Structure searching: what you get is what you wanted.

K. T. Taylor, W. D. Hounshell, J. G Nourse, B. Christie, B. A. Leland

MDL Information Systems Inc
The Need

- Represent chemical structures in computer readable form
- Translate between computer representation and human readable form
- Store structures in a database
- Retrieve the structures selectively, accurately, and reproducibly
The Fundamental Problem

- A molecule is described by a collection of molecular orbitals and atoms
- Difficult to interpret this representation
  - Human and machine
- Representation based on molecular orbitals is not practical
Chemical Structure Identifiers

- **Trivial name**
  - Simple structures
  - Language dependant
- **Systematic name**
  - Complex molecules can be difficult to name
  - Often not unique
- **Classification Systems**
  - Beilstein
  - Derwent
- **Graphical Structure**
  - Universally understood by chemists
What is Benzene?

**Line notation**
- Wiswesser: RH
- MDL LN: C-C=C=C-C=@1
- SMILES: c1ccccc1

**Connection table:**

<table>
<thead>
<tr>
<th>Bond</th>
<th>1 2 3 4 5 6</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1 2 0 0 0 0</td>
</tr>
<tr>
<td></td>
<td>3 4 2 0 0 0</td>
</tr>
<tr>
<td></td>
<td>4 5 1 0 0 0</td>
</tr>
<tr>
<td></td>
<td>2 3 1 0 0 0</td>
</tr>
<tr>
<td></td>
<td>5 6 2 0 0 0</td>
</tr>
<tr>
<td></td>
<td>6 1 1 0 0 0</td>
</tr>
</tbody>
</table>

**Benzene**

- ISIS- 08200115272D
- MUSE00000002
- 71-43-2

**Description:**
Clear, colorless liquid with characteristic aromatic odor.

**Formula:** C₆H₆; **Mol. Mass:** 78.11184;
**M.Pt:** 5.512-5.544°C; **B.Pt:** 80.11 ºC
Graphical Representation

- Well understood by chemists
- Formalizes the molecule in terms of atoms and bonds
  - Based on atoms with a defined valency connected by bonds
  - Very powerful, but an approximation
- Started simple but flexibility needed
  - Variable valency, Bond types, Bond order
  - Organometallics
  - Hydrogen!
- Isomers
  - Geometric, optical, tautomers
- Isotopic variations
- Predates MO theory
Create special atom and bond types as necessary
  - Bond types
    - Tautomer bond
    - Aromatic
    - Cyclic / Acyclic
  - Atom types
    - Chiral center – R / S
    - Trigonal bipyramidyl atom
- Risky
  - Editing structure for new entity or sub-structure query may give incorrect answers
- Use for basic objects
  - Atoms in the periodic table: C / H / O / N / S / Cl / F / Br etc
  - Simple bond types: single / double / triple
Interpret atom environment and bond types at search time
- Bond types
  - Tautomer bond
  - Aromatic
  - Cyclic / Acyclic
- Atom types
  - Chiral center – R / S
  - Coordination number of atom

Enhanced interpretations can be added without invalidating existing data
- Upgrading (large) databases is undesirable

Simplify structural input
- Structural business rules may simplify process
Single entity may have more than one valid representation

Chemists have learned to interpret the graphical representation

More difficult for computer systems to interpret representations
Chemical structure business rules

- Now widely adopted at the corporate level
- Standardize on one representation
  - enforce it a registration
  - Train users in the standardized representation
- Incorporate the business rules in the searching application
Incorporate search capabilities that introduce flexibility in the query

- Similarity searching
  - Very flexible
- MDL’s Flexmatch operator
  - Encode basic structural connectivity
  - Add layers of constraints
    - Parent / Salt
    - Tautomer
    - Isotopic variations
- Tend to be specialist capabilities
Retrieval

- Representation is only one part of the problem
- The query must be matched to the database entry
  - Methylbenzene
    - Examples:
      - Atom numbering depends on drawing history
Approaches

- Select one form for storage and standardize the query
  - Canonical Approach
  - Rendered structure may have unusual orientation

- Register any layout and build intelligence into the search engine
  - Most flexible
  - Preferred orientation is rendered
Substructure searching (SSS) is a mature and well understood operation

- Accuracy depends on quality of the search operation
  - Hits are expected
- Value of the search depends on the richness of query properties available
  - Hits contain features of interest
- Query properties are based on perception
  - Bond type: aromatic / any / cyclic / acyclic / etc
  - Ring bond count / substitution count
  - Atom lists; [O,S] / [Cl,Br,I]
  - Chiral centers; R / S

- Structure conventions are of value
Exact match can be problematic

- Structure conventions for registration and queries are beneficial
  - Nitro group

- Provide a tunable search operator
  - MDL’s Flexmatch search
    - tautomer
    - Parent
    - Isotopic variations
Modern tree-based indexing technologies provide excellent performance for large databases.

Key-based indexing is limited for modern database sizes.
- Secondary benefit: modal fingerprint applications

Hash-based indexing for canonical features.
- Flexmatch keys
Must be able to map query over root-Rgroup boundary

- Database entry:
  - Query
  - Hit
Stereochemical representation

- Remains a major problem for both representation, encoding and searching
- Pure samples that contain only absolute or relative centers can be named and characterized
- This is not reality
  - Chiral centers are seldom optically pure
  - Samples with both absolute and relative centers are common
Traditionally stereochemical qualifiers are stored separately from the structure

- Embedded in structure’s name
- Stored in a separate field in the database
- Information is easily lost or corrupted

Need to track all the stereochemical information and associate it with properties (activity)
A high throughput synthesis workflow

1) screen
2) resolve

1) screen
2) active
3) characterize

1) screen
2) inactive

Mixture

Two Samples

sample 1

'racemic' Type

sample 2a

sample 2b

Relative Type

Absolute Type

sample 3
Typical samples

- Contain a combination of types:

- And purity information is needed
All potentially valid hits are shown

Absolute
Relative
‘Racemic’
Chemical database technology well developed
Representations are rich but not complete
  • Organometallics and transition metal complexes
  • And hydrogens
Registration without matching representation is of little value
Registration without retrieval is of little value
User Interfaces are the key to acceptance
  • Personal preferences for structure representation
  • Use appropriate representation