hide notes

Performance of MolRec at TREC 2011
Overview and Analysis of Results

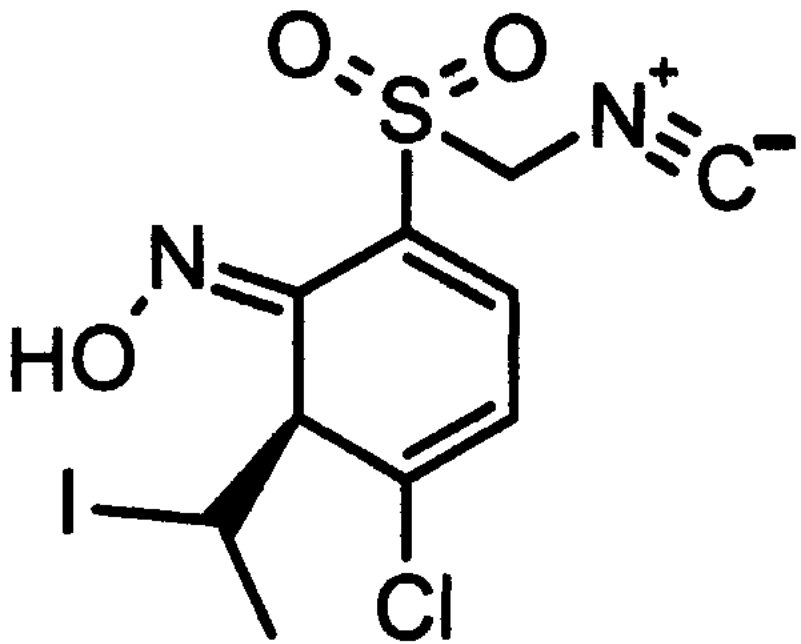
TREC 2011, Gaithersburg, USA

Noureddin M. Sadawi Alan P. Sexton Volker Sorge

School of Computer Science
University of Birmingham, UK

November 2011

Introduction



Recognising molecular diagrams well requires:

- Basic image analysis
- Good shape recognition
- Excellent contextual discrimination of many cases
- Excellent strategy to manage cases

Diagram Elements

Single Planar

Double Planar

Triple Planar

Diagram Elements

Single Planar

Double Planar

Triple Planar

Wedge

Hollow Wedge

Bold

Diagram Elements

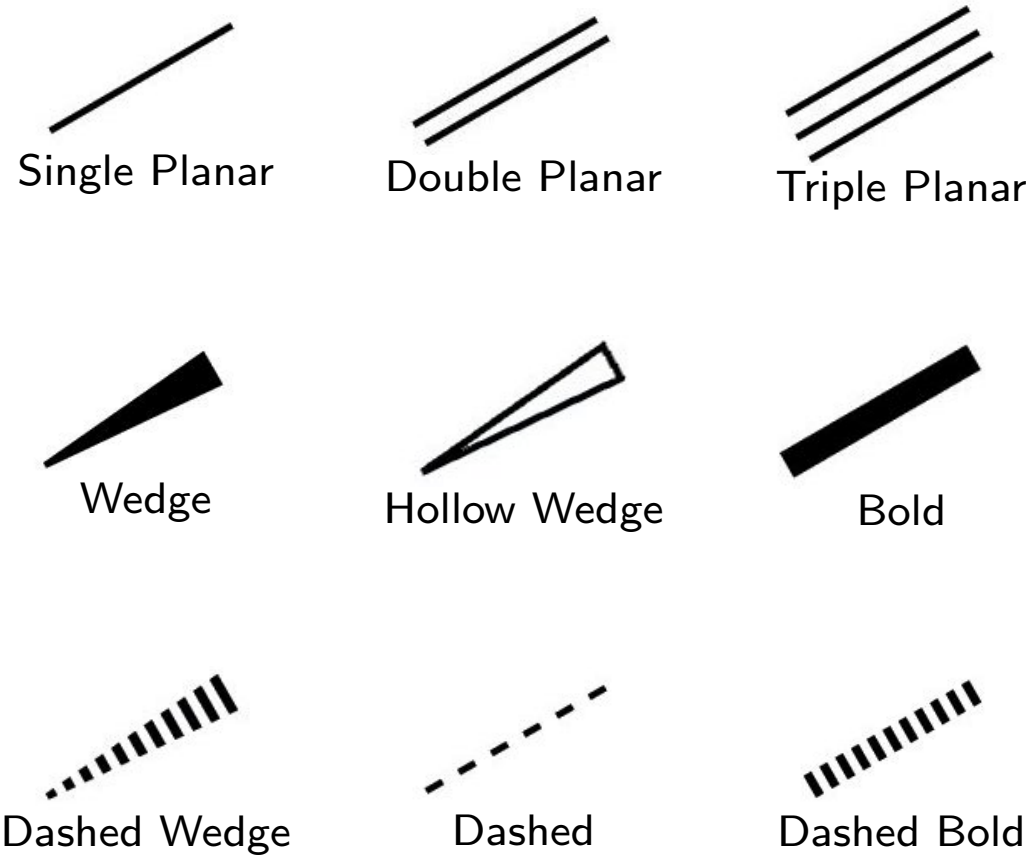
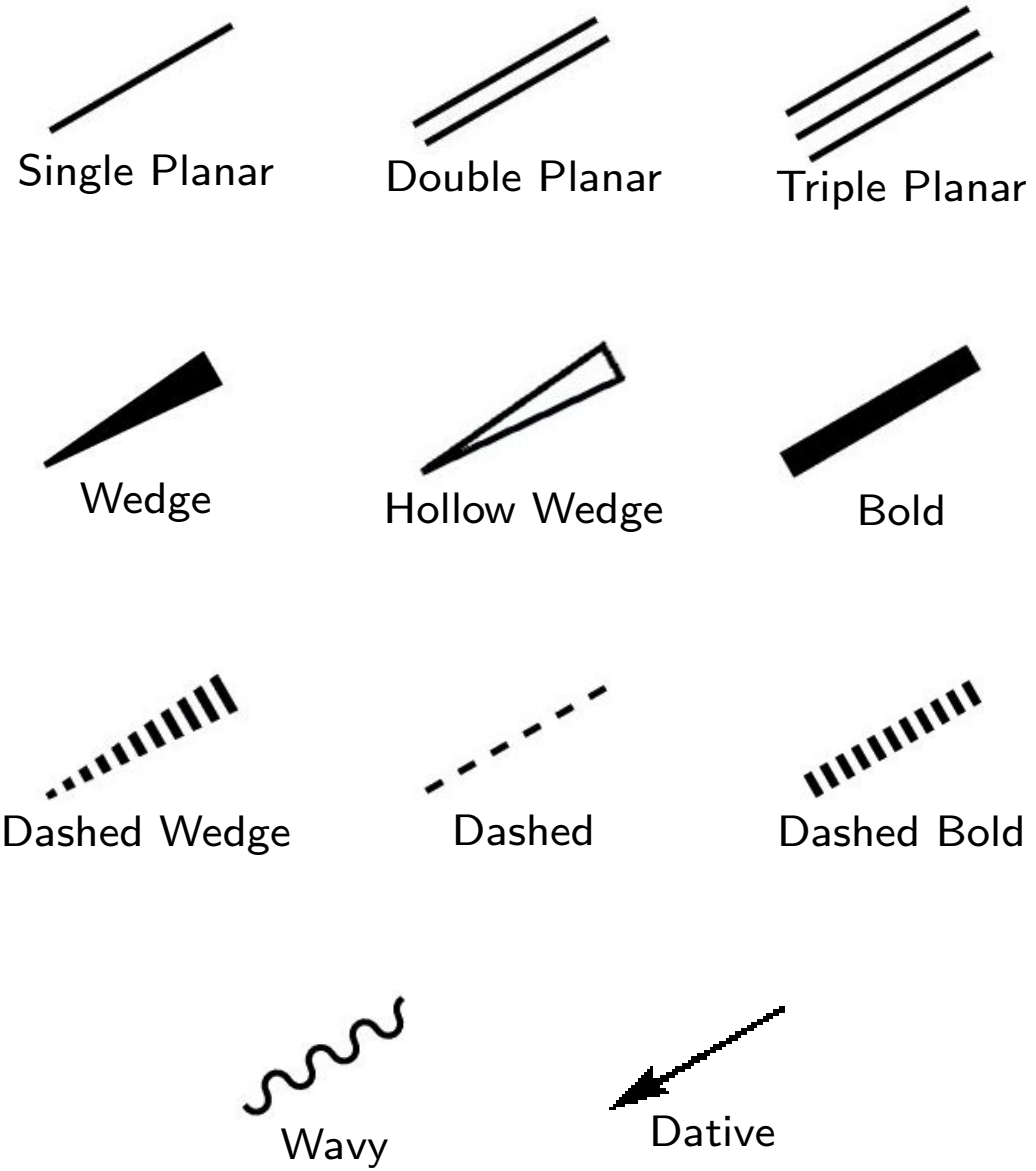


Diagram Elements

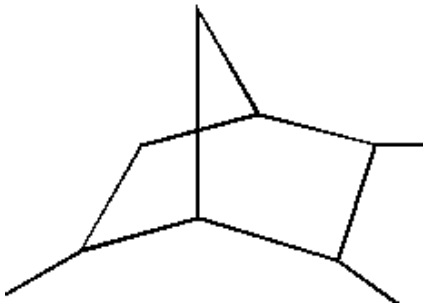


Extra Complexities

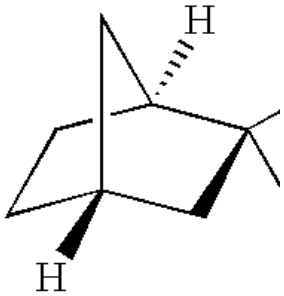
Implicit Nodes in Bond Sequences

Extra Complexities

Implicit Nodes in Bond Sequences



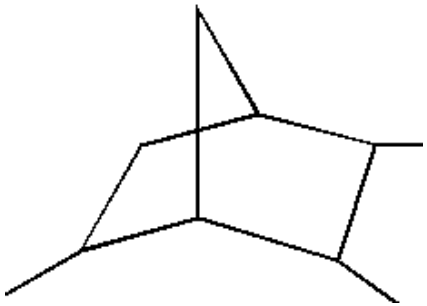
Closed Bridge Bond



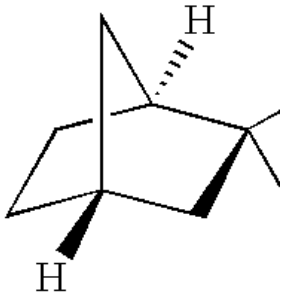
Open Bridge Bond

Extra Complexities

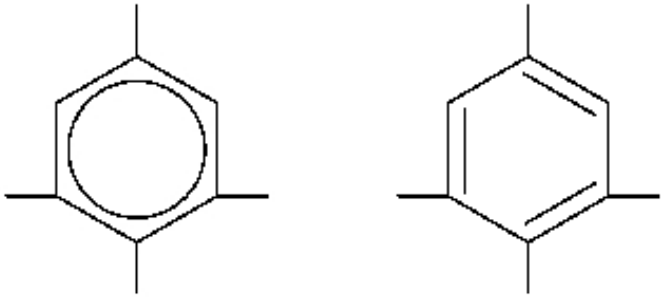
Implicit Nodes in Bond Sequences



Closed Bridge Bond



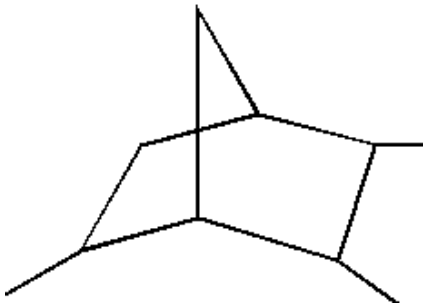
Open Bridge Bond



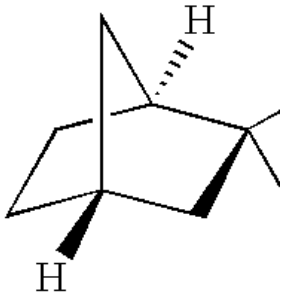
Aromatic Rings

Extra Complexities

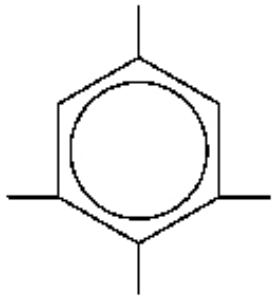
Implicit Nodes in Bond Sequences

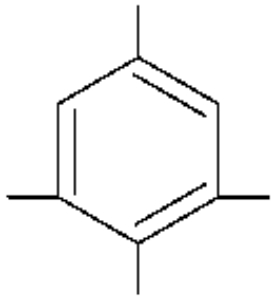


Closed Bridge Bond

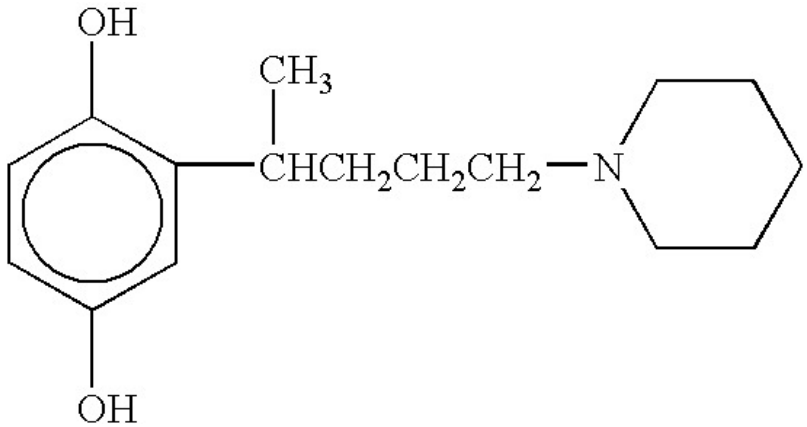


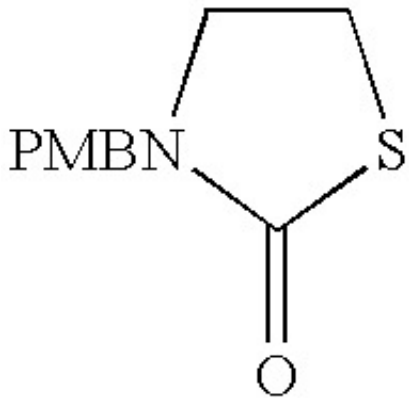
Open Bridge Bond





Aromatic Rings





Superatoms

Implementation Overview

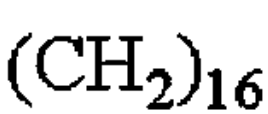
- Denoise, binarise, connected-component analysis
- Simple OCR to identify possible characters
- Group characters, deferring disambiguation to later stages
- Polyline finding: thin then split at multiway junctions
- Douglas-Peucker [DP73] to simplify and split at all junctions
  - Reconnect lines accidentally broken at spurs
  - Wavy lines reduced to zig-zag polyline
- Filled wedge/bold line recognition: "Walking ball"
- Get set of geometric primitives:
  - Line segments, Arrows, Circles, Triangles and Character Groups
- Apply graph-constructing rule based system
- Deal with remaining disambiguations and correct the graph
- Produce MOL file from graph, embedding superatoms

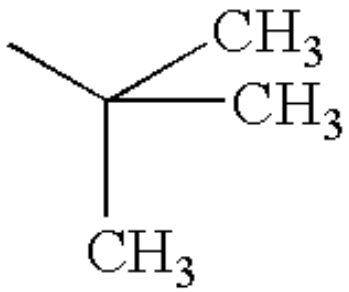
Lessons: Character grouping

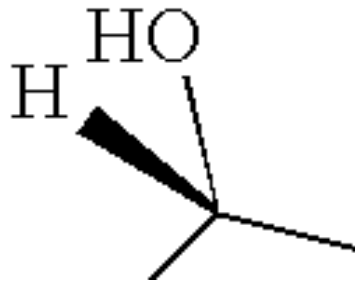
- Simple set of conditions for grouping characters based on:
  - types of characters: letters, digits, symbols
  - geometric relations: vertical, horizontal, diagonal

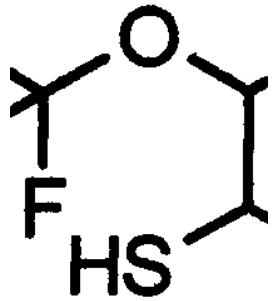
Lessons: Character grouping

- Simple set of conditions for grouping characters based on:
  - types of characters: letters, digits, symbols
  - geometric relations: vertical, horizontal, diagonal
- Works well on:



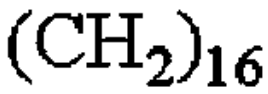


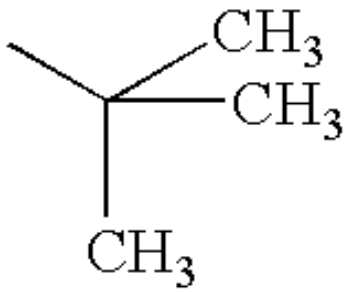


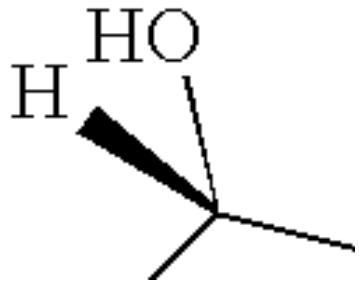


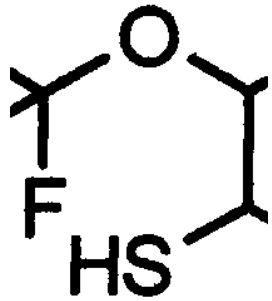
Lessons: Character grouping

- Simple set of conditions for grouping characters based on:
  - types of characters: letters, digits, symbols
  - geometric relations: vertical, horizontal, diagonal
- Works well on:

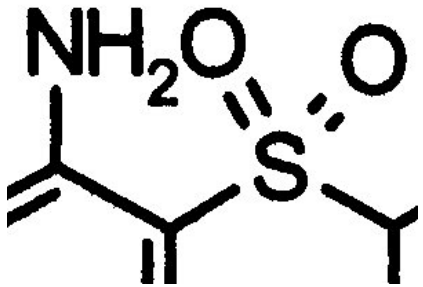






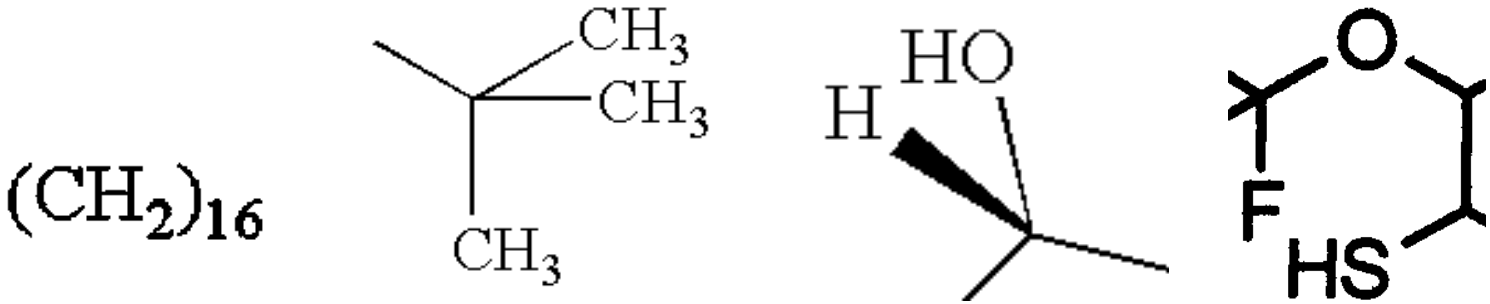


- Fails on:

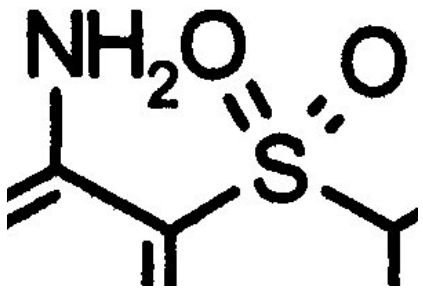


Lessons: Character grouping

- Simple set of conditions for grouping characters based on:
  - types of characters: letters, digits, symbols
  - geometric relations: vertical, horizontal, diagonal
- Works well on:



- Fails on:



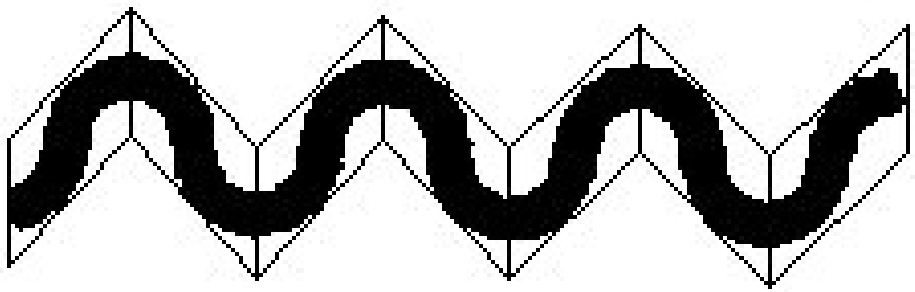
- Needs deeper domain knowledge to improve

Lessons: Line finding

- Tried Hough transforms early on and do not recommend it
- Working on connected components was fast, simple and robust
- Thining leads to many "spurs"; [DP73] cleans them up well
- Having all polylines split into line segments greatly simplifies further processing

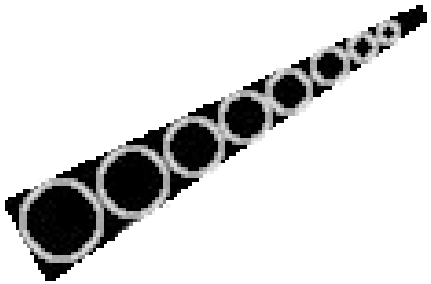
Lessons: Line finding

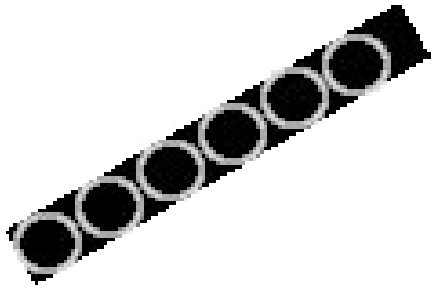
- Tried Hough transforms early on and do not recommend it
- Working on connected components was fast, simple and robust
- Thining leads to many "spurs"; [DP73] cleans them up well
- Having all polylines split into line segments greatly simplifies further processing
- Also simplifies wavy line detection



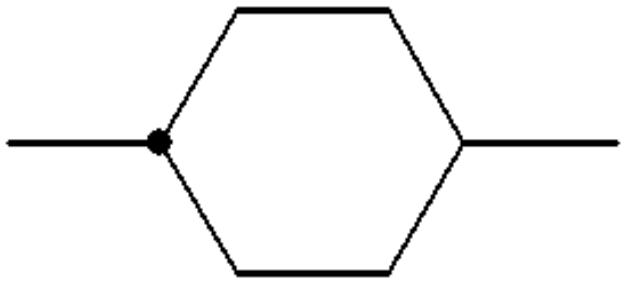
Lessons: Filled Triangle and Bold Line Recognition

- Already have clean skeletons for all bonds, including filled triangles and bold lines
- Find connected components inside which a ball significantly larger than average line width fits.
- Walk ball inside connected component in direction of largest size increase (or minimum size decrease)





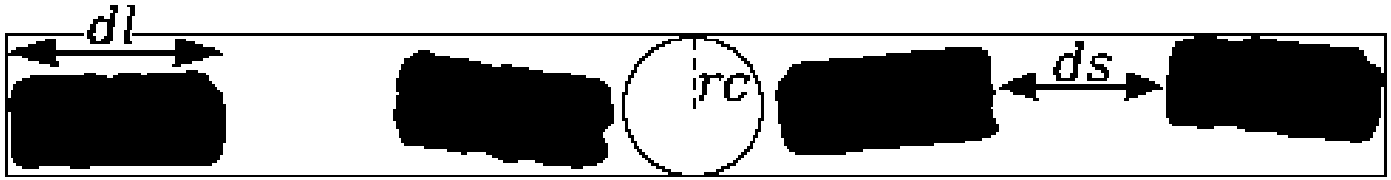
- Simple, fast and robust
- Also works for dotted nodes:



Lessons: Rule Based System for Case Management

- Many different contextual cases, hard to manage interference
- Rule based system forces formal listing of cases
- Easier to check for interference
- Easier to add new cases
- Strictness of rules makes it more robust

Rule Example: Dashed Bonds



A dashed bond is captured using the following conditions (*dl* and *ds* are fuzzy parameters, *rc* is strict):

1. $L = \{l_1, \ldots, l_n\}$, where $n \geq 3$, is a set of line segments,
2. L is approximately collinear with radius of collinearity *rc*
3. Every element of L has approximate length *dl* (dash length)
4. Every pair in L has separation distance of at least *ds* (dash separation) apart
5. Two elements of L, *end elements*, dash-neighbour wrt *ds* precisely one other element of L. All others, *internal elements*, dash-neighbour wrt. *ds* two others

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

Lessons: Large Datasets for Development

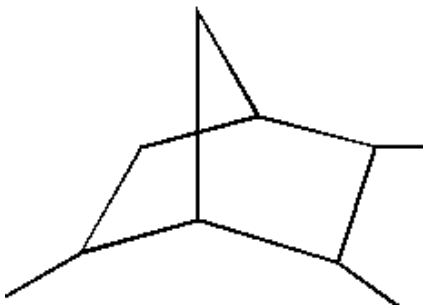
- For testing, experimenting and development, a large dataset is very helpful
- The freely available OSRA dataset is 5735 diagrams
- We scanned 5740 diagrams from the Maybridge catalogue
  - Simple bespoke OCR to extract the Chemical Abstracts Service Registry number from the diagrams
  - Looked up CASRs to get International Chemical Identifiers
  - Converted ICIs to MOL files using OpenBabel
  - Currently investigating copyright issues to make this set publically available

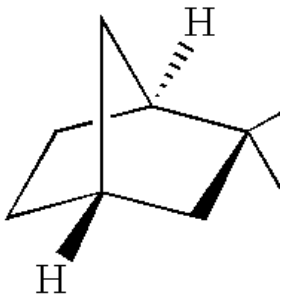
Lessons: Superatom Handling

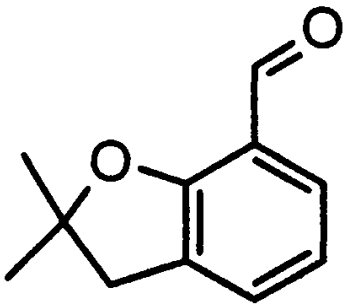
- Superatoms must be handled via a dictionary
- Freely available superatom dictionaries are very limited
- We mined the MOL files of our large datasets to extract all superatom information
- This essentially solved our superatom problems when the superatom has one conection to the molecule
- Permutation problem when the superatom has more than one connection
  - Needs deeper domain knowledge to improve

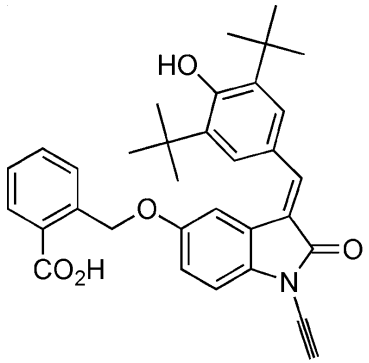
Lessons: Bridge Bonds

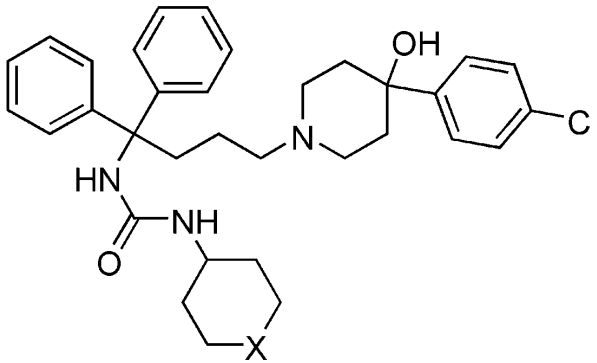
- The crossing points of an open bridge bond is distinctive and easy to recognise
- The crossing point of a closed bridge bond could be the joint of four single planar bonds with an implicit C atom
- Verify/disambiguate by checking if one pair of lines through the joint is part of a non-regular cycle. Compare:











Results on TREC 2011

- 95% and 94.9% correct on two runs of 1000
- Only minor parameter changes between the two runs
- Errors on 50 and 51 diagrams respectively
- 55 diagrams mis-recognised at least once
- 61 causes for the failures

Results on TREC 2011

- Dashed wedge bonds mis-identified: 15
  - 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

Results on TREC 2011

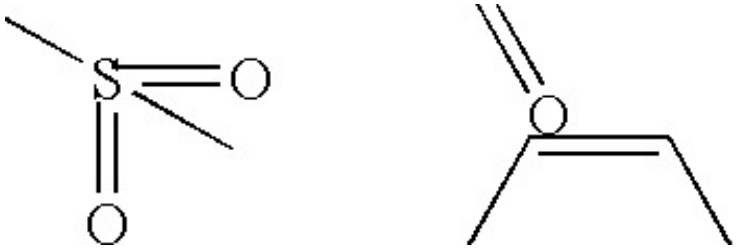
- Dashed wedge bonds mis-identified: 15
  - Short dashes at narrow end mis-recognised as a dashed bond
  - Long dashes at wide end mis-recognised as dashed wedge or dashed bold bond
  - Result: dashed bond mis-recognised as 2 connected bonds

- Incorrect stereochemistry: 10

  - Correct on 158 of 168 stereo bonds with unspecified direction
  - 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

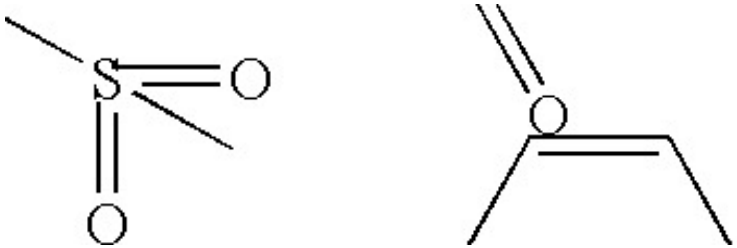
Results on TREC 2011

- Image has touching components: 6

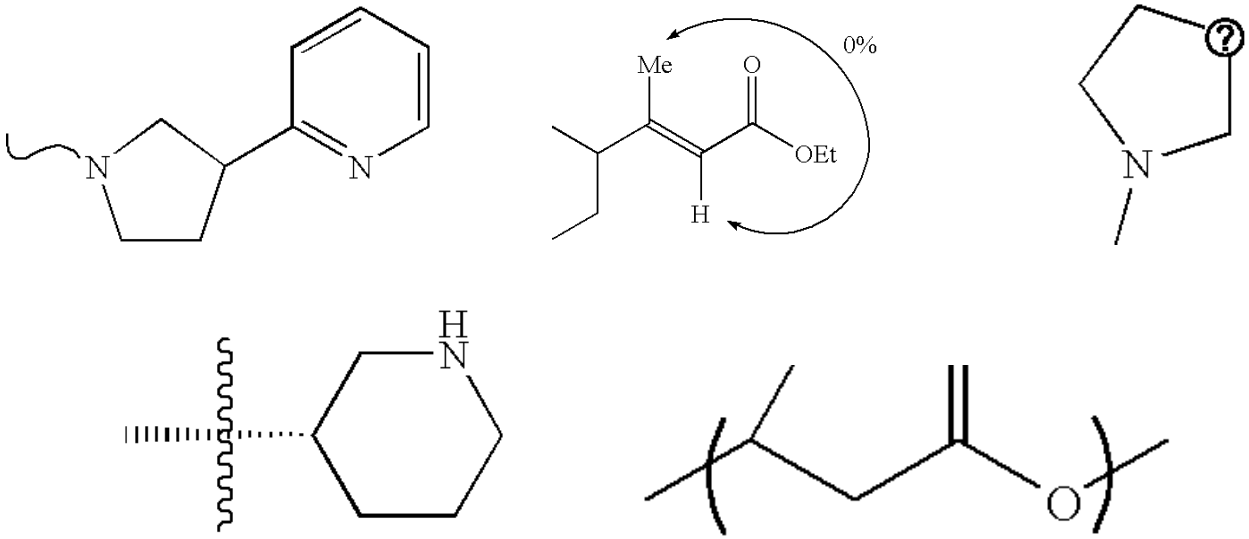


Results on TREC 2011

- Image has touching components: 6

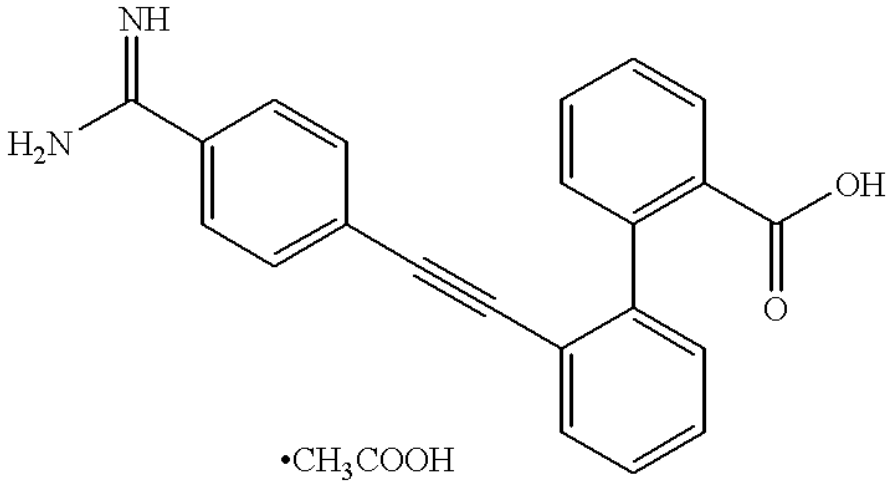


- Unrecognised syntax: 5



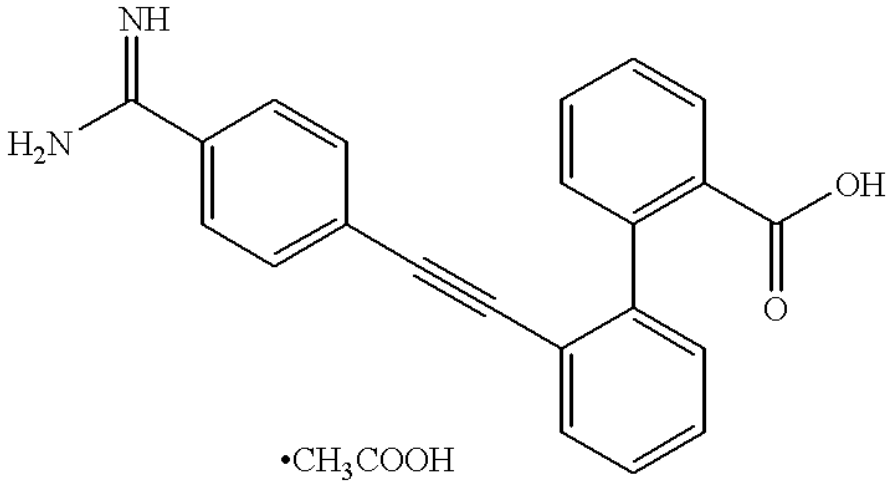
Results on TREC 2011

- Diagram caption confusion: 5

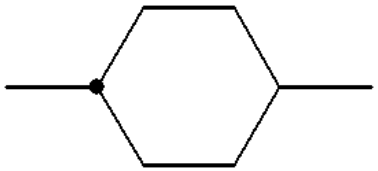


Results on TREC 2011

- Diagram caption confusion: 5

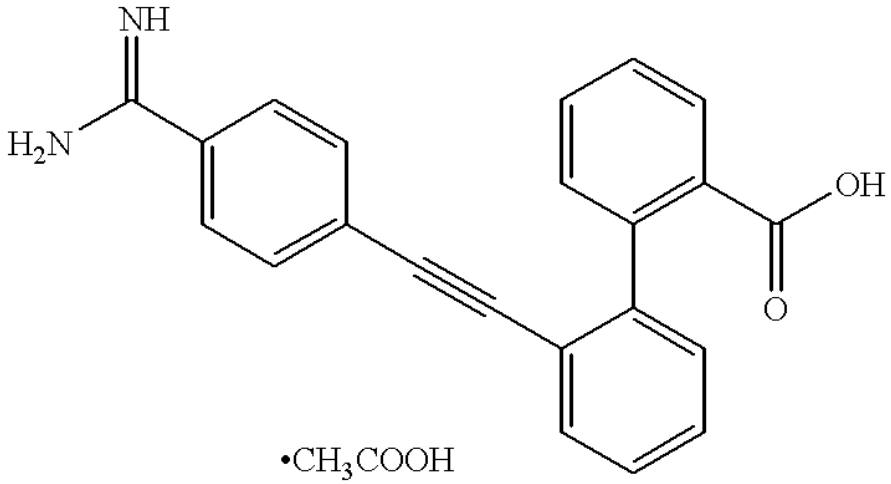


- Dotted joints without 3D Hydrogen Bond: 5
  - Test solutions do not always include 3D Hydrogen bond

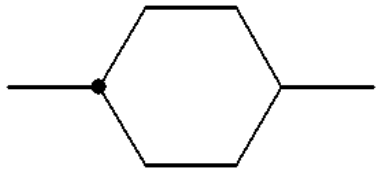


Results on TREC 2011

- Diagram caption confusion: 5



- Dotted joints without 3D Hydrogen Bond: 5
  - Test solutions do not always include 3D Hydrogen bond



- Incorrect character grouping: 5
  - Needs deeper domain knowledge to improve

Results on TREC 2011

- Image has broken characters: 3

Results on TREC 2011

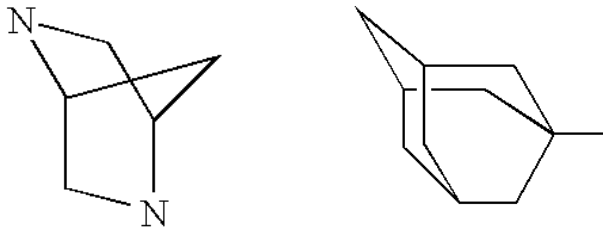
- Image has broken characters: 3

- Incorrect superatom connectivity: 3

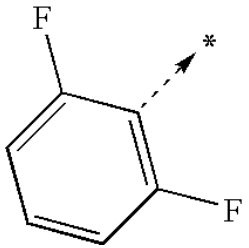
Results on TREC 2011

- Image has broken characters: 3

- Incorrect superatom connectivity: 3
- Problematic bridge bonds: 3



- Unhandled bond type: 1



Conclusions

- MolRec's performance is quite good now
- Some significant improvements can still be made with relatively simple enhancements
- CCA/Thin/Douglas-Peucker approach is very successful
- Breaking all joints (eliminating polylines) significantly simplifies processing
- Filled triangle, bold line and dotted joint detection using "walking ball" is very successful
- Mining superatom data from MOL files is very useful
- Rule based system brings excellent discipline to case handling
- Having large datasets to experiment with is critical
- We really need collaborators who are knowledgeable chemists to make up for our lack of domain knowledge