DOI: 10.1002/zaac.201400374

How can Databases assist with the Prediction of Chemical Compounds?

J. Christian Schön*[a]

Dedicated to Professor Martin Jansen on His 70th Birthday

Keywords: Database; Computational chemistry; Structure prediction; Structure-property relationships

Abstract. An overview is given on the ways databases can be employed to aid in the prediction of chemical compounds, in particular inorganic crystalline compounds. Methods currently employed and possible future approaches are discussed.

1 Introduction

Since its beginning as a systematic science over two hundred years ago, chemistry has had to face the overwhelming richness of the world of chemical compounds that are found in nature or are synthesized by the experimental chemist. Trying to keep track of all the compounds discovered has required a monumental effort, resulting in large book series such as *Gmelin's* "Handbuch der theoretischen Chemie",[1] and its successors, *Gmelin's* "Handbuch der Anorganischen Chemie"[2] and *Beilstein's* "Handbuch der Organischen Chemie",[3] or the more physico-chemically oriented "Landolt-Börnstein"-series.[4] With the advent of powerful computers, it has been a natural step to move these data compilations to electronic databases such as Reaxys,[5] the Cambridge Structure Database,[6] the Inorganic Crystal Structure Database,[7] or the protein data bank,[8] just to name a few. In these databases, one finds an overview over the possible modifications in chemical systems that have been discovered experimentally (plus some theoretically predicted structures).

From a crystal chemistry and materials chemistry point of view, it is of particular interest that experience has shown that there appear to exist some correlations between the structure(s) a chemical compound takes on and (some of) the other physical properties we observe for the compound.[9] Over the years a multitude of such structure-property relationships have been investigated – a quick search for publications with the keyword "structure-property relationship" in the title yielded over 100.000(!) responses – and often documented in appealing graphical form. Applications of such relationships range from molecules,[10,11] in particular biologically active molecules,[12] glasses,[13,14] aerogels,[15] and porous framework materials,[16] polymers,[17] proteins,[18] and thin films,[19] to inorganic (crystalline) solids.[20–23] Of course, there are many different types of properties, for which such relationships can be formulated, depending both on the type of chemical system and the application one has in mind. From the point of view of structure prediction, the "property" of interest is the kinetic (and thermodynamic) stability of a given structure-type in a chemical system. We would love to develop a robust map from a selected set of chemical properties of the atoms comprising a chemical system plus the "known" structure types, i.e. the ones extracted from all the experimentally observed and theoretically simulated chemical compounds, to the degree of kinetic stability of the chemical compound in the specified structure type.

The qualitative and often even quantitative success of employing specific structure – property relationships via interpolation and, to a certain degree, extrapolation, has led to optimistic suggestions that one can establish perfect structure-property relationships, which will produce easy ways for predicting new materials with specific desired properties.[24] In the final consequence, this raises the fundamental question of whether the "design" of chemical materials with specified properties is possible in the first place, or whether we are forced to choose from among the limited number of physically possible compounds on the admittedly rich table provided by nature according to the laws of physics.[25] In practice, we note, however, that there exist numerous examples where the same crystal structure is found, for example, for both insulators and metals, and thus these considerations, and especially the extrapolations involved, need to be taken with a large grain salt. Nevertheless, it is often reasonable to assume that if two compounds belong to the same general class of chemical systems, then they will share some of their properties, and thus some trends in these properties may truly be correlated with the various modifications possible in these systems. This might well allow us to predict some of the properties of not-yet-syn-

* Prof. Dr. J. C. Schön
E-Mail: schoen@fkf.mpg.de
[a] Max-Planck-Institut für Festkörperforschung
Heisenbergstr. 1
70569 Stuttgart, Germany

thesized compounds from the structure type they are predicted to crystallize in.

But this requires us to first take the important step of figuring out, whether a given structure type will constitute a stable modification in a particular chemical system. For practical purposes, this is equivalent to predicting, which possible crystalline modifications can exist in a given chemical system for a reasonable range of thermodynamic boundary conditions. Such modifications can be divided into two groups: those that agree with structure types that have been observed in other chemical systems (which may be either chemically similar to the system under consideration, or actually quite different and only show the same composition), and those structure types that have never been observed in any chemical system or only been realized in chemical systems that are chemically so different from the compound under consideration that "chemical intuition" will not suggest these structure types to us.

Clearly, in the case where a completely new structure type is going to appear during the synthesis of a new compound, a database can only be of very limited assistance in predicting the structure. In that case, we have to turn to fundamental physical principles that tell us that every (meta)stable chemical compound corresponds to a locally ergodic region on the energy landscape of a chemical system.[26–28] At low temperatures such a region corresponds to a local minimum of the potential energy in the space of atom configurations, but at elevated temperatures such a region can be quite large encompassing several or even many local minima. Over the past twenty-five years finding such local minima of the potential energy via global optimization methods has been developed to a certain proficiency, with different research groups proposing different global optimization algorithms (see, for example[28–31] and references cited therein).

However, the number of such local minima grows exponentially with the size of the system. Even if one excludes defect minima and amorphous structures by focusing on only small periodically repeated simulation cells containing up to, say, ten formula units, the number of minima can be quite overwhelming, and the danger that one misses an important modification increases with system size. Furthermore, since the ab initio energy calculations are too time-consuming for large global searches, we often need to employ simplified energy functions for the global search, refining the resulting minima afterwards with more accurate ab initio energy functions using Hartree-Fock- or DFT-based computer codes. In this case, information drawn from databases can be a valuable resource to heuristically guide us to find promising candidates for modifications in a given chemical system.

In this essay, we are going to outline some current and possible future approaches to structure prediction with the help of database analyses, specifically with respect to inorganic crystalline compounds. Many of the concepts and approaches discussed in this essay have already been presented in the literature in some fashion, together with applications. Thus, we are not going to discuss landscape exploration techniques or energy computation methods nor their applications to chemical systems in detail, but focus on providing a general presentation of the use of databases for structure prediction and closely related questions that might inspire the reader to proceed further along the directions mentioned.

2 Energy Landscape, Chemical Similarity, and Structure Prediction

2.1 Energy Landscape

The world of all conceivable atom arrangements for N atoms is called the configuration space ($N \approx N_{\mathrm{Avogradro}}$). A point in the configuration space can be visualized as a vector $\dot{R}$ with 3N coordinates [each atom contributes its position vector $\dot{r} = (x, y, z)$]. For each such configuration, we can compute the potential energy, and the 3N-dimensional hypersurface of the energy over the configuration space is the so-called (potential) energy landscape. As we know from classical mechanics, the dynamics of the chemical system is given by the forces acting on the atoms, i.e. the gradient of the potential energy. Of course, there are also quantum mechanical aspects such as zero-point vibrations, quantum tunneling, or spin degrees of freedom that should be considered, in principle; however, for the purpose of the present discussion, we will assume that the Born-Oppenheimer approximation holds and the effect of such additional features can be assumed to be small.

Clearly, if one picks such a vector $\dot{R}$ at random, the structure associated with it will be a random arrangement of atoms as one finds in the gas or liquid phase, and the ideal crystalline structures listed in databases such as the Inorganic Crystal Structure Database (ICSD) are singular points on the energy landscape that will never be seen by randomly picking points in configuration space. However, most of these randomly selected atom arrangements are quite high in energy, and physics tells us that the chemical system will preferentially occupy those regions of configuration space that are local minima of the free energy $F = E - TS$ at a given temperature (and on a given time scale[26–28]). Specifically, a metastable modification of a chemical compound corresponds to a locally ergodic region of configuration space, i.e. a region containing many (similar) atom configurations, which is locally equilibrated with a low free energy and kinetically stable enough such that the system does not leave this region, on experimental time scales.

Quite generally, the sets of configurations around a minimum or groups of minima on an energy landscape represent the locally ergodic regions at low or intermediary temperatures, respectively, and the kinetic stability of these regions grows with the height of the energetic and entropic barriers surrounding the region. Furthermore, the (crystal) structure that we associate with such a metastable modification is the average over all the configurations in the locally ergodic region; this average is, at least partly, reflected in the so-called thermal ellipsoids derived for the atoms in a crystal structure from the experimental data.

2.2 Structure Prediction

From a physical point of view, structure prediction is therefore equivalent to finding all the locally ergodic regions of a chemical system. Typically, one proceeds by first finding all minima, and then identifying the barriers surrounding them to estimate their stability. Usually, the focus is on the energetically low-lying minima, and thus the first step of the search is equivalent to a global optimization on a highly complex multi-minima energy landscape. Such a procedure is very time-consuming, the computational effort typically growing exponentially with the size of the system. This remains true even if one employs as many simplifications as possible, such as simulation cells containing few formula units with periodic boundary conditions and simple fast-to-evaluate cost functions instead of ab initio energies. As long as we are only interested in crystalline structures for systems where the energy landscape is well-approximated by empirical potentials, this simplification is reasonable, but if one wants to analyze amorphous or defect-controlled compounds much larger simulation cells are needed, of course.

Much progress has been made in developing efficient ways to perform such global optimizations for crystalline chemical systems, but as one knows from databases such as the ICSD, there exist many crystal structures that contain more than only a handful of formula units or cannot be well described by empirical potentials forcing us to employ semi-empirical or ab initio energy calculations instead. As a consequence, even highly refined global search methods cannot guarantee success within reasonable computation times. Furthermore, in general it is not sufficient to find only the global minimum, because in many instances it is a metastable modification, perhaps only realizable as a nano-crystal or thin film, which exhibits the desired properties – a fact well-known in experimental chemistry.[32]

One way to address this problem is by exploiting our knowledge about chemical systems collected in databases. In that way, we can hope to accelerate the search for "good" structure candidates in a not-yet-explored chemical system by extrapolating from other (related) systems, where already some crystalline modifications are known. While the purist might reject such heuristic procedures, it is obvious that one should not refuse any methodology that can assist us in identifying candidates for metastable compounds in a chemical system, even if it is only heuristic and clearly limited. In particular, if one's goal is not to prove the feasibility of a particular prediction methodology but to find all possible modifications for a specific given chemical system – or at least as many as one can in a limited amount of time –, then one should employ every method available. For more information about the many unbiased (i.e. without relying on information beyond the energy function itself) global optimization methods that are currently being employed, we refer to the literature (see references,[28–31] mentioned above), and for the remainder of this essay, we are only going to discuss approaches that rely on the availability of databases that contain large numbers of crystal structures of chemical systems.

2.3 Similarity of Chemical Systems

From an abstract point of view, every chemical system is different, and thus there is no a priori reason to assume that the possible modifications observed in one system are going to exhibit the same or even a similar structure as those found in any other system. And if one defines a "structure" via the exact cell parameters of a crystalline unit cell and the specific atom positions within this cell, the numerical values of these coordinates will be different from those found in any other chemical system. But the experience of the practicing crystallographer, chemist, and materials scientist has shown that among the structures observed in the world of crystalline compounds there are many which differ from each other only by slight changes in atom parameters and/or cell parameters, such that they appear to be the same, at least visually, once one disregards the specific chemical identity of the atoms involved. The crystallographic literature contains many algorithms that have been suggested to quantify this similarity, e.g. based on symmetry,[33] geometry,[34] or topology[35] considerations, each being most useful in specific contexts. But at the moment the decisive issue is that once we subsume all these similar structures under one structure type, we realize that instead of hundreds of thousands of crystalline structures, "only" tens of thousands remain, some of which are found to occur in hundreds of different chemical systems. As an aside we note that one such classification, where one assigns isopointal structure types according to symmetry considerations – with a refinement to isoconfigurational structure types based on additionally using cell parameters and atom coordinates – has recently been introduced in the ICSD.[36] By now most of the structures found in the ICSD have been assigned to one of these types.

From a physical point of view, the occurrence of the same structure type $\acute{T}$ in two different chemical systems means that the energy landscapes of these two systems each possesses a local minimum that belongs to this structure type $\acute{T}$. And if, as not infrequently happens, several of the modifications present in one of the systems are also observed in the second system, then it is a reasonable expectation that the two energy landscapes are going to be quite similar. Then one can hypothesize that many of the other modifications that are only found in one of the two systems are nevertheless capable of existence in the other one.

Clearly, such an educated guess is very helpful in our search for new modifications in a particular chemical system since we can right away perform a local minimization to verify whether this structure type is present as a kinetically stable modification in the second system, which is much faster than the global search necessary otherwise. We note here that this assumption of landscape similarity with the consequence of a highly desirable transferability of the results found for one landscape to another one has also been supported by many global landscape explorations.

One caveat is, of course, that in most cases the global search has been performed only with simplified energy functions. Thus we are comparing only models, i.e. approximate descriptions belonging to similar classes, e.g. hard-sphere two-body interaction models or all-electron ab initio calculations or density functional pseudo-potential models, of the true energy landscapes of the two systems. This might result in misleading similarities due to the inherent features of the models and inter-

Dedicated Cluster

actions. Thus, one must be aware that the true landscapes of the two systems are usually still going to be different in certain aspects: for example, not all minima found in one system will be present in the other, or the energy rankings of those minima present in both systems may be different.

From a chemical point of view, the existence of modifications with the same structure type in two different chemical systems can often be correlated with the chemical similarity of the two systems. Quantities such as differences in electronegativity, cation-anion-(size) ratios, or total valence electron concentration are commonly employed to categorize chemical compounds, and are used to post-dict, and sometimes even to predict, by analogy or "chemical intuition" (if the analogy is not an obvious one), the structure of a new chemical compound. This is a time-honored and very successful approach in experimental chemistry, as most famously demonstrated by the success of the periodic table that still is one of the most valuable organizational tools of the practicing chemist.

In this context, we note that from our experience with the energy landscapes of many chemical systems we have found that even for relatively small simulation cells with few atoms the minima that are associated with known structure types constitute not more than half of the local minima found.[37] But among the lowest-energy minima, this proportion can increase to 90 %, at least for systems with simple A_nB_m-type compositions. Of course, this mainly highlights the fact that, with regard to their synthesis, we have already carefully explored most of the binary systems with simple composition ratios *n:m*. Thus the number of unknown structure types in this group of systems is rather small (especially if one allows for a certain amount of distortions within a structure family). But this also supports the suggestion that the energy landscapes of chemically similar systems can show a relatively high degree of similarity. However, this high similarity is much less frequently found for compounds with more complex compositions.

3 Database aided Structure Prediction

3.1 Structure Prediction for Given Chemical Systems based on Structural Analogy

One obvious way to exploit the putative similarity of the energy landscapes of many chemical systems with the same composition formula consists in replacing the atoms in a known structure type by their analogues in the chemical compound of interest, and then to perform a local minimization and vibrational analysis, in order to check for the kinetic stability of this type in the new system. By repeating this procedure for all appropriate structure types found in the various databases, we can gain a certain overview over the minimum structures of the new system. A more refined way is to use these minima as starting points[38] for a global optimization technique such as the threshold algorithm,[39,40] which explores the configuration space below an energy lid that is accessible from a given energy minimum. By increasing this limiting lid, one can, in principle, globally explore the whole energy landscape.

We note that one important preliminary task required before such a data bank-based search can be performed is the identification of all structure types that are present in the databases, by using one of the comparison algorithms mentioned above. To a certain extent, the classification already provided by the ICSD can be employed as a starting point, but since this classification is based primarily on symmetry one needs to be careful. Hence, one would first classify all known structures into structure families by a geometry-based structure comparison, and then define the center of this group of structures as the structure father or structure type representative.[41] Once this has been achieved, one can take these prototypical structures as starting structures for further global searches or local minimizations, respectively.

Here, we note that for the use in energy landscape exploration and structure prediction, a geometrical comparison criterion is usually most appropriate: the metric that measures distances among neighbors in the real configuration space of atom arrangements is defined by sums over the Euclidean distances between corresponding atoms in the two structures – a geometrical quantity – while topological similarity is based on the bond network, which only rather indirectly correlates with the shape of the energy landscape, and symmetry-based similarity cannot really be mapped to the energy landscape at all. Nevertheless, following the dictum that one should not discard any systematic way to generate structure candidates, one should not hesitate to add all those structure types that have been gained on the basis of topological and symmetry arguments to those starting points that have been selected based on geometrical similarity.

In this context, one should keep in mind that there exist many chemical correspondences between e.g. binary and ternary, or higher, phases. Thus, it might well be that the unknown structure of the new modification or compound of interest in a binary A/B system might be identical to the structure of some known ternary A′/B′/C′ system, if one were to identify A′ with A and B′ and C′ with B! To explore this, we have in the past analyzed the ICSD with regard to similarity between structures and structure types in binary, ternary, quaternary, etc. systems.[41] This procedure will result in a family-relationship-tree that connects every structure not only with other structures of the same type (forming a structure family), but also with structures belonging to related families in the sense described above. This procedure can provide us with additional data bank-based structure types, even if they do not exist, strictly speaking, so far in form of a synthesized compound.

Trying to predict structures by chemical analogy in the way described above has been the oldest method in the literature.[42] If we consider the statistics mentioned above, it is no surprise that there have been quite a number of successes. This has especially been true in the field of high-pressure structures, where the knowledge that two chemically similar compounds have the same structure at standard pressure and that in system one there exists a high-pressure modification with a certain structure type, makes it a relatively easy target to predict that the second system will also exhibit a high-pressure structure with the same structure type. And rules such as the pressure-coordination rule[43] allow us a pretty well-educated guess about the similarity of the structures of chemically analogous systems, e.g. a high-pressure modification of system two exhibiting a structure identical with the one of the standard pressure modification of system one.

From the point of view of exploiting database information during a structure prediction study, this combination of chemical heuristics and hard structural data in the database has been a reasonably successful one, leading many people to the conviction that database mining is all that is needed for successfully predicting the structures of new compounds. In particular groups around *Curtarolo* and *Ceder*[44] have been developing this approach, culminating in the so-called Materials Genome Project.[45] And molecular crystal structure prediction and the prediction of secondary and ternary structures of proteins have also been guided by this kind of data-mining approach (for a review see, for example[27]).

In the case of molecular crystals, statistical analyses have shown that the symmetry groups for about 90 % of all structures of molecular crystals belong to only about ten different space groups, and that many of these structures can be described with only one or two molecules in the asymmetric unit. As a consequence, many search methods for molecular crystal structures exploit this information by e.g. systematically scanning all possible atom arrangements that can be generated by applying these space group symmetries to one or two molecules (where the scanning includes the systematic change of the cell parameters and the orientation of the molecules with respect to the cell axes and to each other).[46]

Similarly, for protein structure prediction, taking the primary sequence of the bases of a protein, and then comparing pieces of these sequences, or the whole sequence, to corresponding pieces of known protein structures, has proven to be quite successful. Following this approach, one can derive the secondary structure elements like the α-helices and β-sheets for the unknown protein, and even establish good guesses regarding the tertiary folded structure.[47–49] Again, the large number of already solved protein sequences and structures serve as a confidence-building foundation for such "structure prediction by analogy". One even has attempted to use statistical methods to estimate, how many "basic protein sequences" and "protein structures" are still missing,[50,51] although there is clearly a self-reinforcing feedback going on: proteins that are similar can be solved with regard to their sequence and structure by similar methods, thus giving too much weight to the "known" structures, while the set of the unknown, and also the unsolved, structures is more likely to contain more not-yet-known structure types than one would expect from extrapolation from the set of known and solved proteins and their structures.

Nevertheless, even with this caveat, it is clear that employing similarity analyses plus chemical intuition will quite likely continue to contribute to our ability to predict the structures of not-yet-synthesized compounds, and especially to the structure solution of synthesized but not-yet fully analyzed compounds. But one should not forget that in the case of bulk solid compounds the lack of simple but relatively rigid bonding rules as one finds in covalent compounds like molecules, makes it much harder to feel fully confident that extrapolation from known cases will yield the truly new structures expected in yet unexplored chemical systems. Again, the statistics is looking too good in some way, because those modifications that are easy to synthesize tend to be the ones that are similar to other chemical compounds whose structures are known. Thus the more complex compounds and the seemingly simple but not-yet-synthesized compounds are statistically more likely to exhibit new structure types never seen so far. These will often even be difficult to post-dict or "explain" after a synthesis, no matter how much one distorts simple structure motifs such as dense sphere packing or basic coordination polyhedra.

3.2 Prediction of Appearance of given Structure Types

After discussing the structure-prediction-by-direct-analogy approach, which is particularly well-suited to the straightforward use of databases, let us turn to some more uncommon ways of their employment. A complementary task where database information can be put to good use is the issue of predicting the appearance of a given structure type $\acute{T}$ (for an example, see reference[52]). Here, the goal is to figure out, which chemical system might support a particular modification that might e.g. fit structurally to a technologically useful layered compound, giving us more ways to fine-tune the properties of an electronic device. In some ways, this task is the dual to the standard structure prediction problem. Up to now, one usually has only relied on chemical intuition and trial-and-error experience, but computational capabilities have increased to the point where they can assist in this task.

Clearly, one can pursue the brute-force approach: Minimize the desired structure type for all chemical systems of correct composition type that would be compatible with the $\acute{T}$ structure type. This is usually quite a computational effort, especially if one keeps in mind that one needs to (a) also verify that the desired modification is kinetically sufficiently stable, and (b) perform at least some global search for every chemical system studied in order to gain an overview of how many competing modifications with lower energy exist in the system. Furthermore, in this brute force approach, one would not really exploit all the pre-knowledge one possesses regarding the structure type, which might be much more extensive than just the overall composition type. Thus, one would like to perform some pre-selection in order to focus on the most likely candidates.

Such a pre-selection should take both structural information about the desired structure type and chemical information about the systems considered into account. For example, certain coordination polyhedra might be present in $\acute{T}$, and thus one would want to first look at those chemical systems which are known to crystallize in structure types that also contain these polyhedra. Scanning the databases for compounds of the right composition exhibiting modifications where these polyhedra have been observed, would thus be a fruitful data-mining approach. Of course, this would require us to automate the search for such coordination polyhedra throughout the whole database, but with an efficient use of scripts and analysis algo-

Dedicated Cluster

rithms such as the Kplot-algorithm[53] this can clearly be achieved, as long as the database is available in a searchable form. In principle, one might even want to include such coordination-polyhedra information for each structure directly in the database itself, for future reference.

Similarly, chemical intuition will be of help by providing some heuristic rules such as the radius-ratio rule for ionic compounds – again, information regarding the ionic radii can be taken either from tables compiled in the past or be generated from a systematic perusal of the structures in the databases. Other kinds of chemical information would be the valence-electron concentration or the number of covalent bonds expected or involved for the structure type $\acute{T}$ – again a systematic pre-scanning of the chemical systems with respect to these attributes could be very helpful for restricting the search range.

Finally, a third way to attack this problem might be called the "reverse approach": In this method, we look at those chemical systems $S^{(k)}$ where a modification with the desired type $\acute{T}_1$ is known to exist, and find all additional structure types $\acute{T}_{i\neq 1}$ that appear as modifications in at least one of these systems. Now, we argue from a possible similarity of the energy landscapes: If in another chemical system $S^{(j\neq k)}$ (where we have not yet observed type $\acute{T}_1$) one of the modifications of structure type $\acute{T}_{i\neq 1}$ exists, then the likelihood is increased that in addition the modification with structure type $\acute{T}_1$ is also a local minimum, at least compared to the likelihood for a randomly selected system that only exhibits the same composition type. Thus, we select such a system $S^{(j)}$, and compute / relax all structure types $\acute{T}_i$ in this system, to see whether $\acute{T}_1$ might be stable in $S^{(j)}$, and how $\acute{T}_1$ compares with the other competing structure types $\acute{T}_{i\neq 1}$ with respect to the energy.

In most cases, we will find that the already known structures in system $S^{(j)}$ are the thermodynamically stable structures, but experience has shown that in many cases the desired structure $\acute{T}_1$ is at least kinetically stable and often quite competitive energy-wise. We note that this approach can e.g. also be used to suggest candidate systems where given high-pressure phases might be likely to exist.

3.3 Prediction of Multinary Phases

In many applications in materials science, one deals with complex multinary phases. Many of these actually constitute solid solutions, and while the prediction of their structure can be addressed by computational means (for a review see, for example[28]), here we will focus only on the case of ordered crystalline modifications where databases like the ICSD can serve as a resource. In general, we can employ such a database for structure prediction of multinary compounds completely analogously to the case we discussed above in the general structure-prediction-by-analogy section. However, there are not really enough such multinary structures available in the present databases to make this approach as reliable as it has been for binary and even ternary compounds; a similar problem is currently encountered in the protein community, where not enough solved RNA-structures are available for simple "prediction by analogy" of new RNA structures.[54] Thus, one needs to find different ways to tackle the prediction of the structures of multinary compounds.

The most promising approach consists in reversing the "family" analysis employed in the prediction by structural analogy, where we constructed e.g. possible A/B structure candidates by a contraction of known A′/B′/C′ structure types. Instead we now consider chemical systems A′/B′/C′, whose structure we expect to be related to some A/B-structure type (based on chemical / structural information, such as, for example, local coordination polyhedra), and assign the atoms of types A′, B′, and C′ to locations of atoms A and B in the known A/B-structure(s). Next, we minimize the energy for all the structure candidates generated in this fashion, and check their kinetic and thermodynamic stability. This approach for predicting structures of multinary phases from related binary or ternary phases by substitution of ions / atoms has a long tradition in experimental chemistry. Nevertheless, it is not clear, to what extent one has, in the past, really systematically exploited all the data available in the databases that can be used to predict new compounds in this fashion.

One issue one should keep in mind, is that there exists an enormous number of ways to perform such substitutions – even if one restricts the search to small unit cells –, and thus often chemical intuition, for what it is worth in this case, needs to be appealed to, in order to control the combinatorial explosion of possible substitutional atom arrangements. Sometimes, simple arguments based on e.g. minimizing the electrostatic repulsion of the ions will lead to a successful reduction in the number of possible configurations that need to be checked regarding their kinetic and thermodynamic stability. But in many cases, the number of candidate structures will still be overwhelmingly large.

Furthermore, as mentioned earlier, the chances of encountering a truly new structure type will increase with the complexity of the chemical system. For example, due to the slightly different ionic radii of the substituting atoms and their dependence on the local coordination by other atoms, it may well be that the real structure of the multinary system will contain a different overall arrangement of coordination polyhedra than those in any of the binary or ternary systems studied so far – or even exhibit completely new coordination polyhedra not yet found for the participating ions in known structures. Thus a deduction of the new structure type for A′/B′/C′ compounds from substituting atoms into one of the known A/B structure types cannot succeed in such a case.

This conclusion is somewhat distressing: it clearly demonstrates the limits of a structure prediction based solely on databases of known structures. However, one way to improve our chances to deal with multinary structures is by creating a new structure database that contains the results of systematic computational explorations of the energy landscapes of ternary and higher chemical systems. Here, we can take a leaf out of the early prediction work on binary systems with empirical potentials,[55] where the general methodology consisted of repeating the global searches for slight variations of the potential parameters about the ones deduced from database information such as the ionic radii. But what we now have in mind is that instead of considering only small parameter variations for one chemical system, one would explore the energy landscapes for a variety of composition types for a large grid of the parameters in the empirical potential.

Since all the "chemical information" about the system that can be taken into account during the global search is encoded in the choice of the potential parameters, covering a large grid of these parameter values is actually equivalent to studying a large variety of chemical systems! In contrast to the case of a single chemical system, where we carefully explored many variations of the potential parameters in the neighborhood of the "chemically sensible" choice, in the procedure proposed here, we are going to scan a large range of physically/chemically feasible situations by choosing a wider grid spacing in parameter space. While we would not "know" from a given set of parameters, which chemical system they best fit, we will obtain many possible "generic" structure types for A′/B′/C′ etc. compounds – the equivalent of the structure representatives we would have extracted from the databases otherwise –, which we then can use as starting configurations in subsequent local optimizations with ab initio codes or highly refined empirical potentials for any chemical system of interest.

3.4 Search for Missing Compounds

So far, we have discussed the use of a database to aid in the prediction of the existence and structure of specific chemical compounds, or, alternatively, the search for possible chemical systems, where a particular crystalline structure is realized. But we can also use a database as a negative-positive screening tool, i.e. we search the database for "holes", and try to use theory and/or experiment to fill these lacunae. What we mean by a "hole" are "missing" chemical compounds, i.e. we want to identify those chemical systems where one would expect a crystalline compound or a specific modification to exist but nothing is registered in any database so far.

Every practicing chemist knows of such "missing" compounds, the most famous being perhaps C_3N_4, which is one of the first compounds, for which a promising three-dimensional structure was predicted about twenty-five years ago;[42] so far only layer-type structures of this composition have been synthesized, however. But there are many other chemical systems, not only multinary ones but even some binary and ternary compounds, which one might expect to exist, but which have not yet been synthesized. So far, such systems have been identified in a more or less haphazard fashion by chemical intuition and personal preference, but by now the computational tools have become efficient enough for a systematic exploration of their energy landscapes once such promising missing compounds have been identified via analysis of various databases.

Of course, one can again follow a brute force route and e.g. for binary systems pick every possible A_nB_m composition registered e.g. in the ICSD for some chemical system, and perform local optimizations for every chemical system for all A/B-structure types found so far. In this fashion, we would be able to identify a plethora of kinetically stable binary compounds that would provide synthesis targets for the experimentalists. But while this might just be possible for binary systems, already for ternary systems it will be extremely expensive computationally, although the Materials Genome project mentioned earlier goes some way in this direction. Furthermore, most of the crystalline modifications predicted will be only marginally kinetically and thermodynamically stable, resulting in a truly momentous heap of essentially dead data.

Preferably, one would want to pre-select candidates by identifying the most promising combinations of chemical system, chemical composition and structural modification, which have not yet been synthesized, and then perform local ab initio energy minimizations for only these candidates, at least at the beginning. Such a preselection step should be rooted in a statistical analysis of the databases available, of course. In a possible first step, we might want to quantify the likelihood of the various bonding situations that appear in the various chemical systems and their compositions, being aware of not only the "average" or "typical" bonding situation but also of the outliers that might be the key to new and unusual structure types.

Once we have determined the range of reasonable compositions for a given chemical system, we can treat each such chemical system in the same fashion as we did for the standard structure prediction for a given chemical compound of interest: Look in the database for modifications that occur in more or less analogous chemical systems with the same composition, and use these as the first candidates for feasible polymorphs in the system under investigation. Finally, one should repeat this process by extending it from the "reasonable" to the "feasible" compositions (the outliers) in the chemical systems. One should note that any extrapolation beyond this would no longer be justified by the analysis of the database – we would either be back at the brute-force level of exhaustively working our way through all feasible chemical systems or have to rely on our personal "chemical intuition" for our choice of system of interest.

Of course, if we proceed like this, the usual caveats apply: if the thermodynamically stable modification in one of the chemical systems we study does not have an analogue among the already synthesized (or perhaps simulated!) crystalline structures, then we are stumped. Furthermore, concerning the thermodynamic and kinetic stability of a given modification, we need to be conscious of the fact that in most ternary or higher chemical systems a modification with a particular composition will not only compete with other modifications with the same composition but also with compounds of neighboring chemical composition in the same overall chemical system. If the modification under consideration is thermodynamically unstable with respect to decomposition in two neighbor phases, it will be crucial to determine its kinetic stability against such a decomposition. But estimating the kinetic stability, while allowing compositional changes and/or a phase separation for a fixed overall composition, is highly non-trivial. Such an estimate is computationally an order of magnitude more expensive than the one for the kinetic stability with respect to a transformation into another modification with the same composition. Still, a systematic perusal of the available databases for "missing" chemical compounds should be of great help in our attempts to explore and understand the world of crystalline chemical compounds.

3.5 Network Structures

Up to now, we have only talked about the whole crystal structure of a chemical compound, and its similarities with other compounds listed in the various databases, and how this information can be directly exploited for the purpose of structure prediction. But there is another aspect of structural similarity frequently employed when trying to understand crystal structures: the local coordination polyhedra around cations and/or anions formed by their counterions, or the presence of complex anions and cations that can similarly be represented by a "rigid" polyhedron. Superficially similar to the local bonding coordination of atoms in molecular chemistry, one often finds the same local building blocks in many compounds that contain the same types of atoms as anions and cations. Earlier, we had already noted how these local coordination can be exploited as characteristic features based on which we select candidate structure from the database by direct analogy.

Frequently, these polyhedra exhibit slight distortions for different compositions, or when different additional cations/anions are present in the structure but are not part of the polyhedra. However, for the present purpose one would only work with the idealized coordination polyhedra; the amount of distortion of these polyhedra will be established in the subsequent local minimization. Once we have investigated all structures in the various databases with respect to the presence of such "typical" building blocks for the atoms comprising the chemical system of interest, we can then, guided by the analogy to molecules, select all, or only the most common, building blocks to generate structure models for new not-yet-synthesized compounds in this chemical system. In the literature, there exist several such "coordination graph" based methods to structure prediction.[56,57] Note that in this approach, we use the database only to identify such polyhedra, but not in the choice and/or generation of the new candidate structures as we did earlier.

While this class of methods seems to avoid the global search we talked about above by straightforwardly yielding many apparently feasible structure candidates for a given chemical system, we quickly realize that there are very many ways one can combine these building blocks, even if one uses certain heuristic rules to restrict the exponentially large number of such polyhedra networks. Furthermore, it is far from obvious and actually quite unlikely that every candidate structure on the energy landscape can be reduced to a network of known building blocks. To amend this problem, we would need to include additional hypothetical coordination polyhedra in the graph generating algorithm. While this would allow us to essentially globally scan the landscape on the one hand, it would also enormously increase the number of candidates generated on the other hand. This greatly reduces the advantage of this method for general structure prediction purposes since one can quickly get overwhelmed by the number of local minimizations required – after all there is no a priori information available, which of the network models are high or low in energy.

3.6 Prediction via Hierarchical Construction

There exists another, related approach that can be employed to generate structure candidates based on local structure elements extracted from databases: the heuristic construction of hierarchical structure models. Here, one would go beyond individual coordination polyhedra and select larger excerpts of known crystal structures as building blocks of a structure candidate. We note that for this kind of structure generation, it is also possible, in principle, to obtain useful building blocks from cluster databases and use such locally optimized clusters to construct larger structures.[58]

There are two types of such heuristic constructions. The first one generates structures that exhibit the same type of interactions within a building block and between neighboring building blocks. These types of models produce network-like structures, and are inspired by e.g. zeolites or polyoxometalates, where so-called secondary building units have been employed to describe, and conversely model and predict the structures of such systems.[59] We remark that, conversely, we can also generate low-density structure candidates for simple chemical systems by replacing building blocks observed in e.g. zeolites by individual atoms.

The second class of structure models is based on different types of interactions within and among the building blocks. Here, the guides are compounds with complex anions, and possible applications are systems containing large clusters such as intercluster compounds. By treating such large building blocks as compact (rigid) units that interact via some effective (van der Waals, metallic, and/or ionic) intraction, one can generate at least approximate structure candidates by insertion of these blocks into the positions of atoms in various known structure types, followed by local structure refinements.

Such a hierarchical approach is often the most effective way to obtain starting points for a prediction of the structure of highly complex compounds. Yet such many-atom-units resemble "atoms" only on a very superficial level, and relying on this analogy to generate new structures by unit-to-atom substitution can be problematic. Both the "shape" of, and the actual interactions among the units are often quite complex such that the true structures might turn out to be rather different from those predicted by a hierarchical construction based on soft- or hard-sphere type models that usually guide our intuition. Furthermore, the local energy minimizations will most likely be both very involved due to the multiple types of interactions and length scales present, and thus rather computationally expensive. But once the computational tools have improved sufficiently to allow us to perform efficient fast local optimizations of such chemical systems, then hierarchical model building can be a sensible way to start the investigation of the energy landscape of highly complex crystalline compounds.

3.7 Structure Prediction by Structure-Property Relationships

Finally, we return shortly to the issue of using the databases to predict materials with specific properties. With the availability of truly gigantic computers and computer farms, one can provide, by direct computation, for every existing chemical compound a database entry that contains not only structural but also many other physical properties. By proposing to employ massive computations of this type, the Materials Genome type projects have set themselves the goal to provide the materials scientist with a large smorgasbord of compounds that might be useful at some point.

But going beyond just accumulating large amounts of data, the grand aim here would be to not only compute various physical properties of interest for all compounds known, but also use this information to automatically set up structure-property relationships. In particular, after computing the properties of many thousands of structures, the database(s) should be large enough that we can go beyond chemical intuition when choosing the structural parameters in these structure-property plots. Instead we would use statistical analyses on a large scale to establish significant relationships among structural, chemical and physical properties of various groups of chemical compounds. In the field of structure-property relationships for molecules and molecule-based systems, such systematic statistical analyses are by now becoming more and more standardized.[60] Similar systematic correlation studies should clearly be possible for crystalline compounds and their properties, too. In particular, machine learning techniques,[61] ranging from clustering analysis to neural networks, should be quite suitable for this task.

In the end, this type of information will then guide us in the choice of which chemical systems will be the most likely ones that exhibit certain physical and chemical properties. In particular, this will be true even for instances when the compound has not yet been synthesized but has only been predicted to exist as a kinetically stable modification. In its turn, the prediction of this modification might have come from the structure prediction analogue of a structure-property plot that now connects kinetic stability, simple or more complex combinations of chemical and physical parameters associated with a chemical system, and known structure types.

While it is tempting to use one's "chemical intuition" to speculate, which kind of such parameters correlate well with the existence of kinetically stable modifications of what structure type in any specific chemical system – the approach that has been employed by experimental chemists since the beginning of chemistry as a science –, we refrain from doing so here. Instead, we suggest one should follow a more systematic approach by e.g. training a neural network with a sufficiently large space of input parameters and hidden nodes to correctly assign to a given chemical system the observed structure types, plus the predicted ones if high-quality predictions are already available for a system. By trying to keep the set of physical and chemical parameters used as input for the neural network as general as possible (within the computational limitations, of course), we should be able to create an unbiased data bank-based structure prediction machine that will nicely complement the global landscape exploration techniques that have been developed over the past two-and-a-half decades.

4 Conclusion

In philosophy and pedagogy, one sometimes considers phase transitions in our understanding: The abstract transition from quantity to quality, as *Hegel* and his successors have discussed,[62] or the deeper understanding of a mathematical theorem gained once we have studied enough examples of its applications, or even in the study of chemistry the profound intuition acquired by the experimental chemist due to the many syntheses he or she has performed. Also in computer or machine learning, such transitions are observed,[63] e.g. during the teaching of a neural network, whose parameters are optimized by feeding it a multitude of teaching examples until the network is ready to correctly analyze and classify instances of input never presented before. Similarly, one often speaks in physics of the new paradigm of emergent properties in complex systems, where there is no longer a simple direct relation between the e.g. macroscopic and/or long-time features of the system and the microscopic and/or short-time aspects of the system that are ruled by e.g. the laws of quantum mechanics.[64] Perhaps the most exciting such dynamical transition is suspected to lie behind the emergence of life and consciousness in biological systems that are governed on the microscopic level by physical and chemical laws.[65,66]

One can propose that this kind of transition will have its counterpart in the field of database applications: Once the amount of information inside the database has grown large enough, then there exists, at least in potentia, the foundation for the emergence of higher order structures among the wealth of information contained in the database. In particular, with regard to chemical databases, we can hope and anticipate that the combination of large amounts of data and sophisticated statistical analysis tools will result in deeper insights into the relationships among chemical systems and their structural, chemical, and physical properties. The final aim would be that with the help of the database we can answer questions about chemical compounds that are not included in the database or not even known to be possible to exist.

Dedicated Cluster

Acknowledgements

My thanks go to *W. Marx* for suggesting the topic of this essay, and to *H. Schier* for valuable comments, and to the Laboratoire d'Analyse et d'Architecture des Systèmes in Toulouse for its hospitality. Many of the ideas presented in this essay are the outcome of discussions with many colleagues and friends, too numerous to list, and I would like to thank them for their exciting conversations and contributions.

References

[1] L. Gmelin, *Handbuch der Theoretischen Chemie*, Verlag Franz Varrentrapp, Frankfurt/Main **1817**.

[2] L. Gmelin, *Handbuch der Anorganischen Chemie*, Universitätsbuchhandlung Karl Winter, Heidelberg **1853**.

[3] F. Beilstein, *Handbuch der Organischen Chemie*, Leopold Voss Verlag, Leipzig **1881**.

[4] H. Landolt, R. Börnstein, *Zahlenwerte und Funktionen aus Physik, Chemie, Astronomie, Geophysik und Technik*, Springer Verlag, Berlin **1950**.

[5] *Reaxys*, http://www.reaxys.com, Elsevier, Amsterdam **2009**.

[6] F. H. Allen, *Acta Crystallogr., Sect. B* **2002**, *58*, 380.

[7] *Inorganic Crystal Structure Database*, http://icsdweb.fiz-karlsruhe.de, ICSD-Fiz-Karlsruhe **2005**.

[8] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, P. E. Bourne, *Nucleic Acids Res.* **2000**, *28*, 235.

[9] R. E. Newnham, *Structure-Property Relations*, Springer Verlag, Berlin **1975**.

[10] J. K. Fink, *Physical Chemistry in Depth*, Springer Verlag, Berlin **2009**.

[11] T. Le, V. C. Epa, F. R. Burden, D. A. Winkler, *Chem. Rev.* **2012**, *112*, 2889.

[12] A. R. Katritzky, U. Maran, V. S. Lobanov, M. Karelson, *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1.

[13] Y. Q. Cheng, E. Ma, *Prog. Mater. Sci.* **2011**, *56*, 379.

[14] G. Calas, L. Comier, L. Galoisy, P. Jollivet, *C. R. Chim.* **2002**, *5*, 831.

[15] H. S. Ma, A. P. Roberts, J. Prevost, R. Jullien, G. W. Scherer, *J. Non-Cryst. Solids* **2000**, *277*, 127.

[16] J. C. Tan, A. K. Cheetham, *Chem. Soc. Rev.* **2011**, *40*, 1059.

[17] W. Hu, *Polymer Physics*, Springer Verlag, Wien **2013**.

[18] A. R. Dinner, S. S. So, M. Karplus, *Proteins Struct. Funct. Gen.* **1998**, *33*, 177.

[19] X. Q. Pan, L. Fu, J. E. Dominguez, *J. Appl. Phys.* **2001**, *89*, 6056.

[20] C. A. Randall, S. F. Wang, D. Laubscher, J. P. Dougherty, W. Huebner, *J. Mater. Res.* **1993**, *8*, 871.

[21] A. Van de Walle, *Nat. Mater.* **2008**, *7*, 455.

[22] H. D. Barke, A. Hazari, S. Yitbarek, *Misconceptions in Chemistry*, Springer Verlag, Wien **2009**.

[23] D. Johrendt, *J. Mater. Chem.* **2011**, *21*, 13726.

[24] M. E. Eberhart, D. P. Clougherty, *Nat. Mater.* **2004**, *3*, 659.

[25] M. Jansen, J. C. Schön, *Angew. Chem. Int. Ed.* **2006**, *45*, 3406.

[26] J. C. Schön, in: *Proceedings of RIGI-workshop 1998*, (Ed.: J. Schreuer), ETH Zürich, Zürich **1998**, pp. 75–93.

[27] J. C. Schön, M. Jansen, *Z. Kristallogr.* **2001**, *216*, 307.

[28] J. C. Schön, M. Jansen, *Int. J. Mater. Res.* **2009**, *100*, 135.

[29] S. M. Woodley, C. R. A. Catlow, *Nat. Mater.* **2008**, *7*, 937.

[30] A. R. Oganov (Ed.) *Modern Methods of Crystal Structure Prediction*, Wiley VCh, Weinheim **2011**.

[31] Y. Wang, Y. Ma, *J. Chem. Phys.* **2014**, *140*, 040901.

[32] M. Jansen, *Angew. Chem. Int. Ed.* **2002**, *41*, 3747.

[33] H. Burzlaff, Y. Malinovsky, *Acta Crystallogr., Sect. A* **1997**, *53*, 217.

[34] R. Hundt, J. C. Schön, M. Jansen, *J. Appl. Crystallogr.* **2006**, *39*, 6.

[35] V. A. Blatov, *IUCr CompComm Newsletter* **2006**, *7*, 4.

[36] R. Allmann, R. Hinek, *Acta Crystallogr., Sect. A* **2007**, *63*, 412.

[37] Z. Cancarevic, J. C. Schön, M. Jansen, *Phys. Rev. B* **2006**, *73*, 224114.

[38] J. C. Schön, *Z. Anorg. Allg. Chem.* **2004**, *630*, 2354.

[39] J. C. Schön, *Ber. Bunsengesellschaft* **1996**, *100*, 1388.

[40] J. C. Schön, H. Putz, M. Jansen, *J. Phys. Condens. Matter* **1996**, *8*, 143.

[41] M. Sultania, J. C. Schön, D. Fischer, M. Jansen, *Struct. Chem.* **2012**, *23*, 1121.

[42] A. Y. Liu, M. L. Cohen, *Phys. Rev. B* **1990**, *41*, 10727.

[43] A. Neuhaus, *Chimia* **1964**, *18*, 93.

[44] S. Curtarolo, D. Morgan, K. Persson, J. Rodgers, G. Ceder, *Phys. Rev. Lett.* **2003**, *91*, 135503.

[45] T. Kalil, C. Wadia, *Materials Genome Initiative for Global Competitiveness*, http://www.whitehouse.gov **2011**.

[46] B. P. van Eijck, W. T. M. Mooij, J. Kroon, *Acta Crystallogr. Sect. B* **1995**, *51*, 99.

[47] P. Y. Chou, G. D. Fasman, *Adv. Enzymol.* **1978**, *47*, 45.

[48] B. Rost, C. Sander, *Proteins* **1994**, *19*, 55.

[49] D. L. Gerloff, F. E. Cohen, *Proteins Struct. Funct. Gen.* **1996**, *24*, 18.

[50] S. Govindarajan, R. Recabarren, R. A. Goldstein, *Proteins Struct. Funct. Gen.* **1999**, *35*, 408.

[51] Y. Wolf, N. Grishin, E. V. Koonin, *J. Mol. Biol.* **2000**, *299*, 897.

[52] D. Zagorac, J. C. Schön, M. Jansen, *Process. Appl. Ceram.* **2013**, *7*, 111.

[53] R.Hundt, *KPLOT*: A Program for Plotting and Investigation of Crystal Structures, University of Bonn, Germany, Version 9.6, 2011, **1979**.

[54] J. Cortes, private communication **2014**.

[55] J. C. Schön, M. Jansen, *Comput. Mater. Sci.* **1995**, *4*, 43.

[56] O. Delgado-Friedrichs, A. W. M. Dress, D. H. Huson, J. Klinowski, A. L. Mackay, *Nature* **1999**, *400*, 644.

[57] B. Winkler, C. J. Pickard, V. Milman, W. E. Klee, G. Thimm, *Chem. Phys. Lett.* **1999**, *312*, 536.

[58] J. C. Wojdel, M. A. Zwijnenburg, T. T. Bromley, *Chem. Mater.* **2006**, *18*, 1464.

[59] C. Mellot-Draznieks, S. Girard, G. Férey, J. C. Schön, Z. Cancarevic, M. Jansen, *Chem. Eur. J.* **2002**, *8*, 4102.

[60] *Advances in Quantitative Structure-Property Relationships*, (Eds.: M. Charton, B. I. Charton), Elsevier, New York **2002**.

[61] I. H. Witten, E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, Elsevier, New York **2005**.

[62] G. W. F. Hegel, *Wissenschaft der Logik*, Edition 1978, Felix-Meiner-Verlag, Hamburg **1812/13**.

[63] A. Giordana, L. Saitta, *Machine Learning* **2000**, *41*, 217.

[64] R. B. Laughlin, D. Pines, *Proc. Natl. Acad. Sci. USA* **2000**, *97*, 28.

[65] R. W. Sperry, *Psychol. Rev.* **1969**, *76*, 532.

[66] T. O'Connor, *Amer. Philos. Quart.* **1994**, *31*, 91.

Received: August 5, 2014
Published Online: September 24, 2014