (*e.g.* ref. 1 and 2).

[1] A. Takada, C. R. A. Catlow and G. D. Price, *J. Phys.: Condens. Matter*, 1995, 7, 8659–8692.
[2] A. Takada, C. R. A. Catlow and G. D. Price, *J. Phys.: Condens. Matter*, 1995, 7, 8693–8722.

Peter Spackman opened discussion of the paper by Scott M. Woodley: About your hashing algorithm for structures: is it actually a hash (*i.e.* a lossy transform) or just a string encoding of the graph of molecular connectivity? If it's a hash, what do you plan to do for potential hash collisions (however unlikely they might be)?

Scott M. Woodley answered: As explained in the main paper, each HASHKEY is a character string for a particular connectivity graph. Yes, actual interatomic distances are lost as we want to be able to determine whether

(a) two configurations are essentially the same and differences (other than connectivity) exist because either different tolerances were used during the optimisation stage or slightly different accuracies with regards to the model employed and

(b) two configurations are the same structural motif even though they are formed of different atom types and therefore have different bond lengths and slightly different bond angles, *e.g.* see Fig. 3 in our main paper, where all configurations shown are considered to have the same structural motif.

We do not consider our solution to be a hash; however, we do intend to implement other approaches to measuring structural similarity between clusters that have been developed by others (as described and cited in the main paper), and allow the user to change the cut-off used in determining the connectivity.

Peter Spackman asked: These sorts of databases would be a useful tool in other areas (*i.e.* for other materials *e.g.* crystals, molecules *etc.*). What sort of standardisation (file formats, structure of submission *etc.*) do you use for the deposition of data?

Scott M. Woodley answered: I also agree that similar databases would be a useful tool in other areas, in fact EPSRC have funded a second project based on producing something similar for surfaces, and further funding is sought for extending the current database so that the environment of the nanoclusters can also be included.

Currently, structural information is uploaded as an xyz file, the atomic coordinates of which are immediately transformed to the reference frame where the cluster's centre of mass is at the origin and the axes of the principal moments of inertia align with the Cartesian axes. The xyz file may contain many configurations, concatenated one after the other, and allows additional information about the cluster (total energy for example) and individual atoms (charge and/or spin, for example). We do not automatically extract this information from the standard output files of materials software, as we have developed our own global optimisation software, KLMC, that automates the process of running third party codes and creating suitably named xyz files.

German Sastre added: I agree with what you say: this database could help to unify chemical knowledge and could help to discover things by analogy. How much human work does it take to upload one entry? Is it to be done by professionals or by the person who published it? How many entries do you have currently?

Scott M. Woodley answered: We have tried to minimise the effort required to upload datasets as ideally we would like the authors of published datasets to upload their dataset(s) into the HIVE. Five of the investigators on the WASP@N project (either listed as a co-author or in the acknowledgments of the main paper) have an extensive track record of publications that report nanocluster structures predicted using global optimisation methods, which we hope will provide the required datasets, or critical mass, for the HIVE to take-off. We have also already uploaded some other datasets after gaining permission from the main author who was not originally part of the WASP@N project. Typically, such authors have perhaps moved away from the field and their xyz files are available (as opposed to only being available as a graphic in their paper).

Anyone can temporarily upload any entry to compare against entries already within the HIVE, whereas permanent entries can only be uploaded by *trusted users* of this project (as this requires level-3 access to the HIVE) who agree to upload their published work for others to use. To apply to become a trusted user requires the applicant to first register online (obtaining automatic level-2 access) and then to send an email to me, Scott M. Woodley, requesting additional access in order to upload their dataset(s) that are already associated with a published journal manuscript. With level-3 access, the user can register the DOI of their manuscript, and then upload one or many xyz files containing one or more atomic structures. To reduce the effort of providing any additional information, there is an initial web form where default values can be provided (for example, the compound and stoichiometry of the set of clusters, and/or the definition of the energy and software used in creating the atomic configurations) for each xyz file. Upon upload, WASP will check whether the uploaded data is consistent with the default values provided and will also warn if any of the uploaded configurations have already been uploaded for the DOI. The uploaded datasets connected to the DOI are only visible once the author uploading these is happy the dataset is correct and makes the *private* DOI *public*. After making the dataset associated with a particular DOI public, the BEE software computes various properties of each configuration, refines each configuration to create another entry, and establishes links between these and other datasets in the HIVE. Importantly, any future changes to published datasets will be documented and these records will form part of the dataset. Originally we had planned that there would be a time limit on how long a DOI can remain private in order to give the owner of the DOI time to upload and check all configurations associated with the DOI (perhaps doing configurations of one cluster size at a time) so that the length of the *record of changes* is minimised; however, as this has not been implemented it requires the author to complete the uploading process by clicking on the "make public" link. During this Faraday Discussion when I checked, there were around 1000 configurations earmarked as public and between 10 and 100 times more that were marked as private. Clearly these need to be made public as soon as possible; moreover, we hope the community will support our efforts and add their datasets so that the HIVE continuously grows with time and becomes progressively more useful.

There is also the question of how we assure the quality of the datasets. Only trusted users have the access level to upload permanent datasets; manually uploaded datasets must come from a publication that has a DOI, simple checks are automatically performed by WASP, and all configurations are refined by the BEE software. Consider, for example, the case of the dataset generated for the main paper, where interatomic potential (IP) local minima (LM) were optimised using the third party code, GULP, in order to minimise the energy of each nanocluster. Before this Faraday Discussion, these were all uploaded and, as a result, each was further refined using an electronic structure approach implemented within FHI-aims. The unique refined configurations are now also available *via* the WASP interface to the HIVE and so one can easily see whether or not the IP LM are reasonable models (assuming the chosen electronic structure approach provides a good representation of these clusters). We leave this analysis as an exercise for anyone who would like to try the WASP interface; after selecting one of our IP nanocluster LM, at the bottom of its webpage there should be two lists, one showing the refined PBEsol configuration the BEE software has generated *via* the call to FHI-aims (which will resemble the IP LM if there is a good match between our IP model and the PBEsol one), and, if the HASHKEY does not

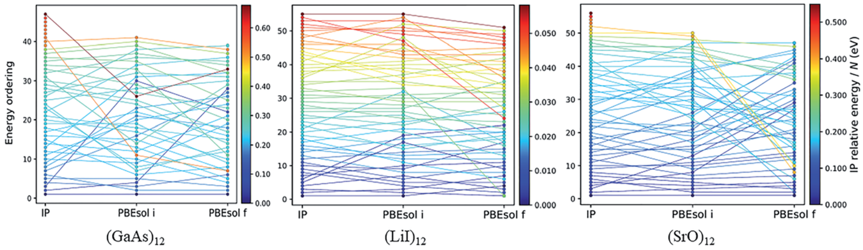


Fig. 3 The energy ranking of the local energy minima reported in the main paper are represented as circles on the left hand side of each of the three panels (one panel per compound), the colour of which indicates the energy difference between the respective local minimum and the global minimum. The circles in the middle and right hand side that are connected to the left hand side are the energies of the left hand side atomic configurations before and after refining the respective configurations using an electronic structure based software, FHI-aims.

change, the PBEsol LM should also be listed as one of the matching structures. The change in ranking when switching between IP and PBEsol can also give an insight into the quality of our interatomic potential parameters in modelling these nanoclusters.

How the rankings of the $(GaAs)_{12}$, $(LiI)_{12}$ and $(SrO)_{12}$ nanoclusters change upon switching between IP and PBEsol is shown graphically in Fig. 3. The first thing to note is that the electronic structure for the original IP atomic configuration for a number of the LM for $(GaAs)_{12}$ and $(SrO)_{12}$ did not readily converge so there is a decrease in the number of LM from IP (far left hand side) to PBEsol energy (middle), and that a number of the IP LM were not stable on the PBEsol landscape, *i.e.* there are fewer PBEsol LM (far right had side) than IP LM. In both cases, the additional IP LM were typically those of the highest energy, greater than 0.5 eV/atom, and not of immediate interest. More importantly, there is good agreement in that the IP global minimum is the lowest PBEsol LM energy found for two of these compounds and is ranked fourth for the third, whereas the new lowest PBEsol configuration was ranked third (of 55 LM) on the IP landscape.

Matthew S. Dyer asked: The unambiguous allocation of edges between nodes on the graphs representing the clusters is clearly important. How is this done (*i.e.* which interatomic distances are chosen to be linked with an edge, and which are not)?

Scott M. Woodley replied: Currently we use a predefined combination of ionic and covalent radii of the elements that the nanocluster is composed of in order to determine the cut-off used to distinguish between atoms that are or are not connected. Typically the cut-off results in edges between nearest neighbours, *i.e.* edges are made between each atom and the atoms within its first coordination shell. In a future version of the WASP interface, we will investigate the possibility of allowing the user to rescale this cut-off and generate a new, perhaps temporary, HASHKEY for the set of configurations already found in a previous search of the HIVE and earmarked for further investigation by the user.

J. Christian Schön enquired: Once you are given the HASHKEY – can you reconstruct the three-dimensional structure from this key?

Scott M. Woodley replied: The HASHKEY was introduced into the WASP@N project as a means to enable a rapid search of the HIVE for matching atomic structures based on connectivity arguments; this can also include the constraint that the nanoclusters must be composed of the same constituent atoms (compound). It was never our intention to reconstruct three-dimensional atomic structures from a HASHKEY, but we have created an alternative SCOTTKEY that is a character string that is formed of element symbols, coordination numbers and the frequency thereof, and thus it is easier to interpret what kind of configuration it represents. We have not attempted to reconstruct three-dimensional atomic structures from a HASHKEY or a SCOTTKEY and currently have no plans to attempt this as, even if it were possible, it would not be a straightforward process and, even for a nanocluster composed of one element, the exact interatomic distances cannot be retrieved without, for example, local optimisation of the original energy function.

J. Christian Schön asked: Concerning "duplicates": if I had performed a study with *e.g.* MgO clusters, and now I would like to find out whether some of the structures have already appeared in the (presumably related) SrO clusters that are in your database, could I find out about this using the tools provided by your database?

Scott M. Woodley answered: Yes, you can use the WASP interface to find whether a structural motif of one compound exists for another compound. If you have published your study on MgO nanoclusters, then we would encourage you to upload these into the HIVE so that the visibility of your work is improved – note that all entries, when displayed, have a hyperlink *via* the DOI of your publication (paper). Searching the HIVE using, for example, the DOI of your paper, you will obtain a list of your MgO nanoclusters (which will look similar to that shown in Fig. 4 of the main paper). Upon selecting any one of your nanoclusters, by clicking on its thumbnail image, will generate a new webpage containing a rotatable ball and stick image of your nanocluster together with details of various properties and, at the bottom of the page, a list of all structures in the HIVE that have the same HASHKEY as your nanocluster. Here, in this list, you would find the equivalent SrO nanocluster, if it is already in the HIVE. Moreover, we have also created the SCOTTKEY for each nanocluster entered into the HIVE, which allows an efficient search based on finding nanoclusters that have a certain structural feature as part of their atomic configuration, *i.e.* one can search for nanoclusters with only three coordinated atoms, or those with at least one six coordinated magnesium or calcium atom, for example.

J. Christian Schön asked: Is the connection to your database possible just *via* the internet, or does one need to get an account on your machine?

Scott M. Woodley answered: WASP@N is a *community* project, sponsored initially by EPSRC for five years, involving partners who either: (a) develop and/or apply global optimisation methods for predicting lowest energy atomic structures of nanoclusters; (b) develop and/or apply materials modelling software; or (c) develop databases. From its conception we agreed that the best way to help our communities, and the general public who may also be interested in what nanoclusters are, is to make the database, or HIVE, and the associated tools developed as part of this project, accessible *via* the internet (the WASP interface). The WASP interface allows four levels of access:

(1) with no registration (a) example nanoclusters can be displayed and (b) an atomic structure can be temporarily uploaded in order to see whether or not it already exists in the HIVE example;

(2) upon registering your name, institution and email address (which is automatically checked) online and agreeing to cite the relevant DOI(s) of any nanoclusters used in your future work and acknowledge use of our website, the HIVE and WASP toolkit is immediately made available;

(3) the registered account can also upload new data into the HIVE;

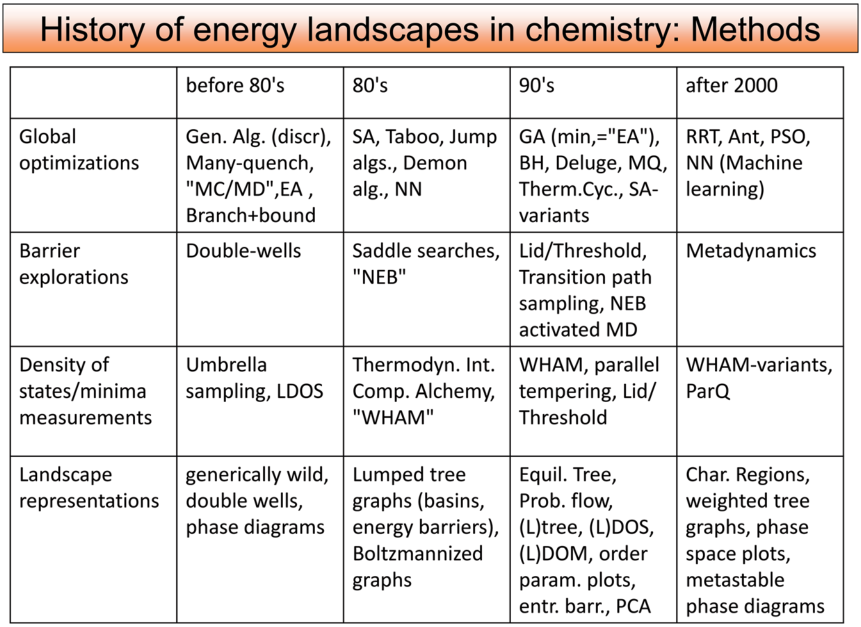
and (4) has admin rights to determine which accounts have either access level 2 or 3, as well as the ability to prioritise which datasets are processed (see the description about links between data entries in main paper).

Even though I am the PI of this project, I still only use the WASP interface when accessing the HIVE; only the actual developers of the WASP, BEE and HIVE software and the hardware support officers need accounts on the machine hosting the software and database; hence, no you do not need an account on our machine.

J. Christian Schön opened a general discussion of the papers from the meeting: Inspired by the talk of Jonas Nyman, with its history of crystal structure prediction for molecular crystals, I would like to show two slides about the history of energy landscape exploration in chemical systems.

Looking at the history of energy landscapes of chemical systems, we can consider it from the methods point of view and the system type perspective. Let us start with the methods (Fig. 4). There are four different types of tasks for which algorithms and general approaches have to be developed: a) the global optimization, *i.e.* the search for local minima; b) the exploration of the barrier structure, in order to determine generalized (energetic + entropic + kinetic + mixed) barriers and the equilibration, escape and transition rates associated with the barriers; c) the determination of the (local) densities of states needed for the computation of (local) free energies; and d) the low-dimensional representation of energy landscapes to help us visualize the properties of the landscape and the time evolution of the chemical system.

Concerning global optimizations, mathematicians, physicists and computer scientists had already before the 1980s developed Monte Carlo[1] and molecular dynamics[2] simulation algorithms that can be combined with stochastic and deterministic quenches (*i.e.* local minimizations), multiple (random) starting point quenches, genetic algorithms[3] (for discrete configuration spaces that could be represented as a binary string), or systematic exploration methods such as



History of energy landscapes in chemistry: Methods

	before 80's	80's	90's	after 2000
Global optimizations	Gen. Alg. (discr), Many-quench, "MC/MD",EA , Branch+bound	SA, Taboo, Jump algs., Demon alg., NN	GA (min,="EA"), BH, Deluge, MQ, Therm.Cyc., SA-variants	RRT, Ant, PSO, NN (Machine learning)
Barrier explorations	Double-wells	Saddle searches, "NEB"	Lid/Threshold, Transition path sampling, NEB activated MD	Metadynamics
Density of states/minima measurements	Umbrella sampling, LDOS	Thermodyn. Int. Comp. Alchemy, "WHAM"	WHAM, parallel tempering, Lid/ Threshold	WHAM-variants, ParQ
Landscape representations	generically wild, double wells, phase diagrams	Lumped tree graphs (basins, energy barriers), Boltzmannized graphs	Equil. Tree, Prob. flow, (L)tree, (L)DOS, (L)DOM, order param. plots, entr. barr., PCA	Char. Regions, weighted tree graphs, phase space plots, metastable phase diagrams

Fig. 4 History of energy landscapes in chemistry: Methods being employed.

branch-and-bound methods.[4] In the 1980s, *e.g.*, simulated annealing,[5,6] taboo searches,[7] jump/bounce algorithms[8] (equivalent to repeated "quench/simulated annealing followed by a large jump or short high-temperature Monte Carlo simulation" cycles), the Demon algorithm[9] (guided multi-walker stochastic quench procedure), and neural network algorithms[10] (for the parameter landscape of the network) were developed. In the 1990s, various simulated annealing variants were added (for a review see *e.g.* ref. 11) such as jumpmove simulated annealing or basin hopping,[12] shadowing simulated annealing (simulated annealing on minima instead of states), sequential multi-quench[13] and thermal cycling[14] (both similar to the bounce algorithm), parallel tempering[15] or the grand deluge algorithm[16] (where a lid is imposed on top of a system that pushes the random walker down into deep-lying minima). After 2000, biology inspired algorithms such as ant-search (originally invented in the 1990s[17]) or particle-swarm optimization (also already invented in the 1990s[18]) became popular in chemistry, a renaissance of neural networks under the name of machine learning took place, and the rapidly growing random tree algorithm (RRT) in many variants[19,20] appeared on the scene.

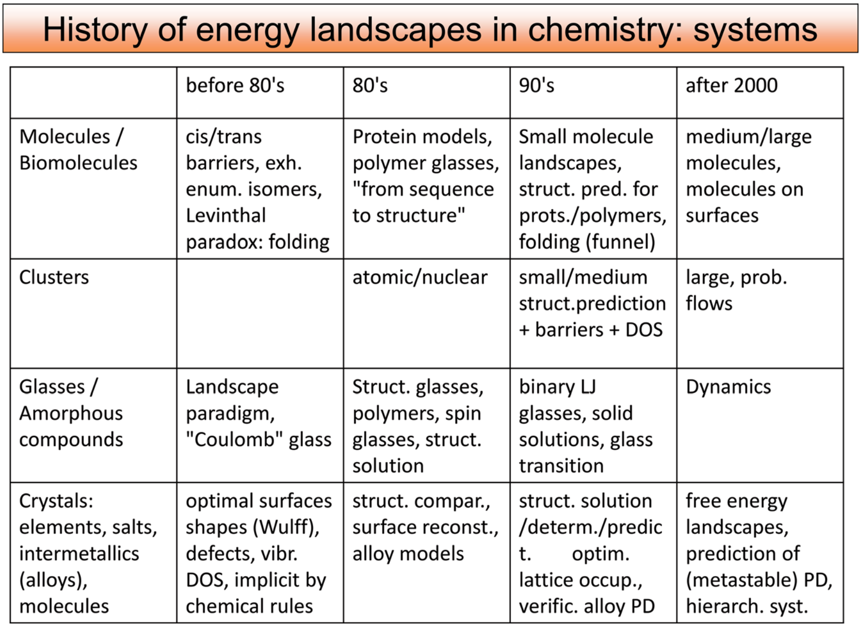
Barrier explorations have not seen so many methods, but still quite a few: before 1980, a major focus was on double-well systems and analytical methods supported by simulations. But in the 1980s people started working on studying the landscape by generalizing the double-well methods to multi-minima systems, and various saddle-search procedures such as slowest-slides[21] or eigenvector-following[22] were employed. There followed a sharp increase in the number of algorithms in the 1990s, starting with the Lid algorithm for exhaustive exploration of pockets of discrete landscapes[23] and the analogous threshold algorithm for continuous ones,[24,25] which allowed the determination of saddle/transition regions and probability flows (as often happens in this field, these algorithms were reinvented a couple of times). Molecular dynamics based methods became very successful such as transition path sampling,[26, 27] the nudged elastic band,[28, 29] or activated molecular dynamics,[30] in order to efficiently cross energy barriers and compute rate constants. In the new millennium, metadynamics[31] was introduced combining molecular dynamics with the taboo and Lid algorithm concepts, as an evolution from the so-called local elevation approach from the 1990s.[32] One should note that the computational effort involved in exploring the barrier structure is typically an order of magnitude larger than the "straightforward" global optimization.

Computation of the density of states and (local) free energies usually requires an even larger computational effort. Before the 1980s, umbrella sampling had been introduced,[33,34] and, of course, the computation of the local density of states around a local minimum in the (quasi-) harmonic approximation was well-established (using empirical/fitted-to-experiment potentials). In the 1980s, direct sampling methods such as the lumped-transition matrix approach[35] were introduced, but also thermodynamic integration/perturbation, *e.g.* ref. 36 (based on early work by Zwanzig[37] and Kirkwood[38]), and its cousin, "computational alchemy",[39] were developed from the umbrella sampling methods, in order to compute the difference in free energy between a known system and the target system. This was supplemented by the first weighted histogram analysis methods,[40] which came into their own in many variants in the 1990s.[41,42] In that time, parallel tempering and subsequently multi-canonical simulations[43] were also introduced that allowed combinations of global optimization tasks with sampling of the density of states. Since 2000, these WHAM-based methods have continued to be developed,[44,45] but also other approaches such as the ParQ-transition-matrix algorithm[46] have been introduced.

Finally, the representation of the energy landscape has evolved during this time. Before the 1980s, we are usually dealing with qualitative sketches of wild-looking landscapes, combining many double-wells, or trying to depict metastable phases in phase diagram pictures. In the 1980s, work on discrete combinatorial optimization problems and spin glasses lead to the introduction of (lumped) tree graphs,[47] where all the states that were connected (*via* the moveclass) by paths below a given energy lid, to a certain local minimum (always the lowest one in the region), were replaced by a single node at the energy of the lowest minimum in the region. The energy lid was raised step by step, and whenever another previously unconnected region could be reached below the new energy lid, another node was created at that energy level, connecting the two previous nodes. As a consequence, a tree graph was generated, where the leaves were the local minima, and the regular nodes of the graph connected to the leaves (and the other nodes) depicted the energy levels where two lower energy nodes or leaves became connected. (This very powerful description could also include temperature *via* the so-called Boltzmannization procedure,[48] where the transition matrix between the nodes was weighted by Boltzmann-acceptance factors analogous to those in the Monte Carlo algorithm.) In particular for glassy systems, the so-called inherent structures were analyzed[49] (the inherent structures are the local minima that can be reached *via* local, usually gradient-based, minimizations along a MD or MC trajectory). In the 1990s these first tree graphs were expanded to multi-lump trees,[50] which incorporated the local densities of states associated with each node and energy band between two energy levels. In addition, probability flows were being depicted between the nodes and leaves of the tree[51].

Supplementing such landscape visualizations, local densities of states and minima were derived, entropic barriers were depicted as so-called return-probabilities[52] (*i.e.* the probability that a random walker does return to its starting basin), and the so-called principal component analysis (originally developed in the 1960s[53]) was used to classify the minima and to develop a coordinate system on which the set of local minima that had been found was most disperse such that *e.g.* order parameters might be derived. Another common way to depict a "cut" through the landscape was to define an order parameter (or a reaction coordinate for the description of chemical processes and transformations), and then to compute *e.g.* an average energy or a "free energy" for a given value of the order parameter. However, in the latter case, one must wonder to what extent it is justified to assume that on the observational time scales on which one could claim that the system possesses a specific value of the order parameter, it is permitted to treat the subregion of the landscape with a given order parameter value as locally ergodic. (If not, we should not be allowed to define a local free energy!) Tree graphs continued to exert their fascination after 2000, with a variety of modifications adding more features to the tree graph and especially the connections between the nodes.[54]A new concept was the idea of characteristic regions,[55] which puts all the states in a given energy range together into a lumped node, if they "see" the same part of the landscape below them, *i.e.* if we perform *e.g.* twenty stochastic quenches from such a point then, within statistical error, they will always reach *e.g.* minimum one with 50%, minimum 2 with 30% and minimum 3 with 20% probability (we would call such a set of states a multi-minima transition region; one should note that basin regions, *i.e.* where nearly all quenches reach only one local minimum can extend to quite high energies compared to the energy of the first saddle point that can be reached from the minimum). In addition, people have started to plot and study phase space (position+momentum space) trajectories and thus construct connected (locally ergodic) regions of phase space.[56]

Turning to the kind of chemical systems with complex landscapes that have been studied in these years, Fig. 5 we find that before the 1980s, mostly (bio) molecules (key word: Levinthal's paradox in protein folding[57]) and glasses were investigated from a landscape point of view, with the landscape paradigm having been introduced by Goldstein[58] in the late 1960s to gain an understanding of what glassy systems might be like and what their dynamics might be. In the field of crystals, crystallographers (and solid state chemists) introduced a variety of rules (usually based on some intuitive understanding of the energy involved, such as the close-packing rule or the radius–ratio-rule) that allowed them to judge the quality of observed or suggested crystal structures, and generate new candidate structures in a more or less systematic way. Of course, organic molecules had already been "designed" on paper by a complex system of rules for quite some time even before 1980,[59] *e.g.* resulting in the systematic enumeration of possible isomers *etc.*,[60] and estimates of *cis–trans*-barriers (analogous to double-well systems). In the 1980s, protein models (first attempts at "from sequence to structure" predictions) and polymer glasses became a mainstay of landscape studies (for an overview up to *ca.* 2000, see *e.g.* ref. 61), as did structural glasses, spin glasses or solid solutions, especially in intermetallic systems. But in



History of energy landscapes in chemistry: systems

	before 80's	80's	90's	after 2000
Molecules / Biomolecules	cis/trans barriers, exh. enum. isomers, Levinthal paradox: folding	Protein models, polymer glasses, "from sequence to structure"	Small molecule landscapes, struct. pred. for prots./polymers, folding (funnel)	medium/large molecules, molecules on surfaces
Clusters		atomic/nuclear	small/medium struct.prediction + barriers + DOS	large, prob. flows
Glasses / Amorphous compounds	Landscape paradigm, "Coulomb" glass	Struct. glasses, polymers, spin glasses, struct. solution	binary LJ glasses, solid solutions, glass transition	Dynamics
Crystals: elements, salts, intermetallics (alloys), molecules	optimal surfaces shapes (Wulff), defects, vibr. DOS, implicit by chemical rules	struct. compar., surface reconst., alloy models	struct. solution /determ./predict. optim. lattice occup., verific. alloy PD	free energy landscapes, prediction of (metastable) PD, hierarch. syst.

Fig. 5 History of energy landscapes in chemistry: Types of chemical systems being studied.

addition, atomic (and also nuclear, in physics) clusters became paradigmatic example systems for the development of global optimization algorithms for structure prediction (for an overview up to *ca*. 2000, see *e.g.* ref. 62). Alloy models and simple comparisons of typical structure candidates (*e.g.* comparing sodium metal in the bcc, fcc and sc structure), were becoming of interest in crystal chemistry and solid state physics, as did studies on the energetics of surface reconstruction.

Energy landscape analysis for chemical systems really came into its own in the 1990s, when structure prediction and energy barrier analysis were applied in all fields of chemistry: small molecule landscapes, structure prediction for proteins and polymers, funnel models for protein folding (such models are very intuitive but also quite tricky because of issues of dimensionality – it is not fully clear whether the landscapes of such molecules really show a funnel structure), prediction of small and medium sized clusters including their energy barriers and local densities of states, computational models of the glass transition, and, of course, crystal structure solution (from raw data), determination (prediction but with direct experimental input restricting the allowed region of the landscape), and full unbiased prediction (only based on the energy landscape). By then, this also included optimal defect occupation structures, verification of alloy phase diagrams, and estimates of the stability of various modifications. In the time since 2000, energy landscapes have become ubiquitous all over chemistry, with even (*pace*, my experimental colleagues!) synthetic chemists getting very excited about the richness of chemical modifications present on the energy landscape. Now, we can study the landscape of even large molecules, or small to medium size molecules on surfaces (for an overview, see *e.g.* ref. 63), explore large clusters and the transition from nanocrystals to bulk materials, follow the dynamics of glassy and amorphous systems, and derive true free energy landscapes of crystalline systems, predict stable and metastable phase diagrams, and predict crystal structures of systems that exhibit complex hierarchical structures.

Of course, this two-slide overview must be very far from complete, and similarly, the references given here can only serve as starting points; *e.g.* we have not touched upon the prediction of phase diagrams (for an overview see *e.g.* ref. 64) or molecular crystals (a good starting point for a dive into the history of this field might be the presentation by J. Nyman (DOI: 10.1039/c8fd00033f). Nevertheless, seeing how many researchers from industry are attending this conference demonstrates the importance of the issue of complex energy landscapes of chemical systems and the prediction of their feasible structures in the "real" world outside academia, and I am very confident that we have by now reached a level of understanding of energy landscapes of chemical systems that will allow us to start assisting our experimental colleagues in their difficult task to synthesize the modifications we have been suggesting to them.

1 N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth and A. H. Teller, *J. Chem. Phys.*, 1953, **21**, 1087.
2 B. J. Alder and T. E. Wainwright, *J. Chem. Phys.*, 1957, 27, 1208.
3 J. H. Holland, *Adaptation in Natural and Artificial Systems*, Michigan Press, Ann Arbor, 1975.
4 E. L. Lawler and D. E. Wood, *Oper. Res.*, 1966, **14**, 555–756.
5 S. Kirkpatrick, C. D. Gelatt Jr. and M. P. Vecchi, *Science*, 1983, **220**, 671–680.
6 V. Czerny, *J. Optim. Theor. Appl.*, 1985, **45**, 41–51.
7 F. Glover, *Interfaces*, 1990, **20**, 74–94.
8 H. Müller-Krumbhaar, *Europhys. Lett.*, 1988, **7**, 479–484.

9 T. Zimmermann and P. Salamon, *Int. J. Comput. Math.*, 1992, **42**, 21–31.
10 J. Hertz, A. Krogh and R. G. Palmer, *Introduction to the theory of neural computations*, Addison-Wesley, Reading, 1991.
11 J. C. Schön and M. Jansen, in *Modern Methods of Crystal Structure Prediction*, ed. A. R. Oganov, Wiley-VCh, Weinheim, 2011, pp. 67–106.
12 D. J. Wales and J. P. K. Doye, *J. Phys. Chem. A*, 1997, **101**, 5111–5116.
13 J. C. Schön and M. Jansen, *Z. Kristallogr.*, 2001, **216**, 307–325.
14 A. Möbius, A. Neklioudov, A. Diaz-Sanchez, K. H. Hoffmann, A. Fachat and M. Schreiber, *Phys. Rev. Lett.*, 1997, **79**, 4297.
15 B. A. Berg and T. Neuhaus, *Phys. Rev. Lett.*, 1992, **68**, 9–12.
16 G. Dueck, *J. Comput. Phys.*, 1993, **104**, 86–92.
17 A. Colorni, M. Dorigo and V. Maniezzo, in *Proceedings of ECAL91 - European Conference on Artificial Life*, Elsevier, Amsterdam, 1991, pp. 124–142.
18 R. Eberhart and J. Kennedy, in *Proceedings of the Sixth International Symposium on Micro machine and Human Science*, IEEE, 1995, pp. 39–43.
19 D. Parsons and J. Canny, in *Proc. Int. Conf. Intel. Sys. Mol. Biol.*, 1994, p. 322.
20 I. Al-Bluwi, T. Simeon and J. Cortes, *Comput. Sci. Rev.*, 2012, **6**, 125–143.
21 A. Banerjee, N. Adams, J. Simmons and R. Shepard, *J. Phys. Chem.*, 1985, **89**, 52–57.
22 D. J. Wales, *J. Chem. Soc., Faraday Trans.*, 1993, **89**, 1305–1313.
23 P. Sibani, J. C. Schön, P. Salamon and J.-O. Andersson, *Europhys. Lett.*, 1993, **22**, 479–485.
24 J. C. Schön, *Ber. Bunsenges. Phys. Chem.*, 1996, **100**, 1388–1391.
25 J. C. Schön, H. Putz and M. Jansen, *J. Phys.: Condens. Matter*, 1996, **8**, 143–156.
26 C. Dellago, P. Bolhuis, F. S. Csaijka and D. Chandler, *J. Chem. Phys.*, 1998, **108**, 1964.
27 P. G. Bolhuis, C. Dellago and D. Chandler, *Faraday Discuss.*, 1998, **110**, 421–436.
28 G. Mills and H. Jonsson, *Phys. Rev. Lett.*, 1994, **72**, 1124.
29 H. Tanaka, *J. Chem. Phys.*, 2000, **113**, 11202.
30 G. T. Barkema and N. Mousseau, *Phys. Rev. Lett.*, 1996, **77**, 4358.
31 A. Laio and M. Parrinello, *Proc. Nat. Acad. Sci. U. S. A.*, 2002, **99**, 12562–12566.
32 T. Huber, A. Torda and W. F. von Gunsteren, *J. Comput.–Aided Mol. Des.*, 1994, **8**, 695–708.
33 G. M. Torrie and J. P. Valleau, *J. Comput. Phys.*, 1977, **23**, 187–199.
34 C. H. Bennett, *J. Comput. Phys.*, 1976, **22**, 245–268.
35 B. Andresen, K. H. Hoffmann, K. Mosegaard, J. Nulton, J. M. Pedersen and P. Salamon, *J. Phys.*, 1988, **49**, 1485–1492.
36 W. L. Jorgensen, *J. Phys. Chem.*, 1983, **87**, 5304–5314.
37 R. Zwanzig, *J. Chem. Phys.*, 1954, **33**, 1420.
38 J. G. Kirkwood, *J. Chem. Phys.*, 1935, **3**, 300.
39 M. Watanabe and W. P. Reinhardt, *Phys. Rev. Lett.*, 1990, **65**, 3301.
40 R. H. Swendsen and A. M. Ferrenberg, *Phys. Rev. Lett.*, 1989, **61**, 2635.
41 S. Kumar, D. Bouzida, R. H. Swendsen, P. A. Kollman and J. M. Rosenberg, *J. Comput. Chem.*, 1992, **13**, 1011–1021.
42 B. Roux, *Comput. Phys. Commun.*, 1995, **91**, 275–282.
43 F. Wang and D. P. Landau, *Phys. Rev. Lett.*, 2001, **86**, 2050.
44 J. Kim, J. E. Straub and T. Keyes, *Phys. Rev. Lett.*, 2006, **97**, 050601.
45 P. Liu and G. A. Voth, *J. Chem. Phys.*, 2007, **126**, 045106.
46 F. Heilmann and K. H. Hoffmann, *Europhys. Lett.*, 2005, **70**, 155–161.
47 K. H. Hoffmann and P. Sibani, *Phys. Rev. A*, 1988, **38**, 4261–4270.
48 P. Sibani, J. C. Schön, P. Salamon and J.-O. Andersson, *Europhys. Lett.*, 1993, **22**, 479–485.
49 F. H. Stillinger and T. A. Weber, *Phys. Rev. A*, 1983, **28**, 2408–2416.
50 J. C. Schön, H. Putz and M. Jansen, *J. Phys.: Condens. Matter*, 1996, **8**, 143–156.
51 M. A. C. Wevers, J. C. Schön and M. Jansen, *J. Phys.: Condens. Matter*, 1999, **11**, 6487–6499.
52 M. A. C. Wevers, J. C. Schön and M. Jansen, *J. Phys.: Condens. Matter*, 1999, **11**, 6487–6499.
53 J. C. Gower, *Biometrika*, 1966, **53**, 325–338.
54 T. Komatsuzaki, K. Hoshino, Y. Matsunaga, G. J. Rylance, R. L. Johnston and D. J. Wales, *J. Chem. Phys.*, 2005, **122**, 084714.
55 M. A. C. Wevers, J. C. Schön and M. Jansen, *J. Phys. A: Math. Gen.*, 2001, **34**, 4041–4052.
56 C. B. Li, A. Shoujiguchi, M Toda and T. Komatsuzaki, *Phys. Rev. Lett.*, 2006, **97**, 028302.
57 C. Levinthal, *J. Chim. Phys. Phys.–Chim. Biol.*, 1968, **65**, 44–45.
58 M. Goldstein, *J. Chem. Phys.*, 1969, **51**, 3728.
59 E. J. Corey, *Pure Appl. Chem.*, 1967, **14**, 19–37.
60 G. Polya, *Acta Mathematica*, 1937, **68**, 145–206.
61 J. C. Schön and M. Jansen, *Z. Kristallogr.*, 2001, **216**, 361–384.
62 D. J. Wales, J. P. K. Doye, M. A. Miller, P. N. Mortenson and T. R. Walsh, *Adv. Chem. Phys.*, 2000, **115**, 1–111.
63 J. C. Schön, C. Oligschleger and J. Cortes, *Z. Naturf. B*, 2016, **71**, 351–374.
64 J. C. Schön and M. Jansen, *Int. J. Mater. Res.*, 2009, **100**, 135–152.

Virginia Burger asked: What do you think is coming next, now that we are approaching 2020?

J. Christian Schön answered: 2020 is a bit too close to go for bold predictions! I would expect future directions in the field of molecular crystals to be strongly influenced by practical (industrial) needs: which information is needed to safeguard intellectual property regarding a system, and to make sure that no unknown polymorph interferes with the applicability of the compound in *e.g.* medicinal applications. Thus, I think estimates of lifetimes/transformation times under various conditions (temperature, humidity, competing medicines that are taken in addition to the one being discussed, *etc.*) will become an important focus, especially since we are talking about shelf-lives of years. (After all, many people keep old medicines in their cabinet and take them out again when a health problem re-appears!) Such estimates will supplement the stability checks performed in the laboratory, and provide information about the mechanism involved (which might be relevant for many other systems, too).

This is not only true for medicine –I recall at a conference five years ago that a company producing detergents and other cleaning agents asked me whether I could help them provide shelf-life estimates. (I did not have the time to take on this problem, but it seemed to be a big issue!) So I would think that structure prediction will branch out to many more chemical systems that have real-life applications.

A much more long-term issue is the one I mentioned earlier: the development of a large tool-box of synthesis modeling techniques to help synthetic chemists to perform target-oriented syntheses in an optimal fashion. For an outline of such a multi-year research program, see ref. 1.

1 J. C. Schön, *Adv. Chem. Phys.*, 2015, **157**, 125–134.

Sharmarke Mohamed opened a discussion of the closing remarks by Artem R. Oganov: It is very promising to see the range of structures and properties that can be predicted using the USPEX code. I was fascinated to see in your slides that multi-parameter optimisations using USPEX can lead to the discovery of novel hard materials and it is also interesting to see that you are letting the code sample all possible binary combinations in the periodic table when searching the range of materials that have a desired property. The composition (*i.e.* assumption of a binary molecular formula of the type MX) appears to be a user-defined parameter in the search algorithm. Can you comment on the feasibility and robustness of USPEX in facilitating the discovery of novel materials with desired properties when the composition of the material (*i.e.* molecular formula) is itself a variable in the parameter space? Clearly the search problem will be an order of magnitude more complex but as you have mentioned in your presentation, assumptions about chemical bonding are not always intuitive or correct so we need to also take into account possible variations in the molecular composition.

Artem R. Oganov responded: Indeed, when you include composition as a search variable, the search problem becomes much harder, but USPEX can deal with it quite efficiently – you can see numerous works that we and users of our code have done using this capability. Since 2010 USPEX has the capability to robustly predict stable compounds, and all you need is to specify the chemical

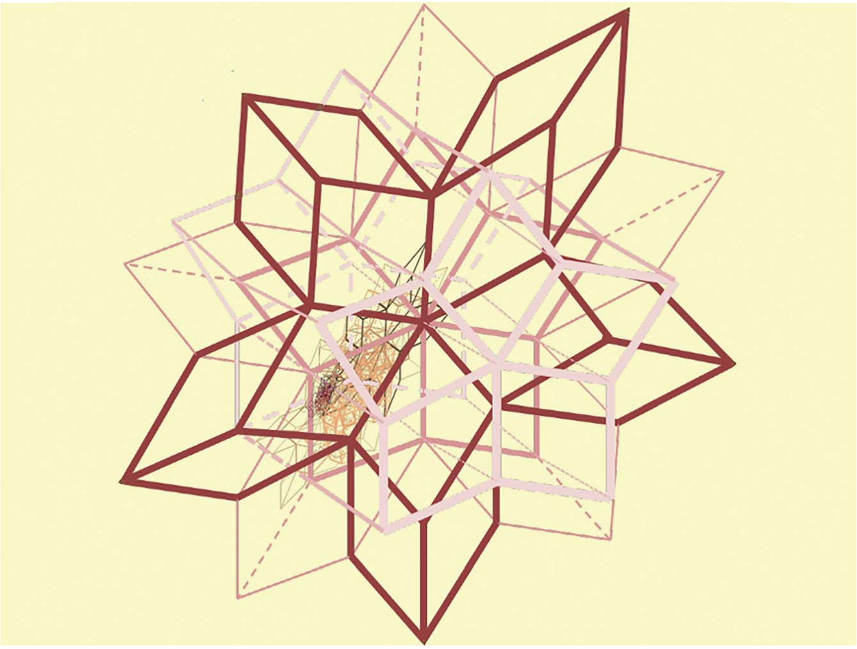


Fig. 6 A stellation of Kepler's rhombic triacontahedron. A rhombic hexecontahedral stellation of the Great Stella class, this non-convex polyhedron has 62 vertices and an internal vertex, with icosahedral symmetry. It comprises 20 golden rhombohedra (all prolate). The stellation is shown enlarged, with self-similarity, by Penrose in 3D. Given that with Shechtman icosahedral symmetry is now regarded as crystalline and that earlier an Al alloy was ascribed the stellar structure, it is reasonable to seek near-icosahedral aluminal species. The fractal nature of the 3D Penrose enlargement and its peculiar characterisation as asymmetric Cantor dust suggests the possibility of a quasi-crystal which may be nowhere dense.

elements. The main limitation is that this works well only for (pseudo)binary and (pseudo)ternary systems (see, *e.g.* an application to a ternary Mg–Si–O system:[1] when the number of compositional search variables becomes too large (>3–4), the number of possible compounds becomes too large).

[1] H. Niu, A. R. Oganov, X. Chen, D. Li, *Sci. Rep.*, 2015, **5**, 18347.

J. Christian Schön asked: In your nice presentation, you highlighted "search" and "energy ranking" as the two major features of crystal structure prediction. Let me add "stability" as the third big challenge that we should take a look at in crystal structure prediction research.

Artem R. Oganov responded: Stability is part of what I mentioned: "ranking" is the sorting of all sampled structures by energy, and "search" is finding the structure with lowest possible energy (*i.e.* the most stable structure). One can also discuss mechanical and dynamical stability, as well as chemical stability.

Alan Hare communicated: So, as regards a quasi-crystal, what does happen if we put exactly 63 Al atoms into a crystal predictor together, either with $Cu_{24} + Fe_{13}$, or with $O_{2+10\alpha+50\gamma}$ + $(OH)_{10+20\beta+200\gamma}$ + $(OH_2)_{40\varepsilon+50\eta}$, say?

Do we indeed see the icosahedritic icosahedron; or (as I've been expecting), a Great Stella stellation: the rhombic hexecontahedral stellation of Kepler's rhombic triacontahedron, perhaps (Fig. 6)?

Or do we get something else, entirely?

Artem R. Oganov communicated in reply: 63 Al atoms should give an fcc structure. Trying 24Cu+13Fe is a good idea, and it's definitely worthwhile.

Conflicts of interest

There are no conflicts to declare.