The k-centre problem for necklaces

Exploring Crystal Structures Effectively

Duncan Adamson
Argyrios Deligkas, Vladimir Gusev, Igor Potapov

January 2023

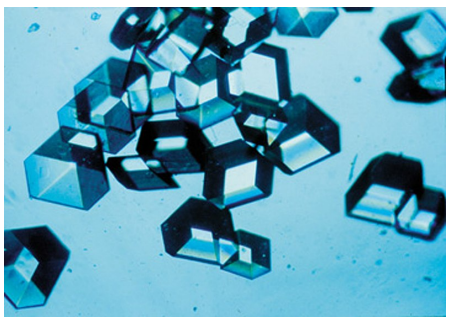
Why Crystals?

- New materials are needed to deal with the challenges of the 21st century, from strong materials for manufacturing to better conductors for electrical systems.
- Crystals are a fundamental, and very common form of matter.
- Importantly, Crystals are **periodic** - meaning that a lot of the properties of a crystaline material can be determined from a relatively small amount of information.

Why Crystals?

- New materials are needed to deal with the challenges of the 21st century, from strong materials for manufacturing to better conductors for electrical systems.
- Crystals are a fundamental, and very common form of matter.
- Importantly, Crystals are **periodic** - meaning that a lot of the properties of a crystaline material can be determined from a relatively small amount of information.
- In general, the problem of predicting crystal structures is undecidable.

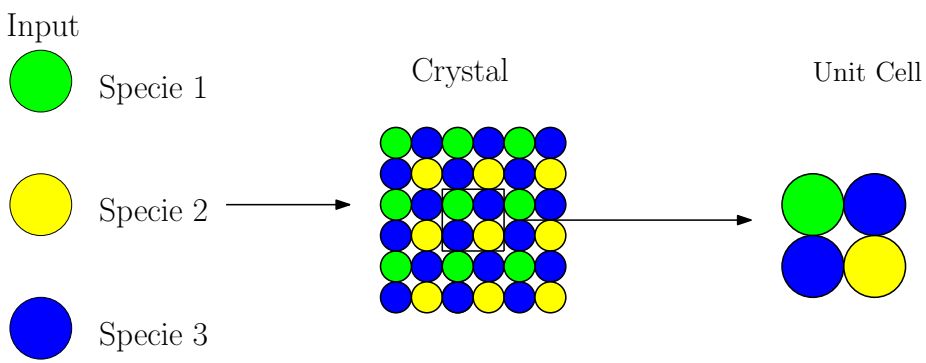
Crystals are everywhere

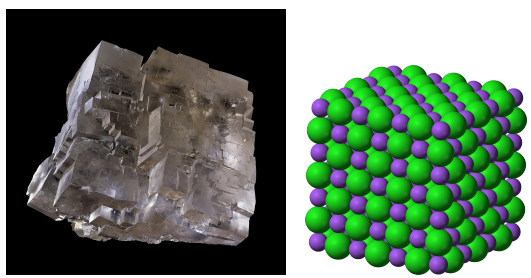


Crystals

Definition (Crystals)

A **Crystal** is a material composed of an (infinitely) repeating **Unit Cell**.

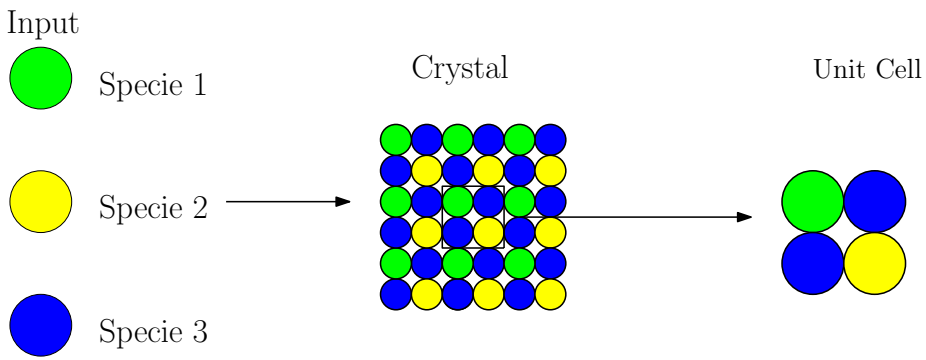


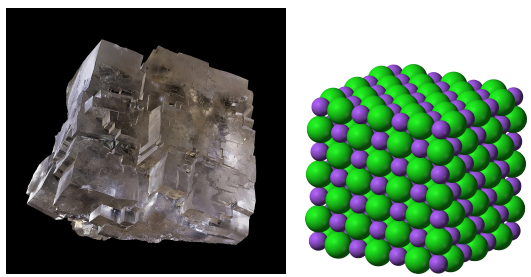


Crystals

Definition (Unit Cells)

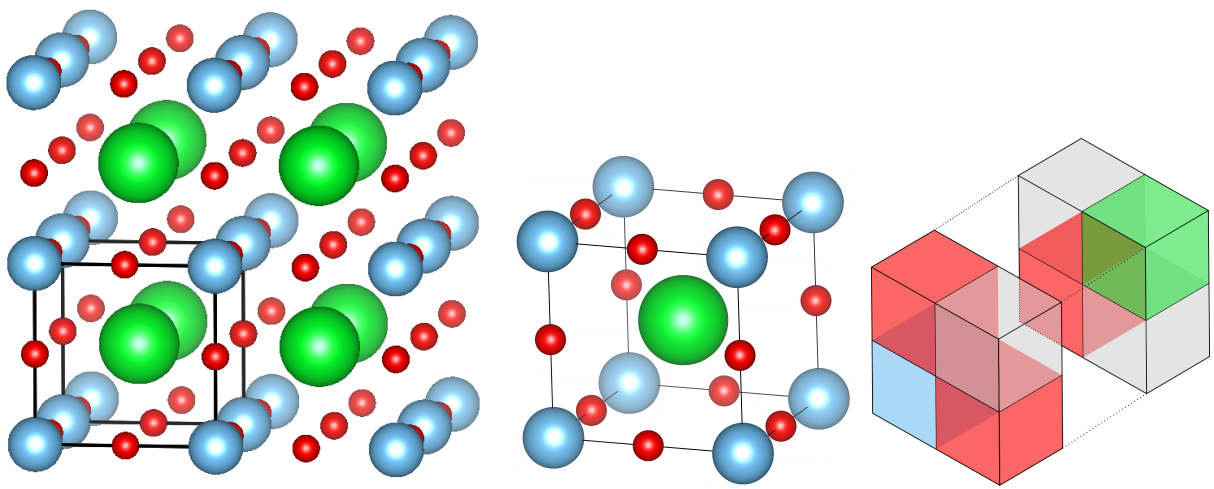
A **Unit Cell** is a contiguous region of space containing some set of **Ions**.





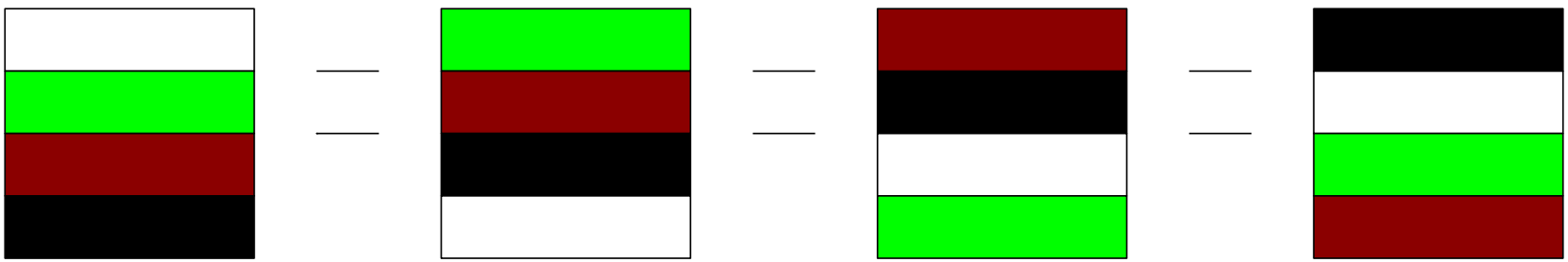
Discrete Crystals

- In this talk we are interested in **Discrete Crystals**, i.e. crystals where every ion is placed on a grid.
- In this model, every cell is either empty, or wholly occupied by an ion (or block of ions).
- For simplicity we assume that each cell can contain only 1 ion, and that each ion can fit into a single cell.



1D Crystals

- In this talk we are going to focus on **1D-Crystals**[1].
- These can be thought of a crystals made from precomputed 3D-blocks, with a high degree of symmetry along two dimensions.
- The main challenge in **uniquely** representing these structures is capturing **translational symmetry**.



[1] C. Collins et al. "Accelerated discovery of two crystal structure types in a complex inorganic phase field". In: *Nature* 546.7657 (2017), p. 280.

Goal

Select a **representative** set of crystal structures from the set of all possible structures of a given size over a given alphabet of "blocks".

Problems

Question

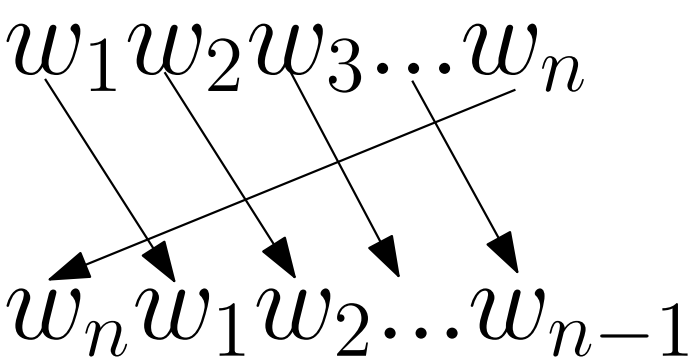
How can we represent crystals **uniquely***?*

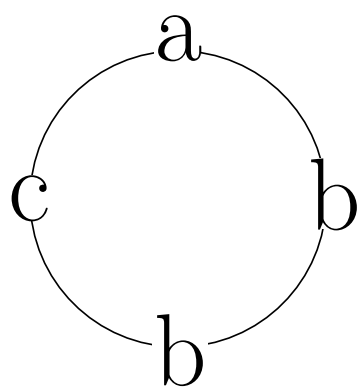
Question

How can we choose a **representative sample** *from this representation?*

Necklaces

- In 1D the problem of representation is solved using **Necklaces**.
- Informally a necklace is a set of words that can be reached from each other by some **translation**.
- The **translation** (or cyclic shift) of a word w by some integer i returns the word w' where $w'_j = w_{j-i \bmod n}$.





abbc
bbca
bcab
cabb

Some Notation for 1D Necklaces

For the remainder of this talk we use the following assumptions:

- Σ to denotes an alphabet, which we assume has size q.
- The **Canonical form** of a necklace ω (denoted $\langle\omega\rangle$) is the **Lexicographically smallest** word $w \in \omega$.
- $\mathcal{N}_q^n$ denotes the set of necklaces of length n over an alphabet of size q, corresponding to the set of all crystal structures of length n over a library of q blocks.

Representative Samples

- To get a set of **representative crystals**, we want to choose a set of crystals that contain as many **local structures** as possible, in order understand the global energy space.
- As energy interaction is strongest at close range, by analysing local structures we can get a good idea about the full energy space.

Question

How do formalise this as a mathematical problem?

Finding Representative Samples

- In order to find a set of representative samples, we turn to the **k-centre problem**.
- Informally, the k-centre problem asks us to select a set k vertices from a graph $G = (V, E)$ minimising the function:

$$\min_{S \subseteq_k V} \max_{v \in V} \min_{s \in S} D(s, v)$$

- Where S is a set of k vertices from V, and $D(s, v)$ is the distance between some pair of vertices in G.

Question

How can we measure the distance between necklaces?

The Overlap Distance for Necklaces

- Following our motivation of comparing **crystal structures**, we need to define a distance that represents structures with similar properties.
- As much of the energy in crystaline structures is due to **local** interactions, we use the number of **shared subwords** as a basis for measuring the similarity.

Definition

The **overlap distance** between two necklaces $\tilde{\mathbf{w}}$ and $\tilde{\mathbf{u}}$ is defined as the number of shared subwords between $\tilde{\mathbf{w}}$ and $\tilde{\mathbf{u}}$, normalised by the total number of subwords.

The Overlap Distance for Necklaces

	word $abab$	word $aabb$	Intersection
1	$a \times 2, b \times 2$	$a \times 2, b \times 2$	
2	$ab \times 2, ba \times 2$	$aa \times 1$ $ab \times 1$, $bb \times 1, ba \times 1$	
3	$aba \times 2, bab \times 2$	$aab \times 1, abb \times 1$, $bba \times 1, baa \times 1$	
4	$abab \times 2, baba \times 2$	$aabb \times 1, abba \times 1$, $bbaa \times 1, baab \times 1$	
$\mathfrak{D}$			?/16

The Overlap Distance for Necklaces

	$abab$	$aabb$	Intersection
1	$\mathbf{a} \times 2, \mathbf{b} \times 2$	$\mathbf{a} \times 2, \mathbf{b} \times 2$	4
2	$\mathbf{ab} \times 2, \mathbf{ba} \times 2$	$aa \times 1$ $\mathbf{ab} \times 1$, $bb \times 1, \mathbf{ba} \times 1$	2
3	$aba \times 2, bab \times 2$	$aab \times 1, abb \times 1$, $bba \times 1, baa \times 1$	0
4	$abab \times 2, baba \times 2$	$aabb \times 1, abba \times 1$, $bbaa \times 1, baab \times 1$	0
$\mathfrak{D}$			6/16

Challenges

- There are $O(q^n)$ necklaces in $\mathcal{N}_q^n$, so explicitly representing the graph is not feasible even for moderate values of n.
- Trying to determine the properties of a necklace as a crystal structure is computationally expensive.
- The graph is highly structured, with some vertices being much better than others.

Our Approach

- Our goal is to select a set of k-centres **maximising** the number of **distinct subwords** that appear in any centre.
- We do this by finding the longest length λ for which every word in Σ^λ can appear at least once in the set of centres.

Observation

Let $S \subseteq_k \mathcal{N}_q^n$ be a set of k necklaces such that every word in Σ^λ appears as a subword in at least one necklace in S. Then, every necklace in $\mathcal{N}_q^n$ is at most $\frac{\lambda^2}{n}$ from the nearest centre in S.

de-Bruijn Sequences

Definition

The **de-Bruijn Graph** of order m over a k-ary alphabet Σ is a directed graph of q^m vertices where each vertex corresponds uniquely to some word in Σ^m. There exists an edge from v to u if and only if $v : u_m = v_1 : u$.

Definition

A **de-Bruijn Sequence** of order m over a k-ary alphabet Σ is a cyclic word w of length q^m such that every word in Σ^m appears exactly once in w. In other words, w corresponds to a Hamiltonian circuit on the de-Bruijn graph.

Example of the de-Bruijn graph and sequence

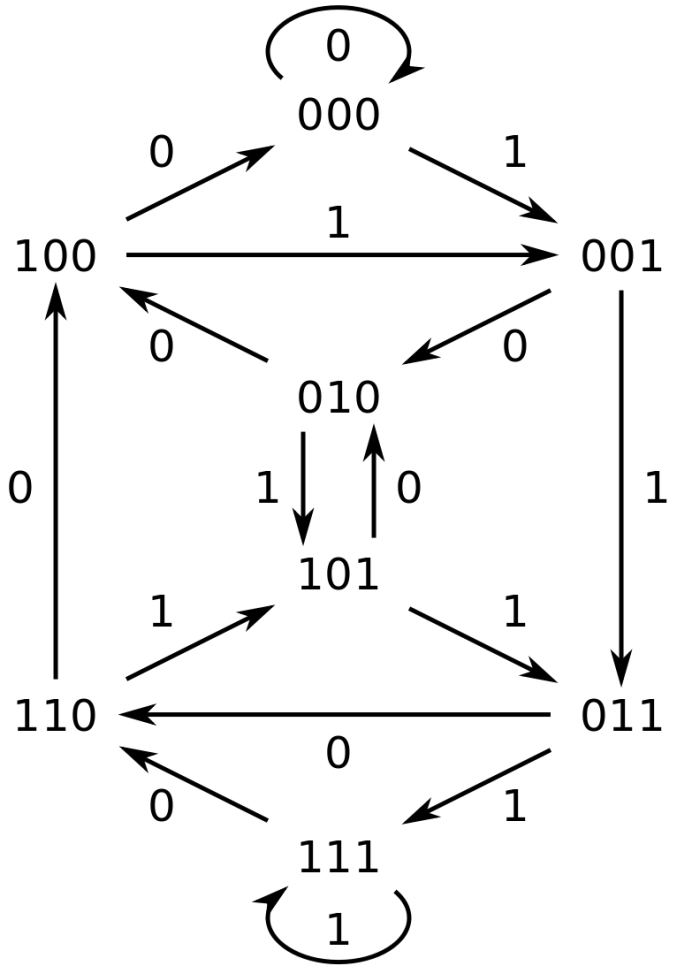


Figure 1: 00010111

de-Bruijn Sequences

- **De-Bruijn Sequences** are a classic combinatorial object.
- There are strong results for generating[2] and decomposing[3].

[2] Fred S. Annexstein. "Generating de Bruijn sequences: An efficient implementation". In: *IEEE Transactions on Computers* 46.2 (1997), pp. 198–200.

[3] T. Kociumaka, J. Radoszewski, and W. Rytter. "Computing k-th Lyndon word and decoding lexicographically minimal de Bruijn sequence". In: *Symposium on Combinatorial Pattern Matching*. Springer International Publishing, 2014, pp. 202–211.

de-Bruijn Sequences

- **De-Bruijn Sequences** are a classic combinatorial object.
- There are strong results for generating[2] and decomposing[3].
- **Idea**. We can use de-Bruijn sequences as a basis to compute a set of centres.

[2] Fred S. Annexstein. "Generating de Bruijn sequences: An efficient implementation". In: *IEEE Transactions on Computers* 46.2 (1997), pp. 198–200.

[3] T. Kociumaka, J. Radoszewski, and W. Rytter. "Computing k-th Lyndon word and decoding lexicographically minimal de Bruijn sequence". In: *Symposium on Combinatorial Pattern Matching*. Springer International Publishing, 2014, pp. 202–211.

One Centre

- When $k = 1$, centre can be computed by finding the largest value of λ for which $q^\lambda \leq n$, giving $\lambda = \lfloor \log_q n \rfloor$.
- This ensures that every word of length λ appears at least once in the centre.

When $k = 1$, this process returns the optimal centre.

Multiple Centres

- When $k > 1$, a slightly more sophisticated approach is needed.
- The main challenge is determining how to **partition** the de-Bruijn sequence.
- **Challenge**. Just partitioning the sequence in to k centres will loose some words.

Multiple Centres

- When $k > 1$, a slightly more sophisticated approach is needed.
- The main challenge is determining how to **partition** the de-Bruijn sequence.
- **Challenge**. Just partitioning the sequence in to k centres will loose some words.
- **Idea**. we need to build some redundancy to the centres.

Multiple Centres

Sequence:	00000100011001010011101011011111
Centre	Word
1	00000100 0110
2	01100101 0011
3	00111010 1101
4	11011111 0000

Figure 2: Example of how to split the de Bruijn sequence of order 5 between 4 centres of length 12. Highlighted parts are the shared subwords between two centres.

Multiple Centres

Theorem

The k-centre problem for $\mathcal{N}_q^n$ can be approximated in $O(n \cdot k)$ time with an approximation factor of $1 + \frac{\log_q(k \cdot n)}{n - \log_q(k \cdot n)} - \frac{\log_q^2(k \cdot n)}{2n(n - \log_q(k \cdot n))}$.

Other Results

- It is NP hard to determine the distance between an arbitrary set of centres and the furthest necklace $w \in \mathcal{N}_q^n$.
- Problem with **multidimensional** necklaces can be solved in $O(k \cdot N^2)$ time within a factor of $1 + \frac{\log_q(kN)}{N - \log_q(kN)} - \frac{\log_q^2(kN)}{2N(N - \log_q(kN))}$, where $N = \prod_{i=1}^{d} n_i$.
- The **online** variant of this problem can be solved within a factor of $1 + \frac{2\log^2(iN) - \log_q^2(i/3d)}{2N - 2\log_q^2(i \cdot N)}$.

Future Work

- Check if this algorithm is optimal by finding a stronger lower bound **or** find a stronger algorithm matching the lower bound.
- Extend these results to other classes of cyclic words.
- Extend these results to other similarity metrics (edit distance, LCS, etc.).