MolRec at CLEF 2012: Chemical Structure Recognition

CLEF 2012, Rome, Italy

Noureddin M. Sadawi   Alan P. Sexton   Volker Sorge

www.cs.bham.ac.uk/~nms|~aps|~vxs

School of Computer Science
University of Birmingham, UK

19 September 2012, CLEF 2012, Rome

Introduction

- Convey information through pictorial representations

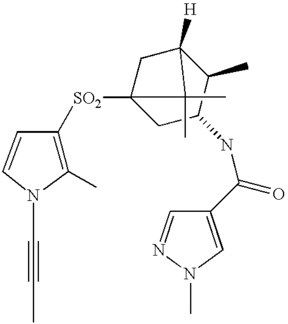
  a picture is worth a thousand words
- Chemical structure diagrams in publications and patents

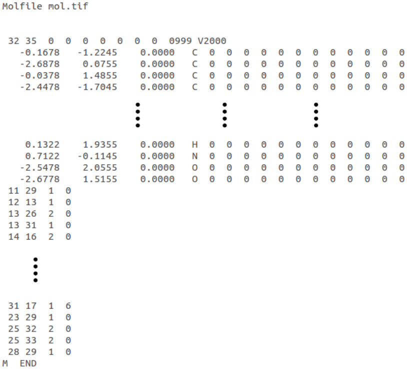
  research articles, patent specs, catalogues, etc.
- Convert images to computer processable/searchable format

  patent search, drug discovery, cancer research, etc
- Simple but powerful approach

  MolRec employs a clearly defined *rule based approach*

Task





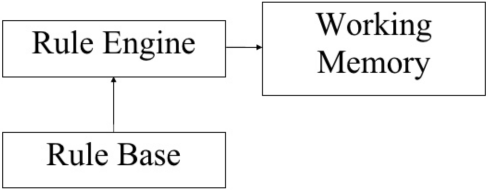
```
Molfile mol.tif


 32 35  0  0  0  0  0  0  0  0999 V2000
   -0.1678   -1.2245    0.0000   C  0  0  0  0  0  0  0  0  0  0  0  0
   -2.6878    0.0755    0.0000   C  0  0  0  0  0  0  0  0  0  0  0  0
   -0.0378    1.4855    0.0000   C  0  0  0  0  0  0  0  0  0  0  0  0
   -2.4478   -1.7045    0.0000   C  0  0  0  0  0  0  0  0  0  0  0  0

                   ⋮              ⋮               ⋮

    0.1322    1.9355    0.0000   H  0  0  0  0  0  0  0  0  0  0  0  0
    0.7122   -0.1145    0.0000   N  0  0  0  0  0  0  0  0  0  0  0  0
   -2.5478    2.0555    0.0000   O  0  0  0  0  0  0  0  0  0  0  0  0
   -2.6778    1.5155    0.0000   O  0  0  0  0  0  0  0  0  0  0  0  0
 11 29  1  0
 12 13  1  0
 13 26  2  0
 13 31  1  0
 14 16  2  0

        ⋮

 31 17  1  6
 23 29  1  0
 25 32  2  0
 25 33  2  0
 28 29  1  0
M  END
```

- Image analysis and shape identification
- Connectivity preservation and contextual discrimination
- Strategy to manage cases
- Suitable output format

Rule-based Systems

- Rule-based Systems (RBS) encode human knowledge
- RBS have three components: a **Working Memory**, a **Rule Engine** and a **Rule Base**
- The Working Memory has a set of facts
- The Rule Base contains rules
- The Rule Engine interprets these rules and applies them as the case may be

More on Rule Engine

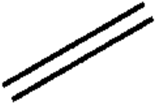


- The rule engine works with the contents of the working memory
- The rule engine continuously accesses the working memory
- A rule is applicable if there exist objects that satisfy its preconditions
- There must be a termination mechanism

MolRec's Overall Procedure

- Vectorisation (Detection of geometric primitives) Character Groups, Circles, Line segments, Triangles and Arrows
- Working memory is a set of primitives
- Rule Base has 18 rules
- Rule Engine (rewrites primitives into a graph)
- Rules are chosen randomly from a rule pool
- Disambiguation and graph correction
- Produce output from graph (e.g. MOL, SMILES)

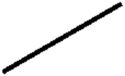
Example Rule: R2. Double Bond



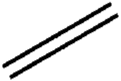
1. $L = \{l_1, l_2\}$ is a set of two line segments,
2. $\forall l \in L : length(l) > wb$ (wedge base)
3. $\forall l \in L : width(l) < bbw$ (bold bond width)
4. $l_1 \parallel_{bs}^{ol} l_2$
5. There is no line segment $l \notin L$ such that $l_1 \parallel_{bs}^{ol} l$ or $l_2 \parallel_{bs}^{ol} l$.

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

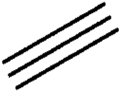
Rules I



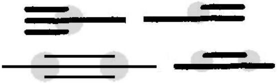
R1.Single



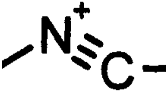
R2.Double



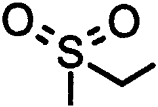
R3.Triple



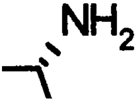
Implicit Nodes



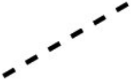
R4. Triple vs Dashed Bold



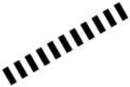
R5. Double vs Dashed Bold



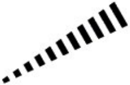
R6. Double vs Dashed Wedge



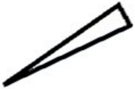
R7. Dashed



R8. Dashed Bold

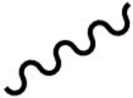


R9. Dashed Wedge

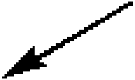


R10,R11. Hollow Wedge

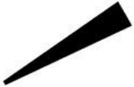
Rules II



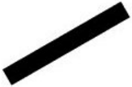
R12. Wavy



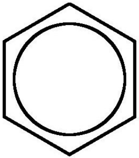
R13. Dative



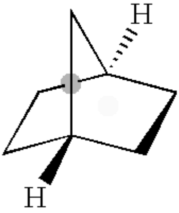
R14. Solid Wedge



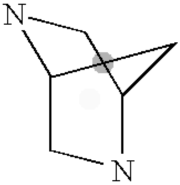
R15. Bold



R16. Aromatic Ring

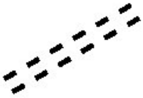


R17. Open Bridge

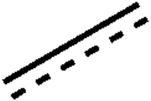


R18. Closed Bridge

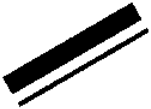
Rules III



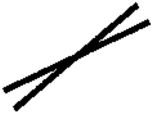
R19. Aromatic



R20. Tautomeric

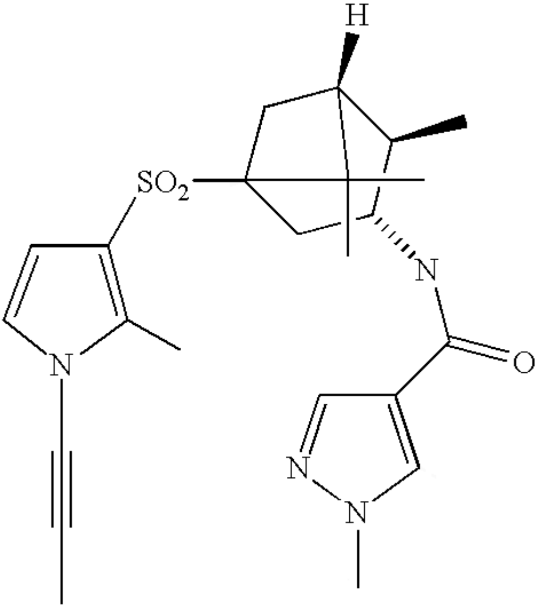


R21. Double Bond with Type 1 Stereo-chemistry

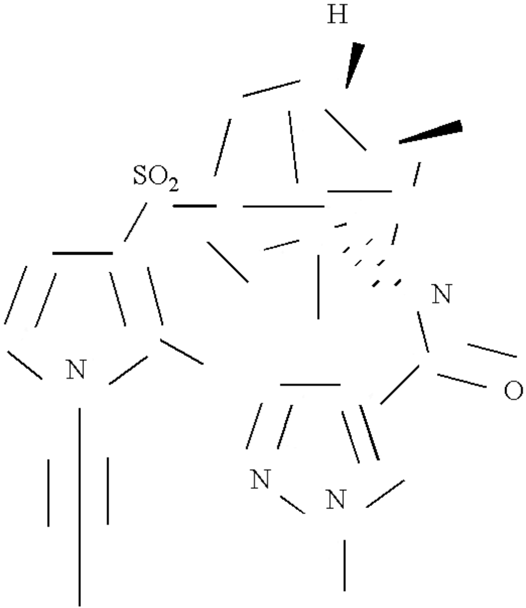


R22. Double Bond with Type 2 Stereo-chemistry

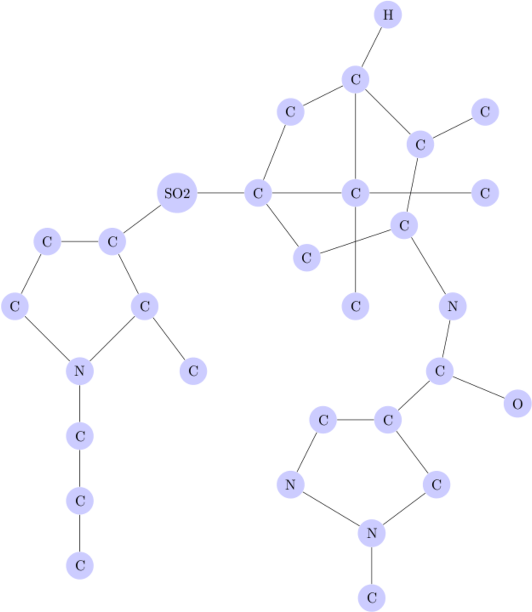
Example Diagram



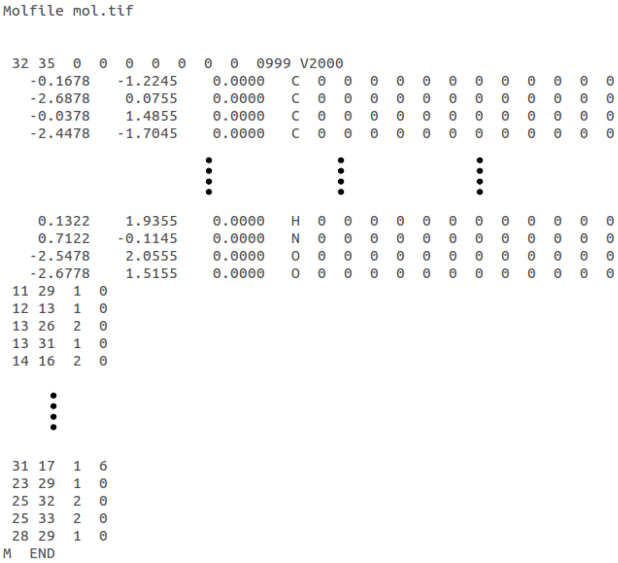
Primitives



Graph



MOL file



```
Molfile mol.tif


 32 35  0  0  0  0  0  0  0  0999 V2000
   -0.1678   -1.2245    0.0000    C  0  0  0  0  0  0  0  0  0  0  0  0
   -2.6878    0.0755    0.0000    C  0  0  0  0  0  0  0  0  0  0  0  0
   -0.0378    1.4855    0.0000    C  0  0  0  0  0  0  0  0  0  0  0  0
   -2.4478   -1.7045    0.0000    C  0  0  0  0  0  0  0  0  0  0  0  0
                           ⋮              ⋮               ⋮

    0.1322    1.9355    0.0000    H  0  0  0  0  0  0  0  0  0  0  0  0
    0.7122   -0.1145    0.0000    N  0  0  0  0  0  0  0  0  0  0  0  0
   -2.5478    2.0555    0.0000    O  0  0  0  0  0  0  0  0  0  0  0  0
   -2.6778    1.5155    0.0000    O  0  0  0  0  0  0  0  0  0  0  0  0
 11 29  1  0
 12 13  1  0
 13 26  2  0
 13 31  1  0
 14 16  2  0

     ⋮

 31 17  1  6
 23 29  1  0
 25 32  2  0
 25 33  2  0
 28 29  1  0
M  END
```

Results

Run	# Recognitions	# Mis-Recognitions	Accuracy
1	832	33	96.18%
2	821	44	94.91%
3	821	44	94.91%
4	832	33	96.18%

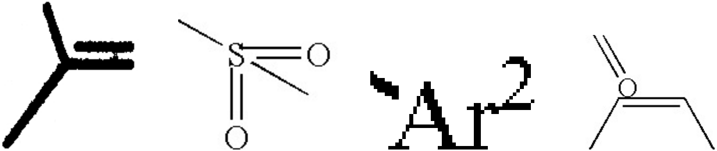
Four Runs on the Automatic Evaluation Set (865 images)

Run	# Recognitions	# Mis-Recognitions	Accuracy
1	44	51	46.32%
2	56	39	58.95%
3	44	51	46.32%
4	54	41	56.84%

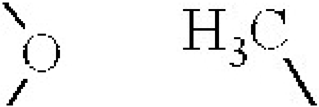
Four Runs on the Manual Evaluation Set (95 images)

Problem Cases

- Touching Components

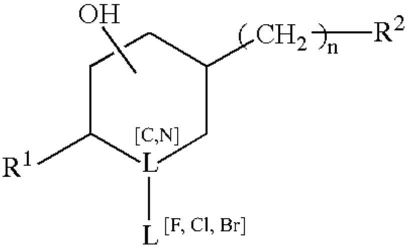


- Broken Components

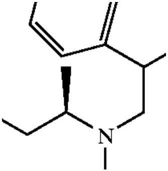


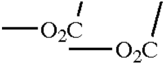
Problem Cases

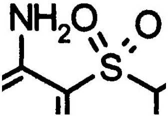
- Markush Structures

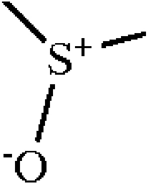


- Grouping Errors



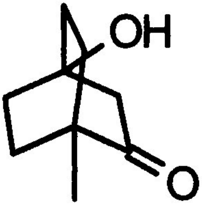


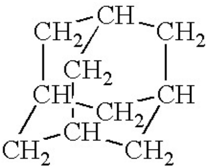




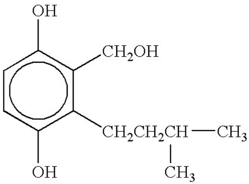
Problem Cases

- Ambiguity

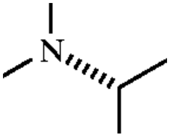




- Other Reasons

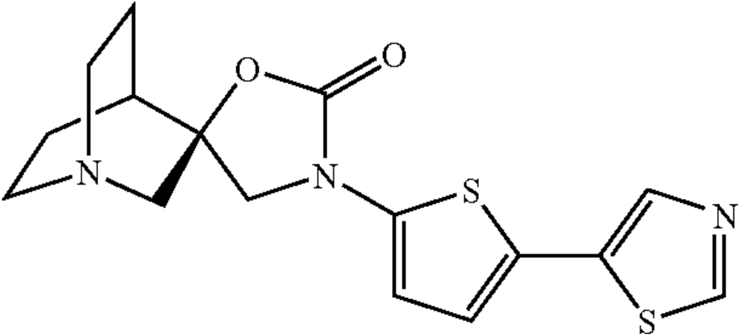


Superatoms

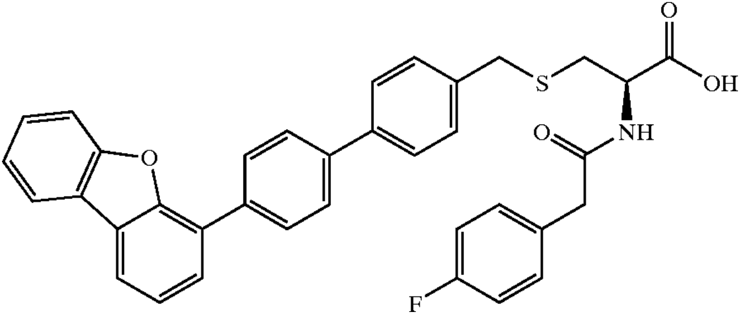


Stereo-centre

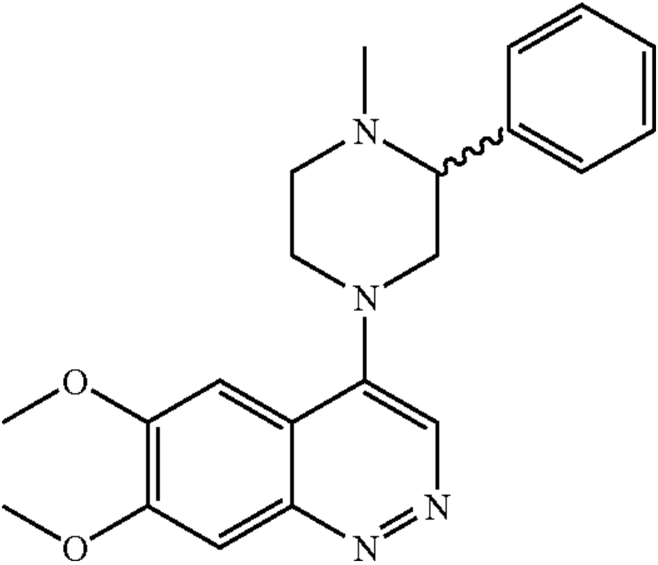
Correct Recognition I



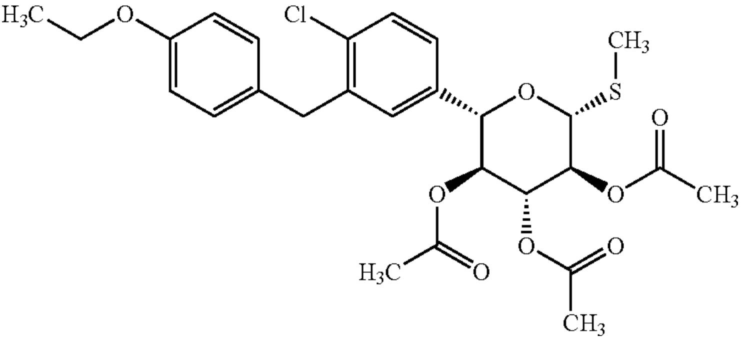
Correct Recognition II



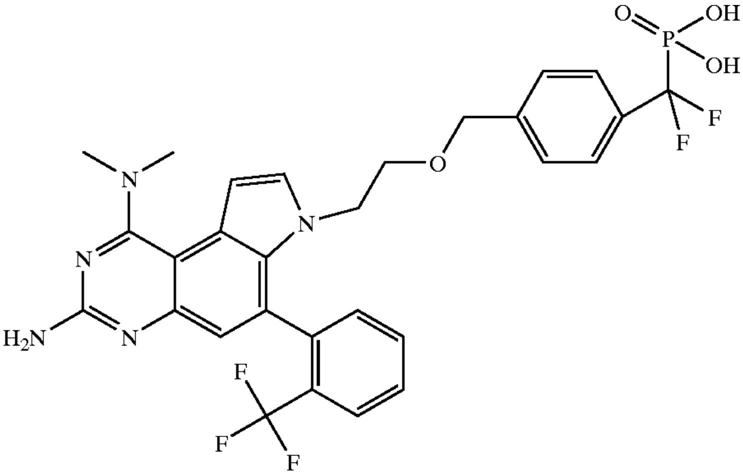
Correct Recognition III



Correct Recognition IV



Correct Recognition V



Summary & Future Work

- The rule based approach means analysis is fast, easily extendible and *flexible*
- Recognition of even complex traditional diagrams works well
- Improved OCR (touching/broken symbol correction) would considerably improve the system
- Handling Markush structures and finding suitable representation
- More domain knowledge to solve connection permutation problem for superatoms
- More domain knowledge to accurately determine stereo-centre
- The larger picture, whole document analysis