Structure prediction and its applications in computational materials design

Qiang Zhu,*[a] Artem R. Oganov,[a,b,c] Qingfeng Zeng[c] and Xiangfeng Zhou[d]

DOI: 10.1039/9781782622703-00219

1 Introduction

Structure is the most fundamental characteristics of a material. X-ray crystallography allows one to determine how atoms are arranged in a molecule and how molecules pack into a crystal. However, it requires a high-quality crystal sample, which is time consuming to prepare and often impossible under extreme conditions.

Theory has been playing a significant role in understanding crystal structures. Pauling summarized the rules for crystal structures of ionic solids,[1] however, similar powerful rules are still lacking for metals. First attempts to use computers to predict crystal structures date back to 1980. Although not very successful at the beginning,[2,3] crystal structure prediction (CSP) began to play an important role nowadays, thanks to many progresses in the last decade.[4–13] Indeed, mathematicians have developed algorithms to solve similar problems. Some of them are quite general and thus could be applied to crystal structure prediction. One can refer to a recent book[14] for a discussion of different methods. In this chapter, we will briefly introduce the modern structure prediction techniques, and review the recent developments in the context of the USPEX method, which is based on the evolutionary algorithm (EA), and has been viewed as a revolution in crystallography.[15] Discussions here follow closely the previous literature,[16–18] with primary focus on the most recent developments.

2 Methodology

2.1 Global optimization methods

Several global optimization algorithms have been devised and used with varying degree of success in structure prediction, for instance, simulated annealing,[4,5] metadynamics,[6,7] genetic and evolutionary algorithms,[8,13] random sampling,[9] basin hopping,[10] minima hopping,[11] and data mining.[12] Most of the methods mentioned above are developed to predict inorganic crystals and nano-clusters. However, the same philosophy can be applied to organic crystals[19–23] and proteins[24] as well.

[a] *Department of Geosciences, SUNY Stony Brook, Stony Brook, NY, USA. E-mail: qiang.zhu@stonybrook.edu*
[b] *Moscow Institute of Physics and Technology, Dolgoprudny city, Moscow, Russia*
[c] *International Center for Materials Discovery, School of Materials Sciences and Engineering, Northwestern Polytechnical Univerisity, Xi'an, China*
[d] *School of Physics, Nankai Univeristy, Tianjin, China*

Strictly speaking, all of the above methods (except metadynamics and data mining) are stochastic methods, thus they possess some inherent randomness. It is not guaranteed the one would obtain the same solution even by starting with the same set of parameter values. The divergence depends on the complexity of the landscape.

One either has to start already in a good region of configuration space (so that no effort is wasted on sampling poor regions) or use a "self-improving" method that locates, step by step, the best structures. The first group of methods includes metadynamics, simulated annealing, basin hopping, and minima hopping approaches. The second group essentially includes evolutionary algorithms. Alternatively, data mining approaches use advanced machine learning concepts and predict the structures based on a large database of known crystal structures.[12] Among all these groups of methods, the strength of evolutionary simulations is that they do not require any system-specific knowledge except chemical composition, and are self-improving, *i.e.*, in subsequent generations increasingly good structures are found and used to generate new structures.

2.2 Energy landscape

Before talking about the prediction of the crystal structure, let us first consider the energy landscape that needs to be explored. The dimensionality of the energy landscape is:

$$d = 3N + 3, \tag{1}$$

where $3N - 3$ degrees of freedom are from N atoms, and the remaining six dimensions are defined by the lattice. CSP is an NP-hard problem, *i.e.*, the difficulty increases exponentially with dimensionality. Yet, drastic simplification can be made if structures are relaxed, *i.e.* brought to the nearest local energy minima. Relaxation introduces intrinsic chemical constraints (bond lengths, bond angles, avoidance of unfavorable contacts). Therefore, the intrinsic dimensionality can be reduced:

$$d^* = 3N + 3 - \kappa, \tag{2}$$

where κ is the number of correlated dimensions, which could vary greatly according to the intrinsic chemistry in the system. For example, the dimensionality drops greatly from 99 to 11.6 for $Mg_{16}O_{16}$, while only slightly from 39 to 32.5 for $Mg_4N_4H_4$.[16] Thereby, the reduced difficulty of the problem (*i.e.*, the number of possible structures) is reduced:

$$C^* = \exp(\beta d^*). \tag{3}$$

This implies that any efficient search method must include structure relaxation (**local optimization**). We also note that all global optimization methods rely on the assumption that the reduced energy landscape will have a well-organized overall shape (Fig. 1), which is often true for chemical systems.[28]

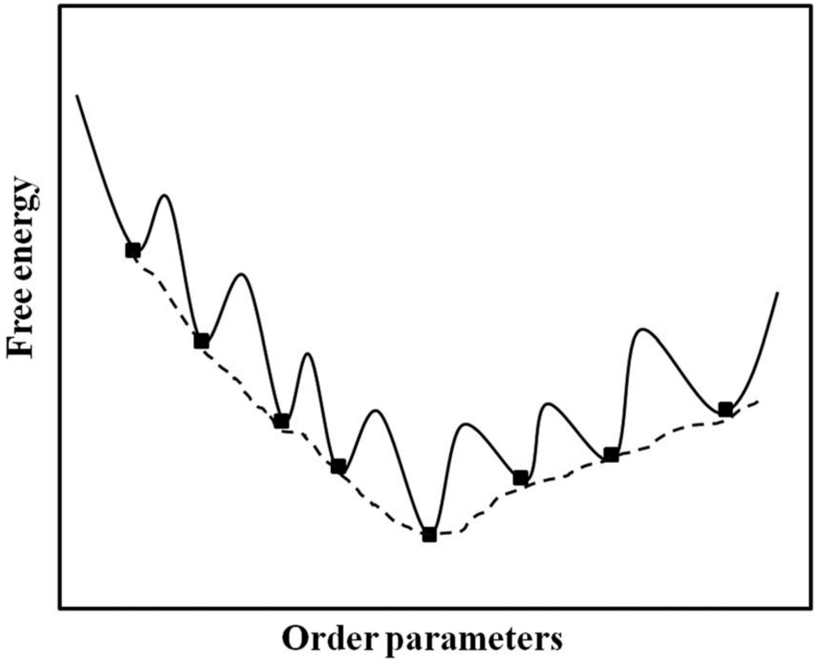


Fig. 1 An illustration of the simplified illustration of energy landscape. The idea of local optimization is to transform the noisy acted landscape (solid line) to a bowl-shaped reduced landscape (dashed line).

2.3 Evolutionary algorithm

Evolutionary algorithms (EA) mimic Darwinian evolution and employ natural selection of survival of the fittest and variation operators including genetic heredity and mutations. It is a stochastic method which is used to solve problems in which there exist many possible solutions for minima. The EA procedure is as shown in Fig. 2:

(1) Initialization of the first generation, that is, a set of structures satisfying the hard constraints are randomly generated;

(2) Perform structural relaxation and determine the quality (fitness) for each member of the population;

(3) Selection of the high-fitness members from the current generation as parents, from which the new generation is created by applying specially designed variation operators;

(4) Repeat steps 2–3 until the halting criteria are achieved;

(5) The above algorithm has been implemented in the USPEX (Universal Structure Predictor: Evolutionary Xtallography) package.[13–17]

2.4 Representation, fitness and variation operators

During evolution, it is important to keep the good structural features from the old generations to the next population. In traditional genetic algorithms, the investigated systems are usually expressed as an array of bits (chromosome), where each bit (gene) represents a different object. This representation behaves like DNA, and is quite convenient for

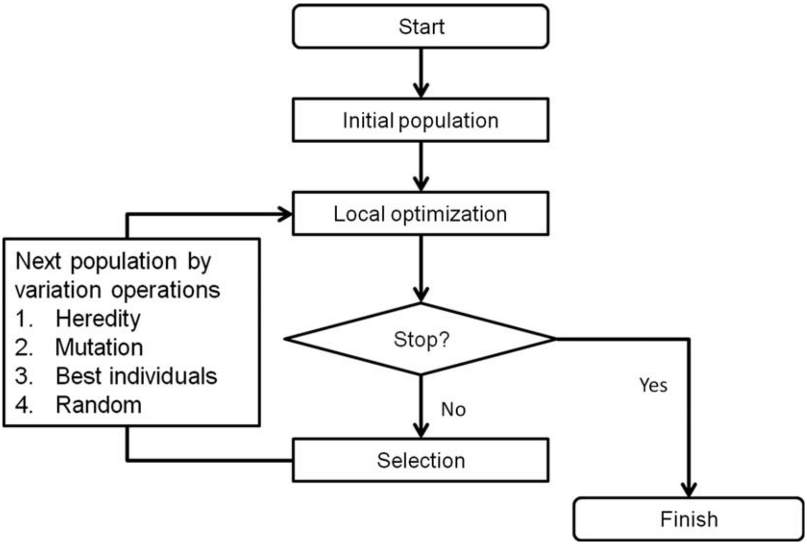


Fig. 2 The EA implemented in the USPEX code for crystal structure prediction.

variation operations (heredity and mutation). However, the disadvantage is that it involves encoding and decoding processes, which make it inconvenient to be applied to chemical systems, and most importantly, structural information is lost rather than transferred from parents to offspring. Deaven and Ho employed a real-space representation, and successfully applied it to the prediction of clusters.[27] The real-space representation in terms of Cartesian or fractional coordinates is more straightforward, and physically more meaningful. And USPEX adopts the **real space representation** as well.

The **Fitness function** mathematically describes the target direction of the global search, which can be either a thermodynamic fitness (to find stable states) or a physical property (to find materials with desired properties).

We rank structures by fitness values, a certain fraction of worst structures are discarded, and the rest are given a chance to be chosen as parents in the **selection** stage. The probability of the survived structures being chosen as a parent increases with the quality of a structure.

An essential step in an EA is to deliver the good genes to the next population, while introducing some variation. In USPEX, it is done *via* **variation operators.**

Heredity is a core part of the EA approach, as it allows communication between different trial solutions or classes of solutions by combining parts from different parents. In USPEX, to generate a child from two parents, the algorithm firstly chooses a plane which is parallel to one lattice plane, and then cuts a slice with a random thickness and random position along the other lattice vector; such slices from two parent structures are then matched to form a child structure. In this process, the number of atoms of each type is adjusted to ensure conservation of chemical composition.

Mutation operators use a single parent to produce a child. *Lattice mutation* applies a stain matrix with zero-mean Gaussian random strains to the lattice vectors; *soft-mode mutation* (which we call *softmutation* for brevity from now) displaces atoms along the softest mode eigenvectors, or a random linear combination of softest eigenvectors; the *permutation* operator swaps chemical identities of atoms in randomly selected pairs of unlike atoms.

A general challenge for global optimization methods is to avoid getting stuck in a local minimum and thus skip the global minimum. To prevent this, the key is to control the diversity of the population, by preventing proliferation of very similar structures and by adding new blood. For the latter purpose, we produce some fraction of each generation with the **random symmetric structures**.

Last, a certain number of best structures in the current generation (**best individuals**) are intentionally transferred to the new generation, and compete with others.

2.5 Fingerprints: a metric of structural similarity

In a global structure search, very similar structures always appear frequently. Duplicate structures do not only create inconvenience in post-processing, but also lead to the situation that the search is ‘trapped’ in some local minimum but not the ground state. Therefore, a technique to measure the similarities between structures is needed. In USPEX, we use the so-called fingerprint function[28] to describe a crystal structure. It is very similar to pair distribution function (PDF), which for an elemental solid is:

$$\mathrm{PDF}(R) = \sum_{i}\sum_{j\neq i}\frac{1}{4\pi R_{ij}^{2}\frac{N}{V}\Delta}\delta(R-R_{ij}) \tag{4}$$

where R_{ij} is the distance between atoms i and j, V is the unit cell volume, N is the number of atoms in the unit cell, and Δ is a bin width (in Å). The index i goes over all atoms in the unit cell and index j goes over all atoms within the cutoff distance from the atom i. The PDF at long distances oscillates around the value $+1$, which is not convenient for our purposes, and we subtract this “background” value for convenience. Generalizing to systems containing more than one atomic type, we introduce fingerprint as a matrix, the components of which are fingerprint functions for A–B type distances:

$$F_{\mathrm{AB}}(R) = \sum_{\mathrm{A}_i,\mathrm{cell}}\sum_{\mathrm{B}_j}\frac{\delta(R-R_{ij})}{4\pi R_{ij}^{2}\frac{N_{\mathrm{A}}N_{\mathrm{B}}}{V}\Delta} - 1 \tag{5}$$

One can measure the similarity between two structures by calculating the cosine distance between two fingerprint functions,

$$d_{ij} = 0.5\cdot\left(1-\frac{F_iF_j}{|F_i||F_j|}\right) \tag{6}$$

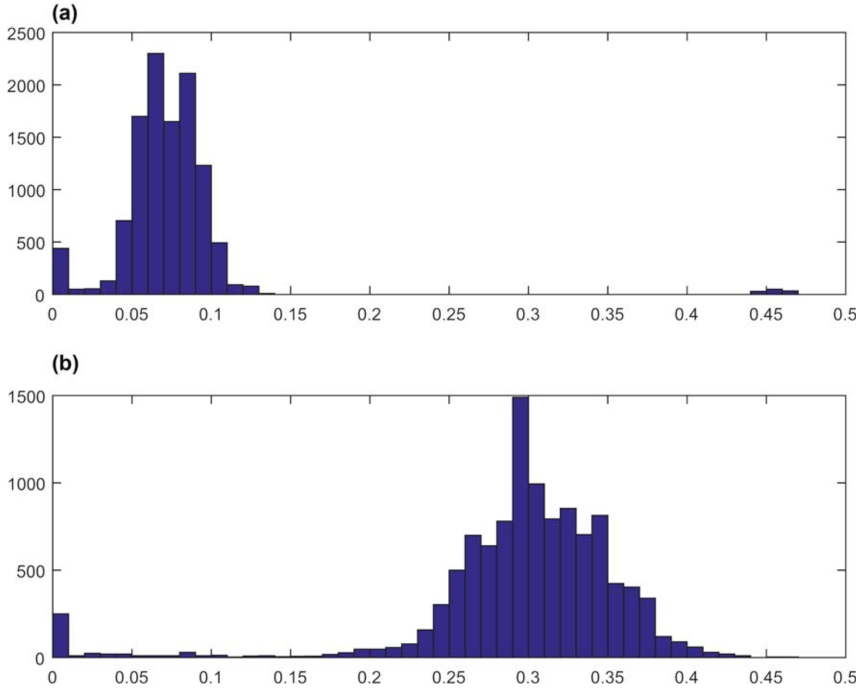


Fig. 3 Fingerprint distance distribution over 1000 structures from a typical USPEX simulation of 4 urea molecule per unit cell, when (a) including all distances (b) excluding intra-molecular distances in the fingerprint calculation for each individual structure.

Using this new crystallographic descriptor, we can improve the selection rules and variation operators above. During the selection process, only one copy of each distinct structure is used, and all its duplicates are killed. Fingerprint theory brings many other benefits (quantification and visualization of energy landscapes, use of ordered fragments of crystal structures, *etc.*).[25] However, it should be noted that the fingerprint function might be revised for different systems, in order to separate structures better. For instance, intramolecular contributions in fingerprint are identical for all different packing of the same molecule and thus decrease the discriminatory power of the fingerprint function (Fig. 3). Therefore, we only consider the intermolecular distances in the computation of the fingerprint function when dealing with crystals made of molecules with the same conformation.[26]

2.6 General choices of EA parameters

In any implementation of EA, the choices of parameters might lead to a different performance of the algorithm. Such parameters include: *population size, number of generations, fractions for each variation operation*. From our experience, a good choice of the population size should be ∼2 times of the number of atoms in the studied system, and the percentages should be 50%–70% for heredity, 20% for different types of mutations, and 15%–30% for random structures, respectively. Typically, such settings would be quite efficient. Therefore, we set another parameter (*stopCrit*) to stop the calculation if the best structure does not change for a given number of generations. We usually set 20 for *stopCrit* for system less than 40 atoms. However, larger values would be needed for larger systems.

A more rigorous way is to assign the fractions of the variation operation according to their performances during the calculation. In the current version of USPEX adopts the following strategy.

(1) At the end of each generation i, we discard identical structures, and select structures according to the fitness ranking;

(2) Within the selected structures, we count the origin of each structure, and obtain the fraction f_i*;

(3) Set $f_{i+1} = (f_i + f_i^*)/2$ for the next generation;

(4) In order to make it more robust, we also set the lower bound for those variation operations which have been proved very important from our experience. Here we put the minimum of 20% for heredity, 10% for random, 10% for softmutation.

According to our tests, the new scheme will generally enhance the searching efficiency by up to 2 times.

3 Recent developments

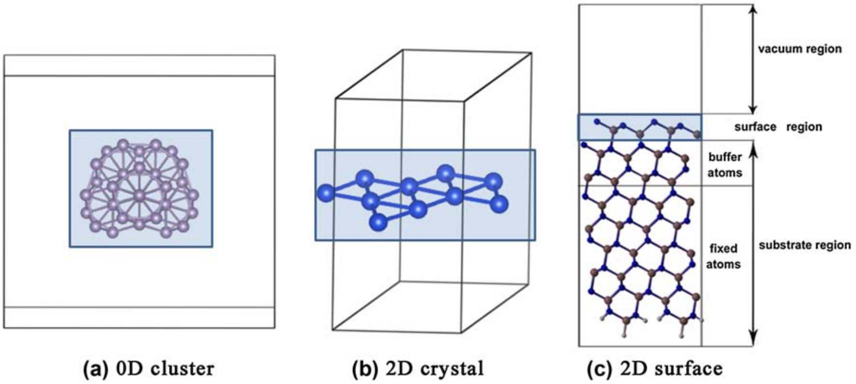
In the past years, we have extended structure prediction techniques to a broad range of systems. In this section, we will discuss the most recent developments. In order to give a complete demonstration, some developments which have been reviewed previously will be also briefly mentioned.

3.1 Choices of fitness functions

Traditional structure prediction is aimed at finding the structure with the lowest energy. Therefore, the fitness function is often defined as the energy or enthalpy. However, one can define the fitness in various ways, based on different applications. For instance, physical properties such as density, hardness and dielectric constants, can be used as the search criterion.[29–32] In these cases, a strategy of hybrid optimization is needed, that is, we search for the global optimum with respect to fitness, considering only structures corresponding to local energy minima. On the other hand, the so-called variable-composition prediction uses a modified energy criterion to evaluate the quality of structures over the whole allowed compositional space. In this case, stability of each individual can be defined as its decomposition energy relative to the easiest decomposition path.[33]

3.2 Low-dimensional systems

Comprehensive extensions of structure prediction in the most recent years are devoted to the low-dimensional systems, which include nanoparticles, two dimensional (2D) crystals, and surface reconstructions. For consistency with most widely used electronic structure codes, we treat



(a) 0D cluster (b) 2D crystal (c) 2D surface

Fig. 4 Two sets of cells in various low-dimensional systems. The small cells (highlighted in shadow) are used for global optimization, while the big cells (the whole structure model) are used for *ab initio* calculation.

the problem with periodic boundary conditions, by adding vacuum to eliminate the interactions with periodic images. Therefore, we have two types of cell representations in EA. Here the small cell represents the structure we want to optimize (excluding vacuum and substrate, this is the cell where variation operators work), and the big cell represents all the structural information (including vacuum and substrate, needed in structure relaxation) (Fig. 4).

3.2.1 Clusters. Several methods have been applied to cluster structure prediction.[11,27] Schönborn *et al.* translated the original version of the USPEX method to predict clusters.[34] However, with some new developments, this can perform even better.[17] The algorithm works as follows. The user gives a list of possible point groups (like, for example, C_2, D_{6h}, *etc.*) and nanoparticles are generated by randomly placing atoms inside the ellipsoid inscribed in the ''small cell'', and then replicating them using the point group symmetry operators. When the cluster is generated or relaxed, we place it in the center of the cell and rotate it so that principal moment of inertia axis with the highest moment is pointed in the *z*-direction. The ''big cell'' is then constructed by adding a certain amount of vacuum in all directions. The thickness of the vacuum region around the cluster is a user-defined parameter; more vacuum means more accurate results, but (for some approaches, such as plane-wave methods) greater computational costs. When interfaced with codes based on local basis set methods, the thickness does not strongly affect the *ab initio* calculation; however, it would be still convenient for performing the variation operations (such as heredity and lattice mutations). Before performing a 'cut-and-splice' heredity, the cut plane is randomly rotated around a random axis that goes through the center of mass of the nano-particle. This idea is similar to random 'shifts' for heredity in crystal structure prediction.

3.2.2 2D crystals. In recent years, 2D or quasi-2D materials have attracted great interests for their fascinating properties. Graphene, a single layer of carbon atoms with honeycomb configuration, has been widely studied due to its novel electronic properties (massless Dirac fermions, *etc.*). It has rapidly become a candidate for the next generation of faster and smaller electronic devices. Besides graphene, other 2D-crystals (such as MoS_2) with excellent properties were also discovered.[35] Quite recently, a systematic strategy for searching for flat 2D-crystals based on particle swarm optimization algorithm was proposed and applied to the B–C system.[36] However, it was later found that constraint of flat configuration will miss a number of meaningful systems. Thus, it was later extended to search for both 2D and quasi-2D materials.[37] In USPEX, we allow the non-planar configurations, and describe the system as a slab with a certain thickness. To initialize, the slabs of the 2D-crystal (small cell) are generated by random plane groups, and then the big cell is constructed by adding vacuum along the *c*-axis. After relaxation, we extract the slab from the big cell and apply the variation operators such as heredity and mutation. This allows one to explore more complex structures.

One must keep in mind that a 2D-crystal is always metastable and if allowed, will grow into a 3D-crystal. In other words, the greater the thickness is allowed, the lower-energy structures can be found. Thus, 2D-crystals give an example of constrained optimization, where the final results are determined by the constraint.

3.2.3 Surfaces. In practical calculations, the surface model includes three parts: vacuum, surface and substrate. Vacuum and substrate regions are pre-specified, while the surface region is optimized by the EA.[38] The number of surface atoms varies from zero to a given maximum number. Meanwhile, the cell size is also variable, in order to explore more complex reconstructions involving multiple unit cells. The fitness function needs to indicate the relative stability of structures with various surface stoichiometries and reconstruction cell sizes. We construct the fitness function based on the surface energy.

$$E_{\text{formation}} = E_{\text{total}} - E_{\text{ref}} - \sum_i n_i \mu_i, \tag{7}$$

where E_{total} and E_{ref} are the total energy of the surface under consideration and of the reference cleaved surface; n_i and μ_i are the number of atoms and chemical potentials for each species. The chemical potential is the energy needed to add or remove one atom from the system, assuming there is a reservoir for each species to equilibrate with. For a simple binary compound (AB), if μ(A) is extremely high, the elemental phase A would condense on the substrate. Therefore, the chemical potentials must satisfy constraints under equilibrium conditions as follows,

$$\begin{aligned} &\mu(\mathrm{A}) \leq \mu(\mathrm{A_0}), \\ &\mu(\mathrm{B}) \leq \mu(\mathrm{B_0}), \\ &\mu(\mathrm{A}) + \mu(\mathrm{B}) = G(\mathrm{AB}). \end{aligned} \tag{8}$$

At 0 K, the Gibbs free energy reduces to the internal energy $E(\mathrm{AB})$. Therefore, the chemical potential is bounded by

$$E(\mathrm{AB}) - \mu(\mathrm{B}_0) \le \mu(\mathrm{A}) \le \mu(\mathrm{A}_0). \tag{9}$$

Thus $E_{\mathrm{formation}}$ can be rewritten as dependent only on $\mu(\mathrm{A})$

$$E_{\mathrm{formation}} = E_{\mathrm{total}} - E_{\mathrm{ref}} - n_{\mathrm{B}}E(\mathrm{AB}) - \mu(\mathrm{A})(n_{\mathrm{A}} - n_{\mathrm{B}}) \tag{10}$$

This method can be employed in different ways (1) fixed number of surface atoms and cell size; (2) fixed number of surface atoms and variable cell sizes; (3) both variable surface atoms and surface unit cells. Here we emphasize the case of variable surface stoichiometry, as illustrated in Fig. 5. First, for two given surface configurations (I and II) which have different numbers of atoms on the same substrate cell, their relative energy differences is a function of the chemical potential $\mu(\mathrm{A})$ according to eqn (7). As shown in Fig. 5, surface I is stable when $\mu_{\mathrm{min}} \le \mu(\mathrm{A}) \le \mu_{\mathrm{eq}}$, while surface II is stable when $\mu_{\mathrm{eq}} \le \mu(\mathrm{A}) \le \mu_{\mathrm{max}}$. For any other unstable configuration, fitness can be viewed as the minimum energy difference compared with the stable configuration. The minimum condition is reached when $\mu(\mathrm{A}) = \mu_{\mathrm{eq}}$. It is useful to express it algebraically. Similar to Qian's work,[39] we define a term E_0 which is invariant to $\mu(\mathrm{A})$:

$$E_0 = E_{\mathrm{total}} - E_{\mathrm{ref}} - n_{\mathrm{B}}E(\mathrm{AB}) \tag{11}$$

Versions (a) and (b) in Fig. 5 contain equivalent information. Stable structures appearing on the phase diagram form a convex hull in energy-composition coordinates. The slope of each section in the convex hull is either the boundary chemical potential, or the equilibrium chemical potential in which stable configurations can coexist. Therefore, we can choose the fitness of a structure to be its distance to the convex hull. The EA search

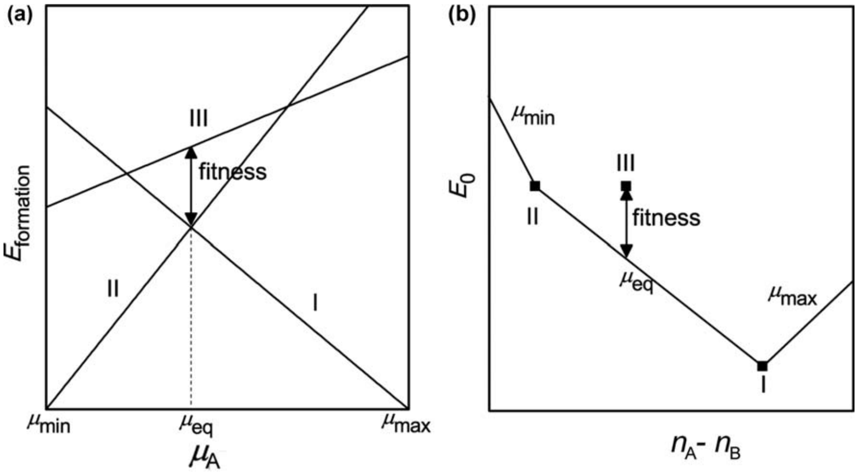


Fig. 5 Illustration of the fitness function used in surface prediction with variable stoichiometry in binary AB system. (a) Phase diagram as a function of $\mu(\mathrm{A})$. (b) Phase diagram as a function of $(n_{\mathrm{A}}-n_{\mathrm{B}})$. The vertices of the convex hull are the stable structures appearing in the phase diagram. The slope of each section is either the boundary chemical potential or the equilibrium chemical potential where stable structures coexist.

then aims to optimize the convex hull. When comparing structures with different surface cell sizes, the energies should be normalized.

3.3 Prediction of molecular crystal structures

The capability to predict molecular crystal structures has been implemented in the USPEX package since 2012.[26] When we adapted EA to organic systems, some concerns need to be addressed.

(1) *Metastablity*. Most of the molecular compounds are thermodynamically less stable than the simple molecular compounds from which they can be obtained (such as H_2O, CO_2, CH_4, *etc.*). This means that a fully unconstrained global optimization approach will produce only a mixture of these simple molecules, instead of the target molecular compounds of interest.

(2) *Weak interactions.* In organic crystals, packing largely depends on the weak inter-molecular interactions, such as hydrogen bonds and van der Waals interactions. These interactions are much weaker and softer than covalent bonds. Therefore, it leads to a very sparse molecular packing and a flat energy distribution in the real space. In this case, a method containing both efficient structural search and accurate energy ranking is needed.

(3) *Symmetry preference.* The distribution of structures over symmetry groups is very uneven. Most organic crystals are found to possess space groups: $P2_1/c$ (36.59%), P-1(16.92%), $P2_12_12_1$(11.00%), $C2/c$(6.95%).[40]

In order to apply EA to organic systems, it is essential to impose constraints, by fixing the bond connectivity and rigid angles; this can be conveniently done when molecular geometry is represented by internal coordinates (bond length, bond angle, torsional angle).[26] Here we introduce two other types of constraints which can be made for different systems.

3.3.1 Linear polymers.

In most of the polymeric crystals, the structure can be viewed as packing of polymeric chains. Provided the chain conformation is known, their packing can be described by (1) relative positions of chains; (2) rotational degrees of freedom associated with the lateral groups; (3) the orientation of the chains. For linear polymers, the mutual orientation of the chains can only be either parallel or anti-parallel. As shown in Fig. 6, we assemble the polymeric chains from the monomers by ensuring the neighboring contacts of these bridging atoms are close to the real situation (in terms of bond length and bond angle). Mathematically, the chain's orientation can be determined by the vector between the geometric centers of two connected monomers, C–C′. Thus we can reorient the linear chain in the (001) or (00-1) direction. In the structure initialization stage, we create a 2D primitive cell in the *a*–*b* plane for the geometric centers, according to the randomly assigned plane group symmetry. Then the *c*-axis is defined by the length of the chain, and the monomers are arranged either up and down around the centers in the 3D unit cell. Accordingly, the rotational axis is always fixed to the *c* direction. This linear chain mode has been applied to study the polymorphism of various systems such

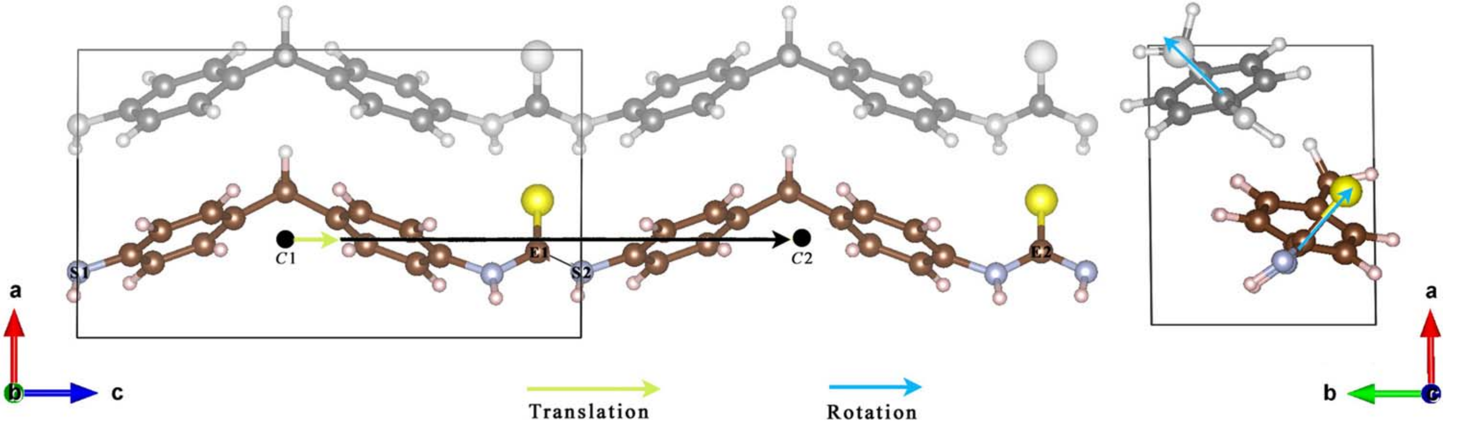


Fig. 6 Structure initialization of linear polymeric crystals. C and C′ are the geometric centers of monomers. The monomers are assembled in such a way that the C–C′ connections are parallel or antiparallel to the *c*-axis of the cell. The operations of translation and rotation will strictly act along *c*-axis.

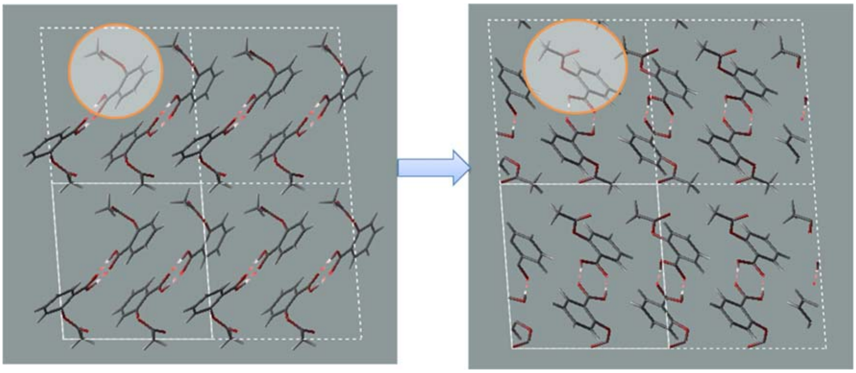


Fig. 7 Illustration of the symmetry-preserving rotational mutation in the case of aspirin. The parent structure has one asymmetric unit ($Z' = 1$) in $P2_1/c$ space group. The child structure is obtained by randomly changing the orientation of one molecule in the parent structure, while the rest molecules are generated according to the symmetry operations.

as poly(vinylidene fluoride) PVDF, and it proved to significantly speed up the searching process.[41]

3.3.2 Symmetry-preserving operations. Based on the fact that the symmetry distribution of organic crystals is very uneven, we design the mutation operators that keep the original symmetry. Since we generate the initial structures with random space groups, we keep the track of the asymmetric units and the corresponding symmetry operations. During mutation, we only perform mutations on those asymmetric units and then reconstruct positions and orientations of the remaining molecules by symmetry operations. As shown in Fig. 7, this operator can efficiently generate structures close to ground state even from a structure with high energy. However, one should use it with caution, as new symmetries are harder to find.

4 Applications

Structure prediction techniques have become increasingly important in materials research. In this section, we will focus on the applications in materials sciences, based on the methodology described above.

In all the calculations described below, global optimizations were done by the USPEX code, and the VASP code[42] was employed for local optimization (*i.e.* structural relaxation), using the PBE exchange-correlation functional[43] and the PAW method.[44] For the soft materials, van der Waals (vdW) dispersion and hydrogen bond are crucial factors in determining the crystal packing. There have been major efforts in improving the accuracy of vdW functional.[45,46] Here we use the Tkatchenko–Scheffler method[47] and optPBE88[48] functional which have proved to give results in satisfactory agreement with experimental data.[49]

4.1 Materials missed in the experiments: CsF_n compounds

Many materials exhibit quite complex chemistry under extreme conditions. Theoretical prediction plays an increasingly important role in this field. A number of new stroichiometric compounds have been predicted[50–56] and even confirmed by experiments.[51] On the other hand, there should be vast opportunities to discover new compounds even under ambient conditions as well. For example, many possible inorganic materials that consist of three or more elements are not studied yet. More surprisingly, chemical space of even a restricted subclass of materials made of two elements is far from being exhausted. Recent studies show a lot of brand new binary compounds which have not been reported yet.[57–60] For instance, our recent work predicted four viable ground-state compounds, with MnB_2, MnB, MnB_4, and another previously never reported MnB_3. Stimulated by the simulation results, the further experiments were able to verify them by annealed samples.[59] Similarly, in the well known Hf–C system, two additional compounds Hf_3C_2 and Hf_6C_5 were predicted to be stable.[57] Therefore, the systematic variable compositional predictions for those materials of interests are in great need.

Here we illustrate the power of the variable-composition prediction by its application to the Cs–F system.[61] Alkali halides *MX* have been viewed as typical ionic compounds, characterized by 1 : 1 ratio necessary for charge balance between M^+ and X^-. It was proposed that group I elements like Cs can be oxidized further under high pressure.[56] We perform a comprehensive study of the CsF–F system at pressures up to 100 GPa, and found extremely versatile chemistry.[61] Our calculation uncovers quite a different scenario (Fig. 8) from Miao's report.[56] A series of

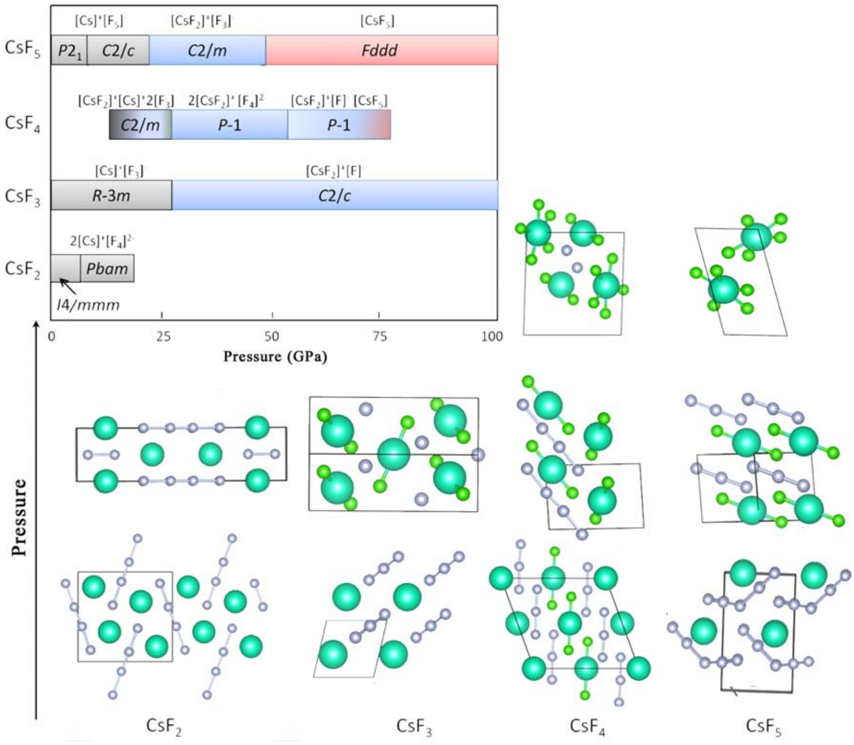


Fig. 8 The pressure-composition phase diagram of the CsF–F system and the corresponding stable crystal structures.

Table 1 Investigated reactions of the CsF–F system at ambient pressure conditions. The wt% gives the weight content of released F_2 gas. $\Delta H^{0\,K}$ and $\Delta H^{300\,K}$ are the calculated enthalpies at $T = 0$ K and 300 K, including the vibrational energies in kJ mol^{-1}. $\Delta S^{300\,K}$ is the corresponding formation entropy in J (K $mol)^{-1}$. T_c is the predicted decomposition temperature at standard atmosphere (1 bar). Note that F_2 is treated as crystalline solid at 0 K.

Reactions	wt%	$\Delta H^{0\,K}$	$\Delta H^{300\,K}$	$\Delta S^{300\,K}$	T_c(°C)
$CsF_2 = CsF + 1/2F_2(g)$	11.1	44.30	37.59	78.25	218
$CsF_3 = CsF + F_2(g)$	20.0	72.24	63.41	152.29	150
$CsF_5 = CsF + 2F_2(g)$	33.3	88.41	76.73	284.96	− 15

CsF_n ($n > 1$) compounds are predicted to be stable already at ambient pressure. Under pressure, 5p electrons of Cs atoms become active, with growing tendency to form Cs (III) and (V) valence states at fluorine-rich conditions. Although Cs (II) and (IV) are not energetically favored, the interplay between two mechanisms (polyfluoride anions and polyvalent Cs cations) allows CsF_2 and CsF_4 compounds to be stable under pressure.

Surprisingly, already at ambient pressure several stoichiometric compounds (CsF_2, CsF_3 and CsF_5) are calculated to be thermodynamically stable. The estimated defluorination temperatures of CsF_n compounds at atmospheric pressure (218 °C, 150 °C, −15 °C, respectively), are attractive for fluorine storage applications. Light halogens, fluorine (F) and chlorine (Cl), at normal conditions are highly reactive and toxic gases. For chemical industry and laboratory use, this presents great inconvenience. Their storage in the gaseous form (even as liquefied gases) is very inefficient, and compressed gas tanks may explode, presenting great dangers. At normal conditions, the volume of 22.4 litres (L) of pure fluorine gas weighs just 36 grams (g), illustrating the dismal inefficiency of storage in this form. To the best of our knowledge, no effective and safe fluorine storage materials are known. Both F and Cl have a huge range of industrial applications, which would benefit from such storage materials, especially if they can achieve high storage capacity, stability and reversibility (Table 1).

4.2 Property optimizations – HfO_2–SiO_2

As we discussed above, crystal structure search can be also property-oriented within the framework of hybrid optimization. In this case, the candidate structures should be locally optimized by energy, and globally selected and operated with respect to the target properties. Following this track, researchers have made significant steps towards to the materials design by properties (including density,[29] hardness,[30,62] band gap,[31,63] and so on). In such studies, the fitness function should be properly defined. Here we illustrate it by the example of searching for high-k dielectric materials.[32]

High-k dielectric materials are important as gate oxides in microelectronics and as potential dielectrics for capacitors. In order to enable computational discovery of novel high-k dielectric materials, we propose a fitness model (energy storage density) that includes the dielectric constant (κ) and an intrinsic breakdown field E_{bd}, and expressing the latter through the bandgap (E_{g}), we obtain the following formula,

$$F_{\mathrm{ED}} = \frac{1}{2}\varepsilon_0 k E_{\mathrm{bd}}^2 = 8.1882\ \mathrm{J\ cm^{-3}} \times k\left(\frac{E_{\mathrm{g}}}{E_{\mathrm{gc}}}\right)^{2\alpha}, \tag{12}$$

where $E_{\mathrm{gc}} = 4.0\,\mathrm{eV}$, the critical bandgap value separating materials into semiconductors and insulators, and ε_0 is the absolute permittivity of the vacuum. With this new fitness descriptor, we can simultaneously account for the dielectric constant, bandgap, and breakdown field during optimization, in a rational and comprehensive way. Remarkably, the same fitness can be used to search for optimal dielectric materials for capacitors and gate oxide materials.

We found a number of high-fitness structures of SiO_2 and HfO_2, some of which correspond to known phases and some of which are new. Our variable-composition searches in the HfO_2–SiO_2 system also uncovered several high-fitness states. The compositional dependences of enthalpy of formation and energy density are illustrated in Fig. 9. The highest FED is shown at each composition. The relationship between compositions and energy density appears to be intriguing (recall that two physical properties E_{g} and k, display quite different variation with respect to composition). A high concentration of HfO_2 does not necessarily result in high energy storage. As an example of a disordered structure, we take $Hf_{0.9}Si_{0.1}O_2$ (Hf_9SiO_{20}) with a relatively large unit cell containing 30 atoms; its dielectric permittivity is relatively high (22.11), but its low E_{g} (3.02 eV) results in very low fitness (33.93 $\mathrm{J \cdot cm^3}$). Ordered phases seem to be superior in terms of their fitness. Among the pseudobinary compounds, the best fitness values are seen for $Hf_{0.5}Si_{0.5}O_2$ ($I4_1/amd$) and $Hf_{0.75}Si_{0.25}O_2$ (I-$42m$); their fitness is three times greater than that of SiO_2 quartz. Clearly, further improvements are possible by considering other systems. The methodology and principles presented here allow a systematic search for such improved materials.

4.3 Nano clusters: B_{36} and LJ_{44}

Baturin *et al.* have applied this method to predict the atomic structure and stability of small silicon nanoclusters passivated by hydrogen.[64] Recently, the B_{36} clusters which was synthesized and theoretically studied by Wang *et al.*[65] Here we applied this method and reproduced the most two stable structures B_{36}–C_{6v} and B_{36}–D_{4h} within 10 generations (30 structures per generation). One can clearly see that B_{36} clusters prefer to adopt high-symmetry structures. For instance, its ground state has six-fold symmetry and a perfect hexagonal vacancy, while the next stable configuration (B_{36}–D_{4h}) has tetragonal symmetry. It is very important to note though, the algorithm does not favor only highly symmetric structures, as shown by tests on the artificial Lenard-Jones systems. LJ_{44} is an example of a cluster with a ground state that has no symmetry. As shown in Fig. 10, the algorithm is still able to identify the ground state even though we start from random symmetric structures. The reason is that

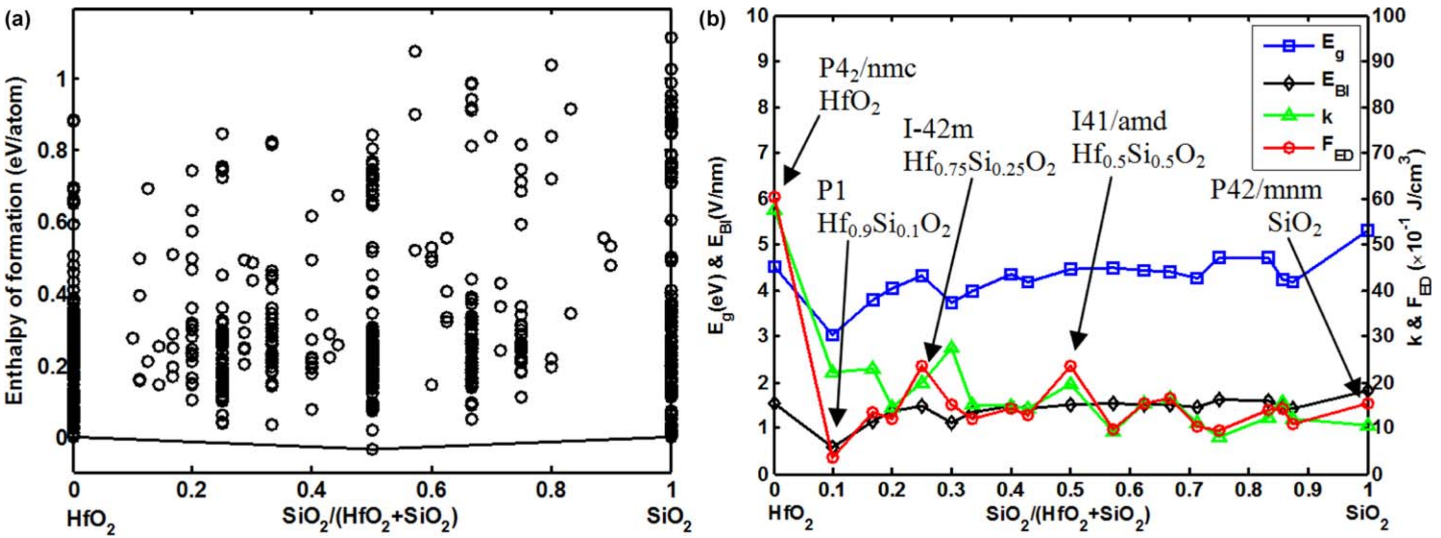


Fig. 9 HfO_2–SiO_2 system: (a) enthalpy of formation, showing stability of hafnon ($HfSiO_4$) and (b) compositional dependence of physical properties.

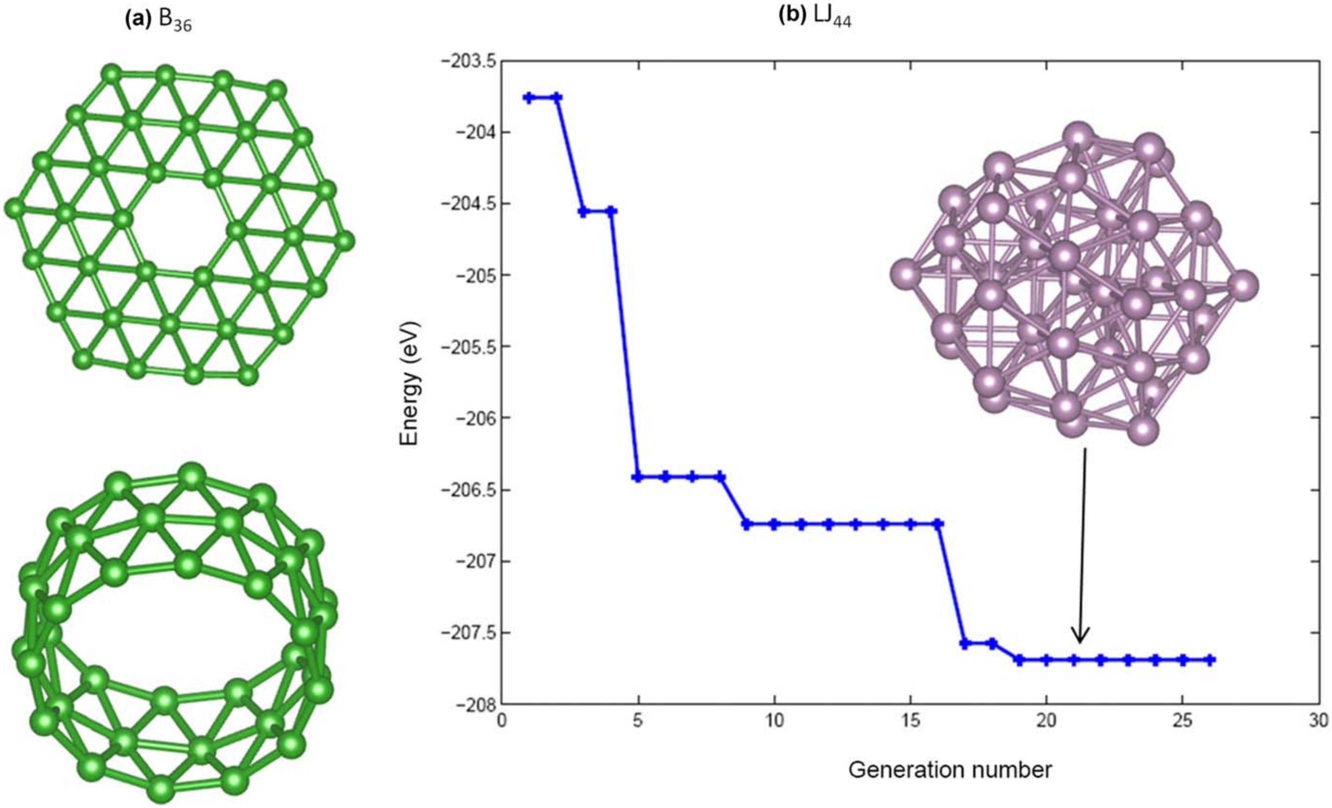


Fig. 10 (a) The best two structures found in B_{36} cluster, C_{6v} and D_{4h}; (b) a typical evolution curve of LJ_{44} cluster in USPEX simulation, showing the lowest energy in each generation.

variation operators in USPEX can break symmetry and enable totally new structures with different symmetry to emerge.

4.4 2D crystals: boron-based Dirac materials

Boron is a fascinating element because of its chemical and structural complexity. Recently, a new class of boron sheets composed of triangular and hexagonal motifs, exemplified by the so-called α-sheet structure, has been identified to be energetically most stable, and argued to be the precursor of B_{80} fullerene.[66] However, the stabilities of both α-sheet and B_{80} were both challenged.[67,68] We explored other potentially stable structures or structures with novel electronic properties. Contrary to the general constructing rules for flat monolayer boron sheets (mixing of triangular and hexagonal patterns),[66] great complexity is uncovered with multilayer structures. The non-flatness of 2D boron sheets enhances its energetic stability and creates novel electronic properties. In particular, we found that a 2D-boron with *Pmmn* symmetry can exhibit anisotropic Dirac cones,[69] after graphene and silicene,[70,71] the third elemental material with massless Dirac fermions. This property may be superior to that of graphene, because transport properties of these Dirac fermions will depend on direction, which gives an additional degree of freedom (with faster-than-graphene and slower-than-graphene directions) for electronic applications (Fig. 11).

4.5 Surfaces

4.5.1 Diamond (100), (111). We firstly studied the known 2 × 1 reconstructions of diamond (100) and (111) surfaces, which are the two most important surfaces for polycrystalline diamond obtained from chemical vapor deposition (CVD). We tried two and six carbon atoms on a 2 × 1 surface cell.[72] Our results are in excellent agreement with those reported in previous literature. The cleaved diamond (100) surface, containing one unsaturated carbon atom with two dangling bonds per unit cell, is unstable. Stabilization is achieved *via* a reconstruction with surface atoms forming one π-bonded C–C dimer per 2 × 1 unit cell. The diamond (111) surface contains two unsaturated carbon atoms with two dangling bonds per 1 × 1 unit cell. Our search also confirmed the model proposed by Pandey, with surface atoms forming Pandey chains along the [011] direction.[73] From the top view [Fig. 12(b)], the Pandey chains further form an extended two-dimensional (2D) network, having the same period as the unreconstructed (111) surface. From the side view, the surface atoms together with the second layer form an alternating (5 + 7)-ring pattern, which is different from the 6-ring pattern in the bulk, but is similar to the structures of *M*-carbon, a metastable carbon allotrope.[9]

4.5.2 GaN–O (1110). We also studied the semipolar GaN (10-11) surfaces in the presence of oxygen,[38] in which we allowed both variable number of surface atoms and variable cell size (restricted to a 2 × 2 or

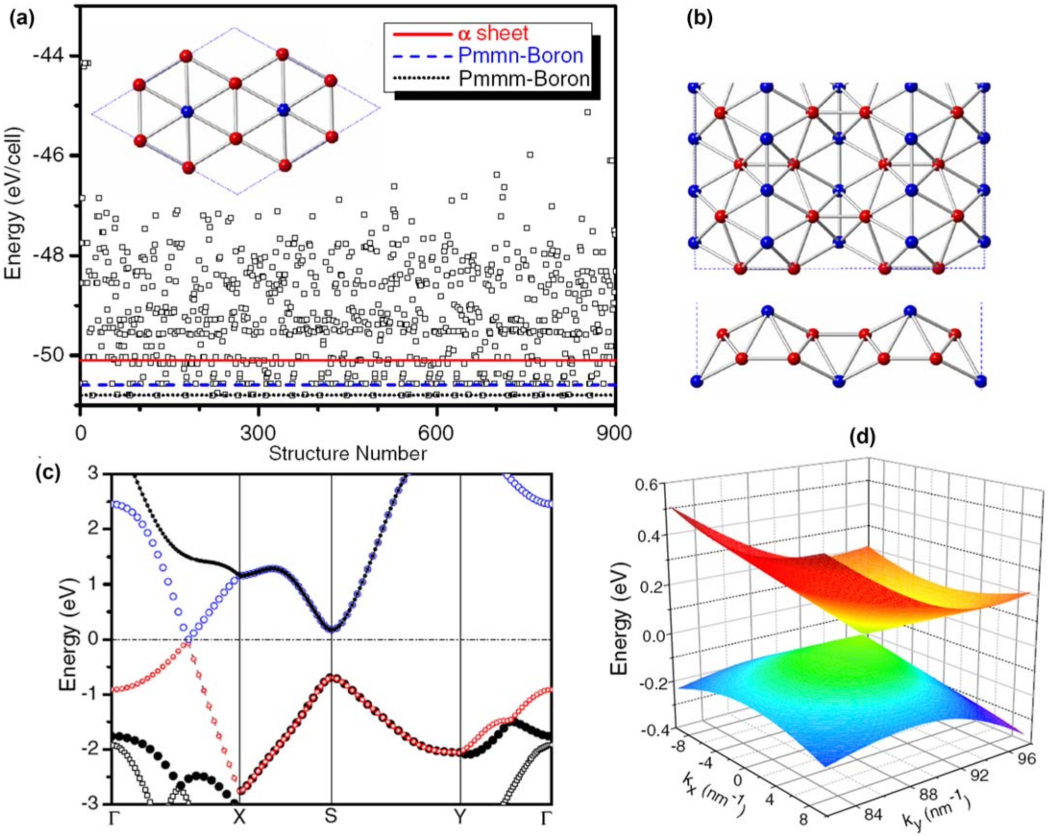


Fig. 11 (a) Enthalpy evolution for an 8 atom 2D boron system during an evolutionary structure search. The insets shows the structure of α-sheet; (b) the top view and side view of *Pmmn*-boron; (c) the band structure of *Pmmn*-boron; (d) the Dirac cone of *Pmmn*-boron.

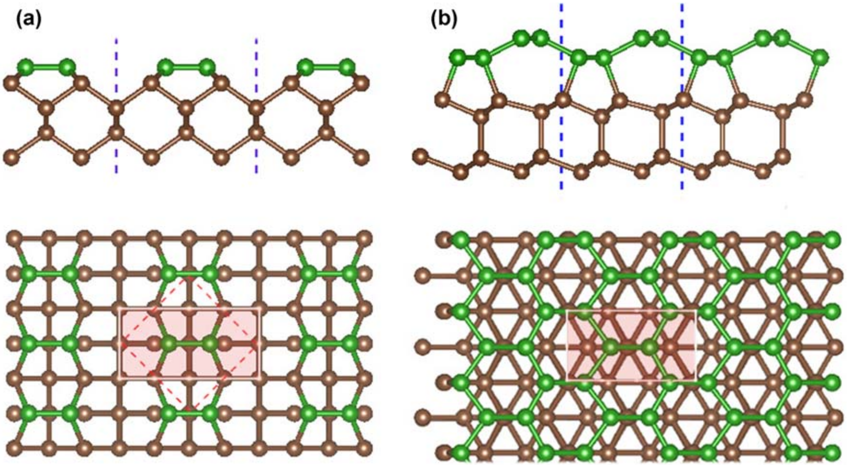


Fig. 12 Reconstructions on diamond (a) 2×1 (100) and (b) 2×1 (111) surface.

smaller surface cell). Figure 13 shows the whole phase diagram as a function of μ(O) and μ(Ga)-μ(N).

Compared to the cleaved surface, structure S1 has two Ga adlayers. Structure S2 has one Ga adlayer. Structure S3 has the top N and half of the second N layer removed. Structure S4 has only the top N layer removed. Structure S5 has an additional N at the bridging position of the two top N atoms. The first four structures were found by Akiyama *et al.*[74] Structure S5 with N_3 trimers is not intuitively obvious, and demonstrates the power of the automated searching by the EA. An analogous Se-trimer has been predicted to be stable on ZnSe (100) reconstructions in Se-rich conditions.[75] Two additional major reconstructions are structures S6 and S7, which appear in presence of oxygen. Compared to the cleaved surface, structure S6 has half of the top N layer removed, and half of the top N and all of the second N layer replaced by O. Structure S7 has the top two N replaced by O. Reconstructions similar to S6 and S7 for the (10-11) surface have been reported.[76]

4.5.3 α-Boron (111) surface. As a neighbor of carbon, boron is in many ways an analog of carbon and its nanostructures. The carbon surface has been thoroughly studied. In contrast, little is known for boron surface due to its exceptional structural complexity. Recently, Amsler *et al.* performed the first study of the reconstruction of the α-boron (111) and predicted several low-energy surface reconstructions by using the minima hopping method. In particular, a metallic reconstructed phase (111)-$I_{R,(a)}$ was predicted to be the most stable configuration, where a conducting boron sheet was adsorbed on a semiconducting substrate, leading to numerous possible applications in nanoelectronics.[77] However, this seems to be in conflict with the general principle the reconstructions usually lower their energies by atomic rearrangement leading to a semiconducting (rather than metallic) surface state.[78] Addressing this contradiction, we found an unexpected surface reconstruction in α-boron (111) using the *ab initio* evolutionary algorithm USPEX

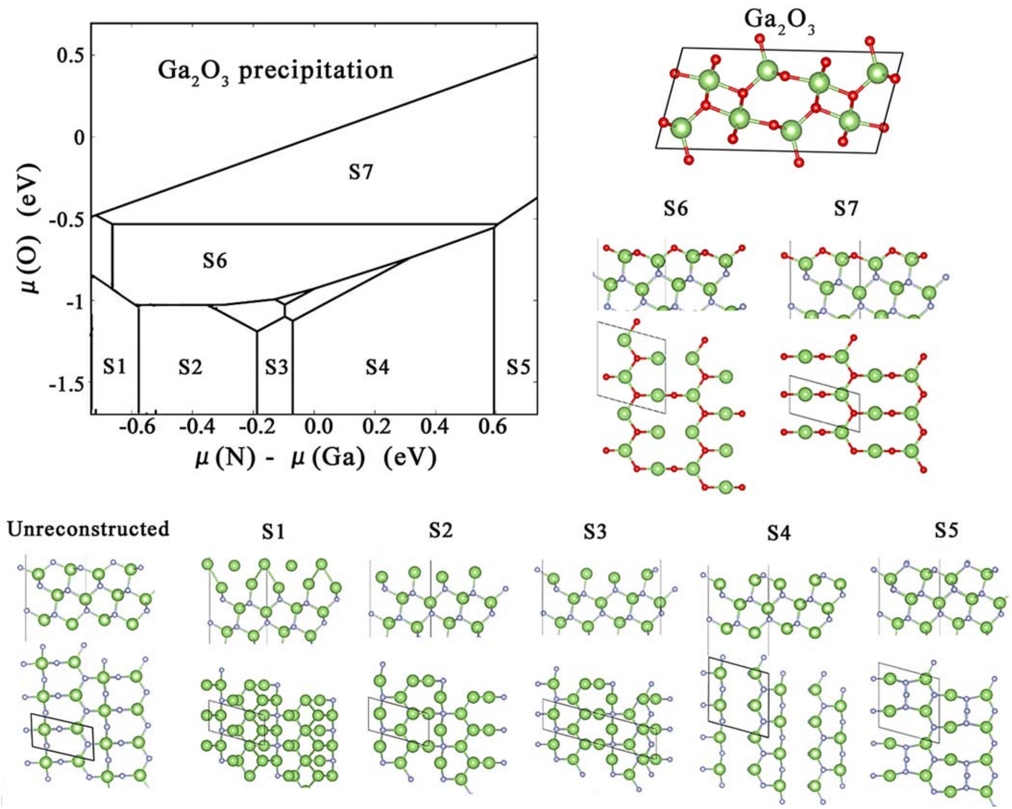


Fig. 13 The phase diagram of GaN (10-11) surface as a function of chemical potentials (μ(O) and μ(N)-μ(Ga)) and the corresponding stable reconstructions at various conditions.

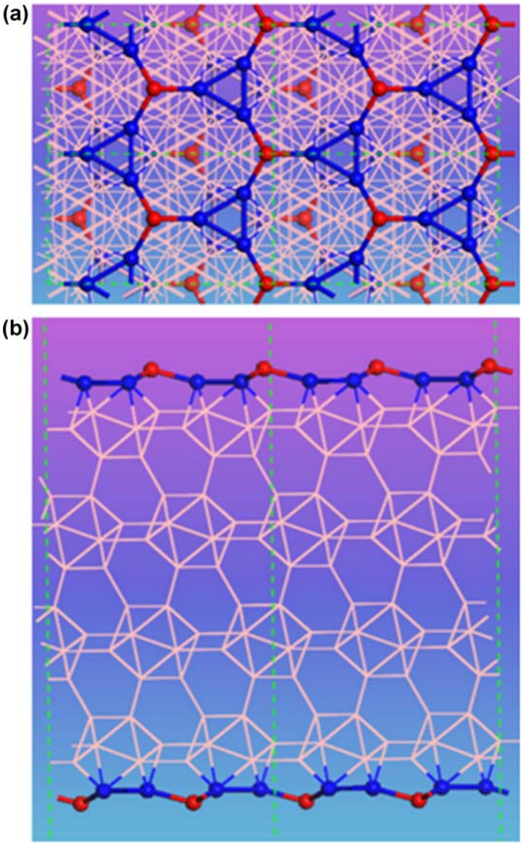


Fig. 14 (a) Projection of the 2 × 2 × 1 supercell of the (111)-$I_{R,(z)}$ structure along the [111] direction. (b) Projection of the 2 × 2 × 1 supercell of the (111)-$I_{R,(z)}$ structure along the [−1 −1 2] direction. The inequivalent surface atoms are shown by different colors.

(see Fig. 14). Our reconstruction has a much lower surface energy and is much simpler than previous predictions.[77] This reconstruction satisfied electron counting rules and is semiconducting.[79]

4.6 Polymers

This capability for prediction of polymeric crystals was developed only quite recently.[41] So far, we have systematically studied ten common polymers. And this module has been used for the design of dielectric polymers.[80] Here we illustrate the application to predict two complex polymers, nylon-6 and cellulose.

4.6.1 Nylon-6. Two crystalline forms of nylon-6 have been experimentally characterized, α and γ. There is a substantial confusion regarding the structure of the α-phase. The earliest reported crystal structure had some incorrect atomic coordinates in Cambridge structural database (CSD entry: LILSUU).[81] Here we used the model suggested in the previous theoretical studies,[82] which is described by the packing of the full-extended chains, possessing eight monomeric units of [–(CH_2)–CO–NH–] per unit cell, while γ-phase has $Z = 4$ based on twisted chains.

We have performed a search with $Z=8$ starting from the full-extended chain, and $Z=4$ starting from the twisted chain, in the hope of finding α- and γ-phase. Indeed, we found that the most stable configuration has a monoclinic symmetry for $Z=4$ (space group $P2_1/c$, $a=4.77$ Å, $b=8.35$ Å, $c=16.88$ Å (fiber axis), $\gamma=121.2°$, in good agreement with experimental results, except that there is a considerable deviation in cell vector b (the direction which largely depends on vdW bonding). In the γ-phase, the antiparallel twisted chains form pleated sheets *via* hydrogen bonds, and the chain directions are opposite in alternating sheets.[41] In our $Z=8$ search, we found the ground state which is very similar to what has been described in the literature.[82] This structure also features nylon sheets joined by H bonds in the antiparallel way.[41]

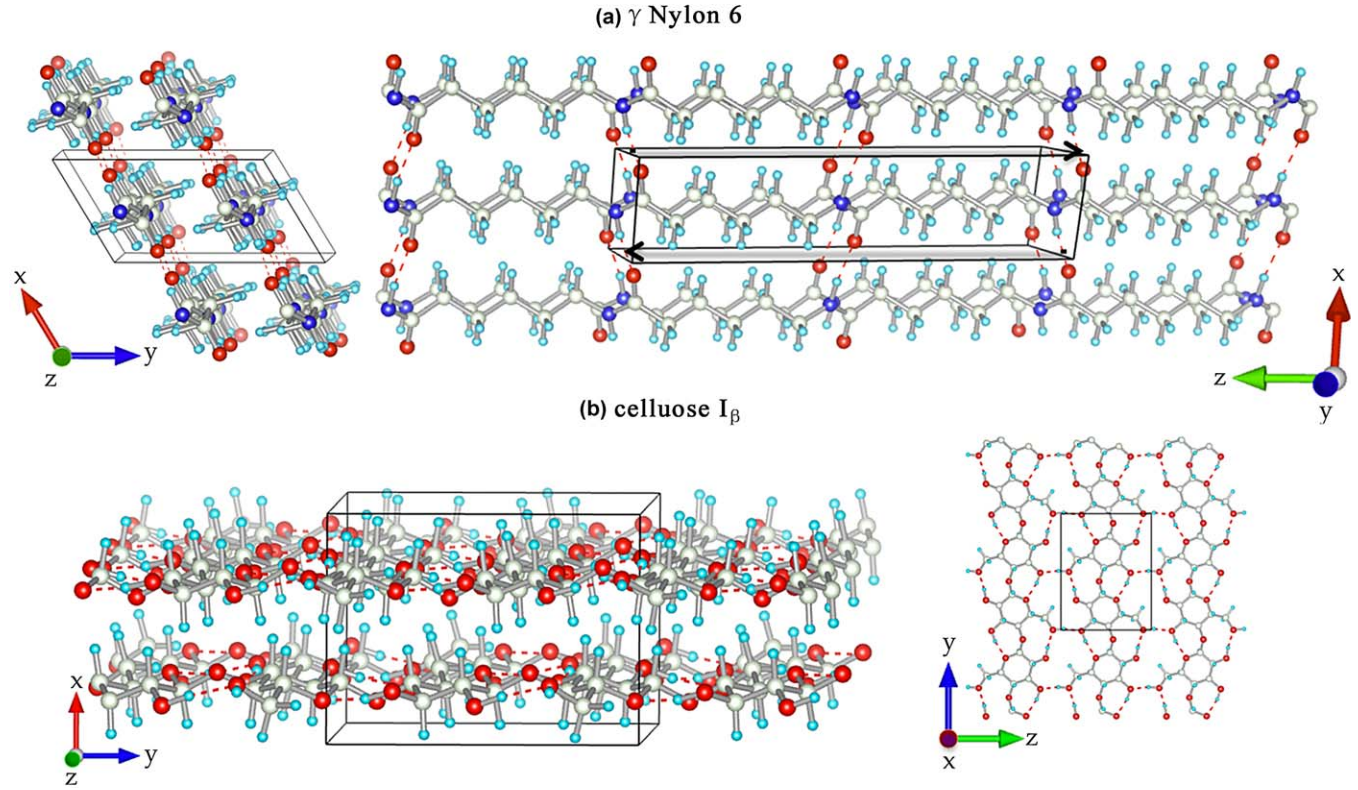
4.6.2 Cellulose. Cellulose is a polymer with repeating D-glucose units $[\text{–}C_6H_{10}O_5\text{–}]_n$. Microfibrils of naturally occurring cellulose correspond to two crystalline forms, I_α and I_β.[83,84] I_α has a triclinic unit cell and crystallizes in $P1$ space group. I_α has a simple unit cell and thus is easy to be predicted. Therefore we focused on the more challenging case of I_β. It was found that I_β is the thermodynamically more stable phase. Starting from the D-glucopyranosyl chains ($Z=4$), we indeed identified I_β as the ground state configuration, and the calculated unit cell parameters agree well with previous reports. As shown in Fig. 15(b), cellulose chains are arranged parallel-up and edge to edge, making flat sheets that are held together by H-bonds. Sheets formed by H-bonded D-glucopyranosyl chains are in the *bc*-plane, while there are no strong H-bonds which are perpendicular to the sheets. Most importantly, the complex hydrogen bond network in the flat sheets is also correctly predicted (Fig. 15(b)).

5 Outlook

We have briefly reviewed the principles of evolutionary algorithms and their application to structure prediction. The USPEX method proved to be a powerful tool enabling reliable and efficient prediction of stable crystal structures. In this chapter, we introduced the recent progress in extending the structure prediction technique to a wide range of problems. Despite its huge success in different fields, the current approach is still limited by the followings:

(1) *Energy accuracy*. So far, most applications in structure prediction are to find the most stable structures with the energy as fitness function. The performance is largely limited by the accuracy of todays's *ab initio* simulations, which for some cases is insufficient. For instance, van der Waals systems needs 1 kJ mole^{-1} accuracy to differentiate the crystal packing, which can be only achieved by using extremely expansive quantum-chemistry methods treatment.[85,86] Despite these significant progresses, it is still not feasible for massive structural searches.

(2) *Free energy versus lattice energy*. In organic crystals, the polymorph energy differences are often quite small. Among them, the lattice energy differences are typically very small. However, vibrational energy


(a) γ Nylon 6

(b) celluose I$_\beta$

Fig. 15 The crystal structures of (a) γ-nylon 6 and (b) cellulose I_β found by USPEX.

differences are often large enough to cause a re-ranking of polymorph stability at room temperature.[87] Therefore, it is particularly needed for organic crystal structure prediction to evaluate the free energy, instead of lattice energy at 0 K. Enhanced and rare-event molecular dynamics sampling techniques (metadynamics[88] and adiabatic free energy dynamics[89]) provide a solution, but choices of order parameter limit their applications for general purpose. The free-energy sampling techniques would be complementary to crystal structure search technique in practice. Still, it is needed to develop more accurate force fields to make it feasible to evaluate the free energy for the case of structure prediction.

(3) *Structural complexity*. We are suggesting USPEX as the method of choice for crystal structure prediction of systems with up to ~300 degrees of freedom (~100 atoms in the primitive cell for non-molecular crystals, and more for molecular crystals), where no information (or just the lattice parameters) is available. Above ~100 atoms per cell runs become expensive due to the "curse of dimensionality". However, some of them are still tractable by using the constraints (such as molecular geometry, lattice constants, *etc.*). Especially in such cases, interaction with experiment is helpful and should be encouraged.

USPEX has been applied to many important problems. Here we highlighted the methodology and some applications in the field of structure prediction. Another closely related subject is how to predict optimal conditions of synthesis of those predicted materials, which requires studies of chemical reactions and phase transition mechanisms. That direction of research is still wide open and we refer the reader to some of the first steps in it.[90–92]

Acknowledgements

Calculations were performed at the supercomputer of the Centre for Functional Nanomaterials, Brookhaven National Laboratory, and the High Performance Computing Center of NWPU. We gratefully acknowledge funding from DARPA (Grants No. W31P4Q1210008 and No. W31P4Q1310005), NSF (No. EAR-1114313 and No. DMR-1231586), the AFOSR (No. FA9550-13-C-0037), CRDF Global (No. UKE2-7034-KV-11), and Government of the Russian Federation (No. 14.A12.31.0003). Q-F Z thanks the National Natural Science Foundation of China (Grants No. 51372203 and No. 51332004), the Foreign Talents Introduction and Academic Exchange Program (Grant No. B08040). X-F.Z thanks the National Science Foundation of China (Grant No. 11174152), the National 973 Program of China (Grant No. 2012CB921900), the Program for New Century Excellent Talents in University (Grant No. NCET-12-0278), and the Fundamental Research Funds for the Central Universities (Grant No. 65121009).

References

1 L. Pauling, *J. Am. Chem. Soc.*, 1929, **51**, 1010.
2 J. Maddox, *Nature*, 1988, 335.

3 A. Gavezzotti, *Acc. Chem. Res.*, 1993, **27**, 309.
4 J. Pannetier, J. Bassas-Alsina, J. Rodriguez-Carvajal and V. Caignaert, *Nature*, 1990, **346**, 343.
5 J. C. Schon and M. Jansen, *Angew. Chem., Int. Ed. Engl.*, 1996, **35**, 1286.
6 A. Laio and M. Parrinello, *Proc. Natl. Acad. Sci. U. S. A.*, 2002, **99**, 12562.
7 R. Martonak, A. Laio and M. Parrinello, *Phys. Rev. Lett.*, 2003, **90**, 075503.
8 S. M. Woodley, P. D. Battle, J. D. Gale and C. R. A. Catlow, *Phys. Chem. Chem. Phys.*, 1999, **1**, 2535.
9 A. R. Oganov and C. W. Glass, *J. Chem. Phys.*, 2006, **124**, 244704.
10 C. M. Freeman, J. M. Newsam and S. M. Levine, *J. Mater. Chem.*, 1999, **3**, 531.
11 D. J. Wales and J. P. K. Doye, *J. Phys. Chem. A*, 1997, **101**, 5111.
12 S. Goedecker, *J. Chem. Phys.*, 2004, **120**, 9911.
13 S. Curtarolo, D. Morgan, K. Persson, J. Rodgers and G. Ceder, *Phys. Rev. Lett.*, 2003, **91**, 135503.
14 *Modern Methods of Crystal Structure Prediction*, ed. A. R. Oganov, WILEY-VCH, Weinheim, 2010.
15 S. L. Chaplot and K. R. Rao, *Curr. Sci.*, 2006, **91**, 1448.
16 A. R. Oganov, A. O. Lyakhov and M. Valle, *Acc. Chem. Res.*, 2011, **44**, 227.
17 A. O. Lyakhov, A. R. Oganov, H. T. Stokes and Q. Zhu, *Comput. Phys. Commun.*, 2013, **184**, 1172.
18 Q. Zhu, A. R. Oganov and X.-F. Zhou, Crystal Structue Prediction and its applications to Earth and Materials Sciences, *Topics in Current Chemisty*, Springer Verlag, 2014.
19 J. P. M. Lommerse, W. D. S. Motherwell, H. L. Ammon, J. D. Dunitz, A. Gavezzotti, D. W. M. Hofmann, F. J. J. Leusen, W. T. M. Mooij, S. L. Price, B. Schweizer, M. U. Schmidt, B. P. van Eijck, P. Verwer and D. E. Williams, *Acta Crystallogr., Sect. B: Struct. Sci.*, 2000, **56**, 697.
20 W. D. S. Motherwell, H. L. Ammon, J. D. Dunitz, A. Dzyabchenko, P. Erk, A. Gavezzotti, D. W. M. Hofmann, F. J. J. Leusen, J. P. M. Lommerse, W. T. M. Mooij, S. L. Price, H. Scheraga, B. Schweizer, M. U. Schmidt, B. P. van Eijck, P. Verwer and D. E. Williams, *Acta Crystallogr., Sect. B: Struct. Sci.*, 2002, **58**, 647.
21 G. M. Day, W. D. S. Motherwell, H. L. Ammon, S. X. M. Boerrigter, R. G. Della Valle, E. Venuti, A. Dzyabchenko, J. D. Dunitz, B. Schweizer, B. P. van Eijck, P. Erk, J. C. Facelli, V. E. Bazterra, M. B. Ferraro, D. W. M. Hofmann, F. J. J. Leusen, C. Liang, C. C. Pantelides, P. G. Karamertzanis, S. L. Price, T. C. Lewis, H. Nowell, A. Torrisi, H. A. Scheraga, Y. A. Arnautova, M. U. Schmidt and P. Verwer, *Acta Crystallogr., Sect. B: Struct. Sci.*, 2005, **61**, 511.
22 G. M. Day, T. G. Cooper, A. J. Cruz-Cabeza, K. E. Hejczyk, H. L. Ammon, S. X. M. Boerrigter, J. S. Tan, R. G. Della Valle, E. Venuti, J. Jose, S. R. Gadre, G. R. Desiraju, T. S. Thakur, B. P. van Eijck, J. C. Facelli, V. E. Bazterra, M. B. Ferraro, D. W. M. Hofmann, M. A. Neumann, F. J. J. Leusen, J. Kendrick, S. L. Price, A. J. Misquitta, P. G. Karamertzanis, G. W. A. Welch, H. A. Scheraga, Y. A. Arnautova, M. U. Schmidt, J. van de Streek, A. K. Wolf and B. Schweizer, *Acta Crystallogr., Sect. B: Struct. Sci.*, 2009, **65**, 107.
23 D. A. Bardwell, C. S. Adjiman, Y. A. Arnautova, E. Bartashevich, S. X. M. Boerrigter, D. E. Braun, A. J. Cruz-Cabeza, G. M. Day, R. G. Della Valle, G. R. Desiraju, B. P. van Eijck, J. C. Facelli, M. B. Ferraro, D. Grillo, M. Habgood, D. W. M. Hofmann, F. Hofmann, K. V. J. Jose, P. G. Karamertzanis, A. V. Kazantsev, J. Kendrick, L. N. Kuleshova, F. J. J. Leusen, A. V. Maleev, A. J. Misquitta, S. Mohamed, R. J. Needs, M. A. Neumann, D. Nikylov, A. M. Orendt, R. Pal, C. C. Pantelides,

C. J. Pickard, L. S. Price, S. L. Price, H. A. Scheraga, J. van de Streek, T. S. Thakur, S. Tiwari, E. Venuti and I. K. Zhitkov, *Acta Crystallogr., Sect. B: Struct. Sci.*, 2011, **67**, 535.

24 J. Moult, K. Fidelis, A. Kryshtafovych, B. Rost, T. Hubbard and A. Tramontano, *Proteins*, 2007, **69**, 3.

25 A. O. Lyakhov, A. R. Oganov and M. Valle, *Comput. Phys. Commun.*, 2010, **181**, 1623.

26 Q. Zhu, A. R. Oganov, C. W. Glass and H. T. Stokes, *Acta Crystallogr., Sect. B: Struct. Sci.*, 2012, **68**, 215.

27 D. M. Deaven and K. M. Ho, *Phys. Rev. Lett.*, 1995, **75**, 288.

28 A. R. Oganov and M. Valle, *J. Chem. Phys.*, 2009, **130**, 104504.

29 Q. Zhu, A. R. Oganov and M. A. Salvado, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 2011, **83**, 193410.

30 A. O. Lyakhov and A. R. Oganov, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 2011, **84**, 092103.

31 Y. Zhang, W. Gao and S. Chen, *Comput. Mater. Sci.*, 2015, **98**, 51.

32 Q. F. Zeng, A. R. Oganov, A. O. Lyakhov, C. Xie, X. D. Zhang, J. Zhang, Q. Zhu, B. Wei, I. Grigorenko, L. Zhang and L. Cheng, *Acta Crystallogr., Sect. C: Cryst. Struct. Commun.*, 2014, **70**, 76.

33 G. H. Johannesson, T. Bligaard, A. V. Ruban, H. L. Skriver, K. W. Jacobsen and J. K. Norskov, *Phys. Rev. Lett.*, 2002, **88**, 255506.

34 S. Schönborn, S. Goedecker, S. Roy and A. R. Oganov, *J. Chem. Phys.*, 2009, **130**, 144108.

35 J. Wang, S. Deng, Z. Liu and Z. Liu, *Nat. Sci. Rev.*, 2015, **2**, 22.

36 X. Y. Luo, J. Yang, H. Liu, X. Wu, Y. Wang, Y. Ma, S. Wei, X. Gong and H. Xiang, *J. Am. Chem. Soc.*, 2011, **133**, 16285.

37 Y. Wang, M. Miao, J. Lv, L. Zhu, K. Yin, H. Liu and Y. Ma, *J. Chem. Phys.*, 2012, **137**, 224108.

38 Q. Zhu, L. Li, A. R. Oganov and P. B. Allen, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 2013, **87**, 195317.

39 G. Qian, R. M. Martin and D. J. Chadi, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 1988, **38**, 7649.

40 C. P. Brock and J. D. Dunitz, *Chem. Mater.*, 1994, **6**, 1118.

41 Q. Zhu, V. Sharma, A. R. Oganov and R. Ramprasad, *J. Chem. Phys.*, 2014, **141**, 154102.

42 G. Kresse and J. Furthmuller, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 1996, **54**, 11169.

43 J. P. Perdew, K. Burke and M. Ernzerhof, *Phys. Rev. Lett.*, 1996, **77**, 3865.

44 P. E. Blochl, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 1994, **50**, 17953.

45 S. Grimme, *J. Comput. Chem.*, 2006, **27**, 1787.

46 J. Klimes and A. Michaelides, *J. Chem. Phys.*, 2012, **137**, 120901.

47 A. Tkatchenko and M. Scheffler, *Phys. Rev. Lett.*, 2009, **102**, 073005.

48 J. Klimes, D. R. Bowler and A. Michaelides, *J. Phys: Condens. Matter*, 2010, **22**, 022201.

49 T. Bucko, D. Tunega, J. G. Angyan and J. Hafner, *J. Phys. Chem. A*, 2011, **115**, 10097.

50 Q. Zhu, D. Y. Jung, A. R. Oganov, C. W. Glass, C. Gatti and A. O. Lyakhov, *Nat. Chem.*, 2013, **5**, 61.

51 W. Zhang, A. R. Oganov, A. F. Gonchanov, Q. Zhu, S. E. Boulfelfel, A. O. Lyakhov, E. Stavrou, M. Somayazulu, V. B. Prakapenka and Z. Konopkova, *Science*, 2013, **342**, 1502.

52 Q. Zhu, A. R. Oganov and A. O. Lyahkov, *Phys. Chem. Chem. Phys.*, 2013, **15**, 7696.

53 G. R. Qian, A. O. Lyakhov, Q. Zhu, A. R. Oganov and X. Dong, *Sci. Rep.*, 2014, **4**, 5606.
54 Y. Liu, A. R. Oganov, S. W. Wang, Q. Zhu, X. Dong and G. Kresse, *Sci. Rep.*, 2014, **4**, 5606.
55 C. Hu, A. R. Oganov, Q. Zhu, G. R. Qian, G. Frapper, A. O. Lyakhov and H. Y. Zhou, *Phys. Rev. Lett.*, 2013, **110**, 165504.
56 M. Miao, *Nat. Chem.*, 2013, **5**, 846.
57 Q. Zeng, J. Peng, A. R. Oganov, Q. Zhu, C. Xie, X. Zhang, D. Dong, L. Zhang and L. Cheng, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 2013, **88**, 214107.
58 X. Yu, G. B. Thompson and C. R. Weinberger, *Acta Mater.*, 2014, **80**, 341.
59 H. Niu, X. Chen, W. Ren, Q. Zhu, A. R. Oganov, D. Li and Y. Li, *Phys. Chem. Chem. Phys.*, 2014, **16**, 15866.
60 X. Cheng, W. Zhang, X.-Q. Chen, H. Niu, P. Liu, K. Du, G. Liu, D. Li, H. Cheng, H. Ye and Y. Li, *Appl. Phys. Lett.*, 2013, **17**, 171903.
61 Q. Zhu, A. R. Oganov and Q. F. Zeng, *Sci. Rep.*, 2015, **5**, 7875.
62 X. Zhang, Y. Wang, J. Lv, C. Zhu, Q. Li, M. Zhang, Q. Li and Y. Ma, *J. Chem. Phys.*, 2013, **138**, 114101.
63 H. Xiang, B. Huang, E. Kan, S. Wei and X. Gong, *Phys. Rev. Lett.*, 2013, **110**, 118702.
64 V. S. Baturin, S. V. Lepeshkin, N. L. Matsko, A. R. Oganov and Y. A. Uspenskii, *Eur. Phys. Lett.*, 2014, **106**, 37002.
65 Z. A. Piazza, H.-S. Hu, W.-L. Li, Y.-F. Zhao, J. Li and L.-S. Wang, *Nat. Commun.*, 2014, **5**, 3113.
66 H. Tang and S. Ismail-Beigi, *Phys. Rev. Lett.*, 2007, **99**, 115501.
67 N. Gonzalez, A. Sadrzadeh and B. I. Yakobson, *Phys. Rev. Lett.*, 2007, **98**, 166804.
68 S. De, A. Willand, M. Amsler, P. Pochet, L. Genovese and S. Goedecker, *Phys. Rev. Lett.*, 2011, **106**, 225502.
69 X.-F. Zhou, X. Dong, A. R. Oganov, Q. Zhu, Y. Tian and H. T. Wang, *Phys. Rev. Lett.*, 2014, **112**, 085502.
70 B. Aufray, A. Kara, S. B. Vizzini, H. Oughaddou, C. Leandri, B. Ealet and G. L. Lay, *Appl. Phys. Lett.*, 2010, **96**, 183102.
71 S. Cahangirov, M. Topsakal, E. Akturk, H. Sahin and S. Ciraci, *Phys. Rev. Lett.*, 2009, **102**, 236804.
72 J. Ristein, *Surf. Sci.*, 2006, **600**, 3677.
73 K. C. Pandey, *Phys. Rev. lett.*, 1981, **47**, 1913.
74 T. Akiyama, D. Ammi, K. Nakamura and T. Ito, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 2010, **81**, 245317.
75 A. Garcia and J. E. Northrup, *Appl. Phys. Lett.*, 1994, **65**, 708.
76 W. J. Lee and Y. S. Kim, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 2011, **84**, 115318.
77 M. Amsler, S. Botti, M. A. L. Marques and S. Goedecker, *Phys. Rev. Lett.*, 2013, **111**, 136101.
78 C. B. Duke, *Chem. Rev.*, 1996, **96**, 1237.
79 X.-F. Zhou, A. R. Oganov, X. Shao, Q. Zhu and H.-T. Wang, *Phys. Rev. Lett.*, 2014, **113**, 176101.
80 V. Sharma, C. Wang, R. Lorenzini, R. Ma, Q. Zhu, D. W. Sinkovits, G. Pilania, A. R. Oganov, S. Kumar, G. A. Sotzing, S. A. Boggs and R. Ramprasad, *Nat. Commun.*, 2014, **5**, 4845.
81 D. R. Holmes, C. W. Bunn and D. J. Smith, *J. Polym. Sci.*, 1955, **17**, 159.
82 Y. Li and W. A. Goddard, *Macromolecules*, 2002, **35**, 8440.

83 Y. Nishiyama, G. P. Johnson, A. D. French, V. T. Forsyth and P. Langan, *Biomacromolecules*, 2008, **9**, 3133.
84 Y. Nishiyama, J. Sugiyama, H. Chanzy and P. Langan, *J. Am. Chem. Soc.*, 2003, **125**, 14300.
85 J. Yang, W. Hu, D. Usvyat, D. Matthews, M. Schutz and G. K. Chan, *Science*, 2014, **345**, 640.
86 S. Wen and G. J. Beran, *Cryst. Growth Des.*, 2012, **12**, 2169.
87 J. Nyman and G. M. Day, *CrystEngComm*, 2015, **17**, 5154.
88 T.-Q. Yu and M. E. Tuckerman, *Phys. Rev. Lett.*, 2011, **107**, 015701.
89 P. Raiteri, R. Martonak and M. Parrinello, *Angew. Chem., Int. Ed.*, 2005, **44**, 3769.
90 Q. Zhu, A. R. Oganov and A. O. Lyakhov, *CrystEngComm*, 2012, **14**, 3596.
91 G. R. Qian, X. Dong, X. F. Zhou, Y. Tian, A. R. Oganov and H. T. Wang, *Comput. Phys. Commun.*, 2013, **184**, 2111.
92 S. Boulfelfel, A. R. Oganov and S. Leoni, *Sci. Rep.*, 2012, **2**, 471.