Chemical Structure Recognition: A Rule-based Approach

DRR 2012, San Francisco, CA USA

Noureddin M. Sadawi Alan P. Sexton Volker Sorge

School of Computer Science
University of Birmingham, UK

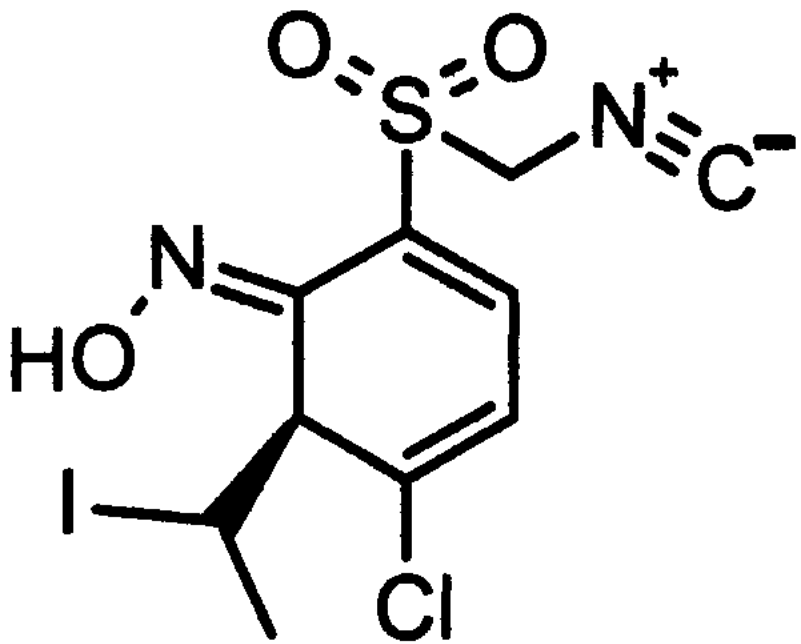
25 January 2012, DRR XIX, San Francisco

Motivation

- Chemical literature contains much information in molecule diagrams
  research articles, patent specs, catalogues, etc.
- Diagram recognition can make information accessible
  visually impaired readers, search, chemical software
- There are a number of academic and commercial approaches
  mostly following a traditional procedural method
  details often only vaguely described

We present a clearly defined rule based approach.

Problem



Recognising molecular diagrams requires:

- Basic image analysis
- Shape recognition
- Contextual discrimination of many cases
- Strategy to manage cases

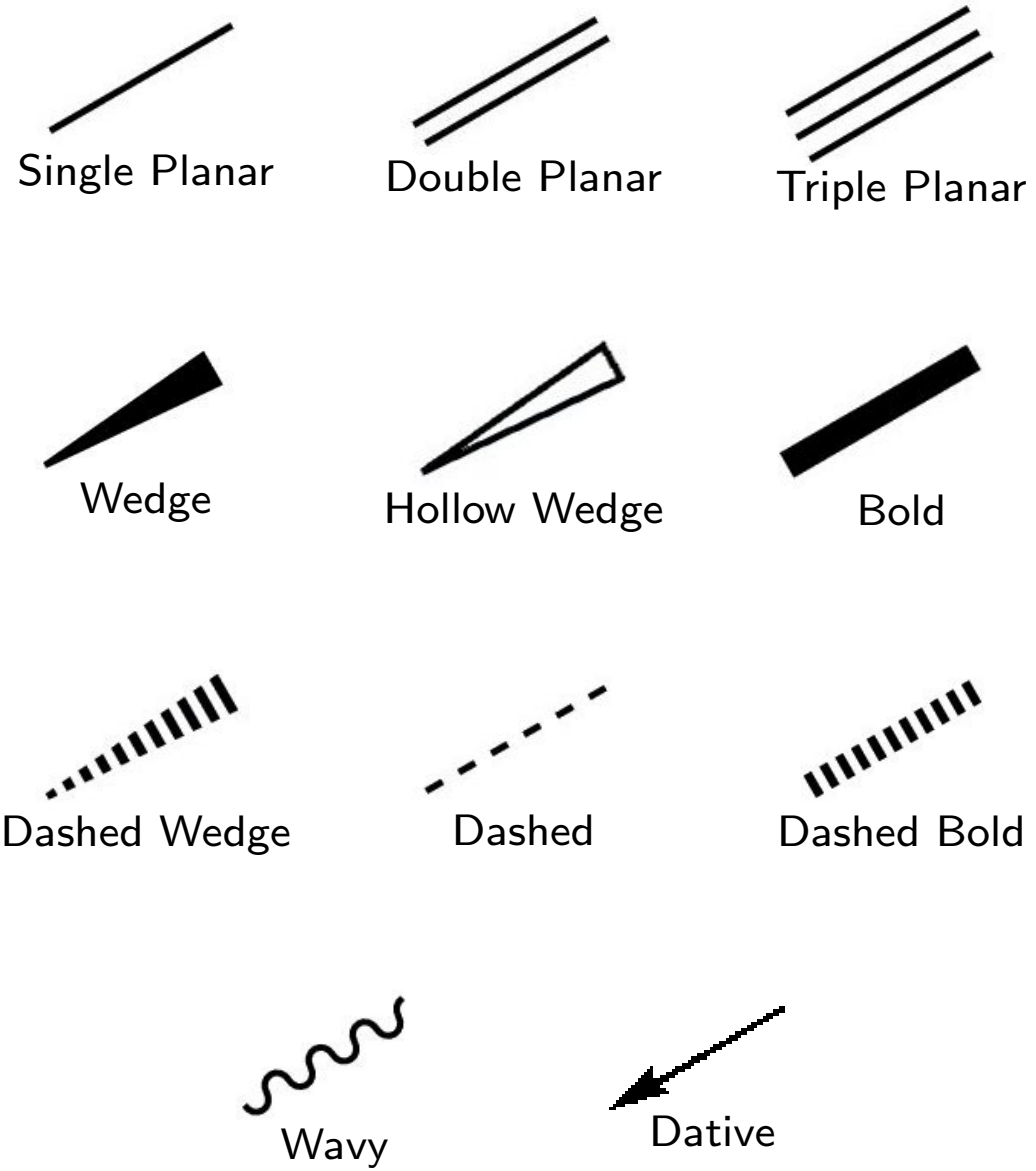
Diagram Elements

Single Planar

Double Planar

Triple Planar

Diagram Elements

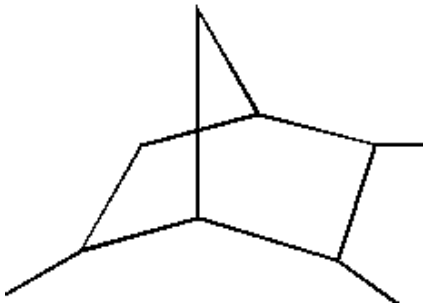


Extra Complexities

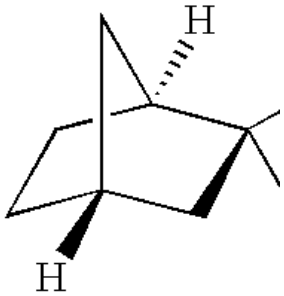
Implicit Nodes in Bond Sequences

Extra Complexities

Implicit Nodes in Bond Sequences



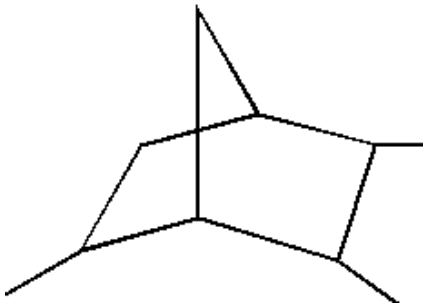
Closed Bridge Bond



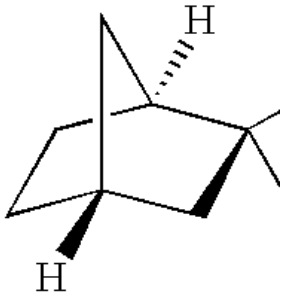
Open Bridge Bond

Extra Complexities

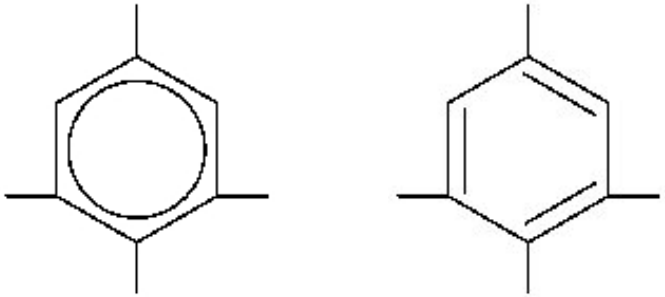
Implicit Nodes in Bond Sequences



Closed Bridge Bond



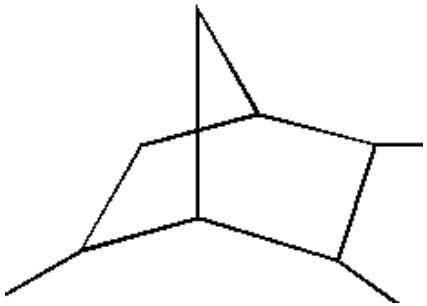
Open Bridge Bond



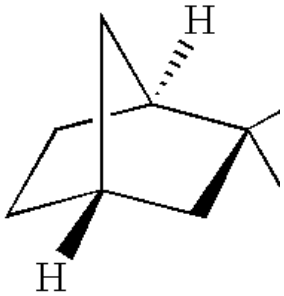
Aromatic Rings

Extra Complexities

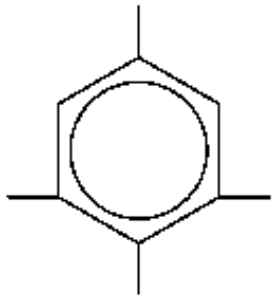
Implicit Nodes in Bond Sequences

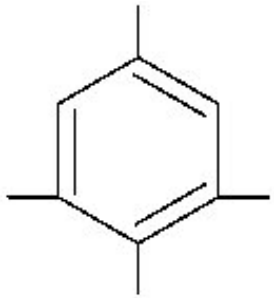


Closed Bridge Bond

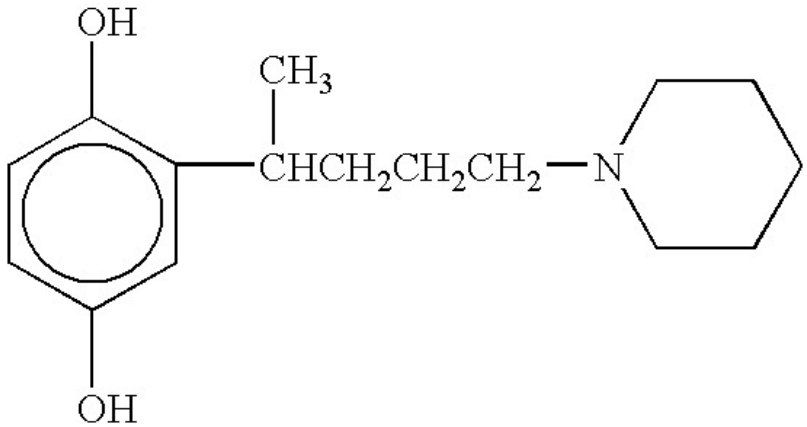


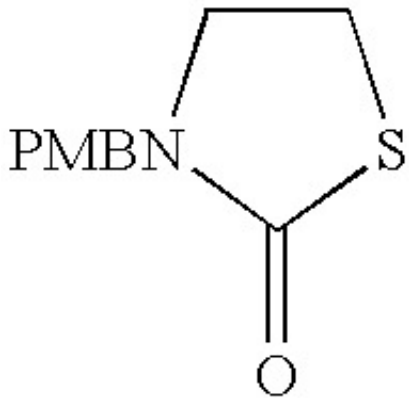
Open Bridge Bond





Aromatic Rings





Superatoms

Overall Procedure

- Preprocessing
  Noise removal, binarise, connected-component analysis
- Simple OCR & Character Grouping
- Detection of geometric primitives:
  Line segments, Arrows, Circles, Triangles, Character Groups
- Rewrite primitives into a directed graph using rule-based System
- Disambiguation and graph correction
- Produce output from graph (e.g. MOL, SMILES)

Rule Based System

Input: Set of well defined geometric primitives:
Line segments, Arrows, Circles, Triangles, [Character Groups]

Output: Directed graph representing a molecule

- 18 rules defined in terms of

  Preconditions: Find and remove appropriate primitives

  Postconditions: Add to output graph and possibly new primitives
- Mutually exclusive pre-conditions (except 2 bridge bond rules)
- Order of application is irrelevant
- A number of fuzzy parameters allow for some training.

Some Concepts and Parameters

- Approximate Collinearity/Dash Neighbouring
  - Radius of collinearity *rc*
  - Dash length *dl*
  - Dash separation *ds*

- Approximately Parallel with separation *bs* and overlap *ol*, $\|_{bs}^{ol}$

Bond Rules I

Single Planar Bond

1. l is a line segment.
2. There is no line segment l' such that $l \parallel^{ol}_{bs} l'$.

Consequence l is a single bond with endpoints of l

Bond Rules II

Double Planar Bond

1. $L = \{l_1, l_2\}$ is a set of two line segments.
2. $length(l_1) > wb$ and $length(l_2) > wb$.
3. $l_1 \parallel^{ol}_{bs} l_2$.
4. There is no line segment $l \notin L$ such that $l_1 \parallel^{ol}_{bs} l$ or $l_2 \parallel^{ol}_{bs} l$.

Consequence Cutting(l_1, l_2) will yield a double bond as well as at most two new line segments.

Bond Rules III

Triple Planar Bond

1. $L = \{l_1, l_2, l_3\}$ is a set of three line segments.
2. for every $l \in L$ we have $length(l) > wb$.
3. $l_1 \parallel^{ol}_{bs} l_2$ and $l_2 \parallel^{ol}_{bs} l_3$ and $l_1 \parallel^{ol}_{2*bs} l_3$.
4. There is no line segment $l \notin L$ such that $l_1 \parallel^{ol}_{bs} l$ or $l_2 \parallel^{ol}_{bs} l$ or $l_3 \parallel^{ol}_{bs} l$.

Consequence Cutting(l_1, l_2, l_3) will yield a triple bond as well as at most four new line segments.

Bond Rules IV

Dashed Bond

1. $L = \{l_1, \ldots, l_n\}$, where $n \geq 3$, is a set of line segments,
2. L is approximately collinear.
3. Every element of L has length of approximately dl.
4. No two elements of L have a separation distance of less than the minimum of ds.
5. Two elements of L, called the *end elements*, dash-neighbour wrt. ds precisely one other element of L. All other elements of L, called *internal elements*, dash-neighbour wrt. ds precisely two other elements of L.

Consequence L forms a dashed bond with endpoints given by the endpoints of the minimal line segment that contains $l_1, \ldots, l_n$.

Results I

- OSRA's benchmark dataset of 5735 images:

	OSRA Succ.		OSRA Failed		Total	
MolRec Succ.	4277	(74.58%)	796	(13.88%)	5037	(88.46%)
MolRec Failed	152	(2.65%)	510	(8.89%)	662	(11.54%)
Total	4429	(77.23%)	1306	(22.77%)	5735	(100%)

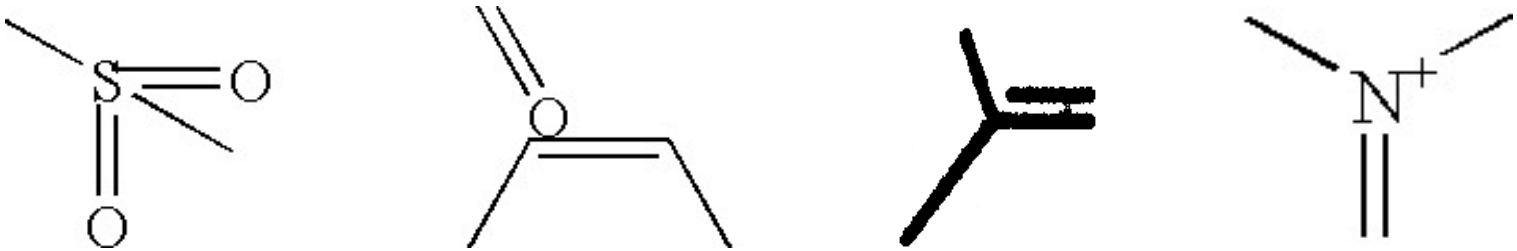
- Our dataset of 5740 images from Maybridge catalogue:

	OSRA Succ.		OSRA Failed		Total	
MolRec Succ.	3823	(66.72%)	981	(17.12%)	4804	(83.84%)
MolRec Failed	335	(5.85%)	591	(10.31%)	926	(16.16%)
Total	4158	(72.57%)	1572	(27.43%)	5730	(100%)

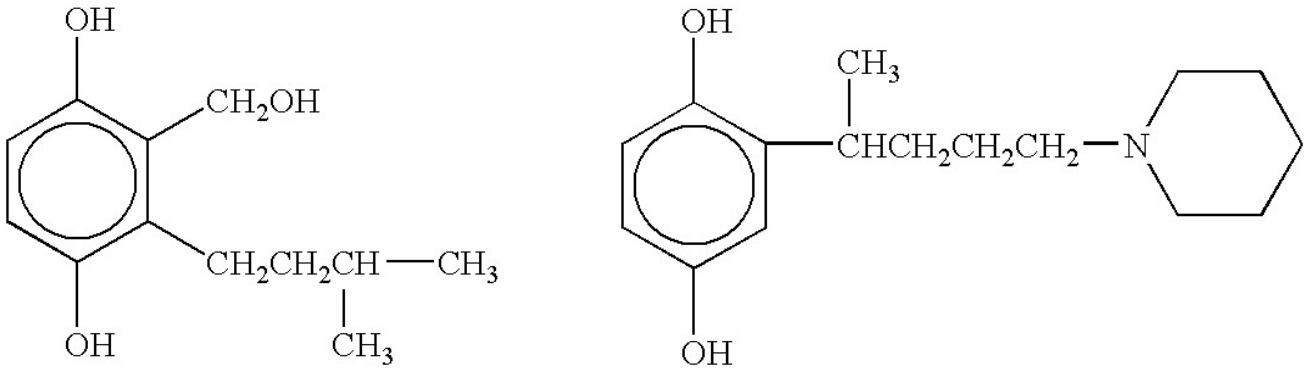
- TREC's Image2Structure Task:
  Achieved highest recognition rate (95%) on 1000 test images.

Reasons for Failure I

- Image has touching components:

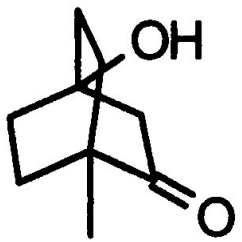


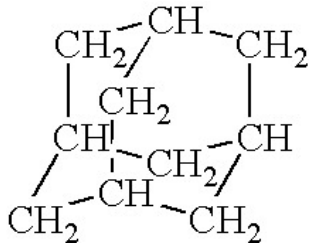
- Incorrect superatom connectivity:



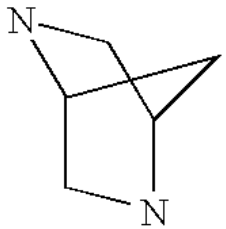
Reasons for Failure II

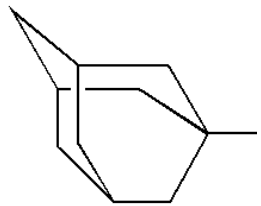
- Ambiguous Images:





- Problematic bridge bonds:





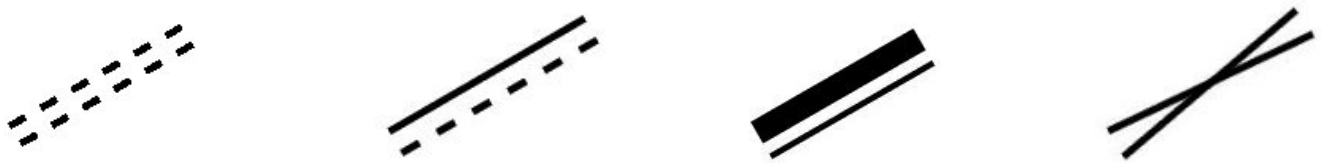
Reasons for Failure III

- Incorrect stereochemistry:

  - Our heuristics for guessing direction based on size of character group/complexity of structure at each end of the bond
  - Needs deeper domain knowledge to improve
- Dashed wedge bonds mis-identified:
  - Short dashes at narrow end mis-recognised as a dashed bond
  - Long dashes at wide end mis-recognised as dashed wedge or dashed bold bond
  - Result: dashed bond mis-recognised as 2 connected bonds

Conclusions and Future Work

- Clearly defined rule based approach to molecule recognition
- Some robust techniques to identify primitives
- First implementation yields good experimental results
- Flexible approach that can be easily extended
  Aromatic bonds, Tautomeric bonds, Double bond with stereo



- Better disambiguation for bridge bonds
- More chemical information to solve connection permutation problem for superatoms