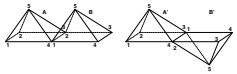
Graphical query interfaces for inorganic crystallographic databases

Hans-Joachim Klein

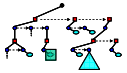
Department of Computer Science
Christian-Albrechts-Universität Kiel, Germany

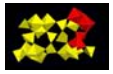
Overview


Problem description


Modelling of polyhedral networks


Graph representation


Topological search


The embedding problem


Geometric similarity

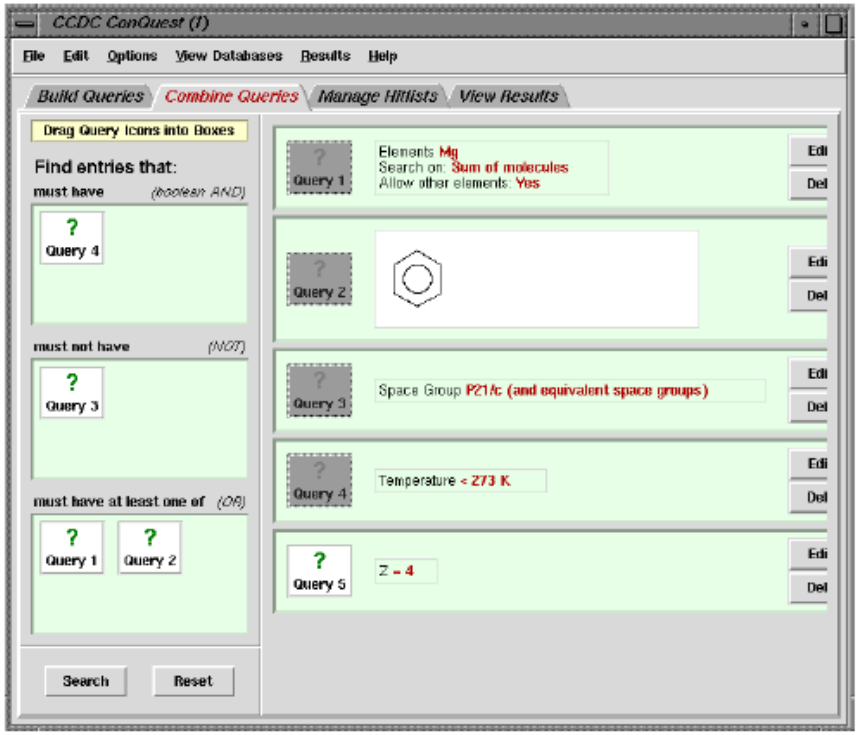

Future work

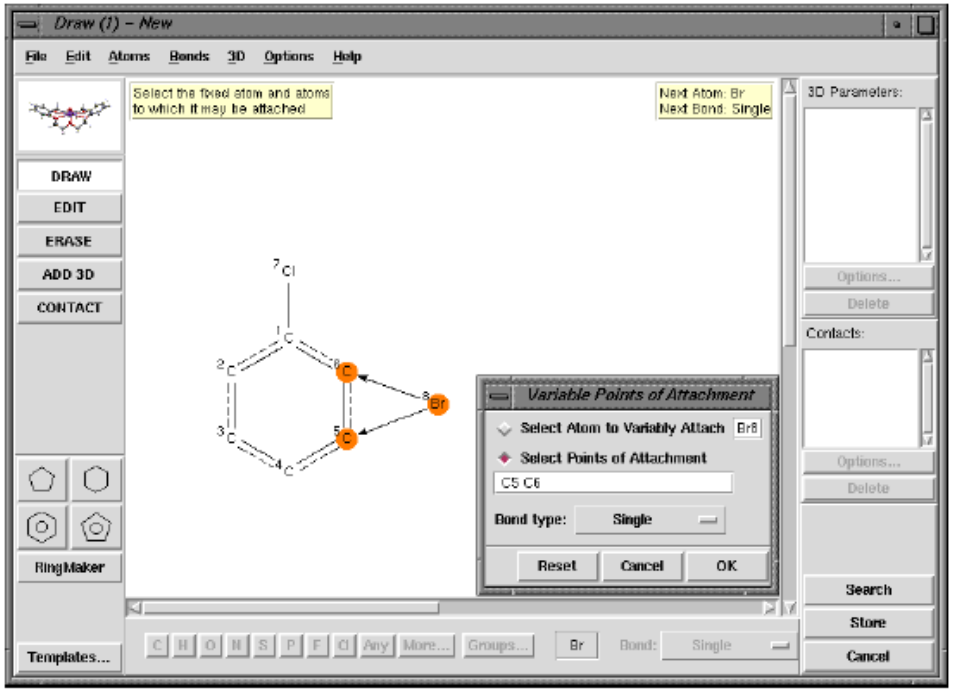
Organic databases:

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.





Inorganic databases:

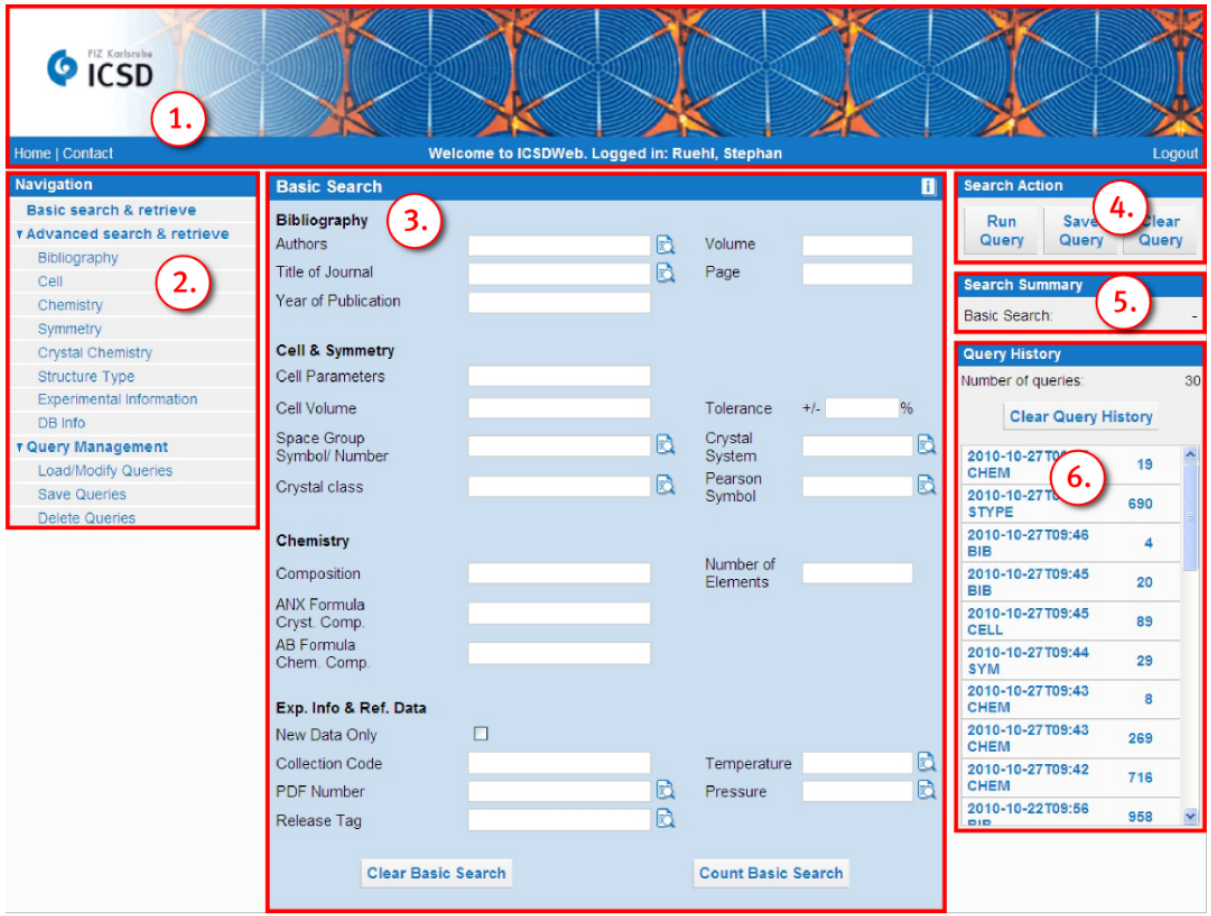
Large variety of chemical elements and patterns ⇒

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.



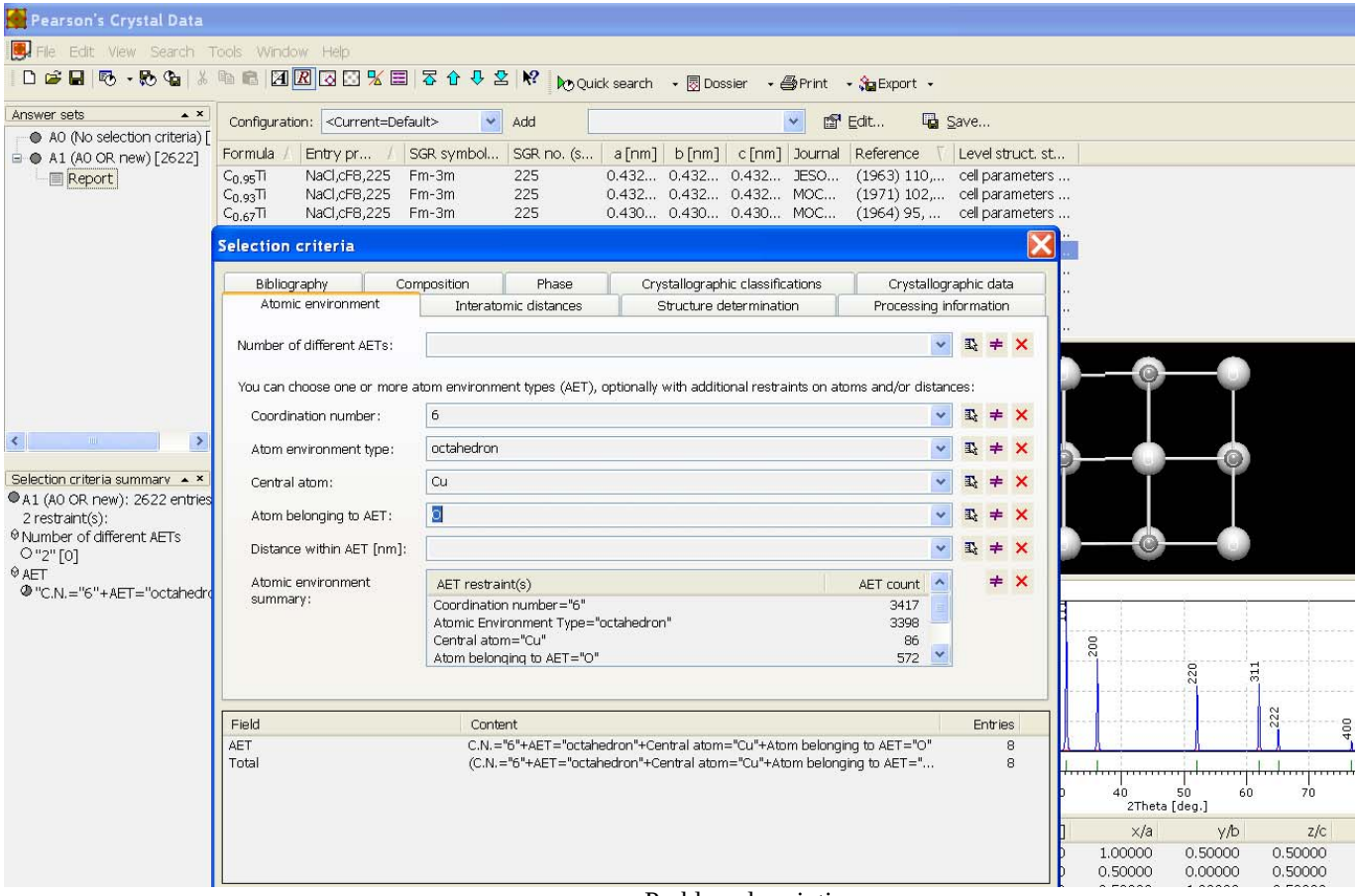
Inorganic databases:

Large variety of chemical elements and patterns ⇒

Problem to fix a suitable set of substructures for indexation.

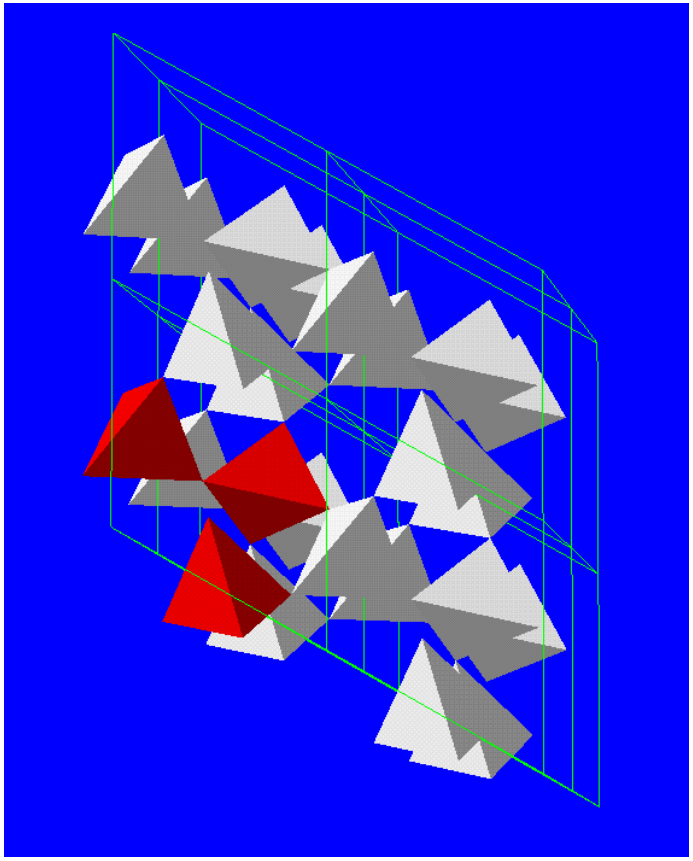
Example: Pearson's Crystal Data; about 212,500 entries.

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

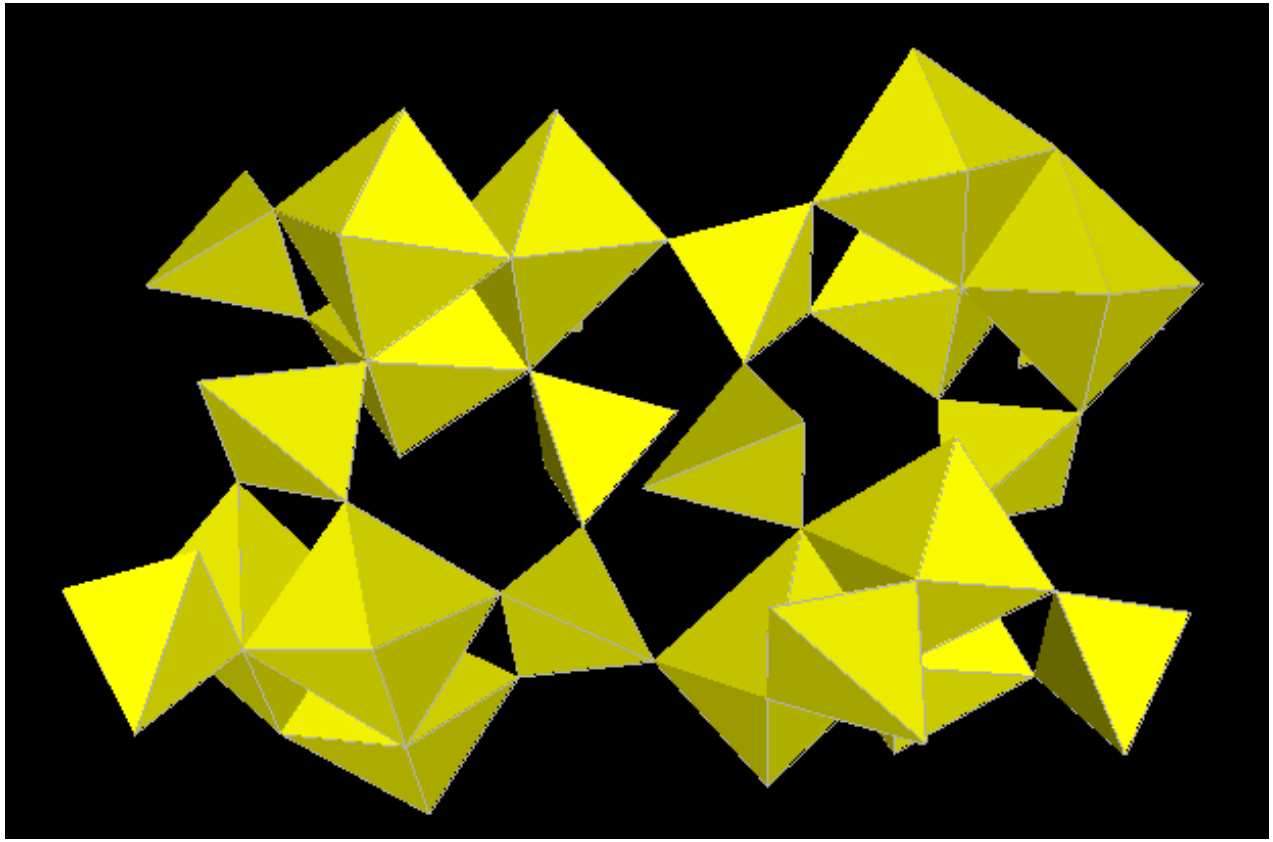


Common level of description: coordination polyhedra and their connections.

Part of the tetrahedral network of α-quartz

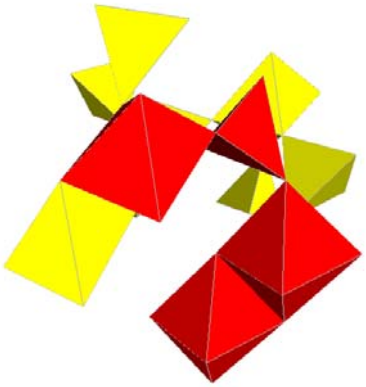


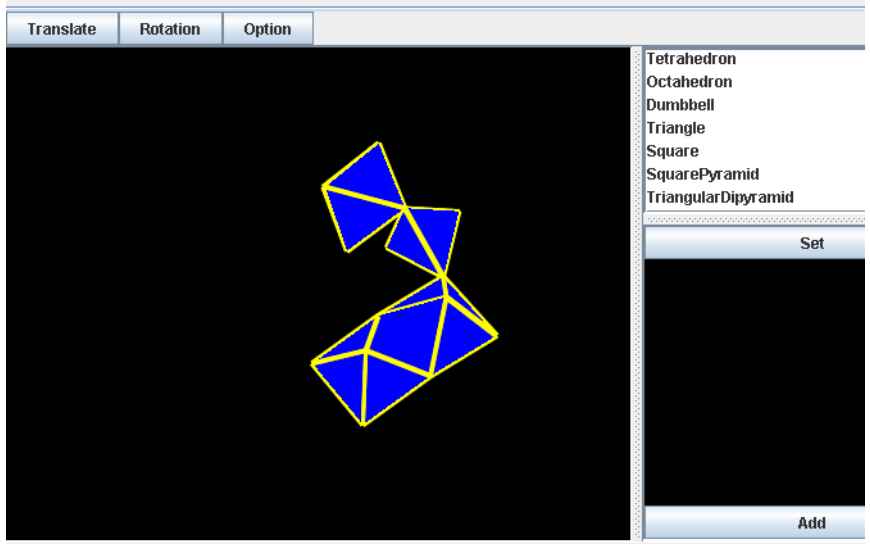
Vertex and edge sharing of octahedra and tetrahedra in zoisite



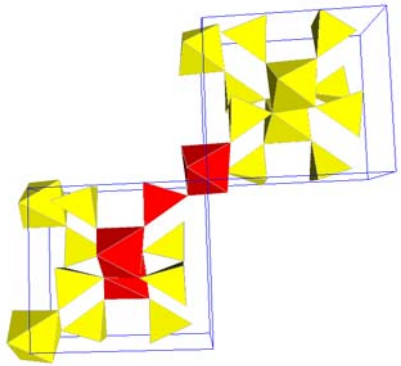
The problem

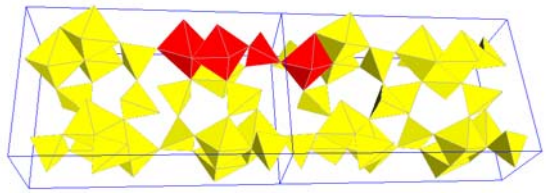
Given the structural pattern of a part of a (real or hypothetical) chemical compound.

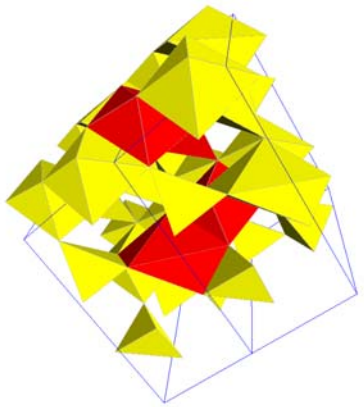


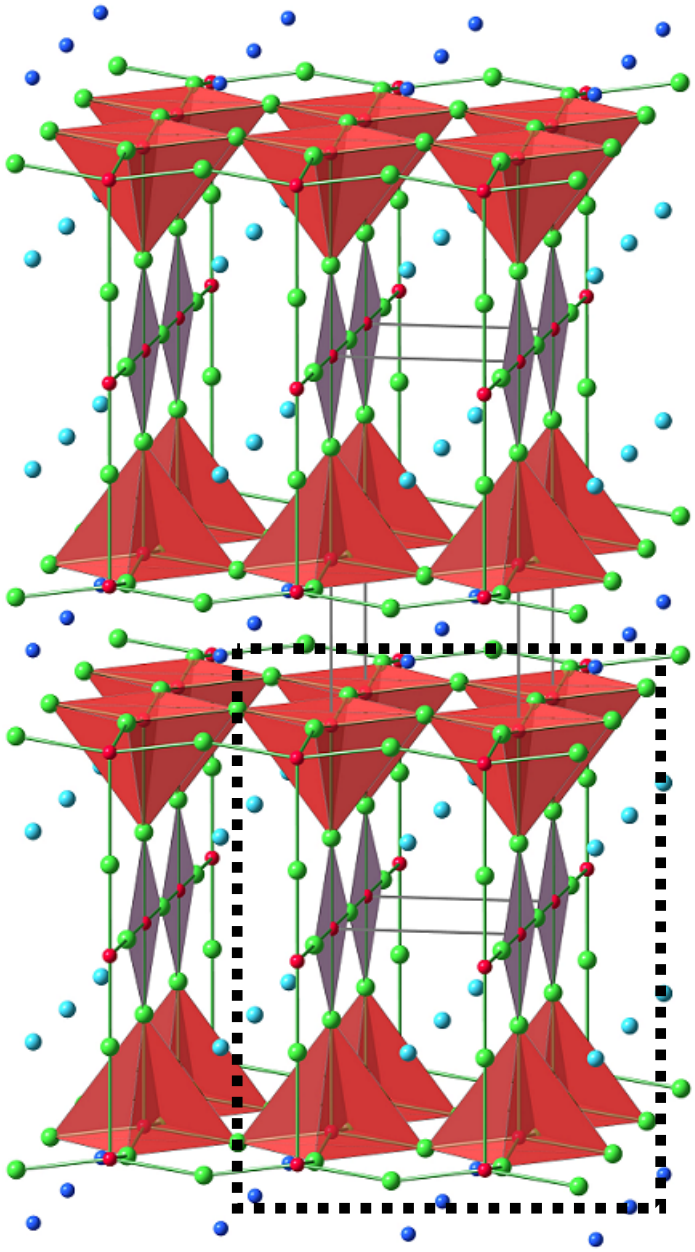


Find all compounds in a given set of models with ***similar*** structural patterns.









+

Combination of graphical and textual query parameters

Cations: Cu

Anions: O

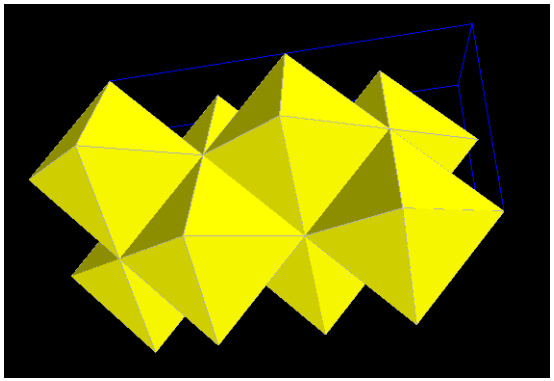
Bonds: 'strong'

Pyramids: inclination > 55°

Difficulties

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

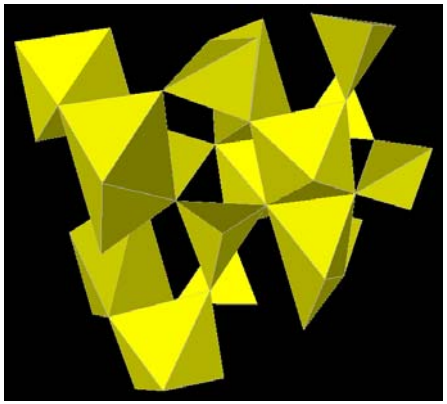
Sodium chloride



Na, Cl

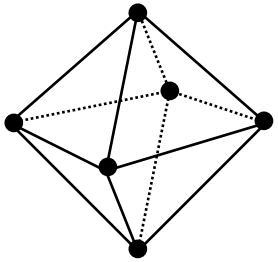
.....

Spodumene

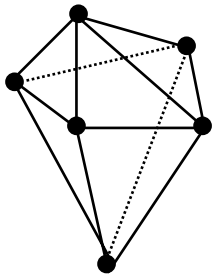


Li, Al, Si, O

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*).

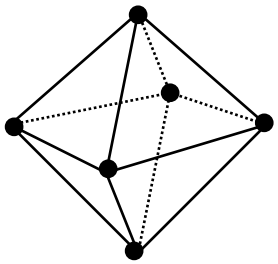
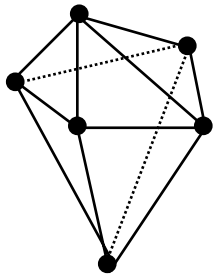
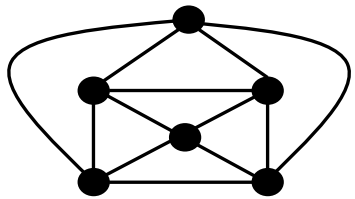
b) Three kinds of connections between polyhedra (*vertex, edge,* or *face*).

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: 

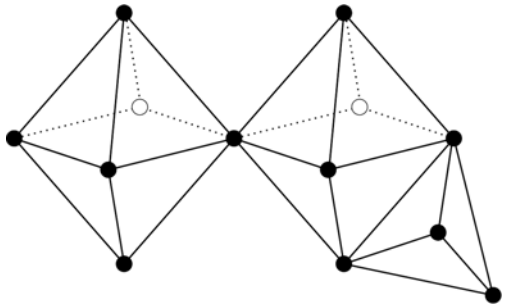
Clusters of polyhedra: Connected units of polyhedra.

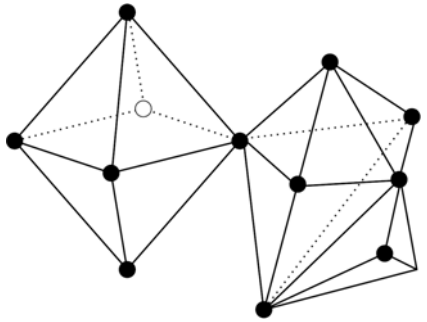
Equivalence of clusters? We are dealing with experimental data!

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

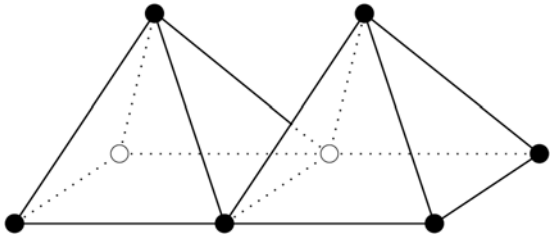
Example:

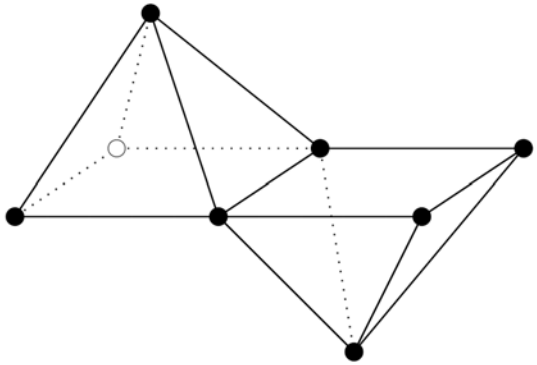
Equivalent:



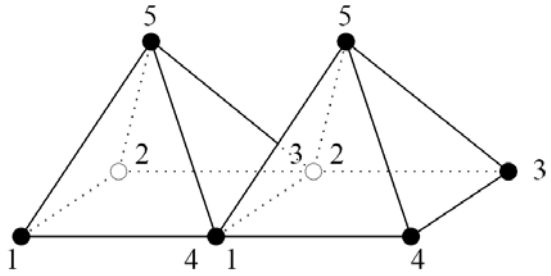


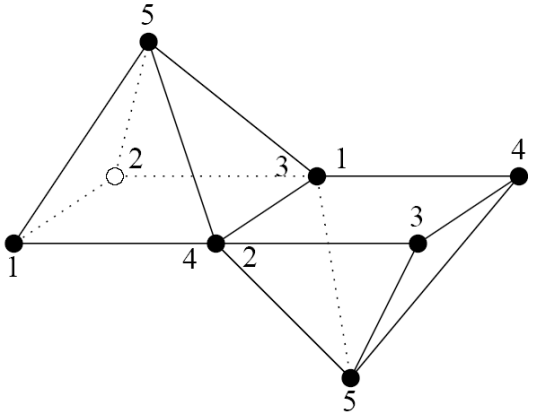
Not equivalent:





Graph representation



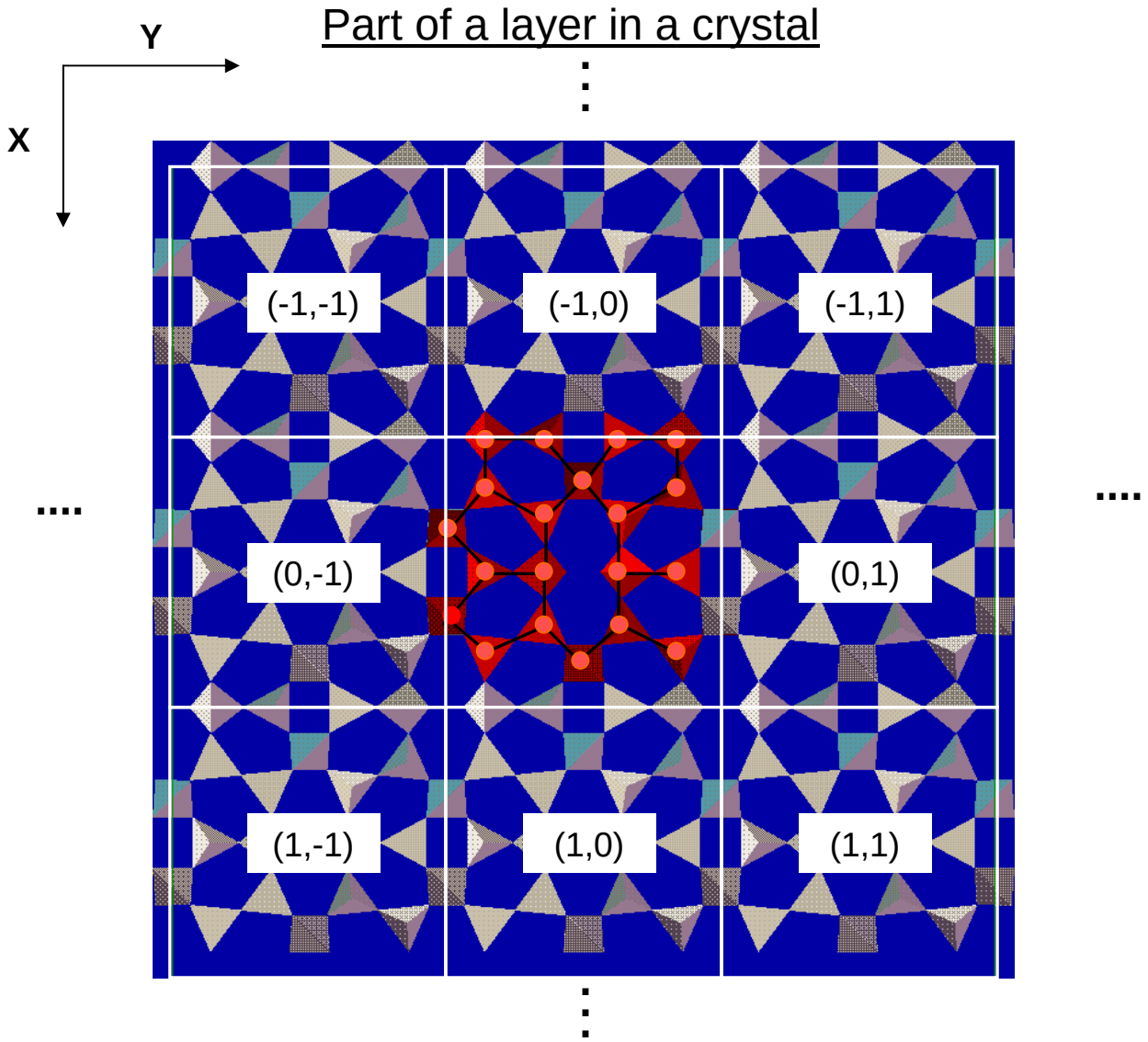


Ordered face representation

1432 154 125 235 345	(4,1)(3,2)	1432 154 125 235 345

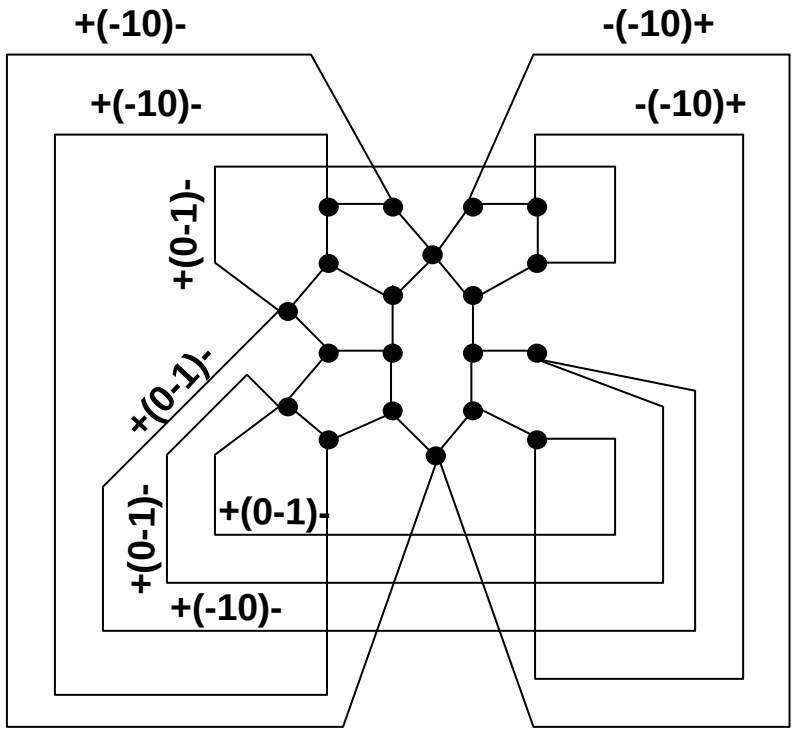
1432 154 125 235 345	(4,2)(3,1)	1432 154 125 235 345

Periodic arrangement of identical cells $\Rightarrow$
finite representation of polyhedra graphs is possible .



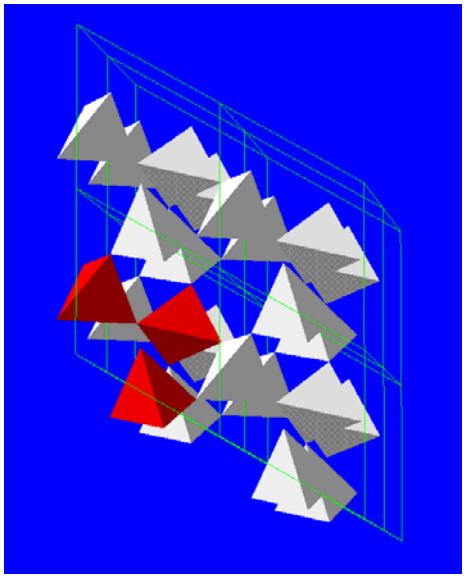
Direction-labeled graph

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

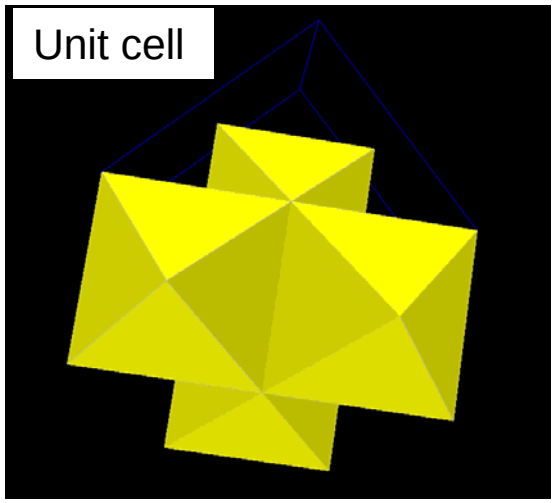


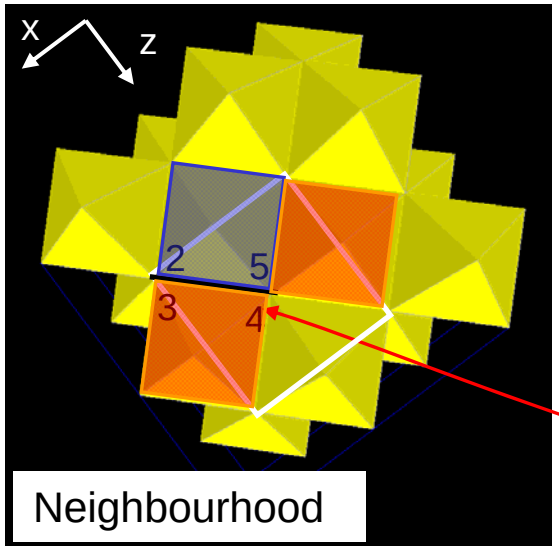
Reduced symmetry-labeled graph

(Klein, 1995)

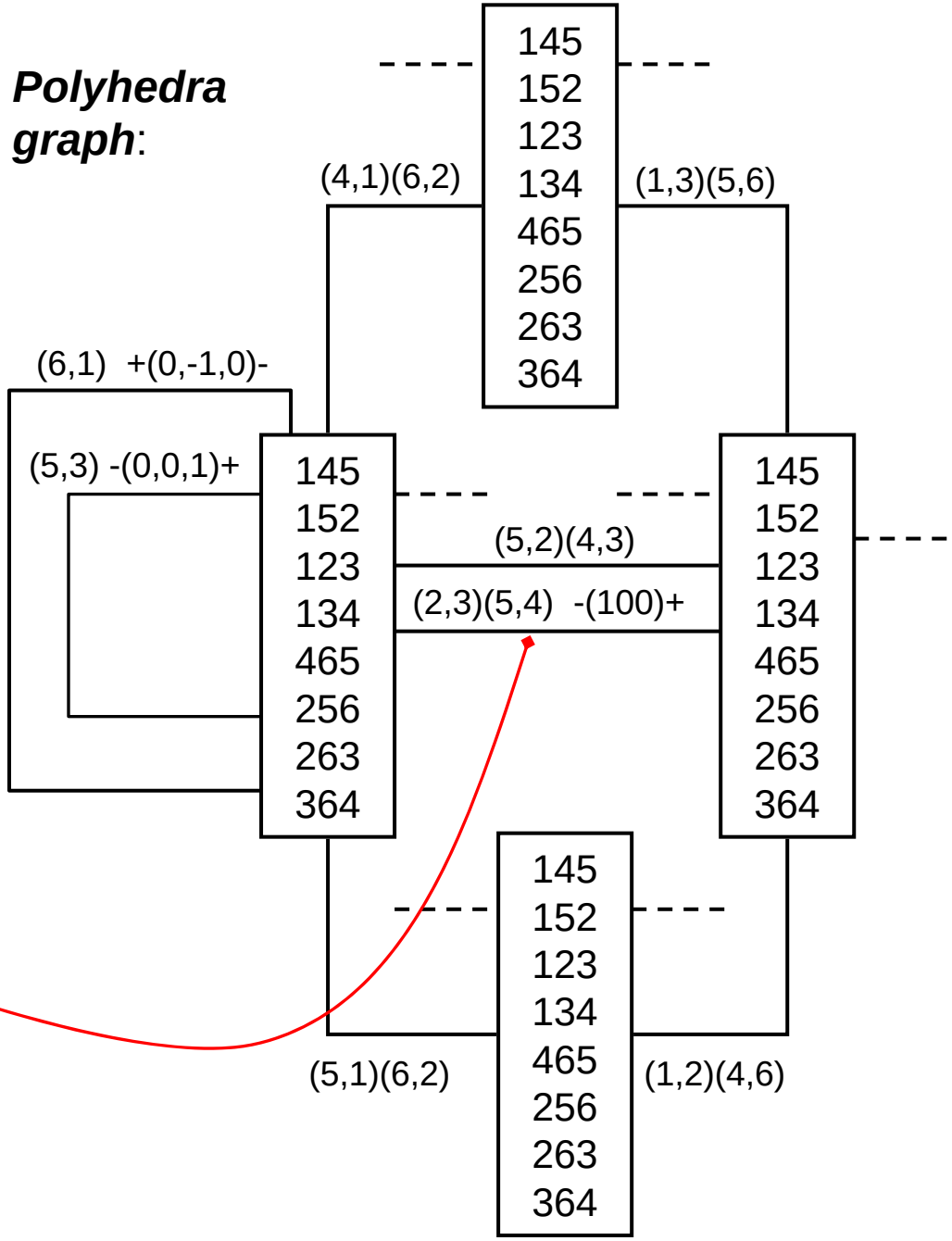


Sodium chloride



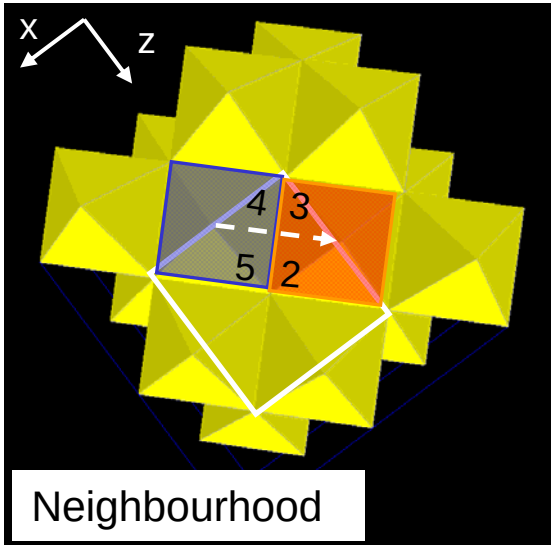


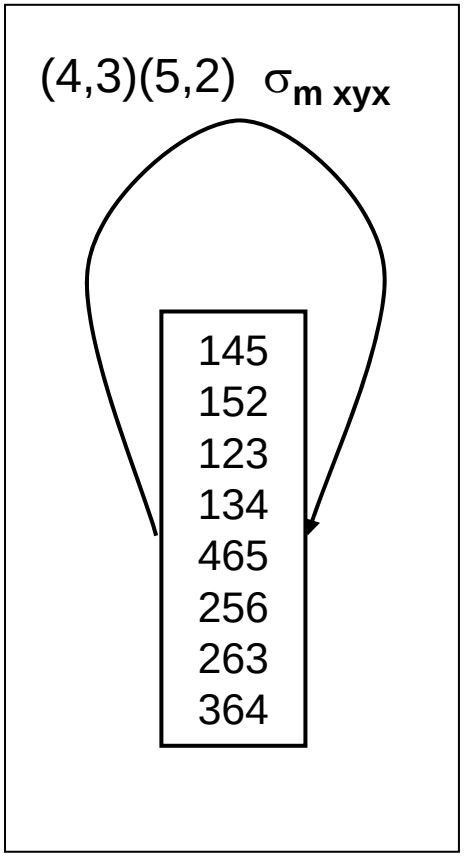
Polyhedra graph:



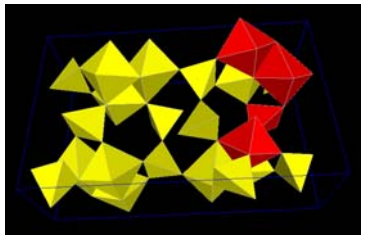
Reduced symmetry-labelled polyhedra graph

Sodium chloride





Topological search



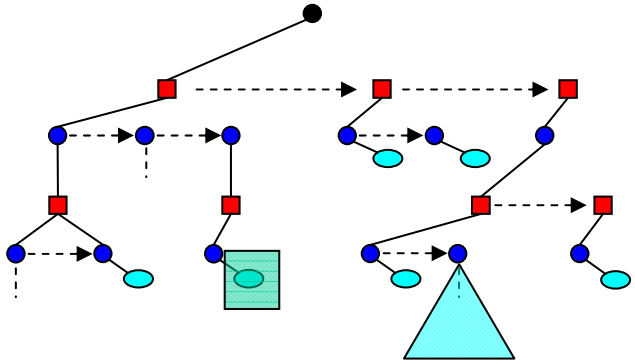
Subgraph isomorphism problem: **computationally hard**.

Preprocessing of model structures in the database.

Indexation of polyhedra graphs

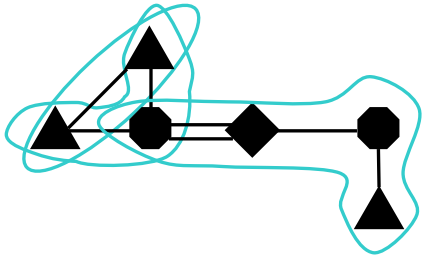
Proceeding

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



Sketch of search:

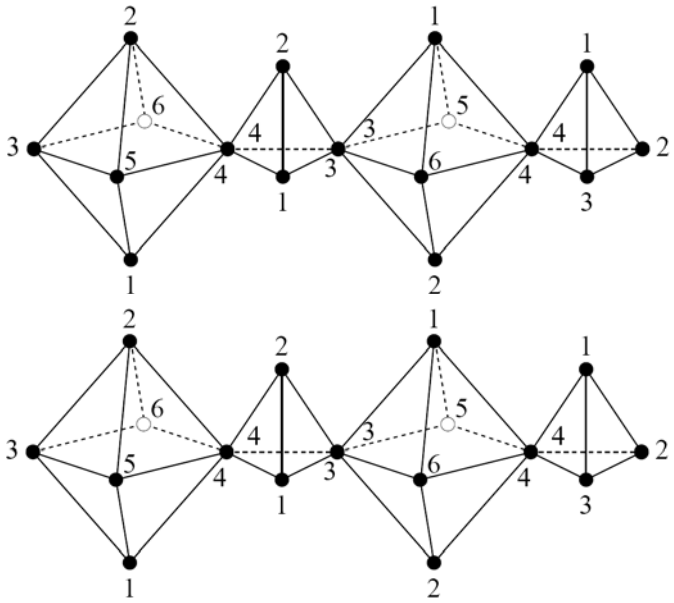
- Compute annotated paths for the input structure.



- Use the index to determine candidate model structures.

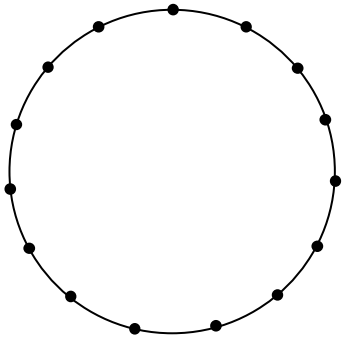
- Try to locate substructures in these model structures having the same path cover as the input structure.

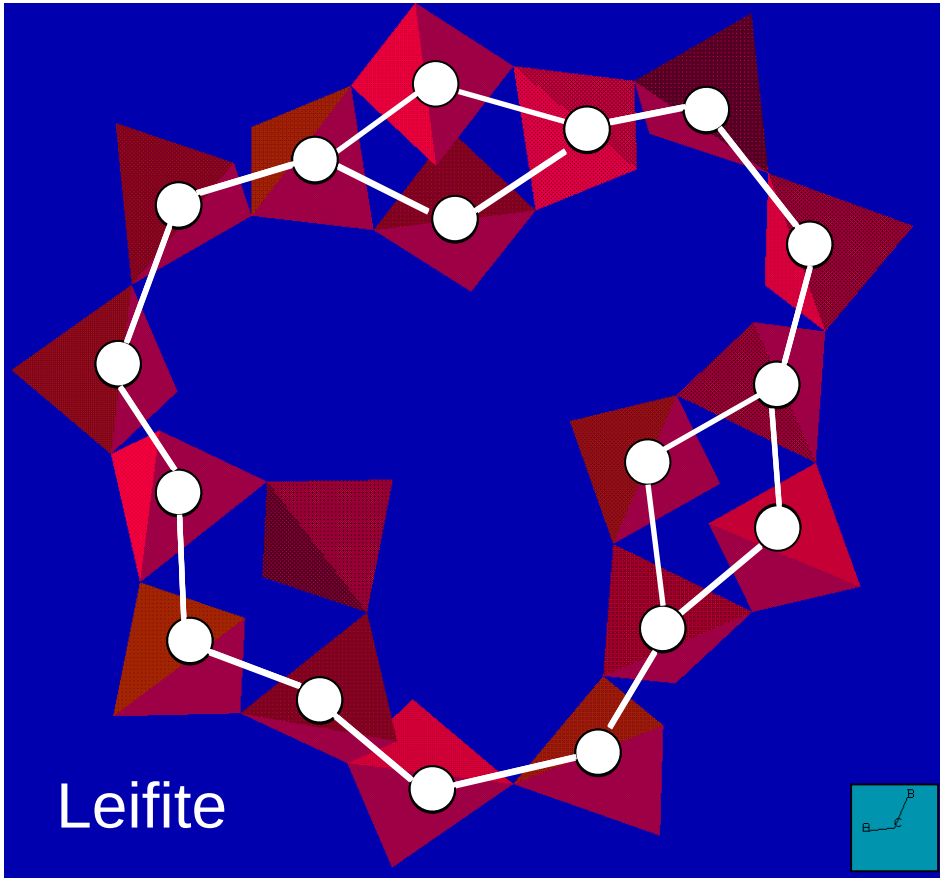
- Check for permutations of polyhedra vertices to get isomorphic graphs.



Answer complexity

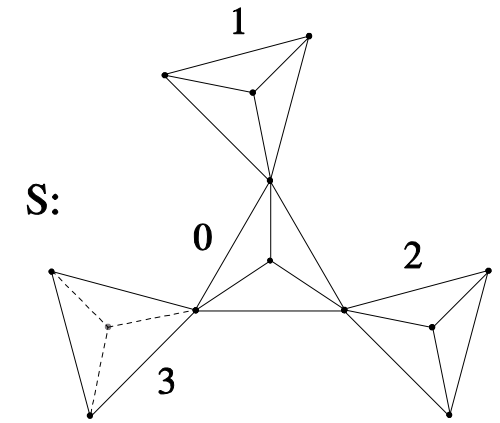
Number of isomorphic substructures in a single model structure?

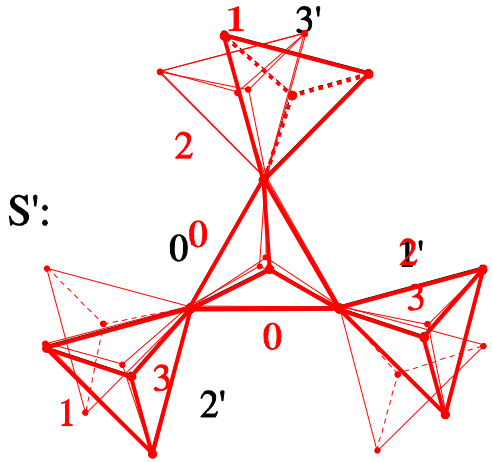




44 not symmetrically equivalent 15-membered rings in leifite.

Embedding

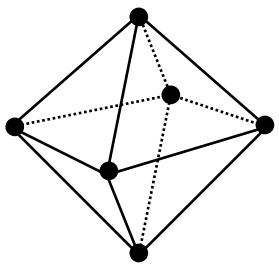


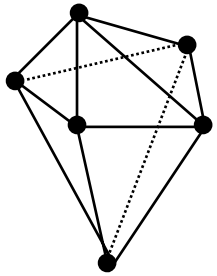


Geometric similarity

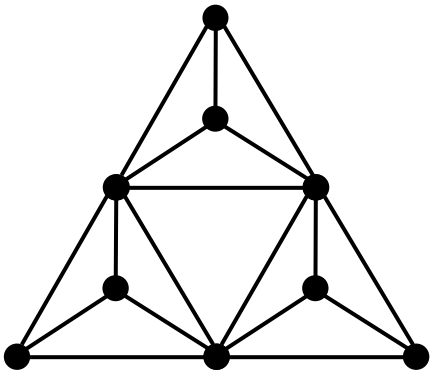
Two levels:

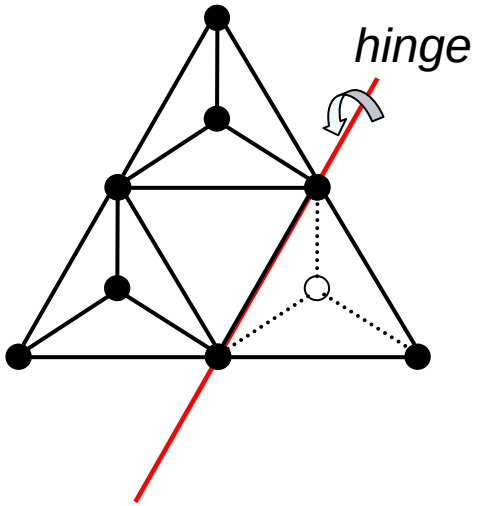
1. Polyhedra

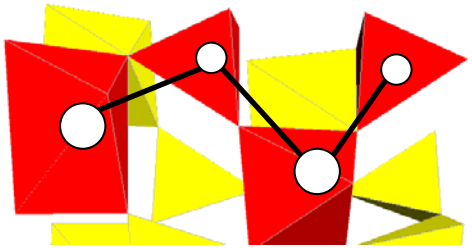


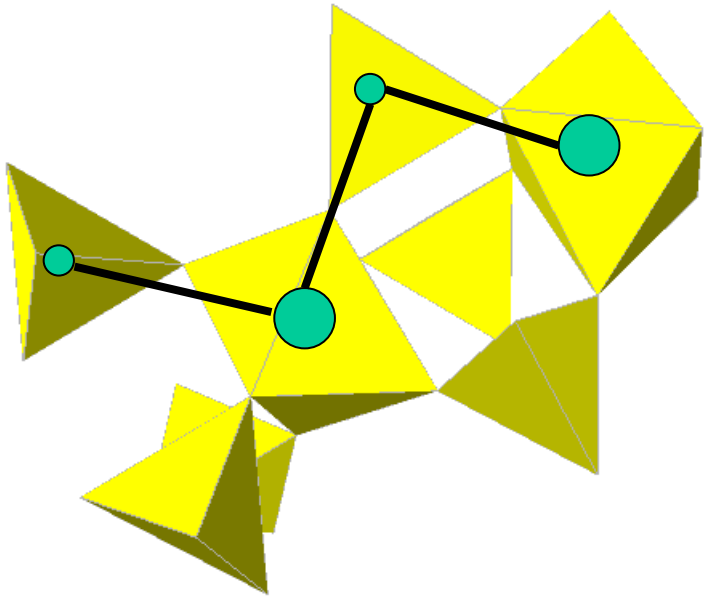


2. Relative positioning of polyhedra.









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:

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

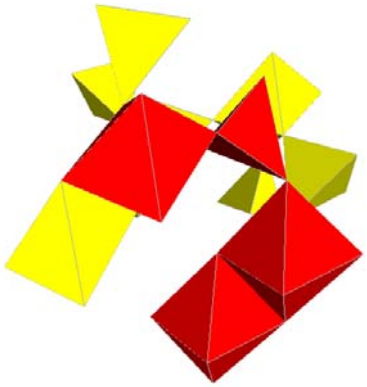
Measuring similarity:

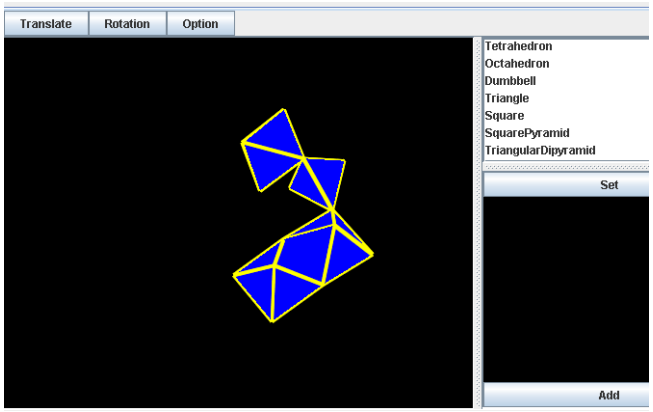
$$\varepsilon := \frac{\sqrt{U}}{n}$$ (**R**oot **M**ean **S**quare)

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

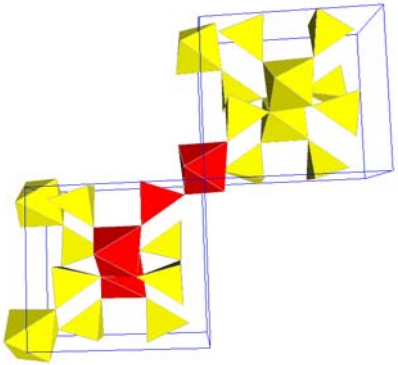
The result

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.

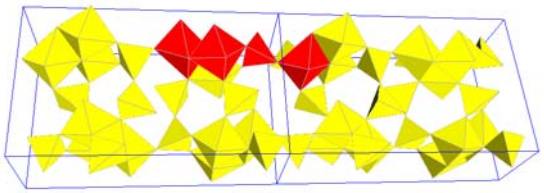




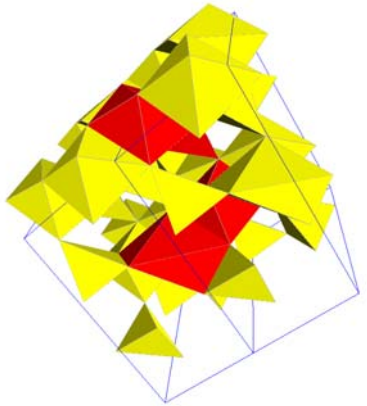
Answer: Compounds with isomorphic structural patterns and their RMS values.



Jadeite: 0.680581



Zoisite: 0.789256

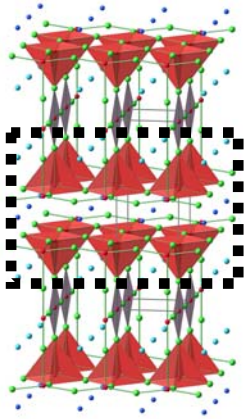


Spodumene: 0.110646

Future work

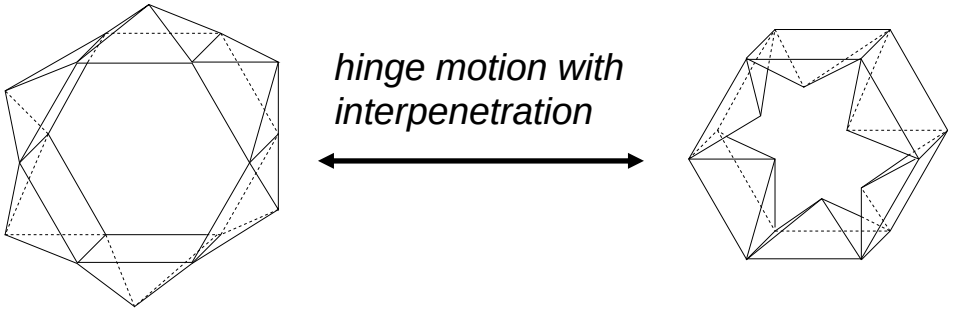
Searching: Improve the embedding algorithm (permutations, symmetries).

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,...?).



Thank you !