crystIT: Complexity and Configurational Entropy of Crystal Structures via Information Theory

Clemens Kaußler, Gregor Kieslich

Submitted date: 01/10/2020 • Posted date: 05/10/2020
Licence: CC BY-NC-ND 4.0

Citation information: Kaußler, Clemens; Kieslich, Gregor (2020): crystIT: Complexity and Configurational Entropy of Crystal Structures via Information Theory. ChemRxiv. Preprint.
https://doi.org/10.26434/chemrxiv.13037993.v1

The information content of a crystal structure as conceived by information theory has recently proved as an intriguing approach to calculate the complexity of a crystal structure within a consistent concept. Given the relatively young nature of the field, theory development is still at the core of on-going research efforts. In this work we provide an update to the current theory, improving the formulas that evaluation of crystal structures with partial occupancies as frequently found in disordered system is feasible. To encourage wider application and further theory development we incorporate the updated formulas in crystIT (crystal structure & Information Theory), an open-source python-based program that allows for calculating various complexity measures of crystal structures based on a standardized *.cif file.

File list (1)

Kristallstruktur_Komplexität_ChemRxiv_01102020.pdf (518.02 KiB) view on ChemRxiv download file
crystIT: complexity and configurational entropy of crystal structures via information theory

Clemens Kaußler and Gregor Kieslich *

Department of Chemistry, Technical University of Munich, Lichtenbergstr. 4, 85748 Garching, Germany.
Correspondence e-mail: gregor.kieslich@tum.de

The information content of a crystal structure as conceived by information theory has recently proved as an intriguing approach to calculate the complexity of a crystal structure within a consistent concept. Given the relatively young nature of the field, theory development is still at the core of on-going research efforts. In this work we provide an update to the current theory, improving the formulas that evaluation of crystal structures with partial occupancies as frequently found in disordered system is feasible. To encourage wider application and further theory development we incorporate the updated formulas in crystIT (crystal structure & Information Theory), an open-source python-based program that allows for calculating various complexity measures of crystal structures based on a standardized *.cif file.

1. Introduction

The definition of complexity is a challenging and similarly fascinating subject, touching different scientific disciplines such as economy, informatics, biology, math and chemistry amongst others. Instead of defining complexity per se, it is in practice easier to ask which system is more complex, nicely showing that the challenge of defining complexity is closely related to the identification of an appropriate scale to measure complexity. Depending on the scientific area and the type of system, different scales have been proposed such as dimension, number of unique components, or simply human observation amongst many others. All of these scales come with their own hurdles such as lack of measurement techniques, definition of unique components and subjectivity, leaving us with the realization that every measuring system for complexity is only useful for a certain observer, in a defined context, for a defined purpose. In this article, the Shannon entropy is used as measuring system as defined by information theory (Shannon, 1948), providing us with a framework to differentiate between the complexity of crystal structures as initially introduced by S. Krivovichev (2014).

Over the years, the term complex has been used various times in the literature for describing crystal structures (Pauling, 1929; Valenzano et al., 2011; Loa et al., 2012) and indeed, parameters such as crystal class, number of different polyhedrons, space groups, number of atoms in the asymmetric unit or in the reduced unit cell are tempting simple measures. Arguably, a combination of these indices is suitable to fully grasp the depth of crystal structure complexity but raising the follow-up question for appropriate weighting factors when interested in a quantitative measure. In turn and without any surprise, J. K. Burdett et al. (1994) came to the conclusion that “complexity is largely a qualitative, frequently intuitive, notion”. Nevertheless, there was and still is interest (and arguably need) to find a quantitative concept to evaluate crystal structure complexity. For instance, Baur et al. (1983) defined topological and crystallographic parsimony indices of crystal structures, and more recently, the number of atoms was used as measure to classify various metallic alloys (Dshemuchadse & Steurer, 2015). Such practical concepts seem suitable for assessing structure complexity within certain material subclasses, but exhibit drawbacks related to a limited discriminating character between simple crystal structures and when interested in comparable measures across different material classes. Other approaches are more abstract, such as the algorithmic complexity descriptions based on work by Chaitin (1975) that was adapted for crystallography by Mackay (2001) and Estevez-Rams & González-Férez (2009). For a more detailed overview over attempts to quantify the complexity of crystal structures, we would like to refer to the review by (Krivovichev, 2014).

In 2014 S. Krivovichev applied the concept of Shannon entropy to crystalline materials, evaluating the information content of a crystal structure. Subsequently he proposed a concise concept that has the potential to capture the full multifaceted challenges of defining the complexity of a crystal structure based on the information content. Recently W. Hornfeck (2020) has suggested a few improvements of the concept, emphasizing the importance of theory development and the current state of research in this relatively young area. Importantly, comparisons between Shannon entropy, crystal structure complexity and configurational entropy can be drawn, opening intriguing opportunities for the assessment of configurational entropy of crystal structures and its change during phase transitions (Krivovichev, 2016). In the long-term, the concept shows large promise to contribute to a general understanding of crystalline matter and properties, where the quantification of configurational entropy of a crystal structure shows the largest potential to close the gap to applied inorganic chemistry.

In this work we follow on from the work by S. Krivovichev, proposing an updated formula that allows for evaluating the...
information content and in turn the complexity of crystal structures with partially occupied sites and defects. The proposed formula as well as recent improvements by W. Hornfeck are incorporated in crystIT (crystal structure & Information Theory), an open source python-based program. crystIT facilitates the application of the approach by non-specialists, the screening of crystallographic databases and method development in general. An intuitive understanding between crystal structure, information content and complexity is then created by applying crystIT to selected research examples.

2. Theory

In 1948 Shannon introduced a concept to determine the information content of a message, today known as Shannon entropy. Motivated by exploring limits in signal processing, data compression and cryptography, the concept has developed to one of the central pillars in information theory. In the following sections, the framework of Shannon’s concept is extended, its application to crystal structures as given by S. Krivovichev, and our and W. Hornfeck’s improvements to the concept.

2.1. Idea & Basics of Information Theory

Following Shannon, the information \( I \) contained in each symbol \( c \) of a message occurring with a probability of \( p_c \) is defined by equation 1.

\[
I = \frac{\text{bit symbol}}{5} = -\sum_i p_c \cdot \log_2 p_c \tag{1}
\]

Looking at a standard example, the message “Hello” is comprised of four distinctive letters, occurring with probabilities of \( p_{\text{H,e,o,l}} = 1/5 \) and \( p_l = 2/5 \). The message’s Shannon information is therefore \( I = 1.9 \) bit symbol\(^{-1} \) and its total information content is calculated by scaling the symbol-wise information content by the number of symbols, i.e. \( I_{\text{total}} = I \cdot 5 \).

2.2. Information Theory & Crystal Structures

By drawing the analogy between a message consisting of symbols \( c \) and the reduced unit cell of a crystal structure comprised of crystallographic orbits \( k \), S. Krivovichev applied Shannon’s formula to calculate the information content as provided by crystallographic data. Following this train of thought, probabilities \( p_c \) are given by the quotient of the crystallographic orbits’ multiplicities \( m_k \) and the number of atoms in the reduced unit cell \( v \), equation 2.

\[
I_c = \frac{\text{bit atom}}{5} = -\sum_k m_k \cdot \log_2 \frac{m_k}{v} \tag{2}
\]

Subsequently, S. Krivovichev established a correlation between the information content of a crystal structure and its perceived complexity, qualifying \( I_c \) as a quantitative and easily conceivable measure of crystal structure complexity, derived from information theory.

Inspecting equation 2 and looking for practical limitations, the question arises by which means partial occupancies can be considered. In its current form eq. 2 is only suitable for calculating \( I_c \) of crystal structures in which each crystallographic orbit is fully occupied by one species. In other words, the information content as calculated by equation 2 reflects the information provided by the decoration of the space group with atomic positions, whereas more information is contained in the specific atoms that are filled into these abstract positions. Materials that adopt partially occupied positions are frequent, such as solid solutions with disordered sites as can be found in alloys and various minerals or for high-temperature disordered phases. The latter becomes important when looking at the relation between the Shannon entropy and thermodynamic entropy, potentially providing insight into the phase transition thermodynamics.

2.2.1. Adaptation of the Information Theory Approach to Partially Occupied Sites

How to include partial occupancies into formula 2? Coming back to the linguistic example as introduced in section 2.1, a repeated string of hellos is presented, in which some words are randomly replaced by their German translation: “HelloHalloHelloHelloHallo”. According to equation 1 and based on the letters’ occurrence probabilities \( p_{\text{H,e,o,l}} = 1/5, p_l = 2/5, p_s = 2/25, p_e = 3/25 \), the information content per character is slightly increased: \( I = 2.1 \) bit symbol\(^{-1} \).

The smallest repeating unit in this string is equal in size to the string itself, and in turn the string is strictly speaking equal in length to its unit cell, which therefore consist of five times as many positions as “Hello”. When now assuming that this string is repeated with an average probability of “Hello” and “Hallo” equal to 3 to 2 but non-periodic distribution, the scenario is better described by

\[
\left[ \text{H}^3 \text{e}^1 \text{l}^1 \text{o} \right]_5
\]

Diffraction as bulk analysis technique to obtain crystallographic information relies on periodicity, returning averages of positions of disordered atoms (in analogy to letters). In other words, we attempt to analyze a five letter unit cell of the repeating pattern “H llo” that is disordered at the second position with probabilities (or occupancies) \( p_s = 2/25 \) and \( p_e = 3/25 \). Application of equation 2 to this disordered unit cell would erroneously result in the same information content as “Hello”, because the calculation is not based on the type of atoms (characters respectively) or occupancies but only their positions. Sufficient for crystal structure description are hence neither solely the atom types (the same element may be involved in entirely different coordination environments) nor the isolated crystallographic orbits (these can be filled partially or by different species).

Therefore we propose equation 3, wherein the sum is formed over distinguishable species \( a \) rather than crystallographic orbits \( k \). A species \( a \) is defined by a unique combination of chemical element or vacancy and crystallographic orbit. Probabilities \( p \) still reflect the chance of encountering a species \( a \) when observing a randomly chosen position in the reduced unit

\[1\] This shows striking parallels to the topological index defined by Baur et al. (1983) but augmented by the implementation of information theory and a finer consideration of topology in the form of crystallographic orbits.
cell; however, to consider fractional occupancies, \( p \) is calculated by the product of the occupied crystallographic orbit’s multiplicity \( m \) with the respective species’ fractional occupancy value \( \text{occ} \), divided by the total amount of positions in the reduced unit cell, \( P \).

\[
I_G = \frac{\text{bit}}{\text{position}} = -\sum_a \frac{m_a \cdot \text{occ}_a}{P} \cdot \log_2 \left( \frac{m_a \cdot \text{occ}_a}{P} \right) \tag{3}
\]

In analogy to the Kröger-Vink notation, fractional vacancies are also considered as individual species \( a \) in equation 3, forming distinctive vacancy-species for every crystallographic orbit that is only partially occupied by atoms (we will expand on this idea in section 2.5). For now, we want to highlight that by including vacancies, the sum over all probabilities is one, as all crystallographic orbits are formally fully occupied and \( P = \sum_a m_a \cdot \text{occ}_a = \sum_a m_a \). For fully occupied orbits, the amount of positions per reduced unit cell \( P \) is equal to the number of atoms \( v \), transforming formula 3 to Krivovichev’s equation (eq. 2). Notably, aliovalently substituted systems are also included in this approach, since there is no difference in structural information content whether the residual space \((1 – \text{occ})\) of a crystallographic orbit is empty or occupied by a different chemical element.

The information content of the whole reduced unit cell is then calculated by multiplication of \( I_G \) with the total amount of positions in the reduced unit cell (equation 4).

\[
I_G_{\text{total}} = \frac{\text{bit}}{\text{red. u.c.}} = I_G \cdot P \tag{4}
\]

For clarity, we consider the lead zirconate titanate (PZT) ceramic \( \text{PbZr}_{0.65}\text{Ti}_{0.35}\text{O}_3 \) as an example. The information relevant to this calculation is provided in table 1 (Mir et al., 2007). A total number of \( P = 10 \) positions are occupied, distributed among three crystallographic orbits since titanium and zirconium ions occupy a shared 2a Wyckoff position. \( I_G \) is calculated by plucking in the given probabilities \( p \) into equation 3, resulting in \( I_G = 1.56 \) bit position\(^{-1}\). Following equation 4, \( I_G_{\text{total}} = 15.6 \) bit red. u.c.\(^{-1}\) is obtained.

Please note that atoms of the same element can be crystallographically nonequivalent. For instance in yttrium barium copper oxide, \( \text{YBa}_2\text{Cu}_3\text{O}_{7-x} \), there are four crystallographically distinguishable oxygen species (Williams et al., 1988) and in turn their summands in eq. 3 are calculated separately. Additionally, the oxygen position at Wyckoff position 1e is only partially occupied \((\text{occ} = 1 - x)\), so that another term is added for the partial vacancy \((\text{occ} = x)\). For \( x = 9\% \) the information content consequently amounts \( I_G = 2.96 \) bit position\(^{-1}\).

### Table 1

Crystallographic information of \( \text{PbZr}_{0.65}\text{Ti}_{0.35}\text{O}_3 \) at 300 K, space group R3c.

| # Species | Element | Wyckoff | Occupancy \( \text{occ} \) | \( p = m_a \cdot \text{occ}_a/P \) |
|-----------|---------|---------|----------------|------------------|
| 1         | Pb      | 2a      | 1.00           | 0.20             |
| 2         | Ti      | 2a      | 0.65           | 0.13             |
| 3         | Zr      | 2a      | 0.35           | 0.07             |
| 4         | O       | 6b      | 1.00           | 0.60             |

#### 2.3. Extension by W. Hornfeck

Recently, W. Hornfeck (2020) pointed out that there are various unrelated structures that share the same \( I_G \) as defined by S. Krivovichev. Yet many of these crystal structures are characterized by different amounts of spatial degrees of freedom of their Wyckoff positions, or “site arities”, which are denoted as \( F \) in context of this paper \((f_w \) for a single Wyckoff position’s arity). In other words, different materials were observed to have the same information content although the occupied Wyckoff positions had different constraints in their \( x, y, z \) coordinates. Therefore, additionally to \( I_G \), he proposed the arity-based configurational complexity, \( I_{\text{conf}} \), in which the sum is formed over individual Wyckoff positions, \( w \), as described in equation 5:

\[
I_{\text{conf}} = \frac{\text{bit}}{\text{position & freedom}} = -\sum_w \frac{f_w}{F} \cdot \log_2 \left( \frac{f_w}{F} \right) \tag{5}
\]

We would like to emphasise that the sum in eq. 5 is over the single Wyckoff positions, whilst different results are obtained when the summation is done over the crystallographic orbits. In order to maintain consistency, W. Hornfeck subsequently renamed S. Krivovichev’s information content \( I_G \) as combinatorial complexity, \( I_G = I_{\text{comb}} \) and defined configurational complexity \( I_{\text{conf}} \) as the strong additive sum of coordinational and combinatorial complexities, see (Hornfeck, 2020) for the mathematical background. Using our updated measure for combinatorial complexity (i.e. \( I_{\text{comb}} \)) and combine it with \( I_{\text{coord}} \), we obtain equation 6:

\[
I_{\text{conf}} = \frac{\text{bit}}{\text{position & freedom}} = -\sum_a \frac{\text{occ}_a \cdot m_a}{P + F} \cdot \log_2 \left( \frac{\text{occ}_a \cdot m_a}{P + F} \right) - \sum_w \frac{f_w}{P + F} \cdot \log_2 \left( \frac{f_w}{P + F} \right) \tag{6}
\]

The use of arities as additional information content leads to a higher discriminating character of the complexity measure, following the chemists’ intuition. At the same time, and looking at this improvement from an information theoretical perspective, it is not obvious why a potential freedom is required to include in the information; the crystal structure can be fully built up without this knowledge, and following a subsequent symmetry analysis the information on the arities is retrieved. We will pick up on this point when discussing redundancies in section 2.6. For now we want to highlight that in the rest of this work we will continue to use S. Krivovichev’s measure \( I_G \). To avoid any ambiguities, both measures are implemented in crystIT.

#### 2.4. Configurational Entropy

Many different areas in chemistry are united in the quest of understanding macroscopic behaviour as a function of microscopic interactions, that is, the identification of structure-composition-property relations. When interested in fundamental principles that underlie the formation of (crystalline) condensed matter (Harper et al., 2019), or temperature and pressure dependent properties (Frenkel, 1999), the entropy \( S \) is an important parameter that ties the macroscopic to the microscopic world.
Statistically, $S$ can be accessed by the Boltzmann formula,

$$S = k_B \cdot \ln \Omega$$

(7)

translating the challenge to $\Omega$ which “counts the ways of finding the internal coordinates of a system for thermodynamically equivalent macroscopic states.” (Fultz, 2010) A typical simplification to approach $\Omega$ and in turn $S$ in crystalline matter is to divide $S$ into several contributions such as $S = S_{\text{conf}} + S_{\text{vib}}$, with $S_{\text{conf}}$ the configurational entropy and $S_{\text{vib}}$ the vibrational entropy.\footnote{There are other potential sources of entropy such as charge carriers, magnetic moments and nuclear spins that we neglect at this point.} This simplification is based on the idea that $S_{\text{conf}}$ is related to the spatial arrangement of atoms which is temperature-independent, whilst $S_{\text{vib}}$ is related to a temperature-dependent contribution describing the movement of atoms which itself is tied to the atomic interactions. Here we only consider the spatial arrangement of atoms and hence the concept of Shannon entropy relates to $S_{\text{conf}}$. In other words, it seems that information theory can provide us with a concept to calculate $S_{\text{conf}}$ of crystal structures.

S. Krivovichev (2016) followed up on this idea, showing that starting from eq. 7 a formula can be obtained in which the information content as provided by crystallographic data negatively contributes to configurational entropy of the structure:

$$S_{\text{cfg}} = \sum_{\text{cfg}}^{\text{max}} - I_G \cdot k_B \cdot N \cdot \ln 2$$

(8)

with $S_{\text{cfg}}$ the configurational entropy, $\sum_{\text{cfg}}^{\text{max}}$ the maximum configurational entropy as obtained when all atoms (positions) are symmetrically equivalent, $I_G$ as obtained from eq. 2, $k_B$ the Boltzmann constant, $N$ number of atoms in the reduced unit cell and $\ln(2)$ as conversion factor between nat and bit. This formula follows the chemist’s intuition that information and entropy are reciprocally related; however, there are some discrepancies of the values that are derived by eq. 8, making this area to an exciting field of active research. One important aspect is related to the scaling of $S_{\text{cfg}}$ to formula units rather than atomic sites. Likewise, the entropy of mixing and its increase along a substitution series should be considered, suggesting that $S_{\text{cfg}}$ itself consists of several contributions.

Here, we only want to point out that the calculation of $I_G$ for the substitution series $\text{Cu}_{1-x} \text{Au}_x$, with $S_{\text{cfg}}^{\text{max}} = 0$ for the pure elements Cu and Au, is in agreement with the entropy of mixing for a binary alloy after multiplication of $I_G$ with $R \cdot \ln(2)$, see figure 1b. This result is motivating for further work in this direction and confirms of how eq. 3 attributes for partially occupied sites. For development purposes, the calculation of $S_{\text{cfg}}$ is implemented in crystIT, see section 8 for details.

2.5. Vacancy

As stated in section 2.2.1, vacancies are considered as individual species for information content calculation according to equation 3. This is in analogy to the Kröger-Vink notation, in which vacant sites are denoted as $V_{M}^{\cdot \cdot}$ or $V_{X}^{\bullet}$ (Kröger & Vink, 1956). In the same manner, in which whitespaces such as ‘ ’ are necessary for the complete description of a language and contribute to its information content, vacancies $V$ are required for the description of defective or disordered structures. Although this approach is entirely logical from an information theory point of view and yields mostly coherent results, the incorporation of vacancies into the calculation has some counter-intuitive consequences.

Let us consider CsPbI$_3$ which adopts a perovskite structure. Above $\approx 150K$ the dodecahedrally coordinated Cs$^+$ cation becomes disordered over two sites, figure 1b (Straus et al., 2020). Intuitively, it would be expected that the information content rises continuously with increasing occupation of the second site; however, a jump in $I_G$ is observed at the temperature at which the first Cs$^+$ occupies the second site ($\text{lim}_{x\rightarrow 0} I_G$). This jump originates from the newly added crystallographic orbit which is immediately filled by a vacancy (1-occ) (figure 1a). Thereafter, the information content behaves as expected.

One could argue in justification of the jump in $I_G$, that the emergence of a new site contains a considerable amount of information. Then again, the space inside the unit cell which is only barely occupied by atoms after the jump did in fact exist before but fully “occupied” by void. This in turn raises the question whether the new crystallographic orbit should already be incorporated into the complexity calculation of the non-disordered perovskite structure - probably not: the ordered crystal structure can be completely described without the added empty site and if this was taken to the extreme, every point in space would need to be included in the calculation which is beyond any meaning.

2.6. Redundancy

Shannon entropy is highest upon equal distribution of atom species among crystallographic orbits or positions (uniform distribution of probabilities $p$), which can be verified by considering the partial derivatives $\frac{\partial}{\partial p_i} \left( - \sum_i p_i \cdot \log_2 p_i \right)$ with boundary conditions of $\sum_i p_i = 1$ and $p_i \in (0,1]$. In turn,
the maximum Shannon entropy per character, $I_{G_{\text{max}}}$, is given by the logarithm of the message’s total number of unique characters $c$, translating to the amount of unique atom species $a$ in the reduced unit cell, which is $A$ (equation 9).

$$I_{G_{\text{max}}} = \log_2 \left( \frac{\text{bit}}{\text{position}} \right)$$  \hspace{1cm} (9)

The redundancy $R$ is then defined as presented in equation 10.

$$R = 1 - I_{G_{\text{norm}}} = 1 - \frac{I_G}{I_{G_{\text{max}}}}$$  \hspace{1cm} (10)

Upon further investigation into the entropy of the printed English language, Shannon (1951) noted that there are different levels in which the entropy of a language can be estimated. Under the assumption of no knowledge about the language and an analysis of its composition entirely based on strings of meaningless letters, as essentially conducted in this work, the redundancy of English is about 50% because of phenomena or “constraints” such as the necessity of letters ‘q’ to be followed by ‘u’, a high tendency of ‘h’ to follow ‘t’ or an overall frequent appearance of the letter ‘e’. Although interesting predictions can be derived from rules found by the purely stochastic approach, an even higher level of redundancy of about 75% is estimated when considering grammatical rules and long range statistical effects in written English. The knowledge of the language therefore enables even better prediction abilities which were demonstrated by Shannon in experiments with native speakers who were supposed to guess missing letters of fill-in-the-blank texts.

In analogy to the application of information theory to crystal structures, Mackay (2001) wrote that “Pauling’s rules reflect chemical experience corresponding to a native knowledge of English in Shannon’s example”. For instance, an (inorganic) chemist can qualitatively construct the crystal structure of β-crystobalite based on chemical intuition and the information it adapts a variation of the diamond structure. Whilst such considerations seem of pure scientific nature at the current state, a large redundancy, particularly when combined with the “chemists’ grammar”, maybe as represented by the Pauling’s rules, might offer new avenues in crystal structure prediction, the identification of ‘wrong’ crystal structures and its subsequent refinement.

Closely related to the topic of redundancy is the question whether all sources of information have been yet included in eq. 3. As mentioned above, W. Hornfeck provided a recent update to the theory through the incorporation of arities, clearly improving on the discriminating character of different complexity measures. Interestingly, with including coordination freedom of the Wyckoff positions only a small increase of approximately 1% is observed for the examples Cs$_2$CuO$_2$, BaAl$_2$, SnF$_4$ and Na$_2$HgO$_2$ that have previously been discusssed by W. Hornfeck. We would like to emphasize here again, that different numbers for $S_{\text{conf}}$ are obtained from crystIT as the summation in eq. 5 is implemented over the Wyckoff positions in crystIT. For clarity and reproducibility, the examples are given in the GitHub repository as test files.

Looking for potential sources of information that are yet not included, it is important to acknowledge that the crystallographic information file is an idealized representation of a real crystal structure that exhibits naturally occuring point defects. Acknowledging the constant efforts from computational scientists in obtaining energies for the formation of point defects, opportunities exist to incorporate these in eq. 3 via statistical approaches in the future. For instance, when knowing the energy that is necessary to create a Schottky defect in NaCl, it is possible to calculate the defect concentration (partial occupancies) as a function of temperature and in turn the temperature dependent complexity is obtained. Given the increasing notion across various materials classes that defects are not independent (Keen & Goodwin, 2015), it remains an open question if such an extension contains significant meaning. Looking at real crystals, the existence of limited crystal volume, i.e. the surface as defect is an important point and fully neglected in this approach and seems only important for particle sizes in the nanoregime; however, this consideration raises interesting questions in the context of complexity measurement of clusters that consist of a defined number of atoms, e.g. the series of neutral and charged gold clusters. The last important point brings us back to the chemical intuition and concerns the meaning of a chemical bond in the context of information theory, intuitively assuming that their existence alters the structure’s information content; however, we came to the conclusion that chemical bonds in crystalline solids are redundant information, as these are unequivocally defined by the type of involved atom species and their positions which are both included in the calculation.

3. The crystIT program

crystIT (crystal structure & Information Theory) is an open source python-based program for calculating the information content of crystal structures.

![crystIT](https://example.com/crystIT.png)

**Figure 2**

Shown is an exemplary output of crystIT, as run in single file mode for K$_2$C$_{60}$. In batch file mode, a *.csv file is generated containing the output data.
The source code provided as ready-to-use python file is freely available3 and is based on the formulas as given in section 2. As input crystIT requires a standardized crystallographic information file (CIF). In single cif mode the program returns the calculated parameters directly into bash, see figure 2 for the output of K3C60 (Stephens et al., 1991). In batch mode a CIF-containing directory is passed to the program and the script outputs a *.csv file containing the different complexity measures. The batch mode is set up for large dataset processing and supports multithreading for better performance. The menu provides access to on-the-fly occupancy editing and options to alter settings regarding symmetry tolerance, recursive sub-directory scanning, the number of threads in batch mode, switching between comma and dot as decimal separator and the output of entropy parameters derived from equation 8.

In attempts to identify potential problems of the program, we observed erroneous space group detection in some cases, which can be circumvented by altering the symmetry tolerance value. We also came across CIF parsing errors in rare cases, which can be fixed by re-exporting the file from VESTA (Momma & Izumi, 2011). For better identification of such cases, error messages are given as output in bash or the *.csv files.

4. Results and Discussion

After describing the mathematical foundation of crystIT, we now proceed to investigate chemical interpretations of IG. By looking at the complexity of crystal structures of some selected examples, the next chapter aims at creating a more intuitive picture between information theory and crystal structure complexity.

4.1. Screening of the Crystallography Open Database (COD)

Krivovichev (2014) has performed a database screen based on crystallographic data as available in the Inorganic Crystal Structure Database. He correlated IG with IGtotal, compared different measures for complexity and evaluated complexity for various substance classes. Therefore, we here focus on the development of complexity with time, using the full COD-dataset as of June 2020. The complete dataset consists of approximately 440,000 CIFs (60 GB). The dataset was batch-processed in about six hours by using crystIT on a single workstation, demonstrating the scalability and robustness of the program.

Database screening studies heavily rely on the quality and amount of data entries. Therefore, an initial assessment of the numbers of published structures as a function of year is important, see figure 3a. The overall exponential increase of available crystal structures is testimony to the growing amount (and efficiency) of research capabilities, which affects the field of crystallography as indispensable analysis tool for synthetic chemistry in various areas. A sharp decrease in the amount of structures per year between 2014 and 2020 reflects a delay between publication of the crystal structures and incorporation to the database. Given that the number of structures is still reasonably large, it can be assumed that the database entries are sufficient for qualitative trend evaluation of crystal structure complexities.

In the next step, the development of the annually averaged IG is assessed, see figure 3b. Broadly, three periods can be identified: (i) the early years just after the discovery of X-rays by W. Röntgen where only little increase in the information content of the reported crystal structures is observed, (ii) the 60s to 80s with a noticeable increase in average IG from around two to three bit per atom which might be caused by the increasing availability of computers and lab diffractometers, and (iii) the 90s until today, where the average information content IG skyrocketed. The very sharp increase in the 90s is difficult to attribute to a single factor; however, the development of computer technologies, the rise of synchrotrons as highly brilliant light sources for X-ray diffraction, availability of neutron sources and advances in detector technologies are all important aspects. The current average information content resides at around six bit per atom.

Interestingly, the contour plot of IG distribution versus time, see figure 3b background, not only discloses that the determination of crystal structures IG > 5 bit per atom was not feasible until 1990 but also shows that the discovery of less complex crystal structures, i.e. structures with IG < 4 bit per atom, has suddenly abated and virtually ground to a halt around 2010. Furthermore, structures with IG > 9 bit per atom are still uncommon. It will be interesting to see how this further develops over time.

The most complex structures found in this screening have complexities of around IG ≈ 11.5 bit atom−1 and were discovered within the past five years. Many of these structures indeed seem highly complex, such as supramolecular arrays of helical oligoamides which self-assemble around a linear, rod-like oligocarbamate (Wang et al., 2017). Additionally, there are various coordination cages and multimetallic complexes with large IG values; however, there are also examples which have been assigned a large IG due to large unit cells in which assemblies of smaller subunits such as an eight-fold poly-catenated hydrogen-bonded and π-stacked framework of 1,3,5-

3 see section 8 and in the GitHub repository github.com/GKieslich/crystIT for further details

Figure 3
Information content screening of the Crystallography Open Database data set (June 2020). (a) shows the number of structures as available through the COD as a function of their year of publication, with a sharp increase in the late 90s. In (b) the annual average information content IG of the crystal structures contained in COD, with the contour plot in the background representing the frequency. Notably, since 2000 it seems that published structures show an ever increasing information content and the publication of “simple” structures has virtually stopped.
tris(4-carboxyphenyl)benzene (Zentner et al., 2015) can be observed. Such examples that are clearly composed of subunits show an intrinsically large compressibility. The algorithmic complexity approach brought forward by Chaitin (1975) would improve on this discrepancy between calculated and perceived complexity, but a new complexity descriptor for molecules as well as a measure for the three-dimensional molecular alignment in the reduced unit cell would need to be generated. This once again highlights the difficult implementation of an algorithmic approach. Nonetheless, with awareness for the redundancy drawback, the Shannon entropy as implemented in equation 3 has proven to be a fairly reliable complexity measure. Also, it should not go unmentioned that the information content of proteins exceed those of any structures discussed herein, even though the redundancy is expected to be very high - a few amino acids are sufficient as basic building blocks for proteins.

4.2. Silicon Carbide Polytypes

In a footnote, Pauling (1929) mentioned that by variation of the order of closed packed ABC layers infinitely many combinations “with ever increasing complexity” are possible. Inspired by this note and motivated to test $I_G$ against Pauling’s statement, we chose silicon carbides as the next example. Many different polytypes of silicon carbide are known which only differ in the order in which $C_{1/2}$-$Si$-$C_{1/2}$ slabs rotate around a $C_3$ axis, giving rise to (hypothetically) infinitely large unit cells (Parthé et al., 1993). For instance SiC $2H$ has an AB order, SiC $4H$ ABAC, SiC $6H$ ABCACB and so on (cf. figure 4b).

Indeed, the information content and the calculated complexity is rising with the number of layers (fig. 4a: $I_G$ approximately logarithmic and $I_{G,\text{total}}$ slightly faster than linear); however, it is also clear that the information content of rhombohedral polytypes rise at a lower rate than those that can be described by a hexagonal lattice (note that rhombohedral lattices are typically observed when the number of layers is a multiple of three). Looking for the origin of this phenomenon, it can be observed that in the rhombohedral cell there are two additional lattice points compared to a hexagonal Bravais lattice. In turn, only two thirds of the crystallographic orbits required for the description of SiC in the hexagonal case are necessary when describing SiC rhombohedral. Upon closer inspection, a kink in the $I_G$ development of SiC $Nh$ is visible between eight and ten layers. This is also related to different relative amounts of crystallographic orbits that must be defined depending on the space group (hexagonal P63mc: one per layer; trigonal P3m1: two per layer).

Although the general trend is in agreement with the intuitive understanding of crystal structure complexity, the differences related to the rhombohedral and hexagonal series is at minimum counterintuitive. In the context of information theory, however, this result is expected, and at the current state of research, seems to be an intrinsic artifact when using crystallographic orbits as measure for complexity calculations of crystal structures.

### 4.3. Ruddlesden-Popper Series

The series of Ruddlesden-Popper (RP) oxides is another interesting example and conceptually related to silicon carbides through the idea of increasing complexity via the incorporation of layers with varying repetition units. The structure of RP oxides are built from 2D slabs of perovskite unit cells with unit cell thicknesses $n$. These slabs are sandwiched between rock salt (AX) layers to form RP oxides with the general formula $A_{n+1}B_nX_{3n+1}$. Importantly, for $\lim_{n \to \infty}$ RP, the perovskite structure $ABX_3$ is obtained.

Similar to the silicon carbide example an infinite number of crystal structures can in principle be envisioned based on the variation of $n$. Intuitively we therefore expect an increase of complexity with increasing $n$, although experimentally, known RP examples do not exceed $n = 3$. We chose $Sr_{n+1}Ti_{n}O_{3n+1}$ as well as $Rb_{n+1}Cd_{n}Cl_{3n+1}$ and calculated the series’ complexities (see table 2). All RP phases in table 2, including the prototypical perovskite when treated as $n = 0$, demonstrate a direct proportionality between complexity and amount of layers: $I_G \propto n$ ($R^2 = 0.998$). In contrast to the $ABX_3$ compounds with fluorides, which typically adapt the perovskite structure, RP examples do not exceed $n = 3$.

![Figure 4](image_url)

**Figure 4**

Analyzing the complexity of various silicon carbides as a function of number of layers. (a) Complexity is plotted as a function of layers of various silicon carbide polytypes, and (b) shows SiC $4H$ as viewed along the b-axis with labelled ABAC layering; Si in blue and C in black. Depending on the number of layers in silicon carbides a 6-fold or 3-fold axis is present, which is reflected in the complexity measure and shows the close relation between complexity and symmetry.

| $n$ | Ruddlesden-Popper | Related Compounds | $I_G$ [bit atom$^{-1}$] |
|-----|-------------------|-------------------|------------------------|
| 0   | SrTiO$_3$         | -                 | 1.37                   |
| 0   | RbCdCl$_3$       | -                 | 2.32                   |
| 1   | Sr$_2$TiO$_4$    | Ca$_2$CuCl$_2$O$_4$ | 1.95                   |
| 1   | Rb$_2$CdCl$_4$   | La$_2$CuO$_4$    | 1.95                   |
| 1   | -                 | Ca$_2$CuO$_4$    | 1.92                   |
| 2   | Sr$_3$Ti$_2$O$_5$ | Ba$_3$In$_2$Cl$_2$O$_6$ | 2.42                  |
| 2   | Rb$_2$CdCl$_4$   | -                 | 2.42                   |
| 3   | Si$_2$Ti$_2$O$_7$ | -                 | 2.91                   |

Table 2

Complexity calculations of some selected Ruddlesden-Popper phases with focus on series based on SrTiO$_3$ and RbCdCl$_3$. Variations from canonical Ruddlesden-Popper series: $^a$oxychloride, $^b$distorted, $^c$anion deficient. CIFs sourced from: $^a$(Al-Shakarchi & Mahmood, 2011), $^b$(Natarajan et al., 1978), $^c$(Miwa et al., 2007), $^d$(Kruglik et al., 1989), $^e$(Lukaszewicz, 1959), $^f$(Villars & Cenzual, 2012a), $^g$(Villars & Cenzual, 2012b), $^h$(Grande & Müller-Buschbaum, 1977), $^i$(Grande et al., 1977), $^j$(Teske & Müller-Buschbaum, 1971), $^k$(Teske & Müller-Buschbaum, 1969), $^l$(Gutau & Müller-Buschbaum, 1990).
RbCdCl$_3$ crystallizes in a structure containing double rutile-like columns of CdCl$_6$ that are linked by Rb atoms (Natarajan et al., 1978). Therefore, a different complexity compared to the rest of the Ruddlesden-Popper phases is obtained for RbCdCl$_3$, nicely showing that $I_C$ depends on factors beyond the empirical formula. Surprisingly, other related compounds that do not satisfy the general RP-formula, but crystallize in similar structures such as the oxyhalide Ca$_2$CuCl$_2$O$_4$ or distorted variations, e.g. La$_2$CuO$_4$, show complexities equal to those of the canonical Ruddlesden-Popper phases. Even the complexities of anion-deficient M$_2$CuO$_3$ with M = Ca$^{2+}$ or Sr$^{2+}$ do not differ much.

The Ruddlesden-Popper series is therefore a beautiful example in which the intuitive understanding of complexity is well-matched by the complexity values calculated from information theory.

### 4.4. Perovskite Tilt Systems

Having a closer look at the iconic material class of perovskites, it is interesting to look for correlations between tilt systems and complexity as represented by $I_C$. For the classification of perovskite tilts after the Glazer (1972) notation we would like to refer to some insightful book chapters and reviews (Shimakawa, 2017; Woodward, 1997).

A selection of tilt phases for NaNbO$_3$ as phase rich example is given in table 3. Intuitively, we would assign the highest complexity to the phase with three tilts of different magnitudes. Our intuition is challenged when considering the a $\cdot$ a $\cdot$ a$^-$ tilt system. Although representing three activated tilts, the tilts are of the same magnitude and direction (as required through symmetry). In turn, one can argue that a $\cdot$ a $\cdot$ a$^-$ and a$^0$a$^0$a$^0$ are of similar complexity given that the number of different tilt angles is equal. Complexities obtained by information theory confirm this perspective, see table 3. Furthermore, it seems that the trend as expected from intuition holds for other examples such as KKnF$_3$ or CaTiO$_3$.

Therefore, the perovskite phases nicely highlight the subtle differences between symmetry and complexity, a difference that was not so clear from the example of silicon carbides. We would like to note that this is far away from an exhaustive study and it will be interesting to see how complexities of perovskites develop when considering examples with Jahn-Teller active B-site cations or other structural distortions; however, this is beyond the scope of this study.

### Table 3

| Compound       | space group | tilt          | $I_C$ [bit/atom$^{-1}$] |
|----------------|-------------|---------------|-------------------------|
| NaNbO$_3$      | P$m_3m$     | a$^0$a$^0$a$^0$ | 1.37                    |
| NaNbO$_3$      | R$3c$       | a$^-$a$^-$a$^-$ | 1.37                    |
| NaNbO$_3$      | P$4/nmb$    | a$^0$a$^0$a$^0$ | 1.92                    |
| KMnF$_3$       | P$4/nmb$    | a$^0$a$^0$a$^0$ | 1.92                    |
| CaTiO$_3$      | P$mnm$      | a$^-$b$^-$b$^-$ | 1.92                    |
| NaNbO$_3$      | C$mcn$      | a$^0$b$^0$    | 2.52                    |

Crystallographic information was obtained from: $^a$(Barth, 1925), $^b$(Seidel & Hoffmann, 1976), $^c$(Darlington & Knight, 1999), $^d$(Asbrink & Waskowska, 1994), $^e$(Buttner & Maslen, 1992), $^f$(Darlington & Knight, 1999).

### 5. Concluding Remarks

In conclusion, we introduced an update to the Krivovichev measure of crystal structure complexity to crystal structures with partial occupancies. For better applicability by non-specialists and theory development in the future, we have incorporated the concept into crysIIT, a python-based program that allows for calculating the complexity of crystal structures based on cif files. Looking at the discussed examples, we can observe a few counter-intuitive consequences to the utilization of crystallographic orbits for complexity calculations. For instance, we can find a pronounced space group dependency as observed for silicon carbides, and discontinuous behaviour of $I_C$. It is thus evident that further progress is necessary in this direction, either to elucidate these phenomena, or to provide further adjustments to the calculations. The general tenor is therefore that theory development is at the heart of on-going research activities. It is important to remind us that the outcomes are only as reliable and accurate as the source of information, in this case the reliability of the crystallographic data as provided through the crystallographic structure file.

In attempts to identify the potential of the approach, a breakthrough in the calculation of configurational entropy based on crystallographic data has clearly the potential to bring the concept of Shannon entropy closer to applied material sciences. Potential research directions might be a more quantitative analysis of calorimetric data to extend our understanding of phase transition thermodynamics in inorganic materials and coordination polymers alike. Likewise, in the text we mentioned the calculation of complexities of clusters based on information theory - but why stop at periodic matter, the elucidation of quasicrystals’ complexities seems a difficult but scientifically intriguing future task.

### 6. Utilized Software and Databases

The crystal structures for the calculations were either obtained from the Crystallography Open Database (COD) (Gražulis et al., 2009; Gražulis et al., 2012; Gražulis et al., 2015; Merkys et al., 2016) or the Cambridge Crystallographic Data Center (CCDC) (Groom et al., 2016) in form of crystallographic information files (CIFs). Some CIFs had to be generated from publications. The CIF generation process as well as the creation of crystal structure images were performed in Visualization for Electronic and Structural Analysis (VESTA) software suit (Momma & Izumi, 2011).

The provided Python (Van Rossum & Drake, 2009) program requires the Atomic Simulation Environment (ASE) library (Larsen et al., 2017) for CIF-parsing as well as Spglib (Togo & Tanaka, 2018), PyXtal (Fredericks et al., 2019) and numPy (van der Walt et al., 2011) for symmetry calculations.

### 7. Acknowledgements

CK and GK would like to acknowledge the scientific exchange with S. Krivovichev, with emphasis on the fruitful exchange regarding configurational entropy. Additionally, GK would like to thank the FCI and the DFG for financial support as pro-

---

Kaufler, Kieslich - Crystal Structure Complexity by Information Theory 2020
vised through the Liebig-Fellowship scheme and the SPP1928 (COORNET) startup fund.

8. Quick-Start Guide to the Python Program

crstIT as open-source program can be downloaded free of charge from www.github.com/GKieslich/crstIT together with an elaborate readme file. crstIT is written in Python 3, therefore compatible across multiple platforms. Package dependencies are described in section 6 and in the readme file.

As input, it requires a valid path to either a crystallographic file (CIF) or a directory containing CIFs (batch mode). Depending on the input, it either outputs the information parameter directly to the bash (single file) or creates a characterized-separated value file (CSV) in the directory (batch mode).

The settings can be accessed by typing ‘s’ and confirming with enter. By activating the recursive subdirectory scan (‘r’), subfolders are scanned in batch mode. The maximum number of threads for multiprocessing in batch mode is automatically set to the maximum number of available threads but can be adjusted by integer input. The occupancy options (‘o’) allow for on-the-fly occupancy editing in single file processing. A float input changes symprec which defines the tolerance in cartesian coordinates for Spglib to find symmetry: |x′ − x| < symprec. Entropy calculation is activated with ‘s’ and the decimal separator can be toggled between dot and comma by typing ‘d’. Finally, the menu is exited with ‘e’.

References

Al-Shakarchi, E. K. & Mahmood, N. B. (2011). J. Mod. Phys. 02(11), 1420–1428.

Asbrink, S. & Waskowska, A. (1994). Eur. J. Solid State Inorg. Chem. 31, 747–755.

Barth, T. F. W. (1925). Nor. Geol. Tidsskr. 8(9).

Baur, W. H., Tillmanns, E. & Hofmeister, W. (1983). Acta Crystallogr. B. 39(6), 669–674.

Burdett, J. K., Marnie, C. & Mitchell, J. F. (1994). Inorg. Chem. 33(9), 1848–1856.

Butnariu, R. H. & Måslen, E. N. (1992). Acta Crystallogr. B. 48(5), 644–649.

Chaitin, G. J. (1975). J. ACM, 22(3), 329–340.

Darlington, C. N. W. & Knight, K. S. (1999). Acta Crystallogr. B. 55(1), 24–30.

Deshmukhe, J. & Steurer, W. (2015). Inorg. Chem. 54(3), 1120–1128.

Estevez-Rams, E. & Gonzalez-Ferez, R. (2009). Z. Kristallogr. Cryst. Mater. 224(3), 179–184.

Fredericks, S., Sayne, D. & Zhu, Q. (2019). PxTaal: a Python Library for Crystal Structure Generation and Symmetry Analysis. URL: https://arxiv.org/abs/1911.11123

Frenkel, D. (1999). Physica A, 263(1), 26–38.

Fultz, B. (2010). Prog. Mater. Sci. 55(4), 247–352.

Glazer, A. M. (1972). Acta Crystallogr. B. 28(11), 3384–3392.

Grande, B. & Müller-Buschbaum, H. (1977). Z. Anorg. Allg. Chem. 429(1), 88–90.

Grande, B., Müller-Buschbaum, H. & Schweiter, M. (1977). Z. Anorg. Allg. Chem. 428(1), 120–124.

Gražulis, S., Chatteigner, D., Downs, R. T., Yoko, A. F. T., Quiróz, M., Lutterotti, L., Manakova, E., Butkus, J., Moeck, P. & Le Bail, A. (2009). J. Appl. Crystallogr. 42(4), 726–729.

Gražulis, S., Daškevič, A., Merkys, A., Chatteigner, D., Lutterotti, L., Quíroz, M., Serebryanyaa, N. R., Moeck, P., Downs, R. T. & Le Bail, A. (2012). Nucleic Acids Res. 40(1), D420–D427.
Villars, P. & Cenzual, K., (2012b). Sr₄Ti₃O₁₀ Crystal Structure: Datasheet from “PAULING FILE Multinaries Edition” in SpringerMaterials.
van der Walt, S., Colbert, S. C. & Varoquaux, G. (2011). Comput. Sci. Eng. 13(2), 22–30.
Wang, X., Wicher, B., Ferrand, Y. & Huc, I. (2017). J. Am. Chem. Soc. 139(27), 9350–9358.

Williams, A., Kwei, G. H., Von Dreеле, R. B., Raistrick, I. D. & Bish, D. L. (1988). Phys. Rev. B, 37(13), 7960–7962.
Woodward, P. M. (1997). Acta Crystallogr. B, 53(1), 32–43.
Zentner, C. A., Lai, H. W. H., Greenfield, J. T., Wiscons, R. A., Zeller, M., Campana, C. F., Talu, O., FitzGerald, S. A. & Rowsell, J. L. C. (2015). Chem. Commun. 51(58), 11642–11645.
