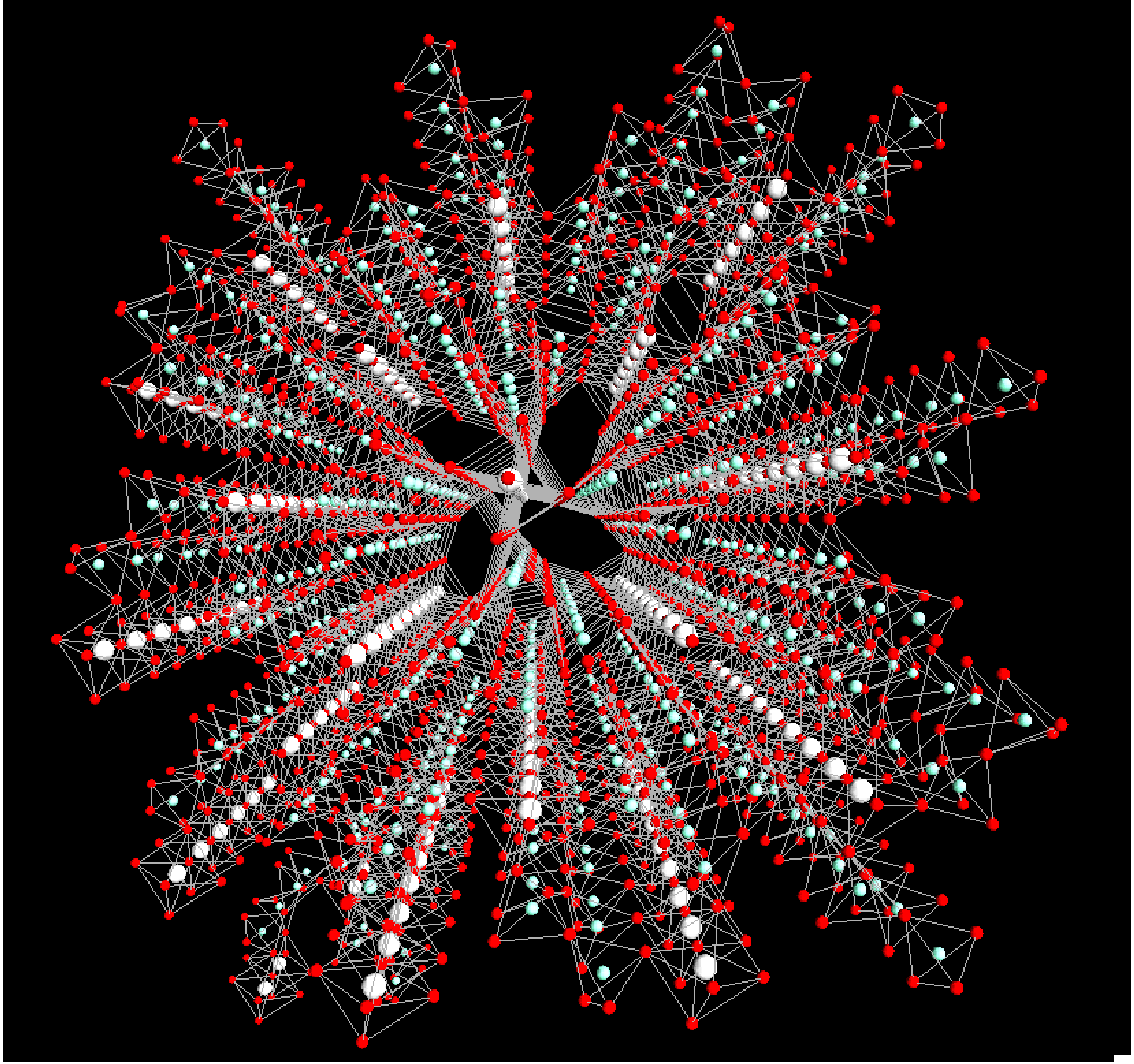
Searching and ranking similar clusters of polyhedra in inorganic crystal structures

Hans-Joachim Klein
Institut f. Informatik
Christian-Albrechts-Universität Kiel
Germany



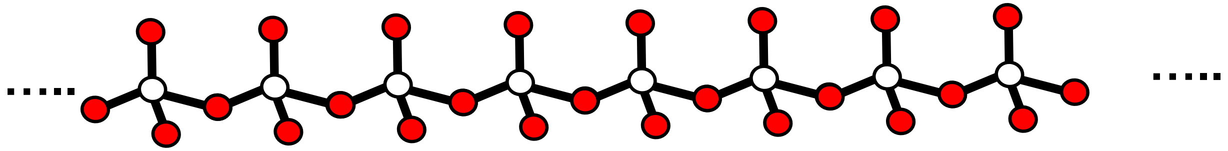
Definition:

A ***crystal*** is an anisotropic homogeneous body consisting of a three-dimensional periodic ordering of atoms, ions, or molecules.

direction-dependent physical properties

parallel directions: same behaviour

Three-dimensional, periodic: basic units (atoms, ions, molecules), repeating in all directions.



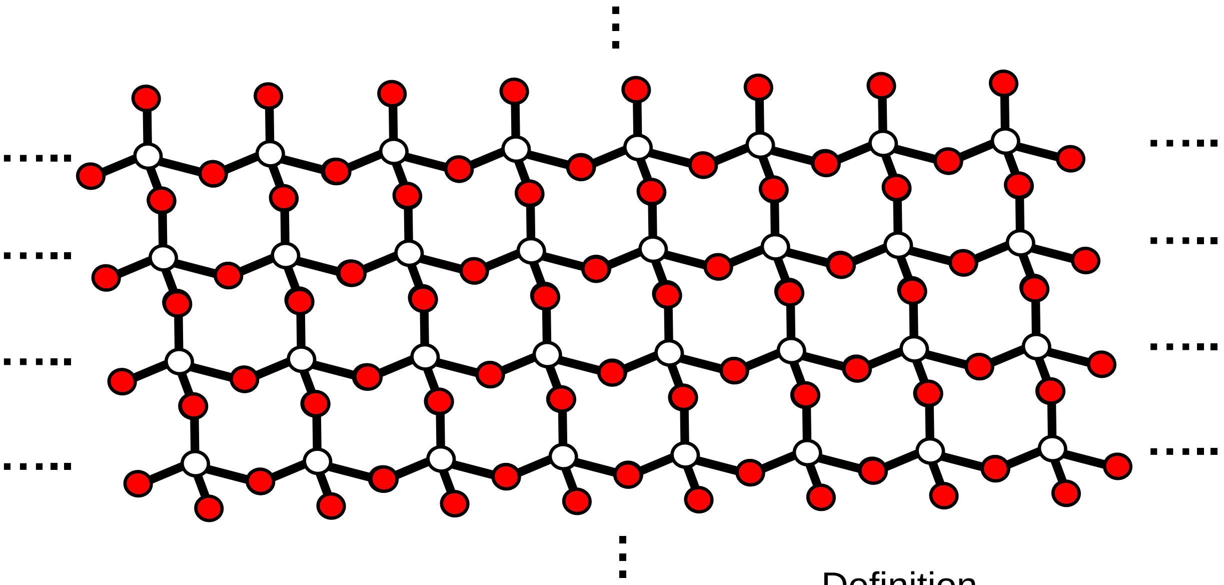
Definition:

A ***crystal*** is an anisotropic homogeneous body consisting of a three-dimensional periodic ordering of atoms, ions, or molecules.

direction-dependent physical properties

parallel directions: same behaviour

Three-dimensional, periodic: basic units (atoms, ions, molecules), repeating in all directions.



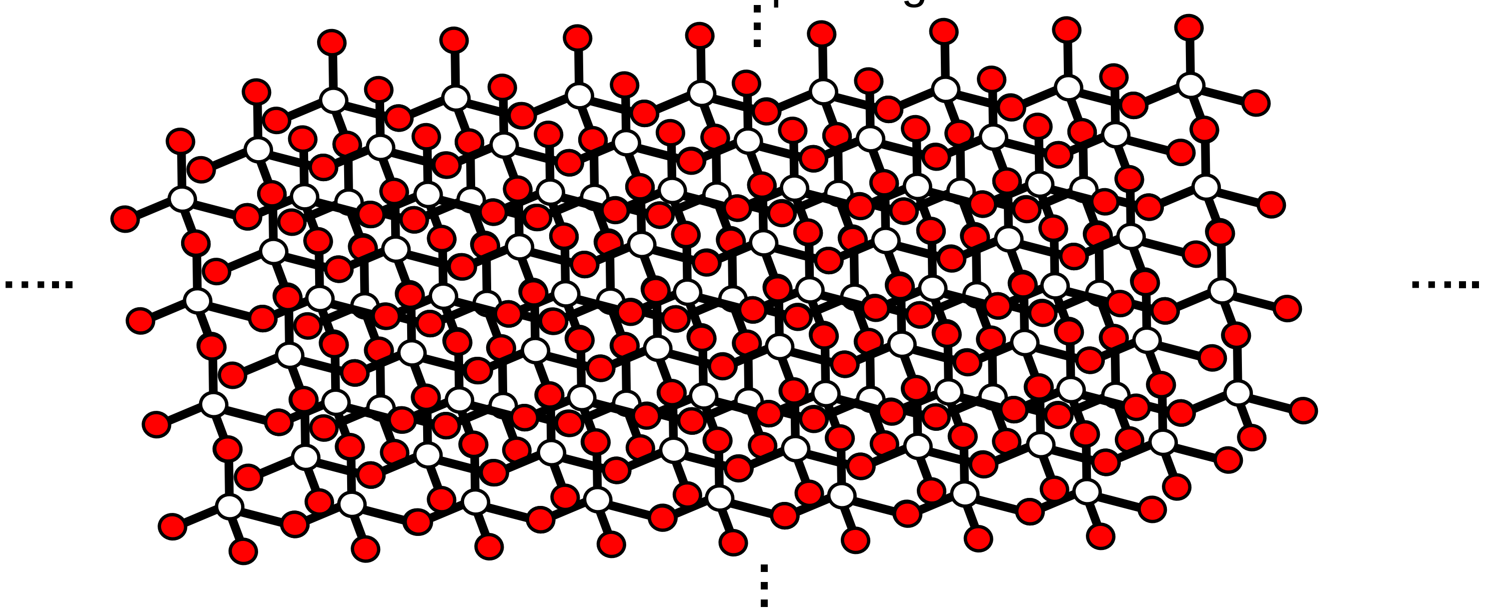
Definition:

A ***crystal*** is an anisotropic homogeneous body consisting of a three-dimensional periodic ordering of atoms, ions, or molecules.

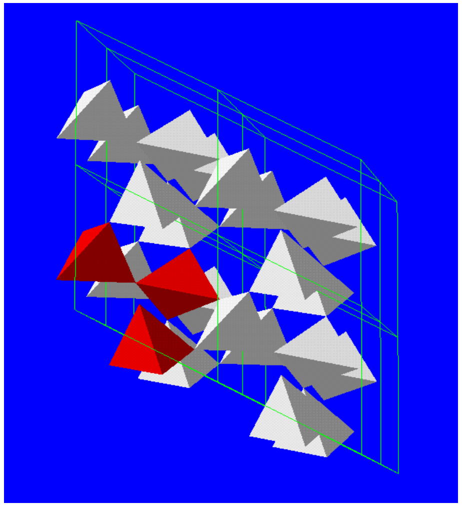
direction-dependent physical properties

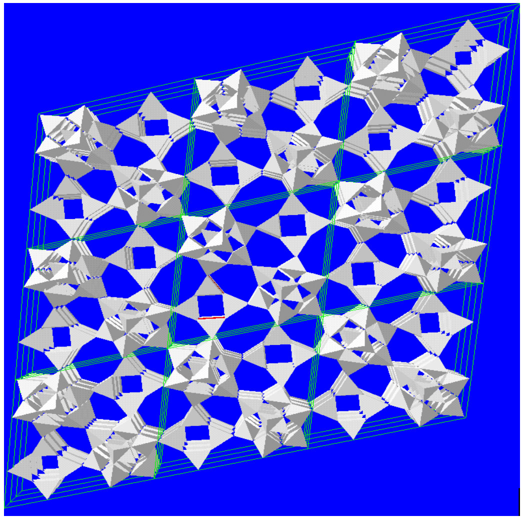
parallel directions: same behaviour

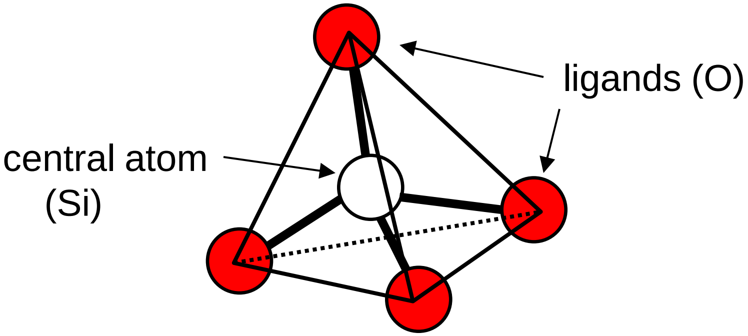
Three-dimensional, periodic: basic units (atoms, ions, molecules), repeating in all directions.

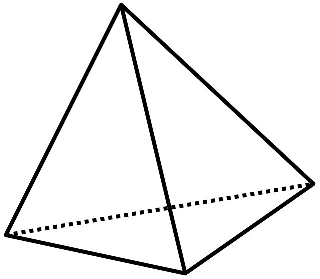


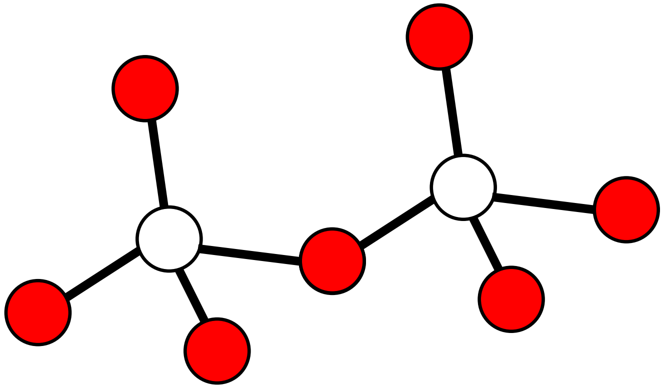
Abstraction by graphs
(the simple case)

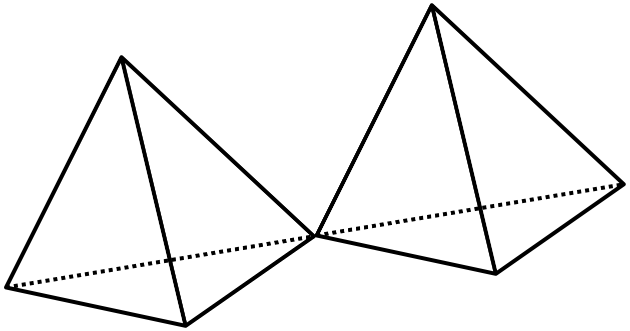


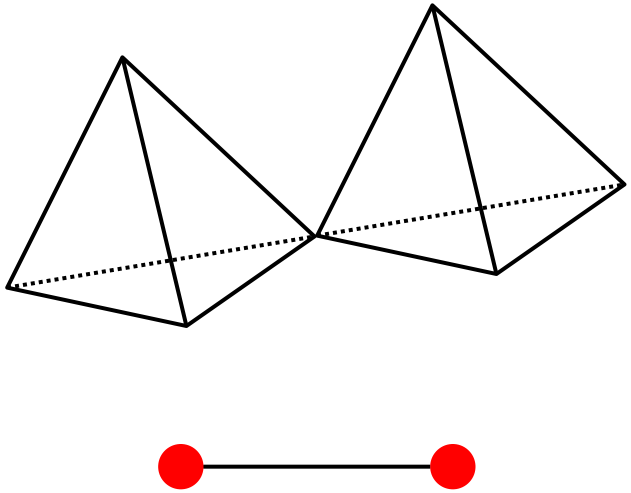




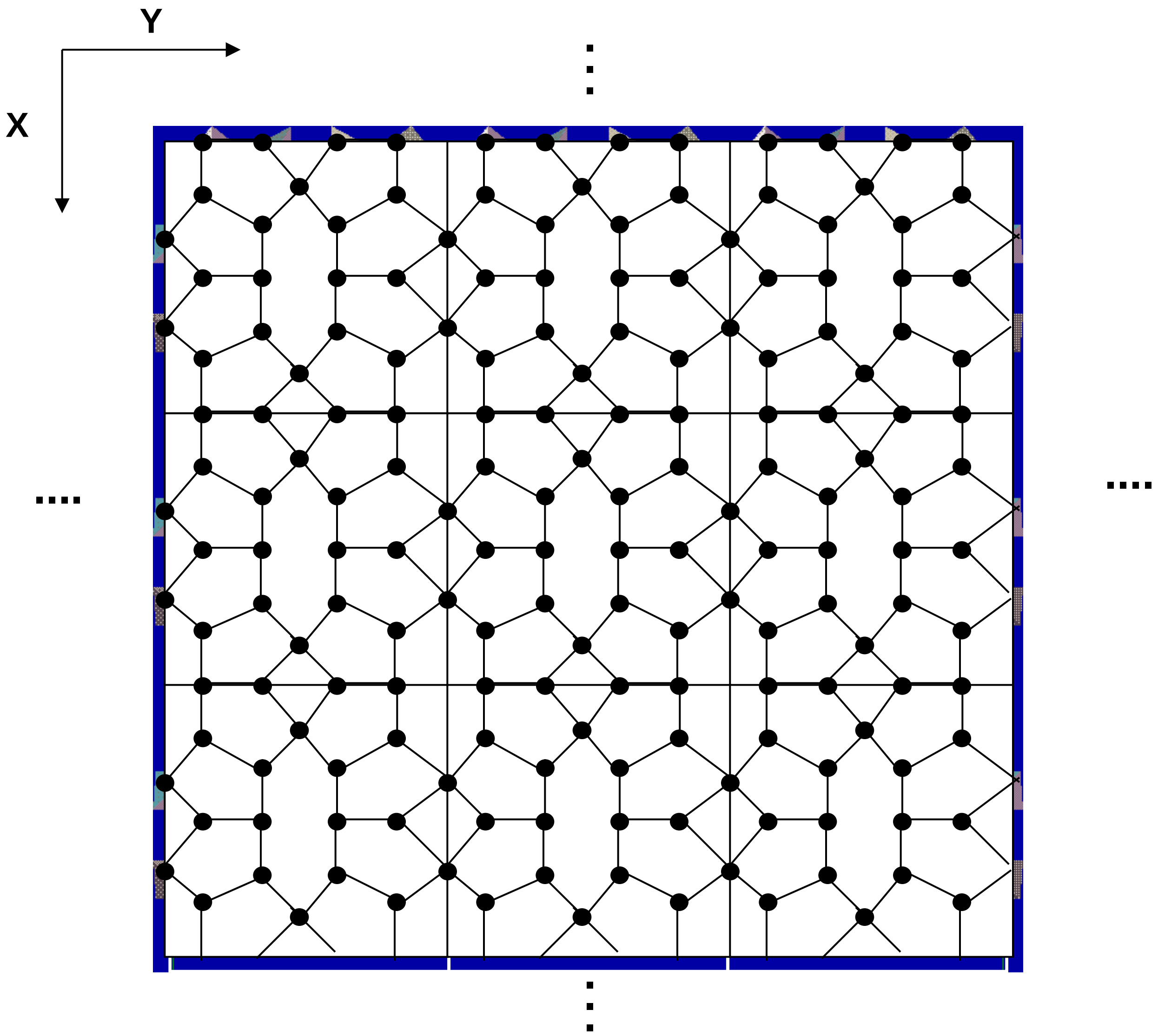




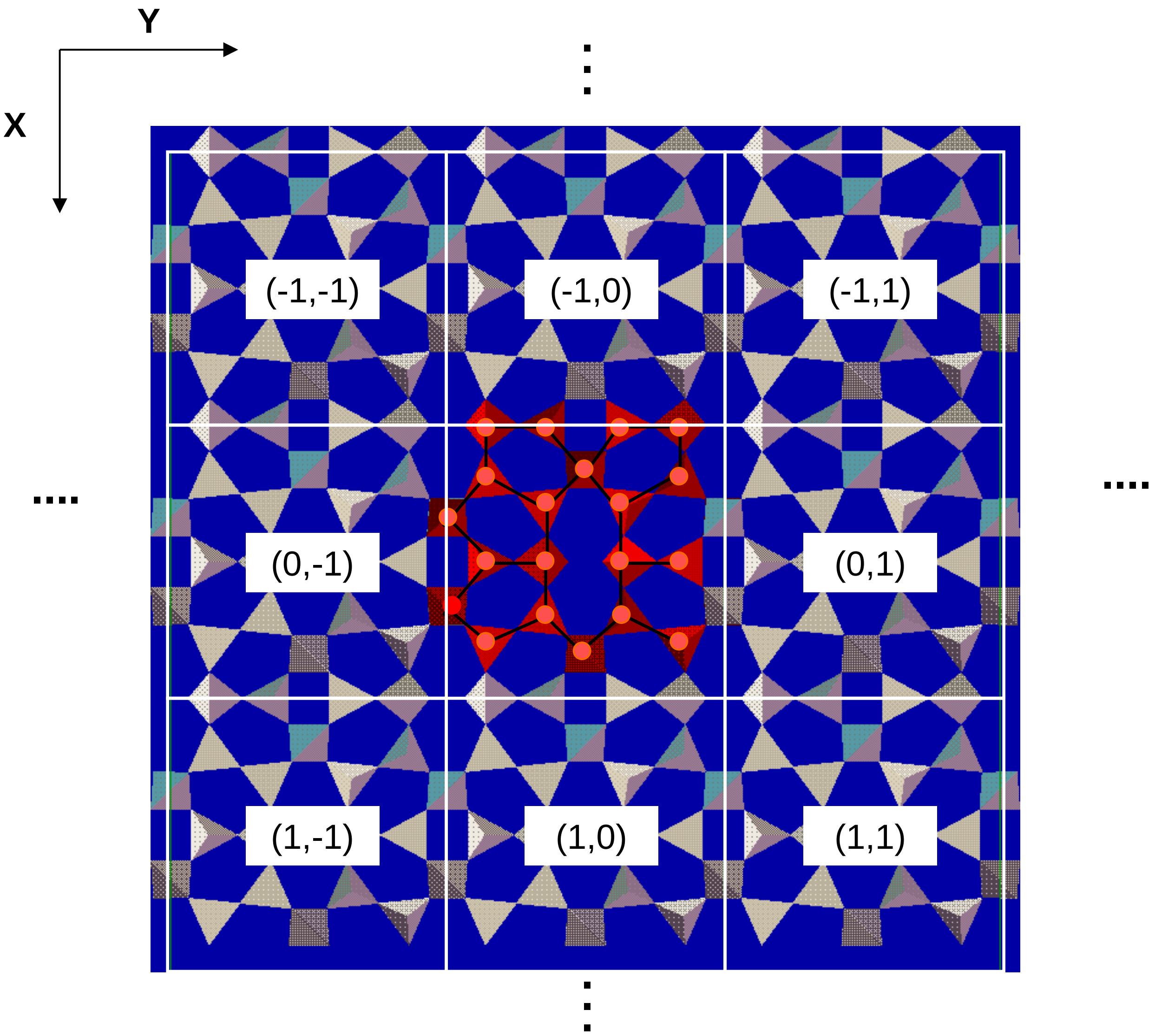


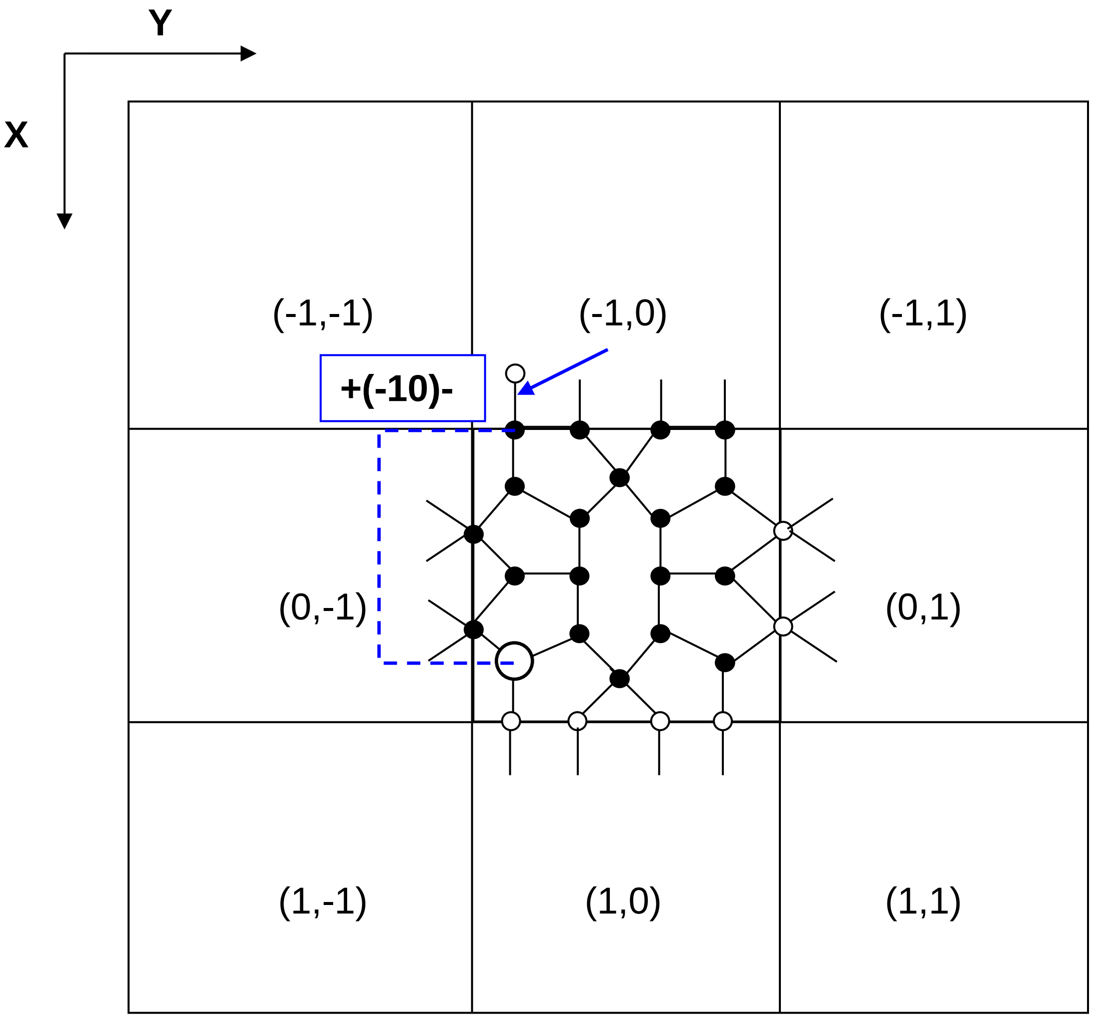


Periodicity



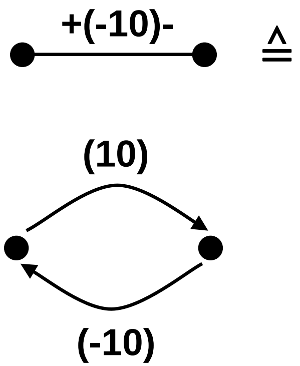
Part of a layer in semenovite

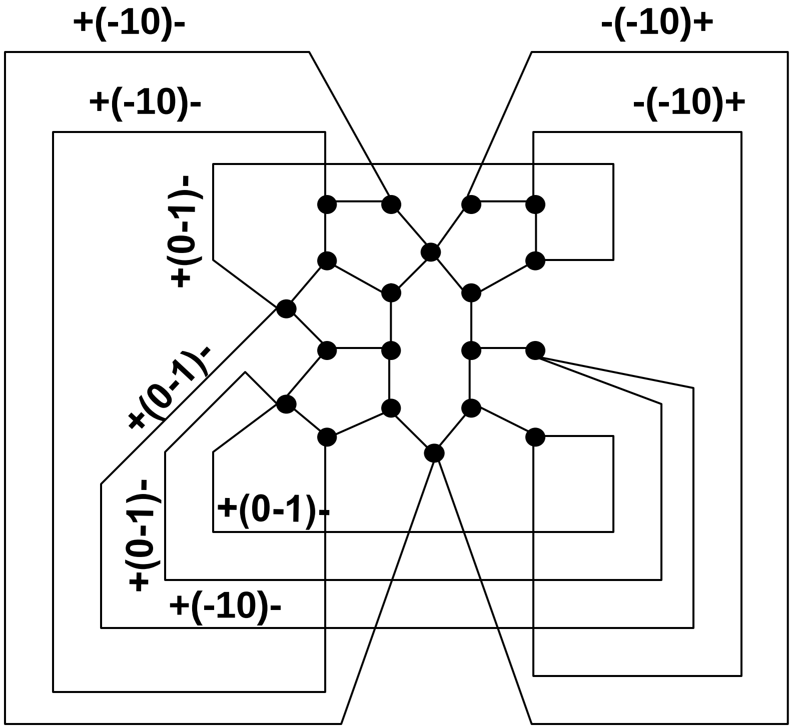



Y
X
(-1,-1)
(-1,0)
(-1,1)
+(-10)-
(0,-1)
(0,1)
(1,-1)
(1,0)
(1,1)

Labeled quotient graph / Direction-labeled graph

Chung/Hahn/Klee (1984), Goetzke/Klein (1987)

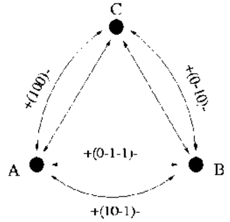




Some properties of direction-labeled graphs

- n-colourability is decidable for n $\leq$ 2 but undecidable for n > 2.
- Decomposition into fundamental chains (Liebau method) is NP-complete.
- Isomorphism problem?

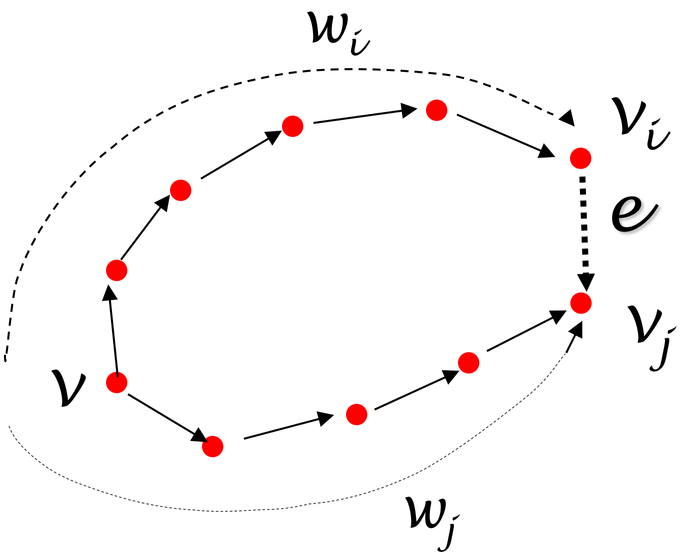
Dimensionality



$$\Delta(v_1,e_1,v_2,\ldots,v_{l-1},e_{l-1},v_l) \quad =_{df} \sum_{i=1}^{l-1} \delta(e_i)$$ (path direction)

$DR(G_{dl}) =_{df} <\{\Delta(c) \mid c \text{ Zyklus in } G_{dl}\}> \cup \{(0,0,0)\}$ (set of directions of repetition)

Dimension of a graph G_{dl}: Rank (dimension) of $DR(G_{dl})$



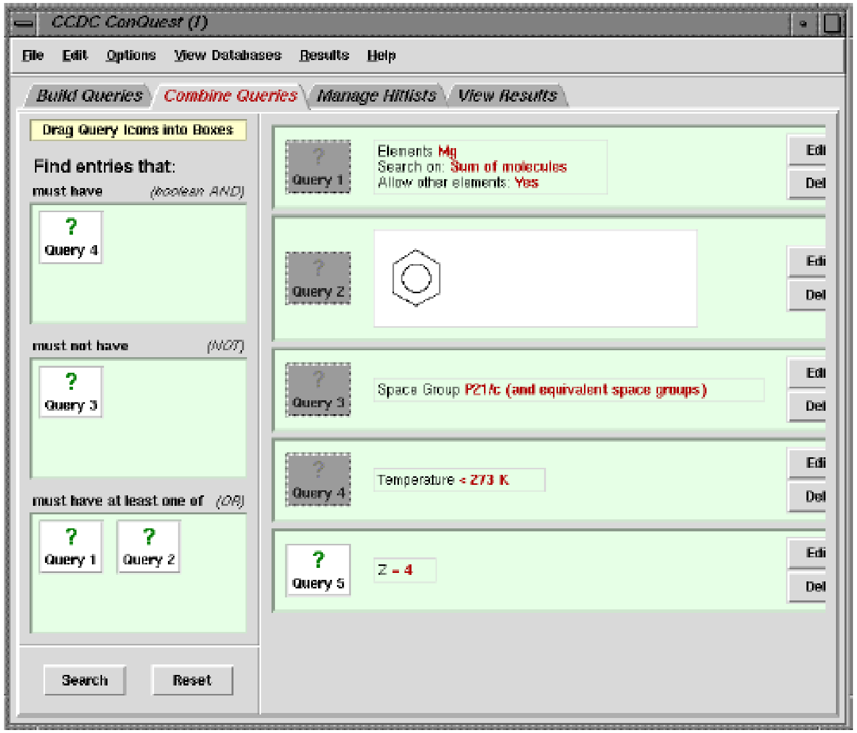
$<\{\Delta(w_i) + \delta(e) - \Delta(w_j) \mid e \in E,\ \alpha(e) = v_i,\ \omega(e) = v_j\}>$

Organic databases:

Relatively small number of rigid substructures, mapping of 2D chemical structure onto 3D molecular structure can be used for substructure searches.

Example: **C**ambridge **S**tructural **D**atabase (organic and metal-organic compounds); more than 500,000 single-crystal X-ray structures.

Search facilities: Structural search by drawing all or part of a molecule, selection by providing chemical, bibliographic, ... information.



Drag Query Icons into Boxes

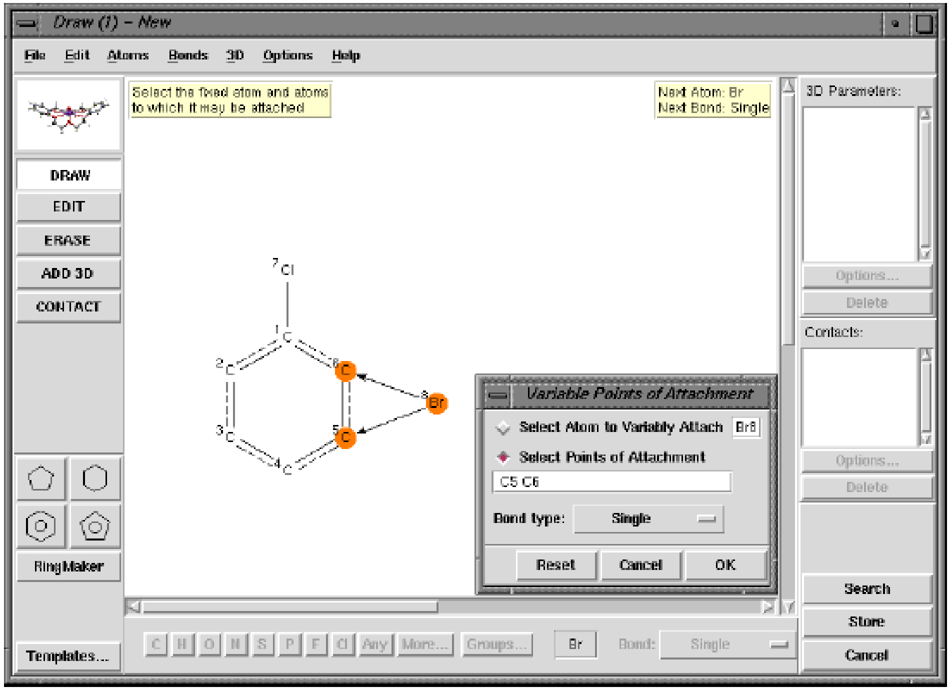
Find entries that:

Elements Mg
Search on: Sum of molecules
Allow other elements: Yes

Space Group P21/c (and equivalent space groups)

Temperature < 273 K

Z = 4



Select the fixed atom and atoms to which it may be attached

Next Atom: Br
Next Bond: Single

◇ Select Atom to Variably Attach Br8

◆ Select Points of Attachment

C5 C6

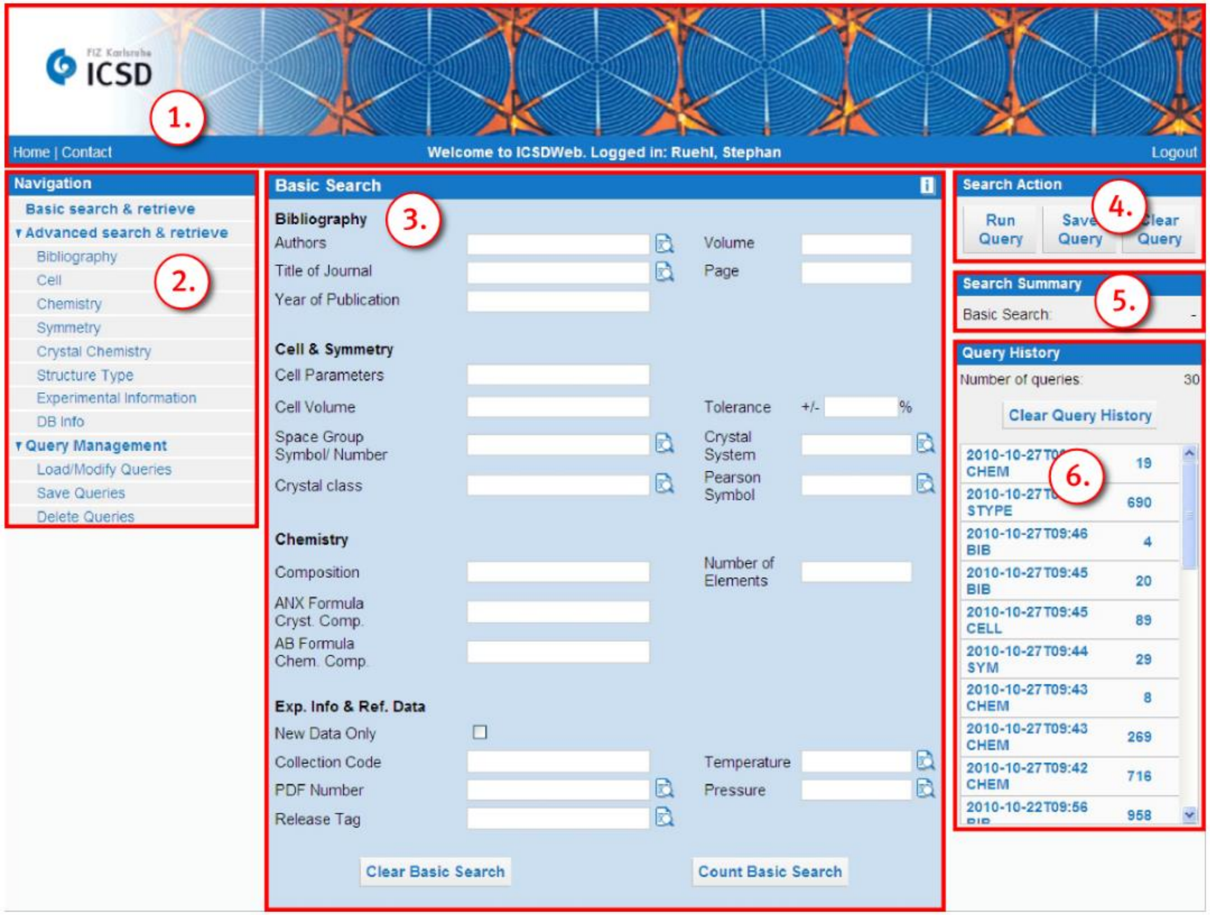
Bond type: Single

Inorganic databases:

Large variety of chemical elements and patterns ⇒
Problem to fix a suitable set of substructures for indexation.

Example: **I**norganic **C**rystal **S**tructure **D**atabase; more than 140,000 entries.

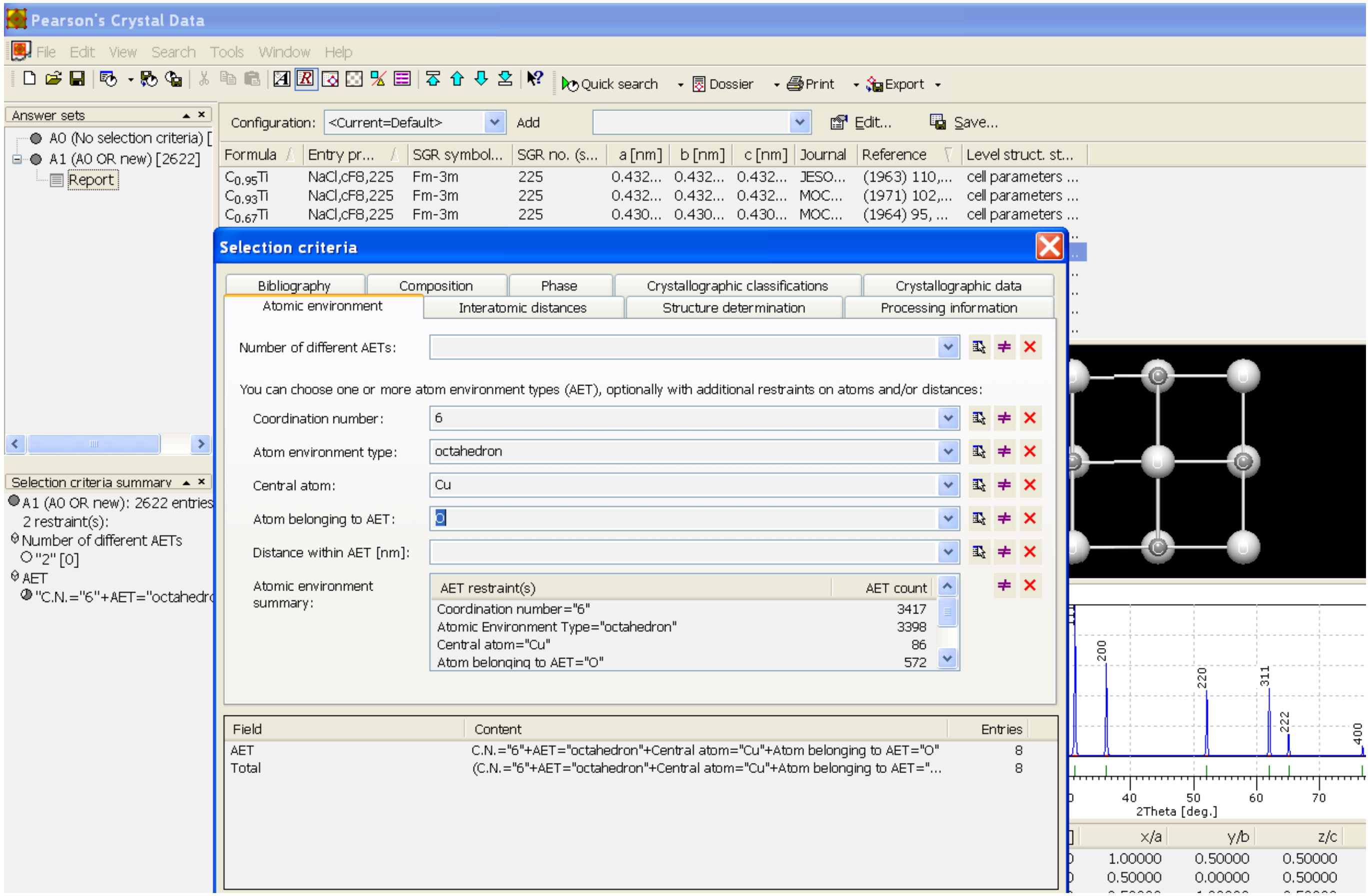
Search facilities: Selection in the categories
cell, chemistry, symmetry, crystal chemistry, structure type, bibliography.



Inorganic databases:
Large variety of chemical elements and patterns ⇒
Problem to fix a suitable set of substructures for indexation.

Example: Pearson's Crystal Data; about 212,500 entries.

Search facilities: interatomic distances, phase information, chemical composition, atomic environment (coordination number, atom coordination), ...



Configuration: <Current=Default> Add Edit... Save...

Formula	Entry pr...	SGR symbol...	SGR no. (s...	a [nm]	b [nm]	c [nm]	Journal	Reference	Level struct. st...
$C_{0.95}Ti$	NaCl,cF8,225	Fm-3m	225	0.432...	0.432...	0.432...	JESO...	(1963) 110,...	cell parameters ...
$C_{0.93}Ti$	NaCl,cF8,225	Fm-3m	225	0.432...	0.432...	0.432...	MOC...	(1971) 102,...	cell parameters ...
$C_{0.67}Ti$	NaCl,cF8,225	Fm-3m	225	0.430...	0.430...	0.430...	MOC...	(1964) 95, ...	cell parameters ...

Selection criteria summary
A1 (A0 OR new): 2622 entries
2 restraint(s):
Number of different AETs
"2" [0]
AET
"C.N.="6"+AET="octahedr

Number of different AETs:

You can choose one or more atom environment types (AET), optionally with additional restraints on atoms and/or distances:

Coordination number: 6
Atom environment type: octahedron
Central atom: Cu
Atom belonging to AET: O
Distance within AET [nm]:

Atomic environment summary:

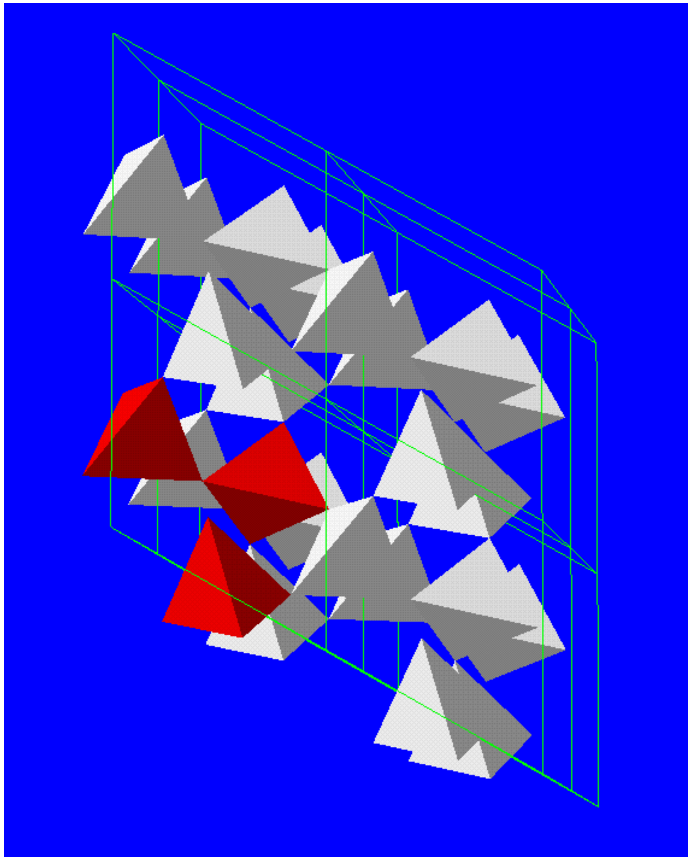
AET restraint(s)	AET count
Coordination number="6"	3417
Atomic Environment Type="octahedron"	3398
Central atom="Cu"	86
Atom belonging to AET="O"	572

Field	Content	Entries
AET	C.N.="6"+AET="octahedron"+Central atom="Cu"+Atom belonging to AET="O"	8
Total	(C.N.="6"+AET="octahedron"+Central atom="Cu"+Atom belonging to AET="...	8

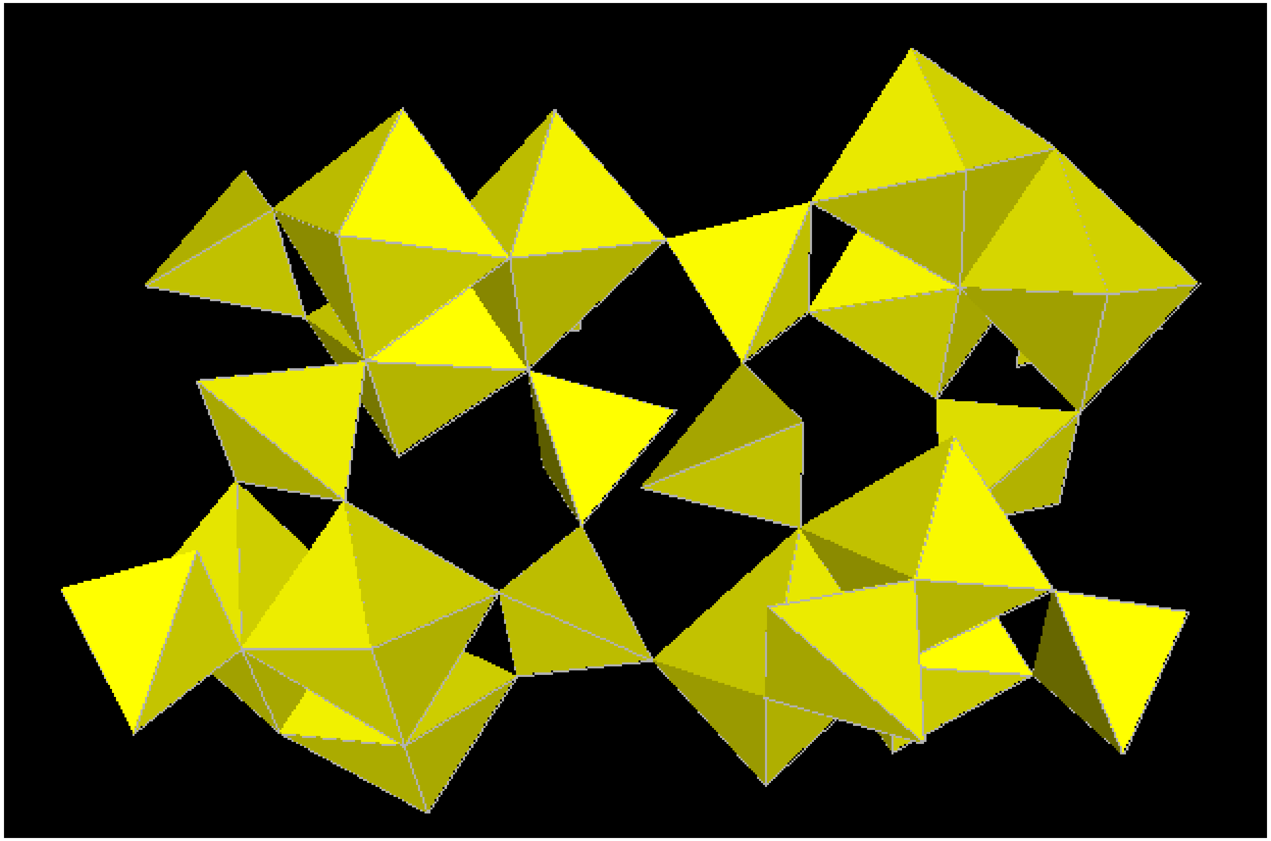
x/a	y/b	z/c
1.00000	0.50000	0.50000
0.50000	0.00000	0.50000

Common level of description: coordination polyhedra and their connections.

Part of the tetrahedral network of α-quartz

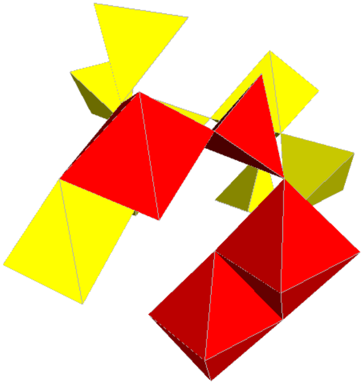


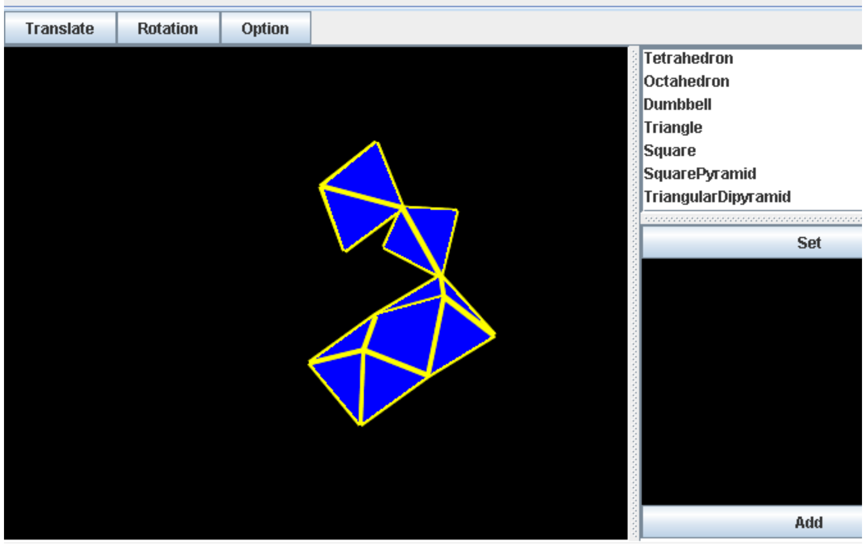
Vertex and edge sharing of octahedra and tetrahedra in zoisite



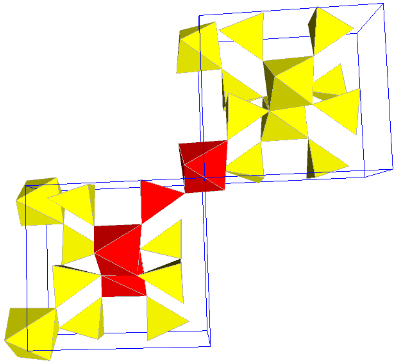
The problem

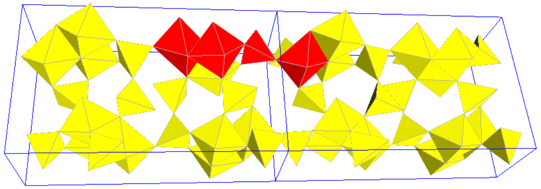
Given the structural pattern of a part of a (real or hypothetical) chemical compound.

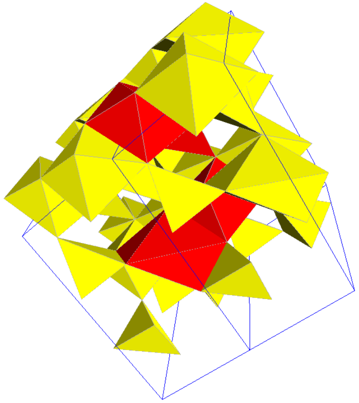




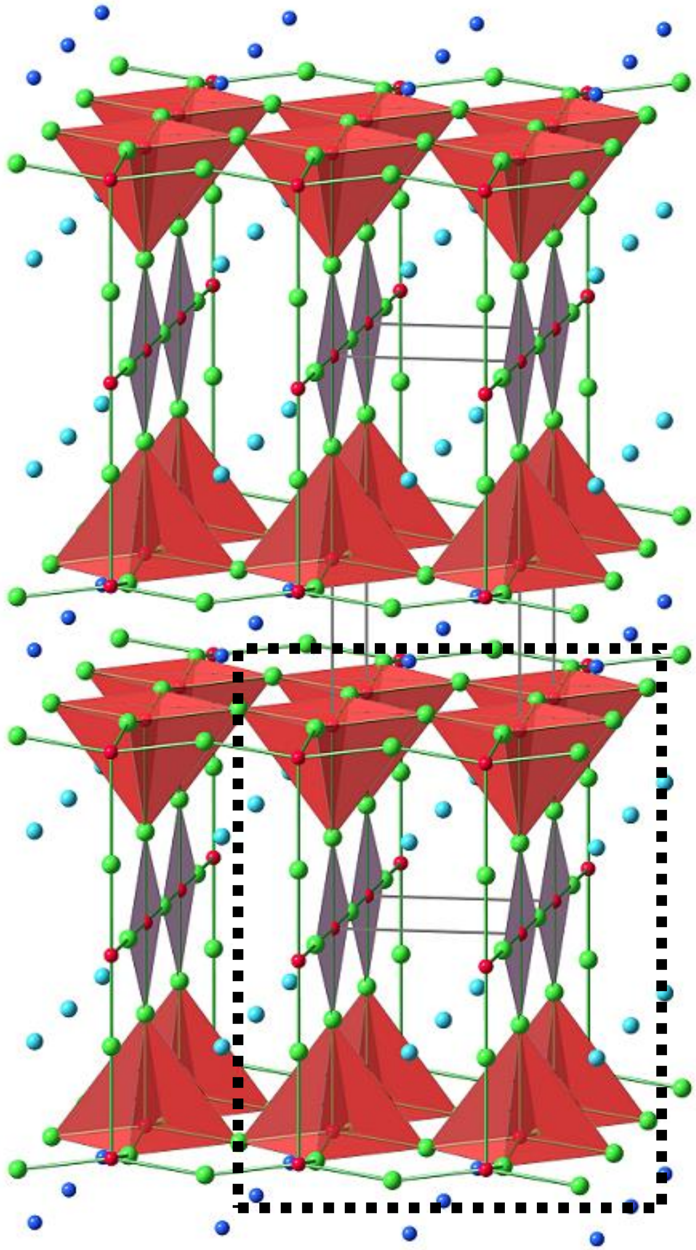
Find all compounds in a given set of models with ***similar*** structural patterns.







Combination of graphical and textual query parameters



+

Cations: Cu

Anions: O

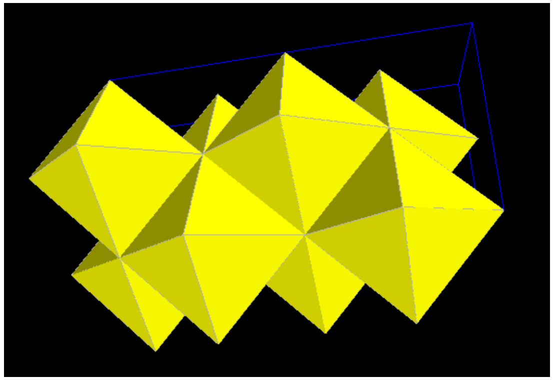
Bonds: 'strong'

Pyramids: inclination > 55°

Difficulties

- Definition of coordination polyhedra (linear/quadratic gap, bond valence > 0.02 vu, ... ?).

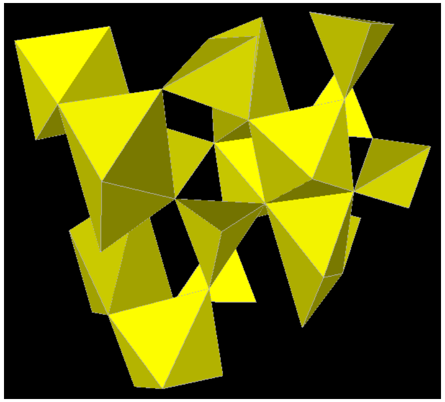
Sodium chloride



Na, Cl

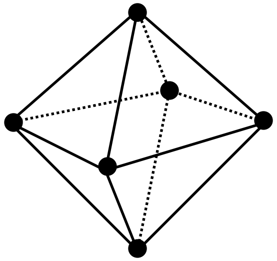
.....

Spodumene

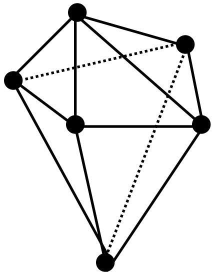


Li, Al, Si, O

Regular convex polyhedra



as well as distorted polyhedra



Similar: The use of the term 'similar', ..., arises from the inherent difficulty in defining *a priori* limits on the similarity of geometrical configurations or physical/chemical characteristics.

(In: Terms that define different degrees of similarity between inorganic structures, *Nomenclature Commission, International Union of Crystallography*)

Search for substructures should be flexible!

Two phases:

1. Determine ***topologically equivalent*** substructures.
2. Check possible embeddings for ***geometrical conformity***.

Modelling of polyhedral networks

Problems to be solved:

a) Large variety of polyhedra (*regular* or *distorted*).

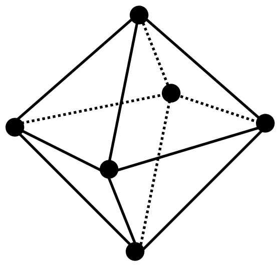
b) Three kinds of connections between polyhedra (*vertex, edge,* or *face*).

c) Infinite periodic structures.

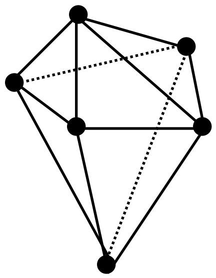
Topological view of convex polyhedra:

Three-connected planar graphs (Theorem of Steinitz).

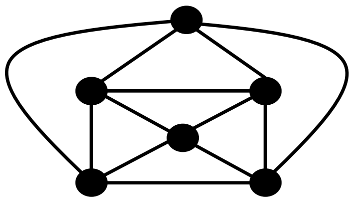
Equivalence of polyhedra: Isomorphic topological views (allowing distortions).



and



are topologically equivalent:



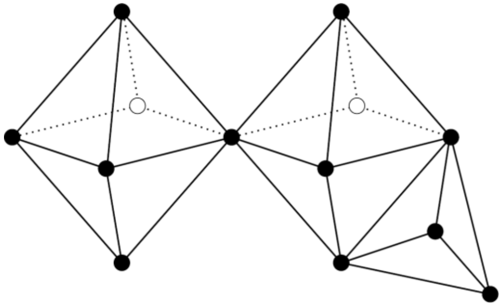
Clusters of polyhedra: Connected units of polyhedra.

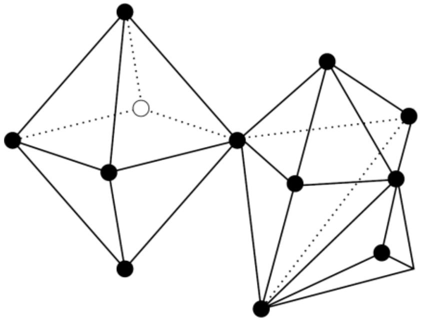
Equivalence of clusters? We are dealing with experimental data!

Transformation should be possible without breaking connections between polyhedra (but allowing distortions).

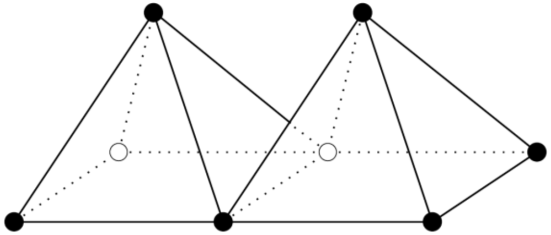
Example:

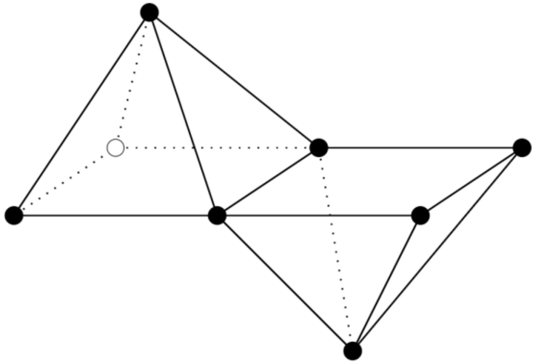
Equivalent:



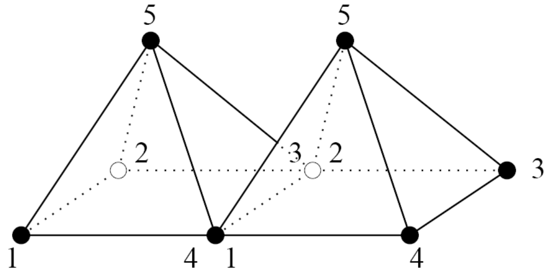


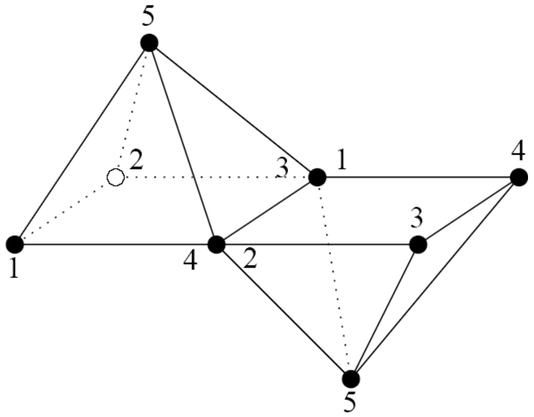
Not equivalent:



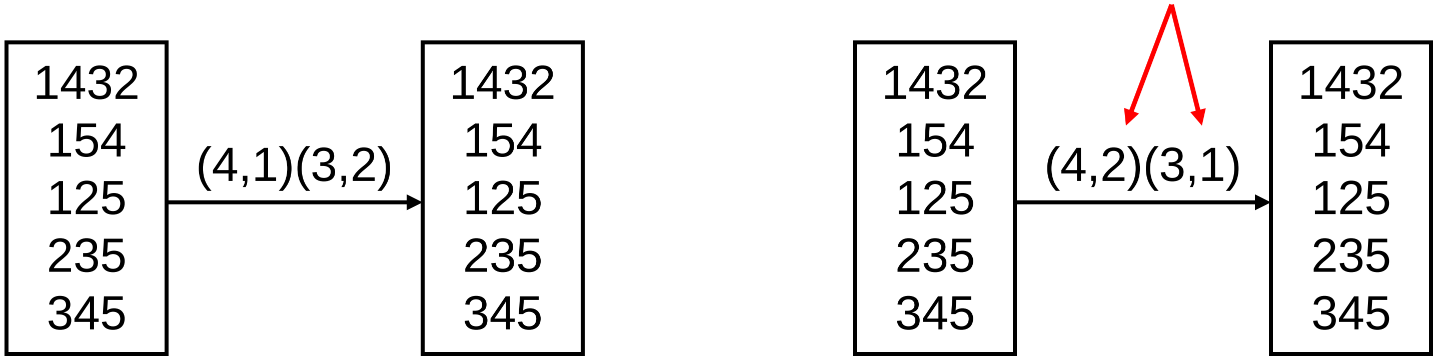


Graph representation





Ordered face representation



1432
154
125
235
345

(4,1)(3,2)

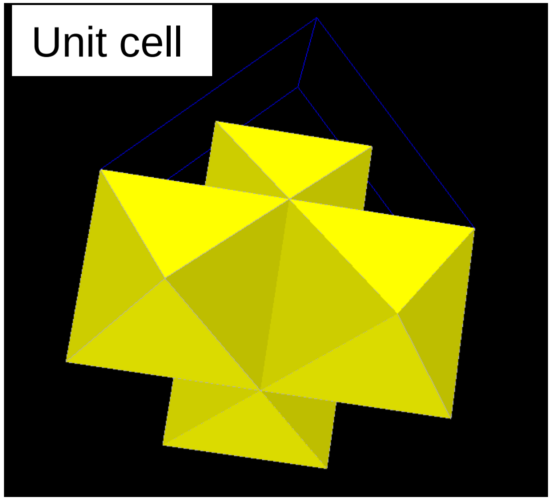
1432
154
125
235
345

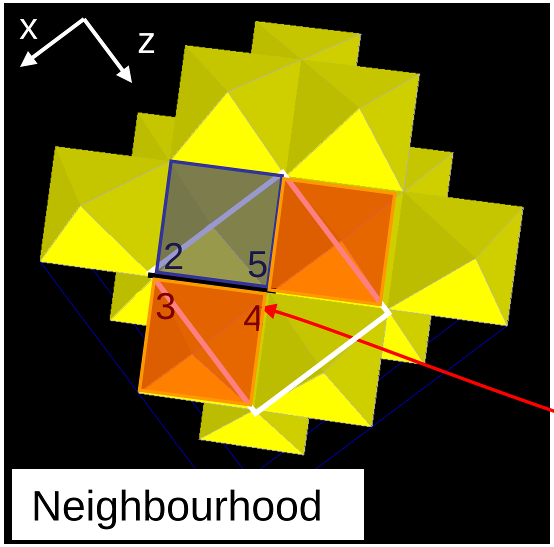
1432
154
125
235
345

(4,2)(3,1)

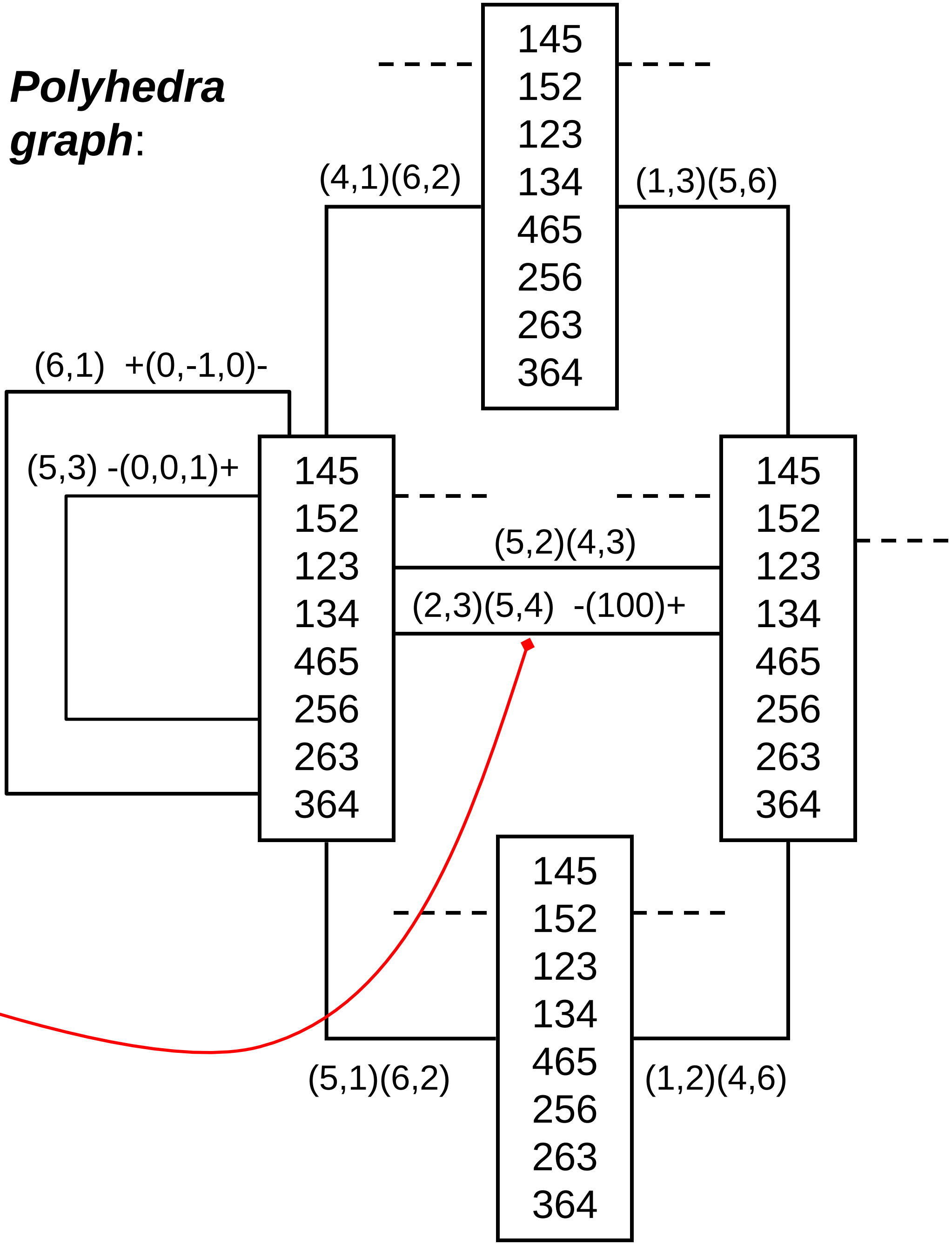
1432
154
125
235
345

Sodium chloride





***Polyhedra graph*:**


(4,1)(6,2)

(1,3)(5,6)

(6,1) +(0,-1,0)-

(5,3) -(0,0,1)+

(5,2)(4,3)

(2,3)(5,4) -(100)+

(5,1)(6,2)

(1,2)(4,6)

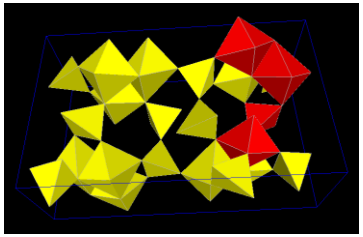
145
152
123
134
465
256
263
364

145
152
123
134
465
256
263
364

145
152
123
134
465
256
263
364

145
152
123
134
465
256
263
364

Topological search



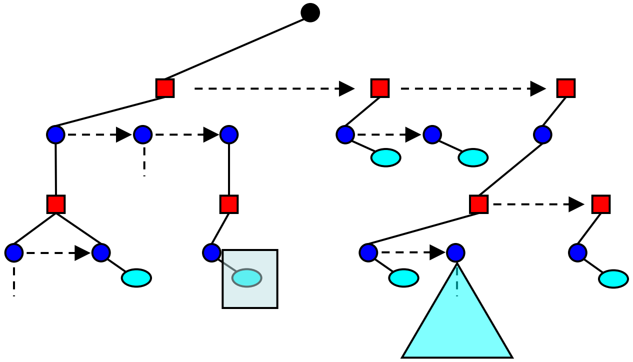
Subgraph isomorphism problem: **computationally hard**.

Preprocessing of model structures in the database.

Indexation of polyhedra graphs

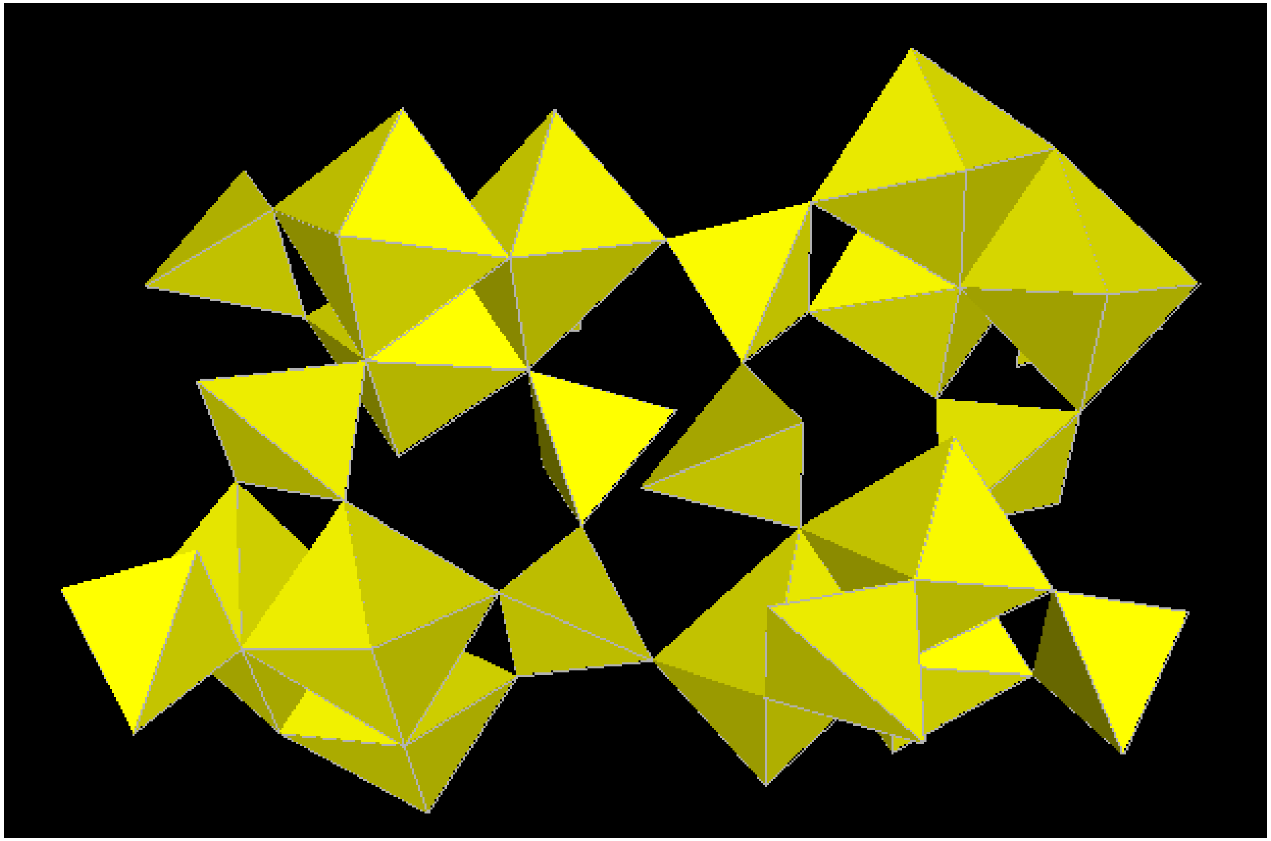
Proceeding

- Consider paths up to some fixed limit length.
- Extract information relevant for topological search.
- Organize this information as an index.

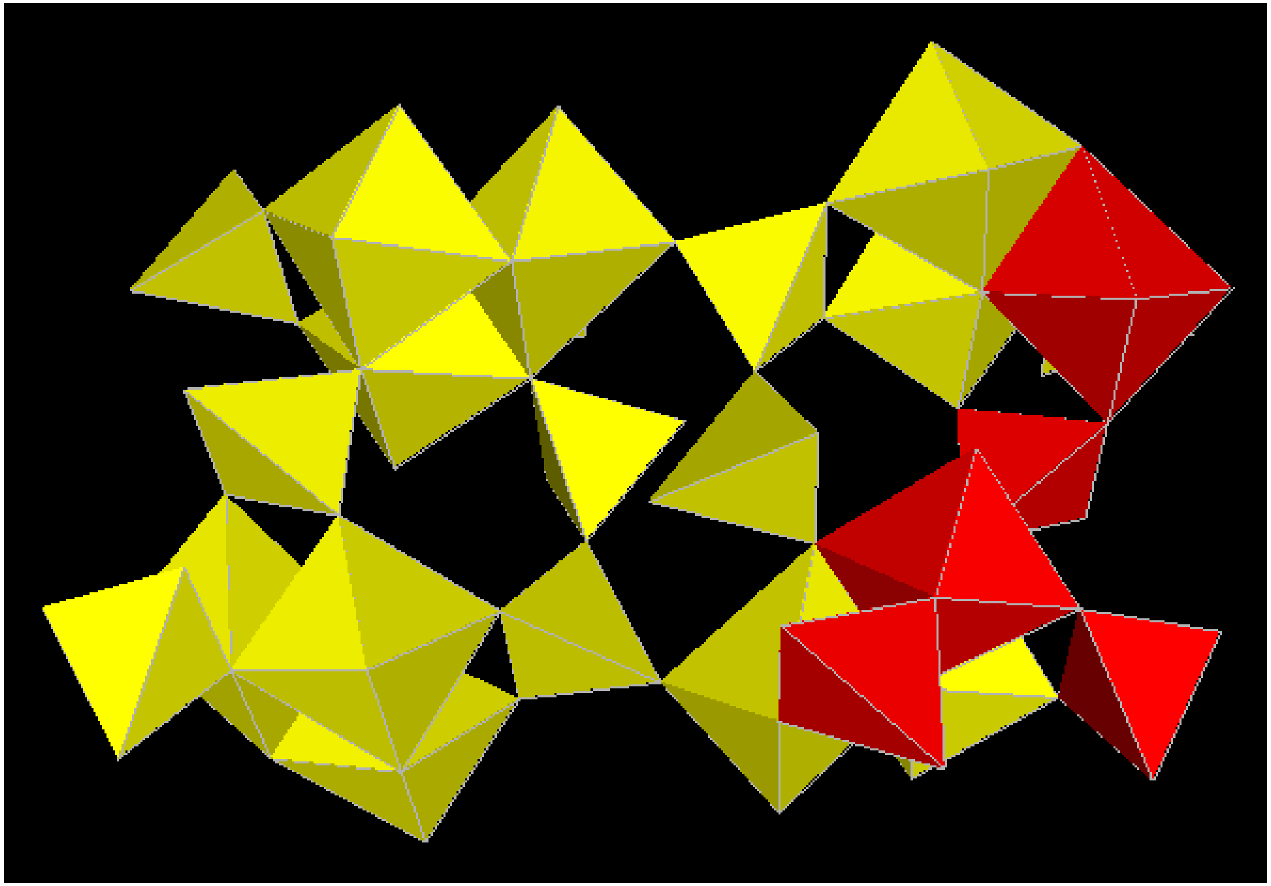


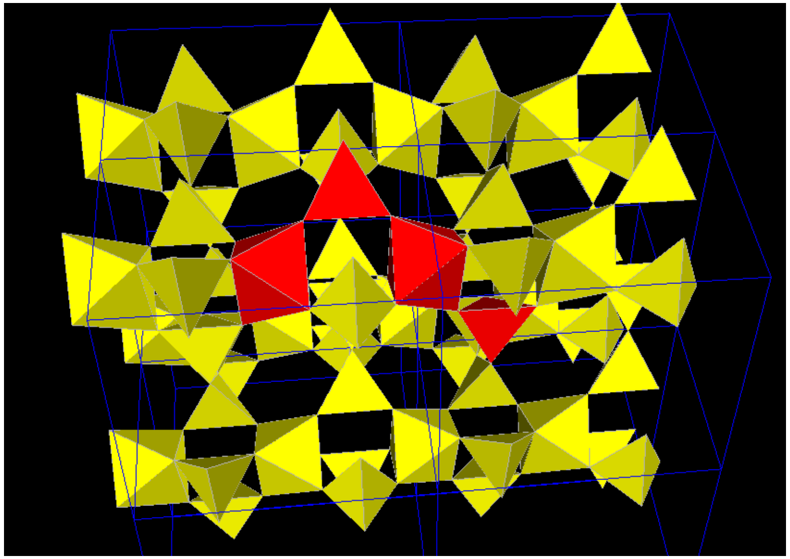
Topological search

Retrieval of substructures



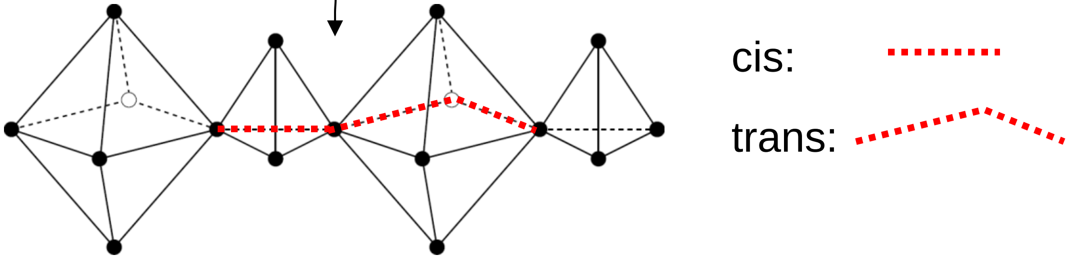
Substructure marked for search





topological information

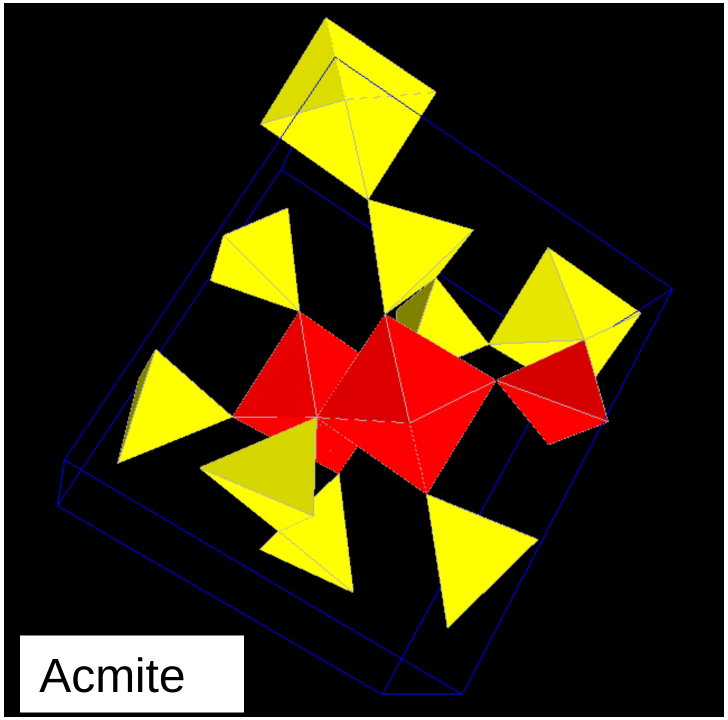
Search chain

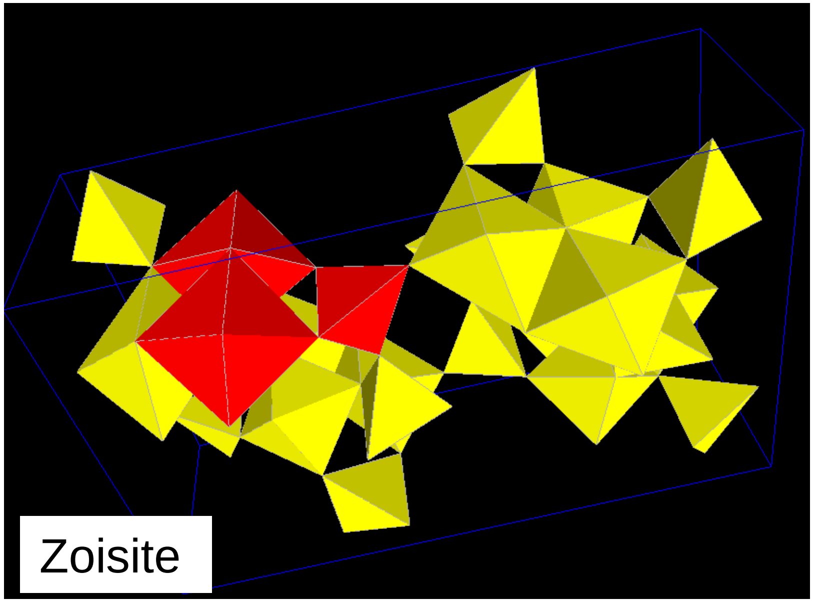


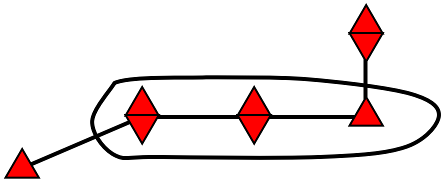
Polyhedron	Immediate neighbours	Connection
octahedron	6	1
tetrahedron	4	1, cis
octahedron	6	1, trans
tetrahedron	4	-

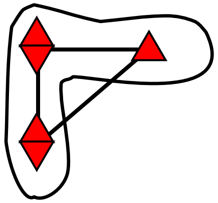
max, extendable, no precycle

coding





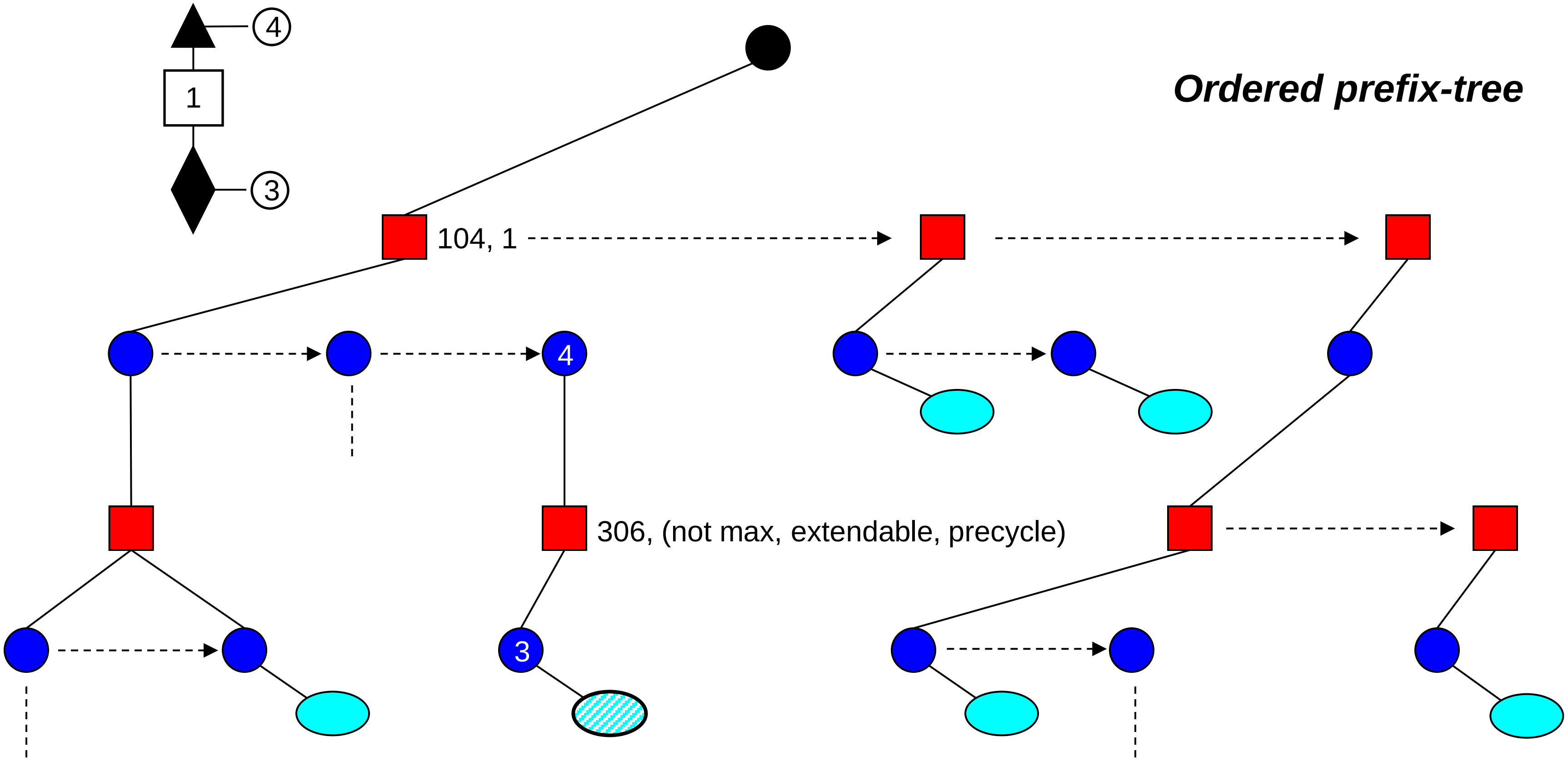




Precycle

jump

Organization of search chains



Ordered prefix-tree

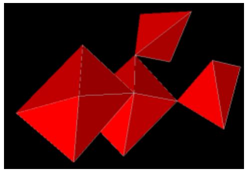
🟥 **p-node**: Polyhedron identifier, kind of connection to successor or general information on search chain (in case of last polyhedron).

Exact match

● **l-node**: Number of immediate neighbours.

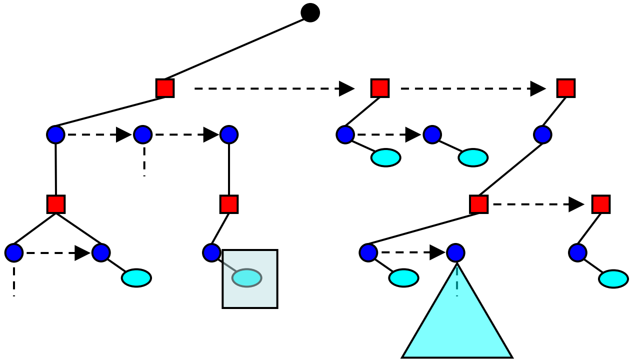
⬭ **i-node**: Ordered list of identifiers of model graphs.

Matching: ≤

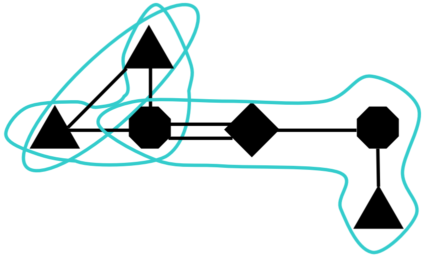


Determination of isomorphic substructures

1. Compute search chains for the input structure.
2. Collect candidates by inspecting the prefix tree.

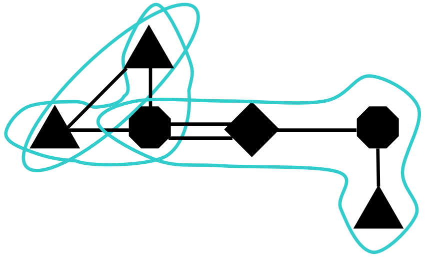


3. Determine an edge covering of the input structure.
4. For every candidate model graph compute instances for all chains from this covering.
5. Try to find a subset of instances such that the corresponding subgraph is isomorphic to the input structure up to vertex labels.



Sketch of search:

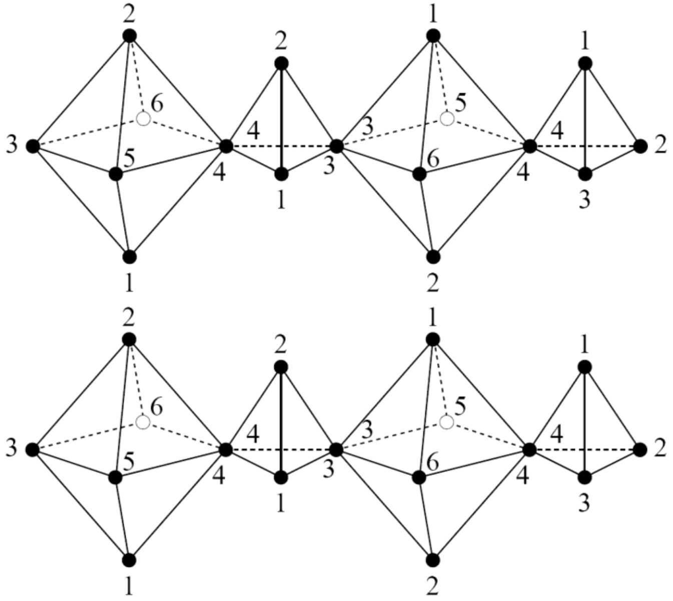
- Compute annotated paths for the input structure.



- Use the index to determine candidate model structures.

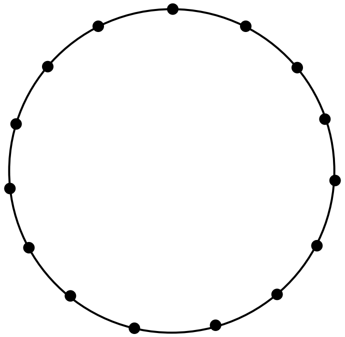
- Try to locate substructures in these model structures having the same path cover as the input structure.

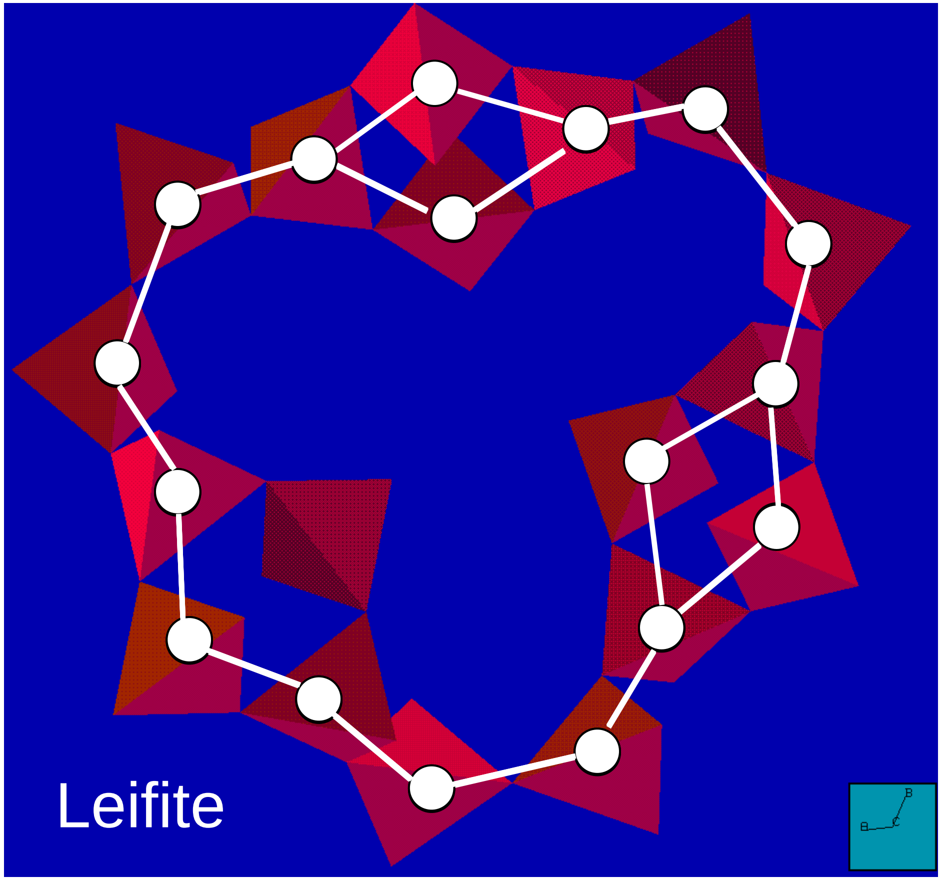
- Check for permutations of polyhedra vertices to get isomorphic graphs.



Answer complexity

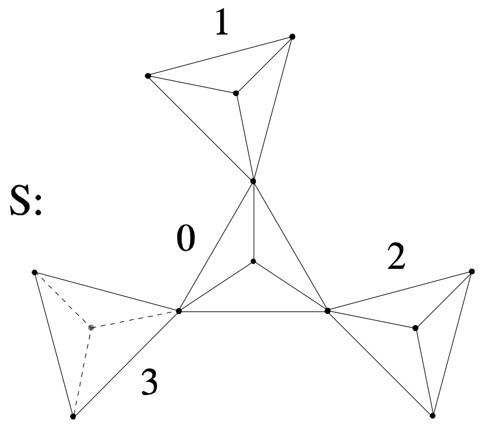
Number of isomorphic substructures in a single model structure?

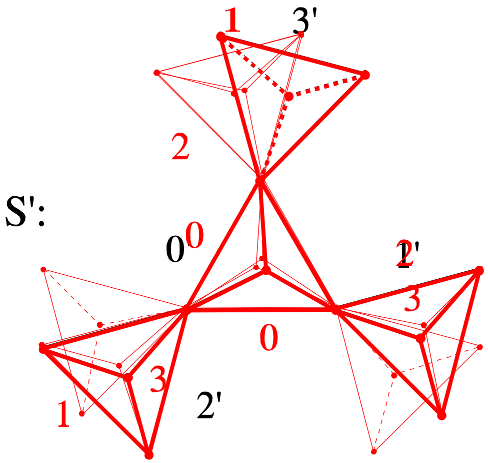




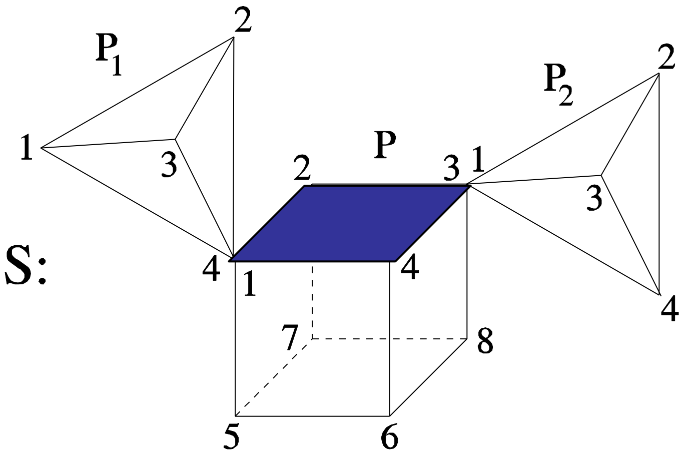
44 not symmetrically equivalent 15-membered rings in leifite.

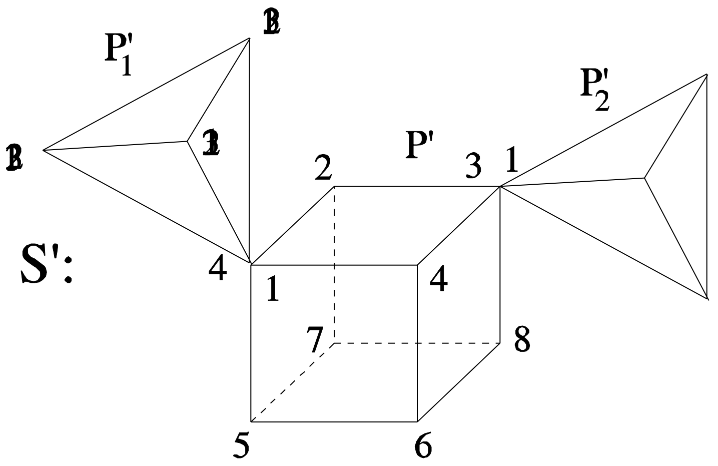
Embedding





Permutations





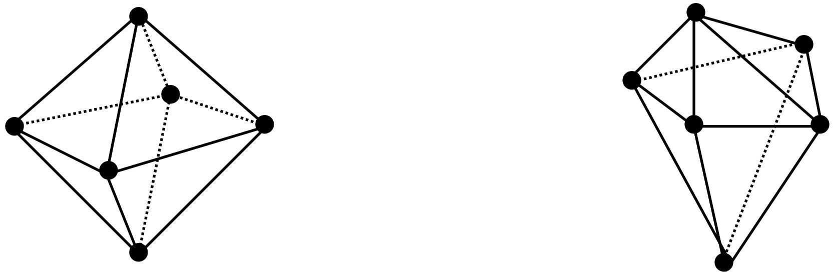
1234
1465
3864
2783
1572
5687

1234
1465
3864
2783
1572
5687

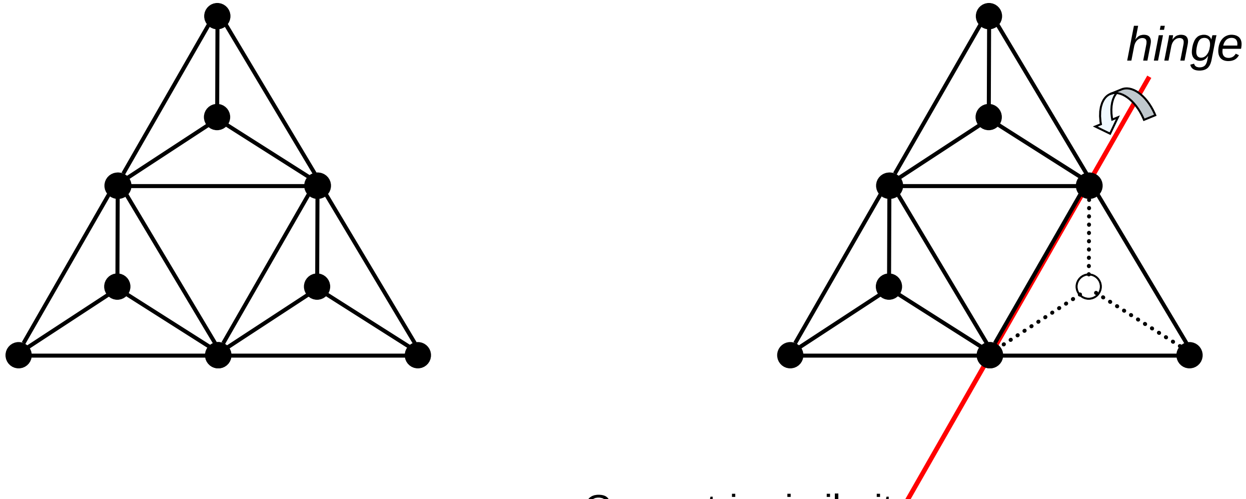
Geometric similarity

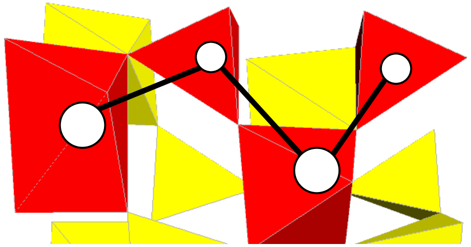
Two levels:

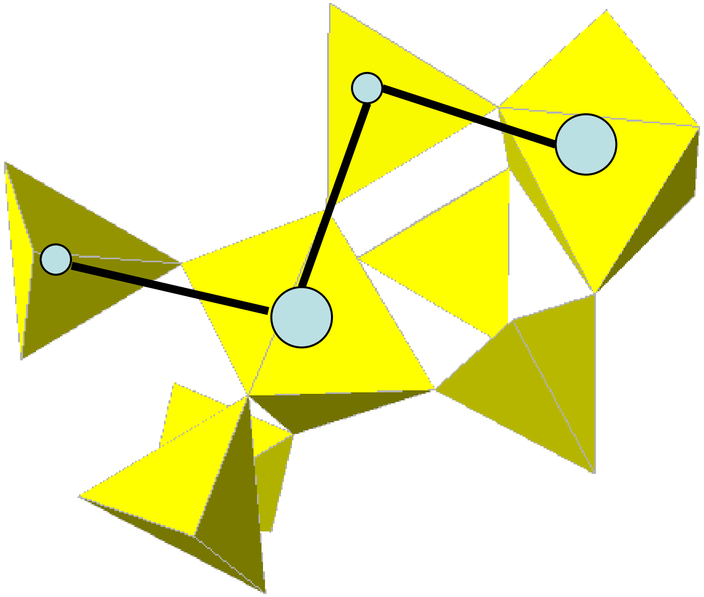
1. Polyhedra



2. Relative positioning of polyhedra.







To solve: *The problem of absolute orientation.*

$C_S : \{c_1,...,c_n\}$, $C_{S'} : \{c_1',...,c_n'\}$

sets of the coordinates of the central atoms of isomorphic structures S and S', resp.

Consider C_S and $C_{S'}$ as rigid subsets of $\mathbb{R}^3$.

Look for a motion *T* in the group of proper Euclidean motions solving the following least-squares problem:

$$U := \sum_{i=1}^{n} \|c_i' - T(c_i)\|_2^2 = min$$

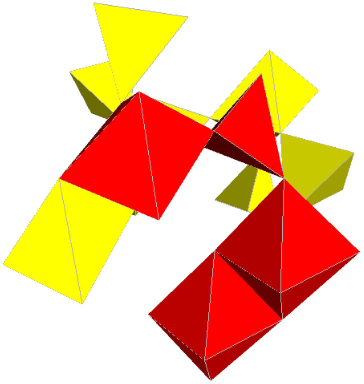
Measuring similarity:

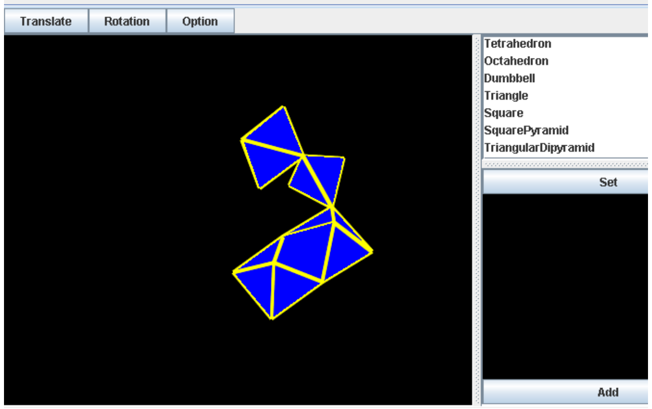
$$\varepsilon := \frac{\sqrt{U}}{n} \qquad \text{(\textbf{R}oot \textbf{M}ean \textbf{S}quare)}$$

Implementation: Closed-form solution using unit quaternions (algorithm of B.K.P. Horn, 1987).

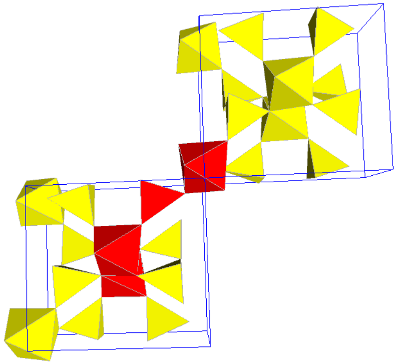
The result

Given: A structural pattern of a part of a chemical compound (real or hypothetical) and a database with structure data (including polyhedra graphs) and index.

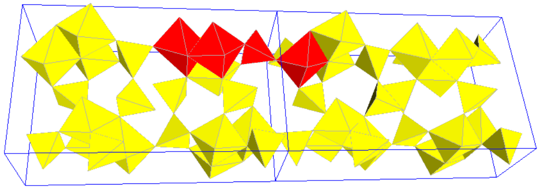




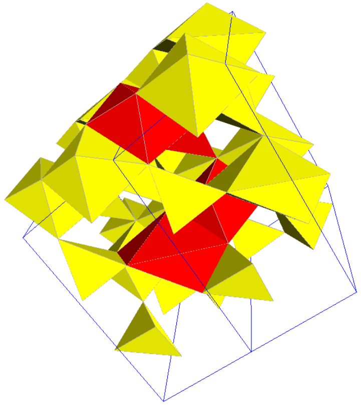
Answer: Compounds with isomorphic structural patterns and their RMS values.



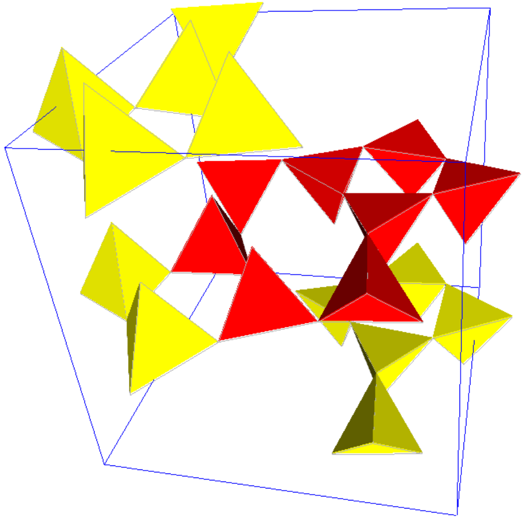
Jadeite: 0.680581



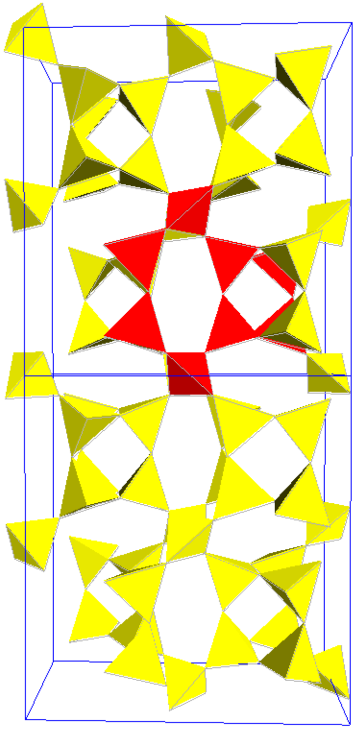
Zoisite: 0.789256



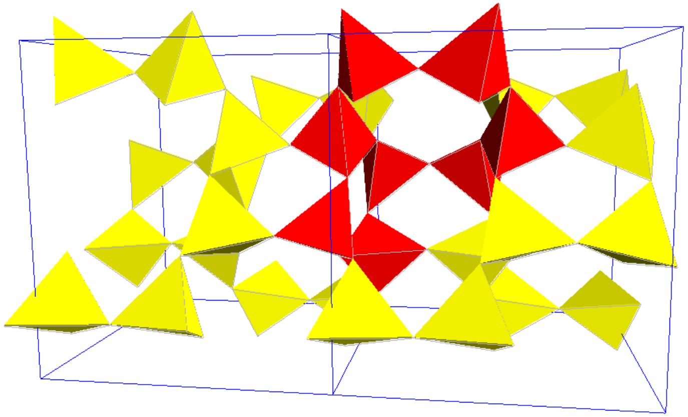
Spodumene: 0.110646



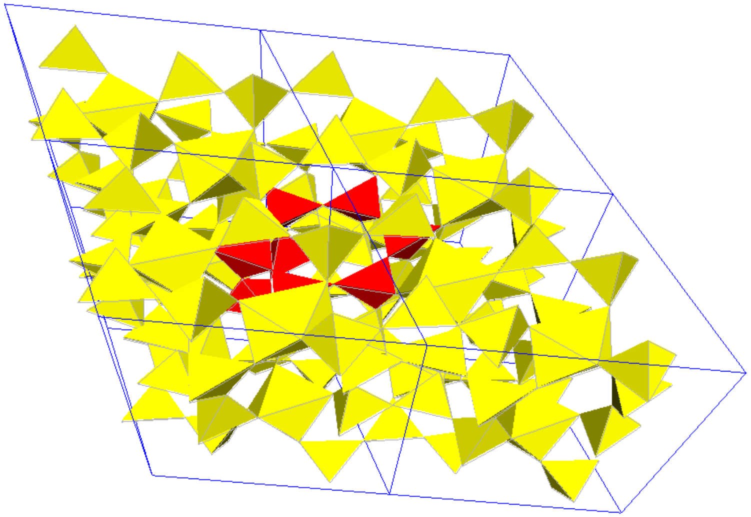
Search structure in aminoffite



Epididymite:
RMS 0.162449



Paracelsian:
RMS 0.485688



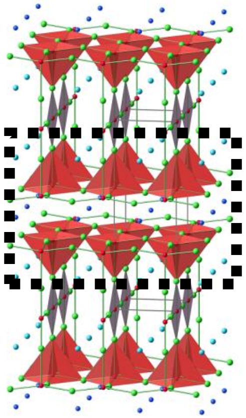
Merrihueite:
RMS 0.711812

Geometric similarity

Future work

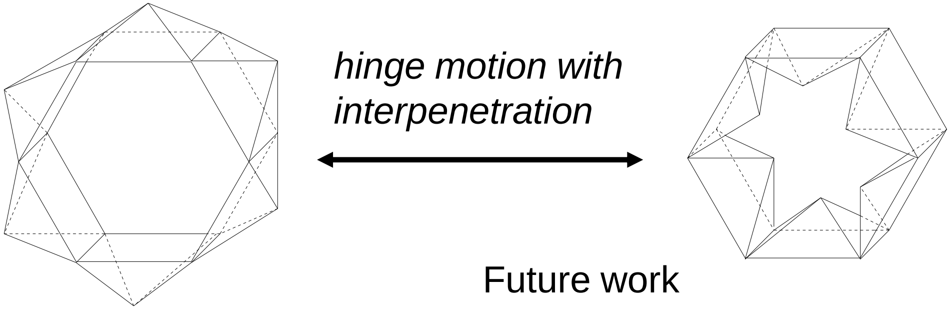
Searching: Improve the embedding algorithm (permutations, symmetries).

Allow more than one connected component:



Ranking: Include measures of distortion in the description of coordination polyhedra.

General: Investigate the realization space of polyhedra graphs (subspace of generalized hinge motions, generators,...?).


hinge motion with interpenetration