Semantic Analysis of Chemical Patents

David Jessop

dmj30@cam.ac.uk

UNIVERSITY OF CAMBRIDGE

Unilever Cambridge

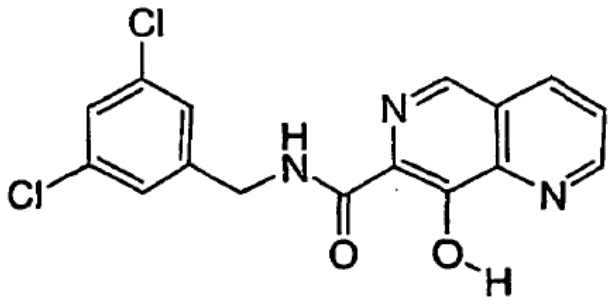
Centre for Molecular Informatics

Chemical Patents

EXAMPLE 1

N-(3,5-dichlorobenzyl)-8-hydroxy-1,6-naphthyridine-7-carboxamide

[0135]



Step 1: Preparation of 3-{[Methoxycarbonylmethyl-(tol
propyl ester

[0136] Isopropyl 3-(hydroxymethyl)pyridine-2-carbox
Chem. 1989, 32, 827), methyl N-[(4-methylphenyl)sulf
1.5 mol) were dissolved in dry THF (3000mls) and coole
(267.6 g, 1.5 mol) was dissolved in dry THF (250 mls)

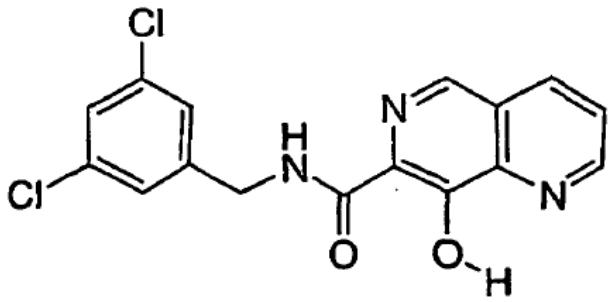
- ≈ 70 European chemistry patents per week
- Hundreds of pages in length
- Hundreds of reactions per document
- Not machine readable!

Semanticizing Reactions

EXAMPLE 1

N-(3,5-dichlorobenzyl)-8-hydroxy-1,6-naphthyridine-7-carboxamide

[0135]

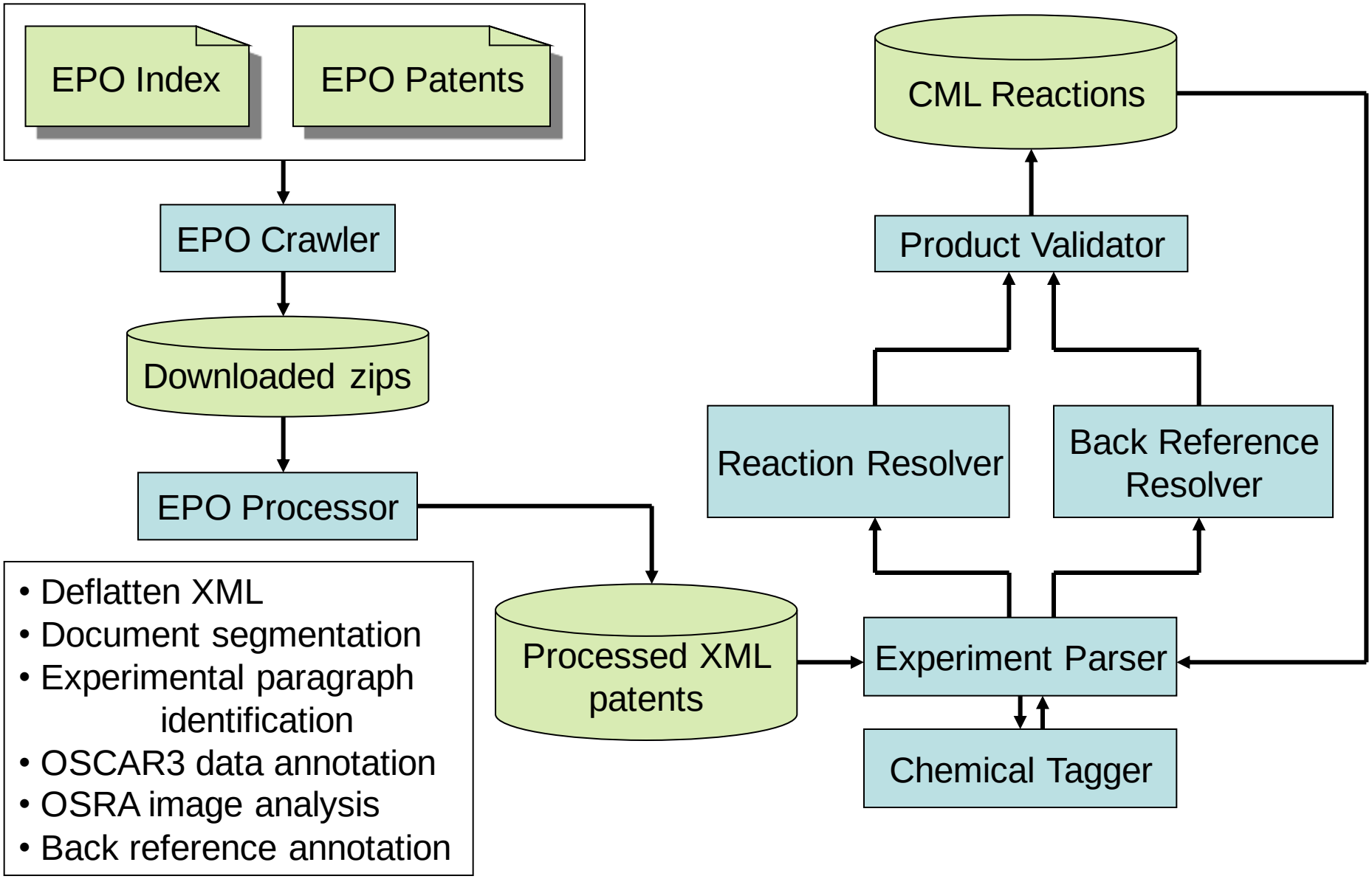


Step 1: Preparation of 3-{[Methoxycarbonylmethyl-(tol
propyl ester

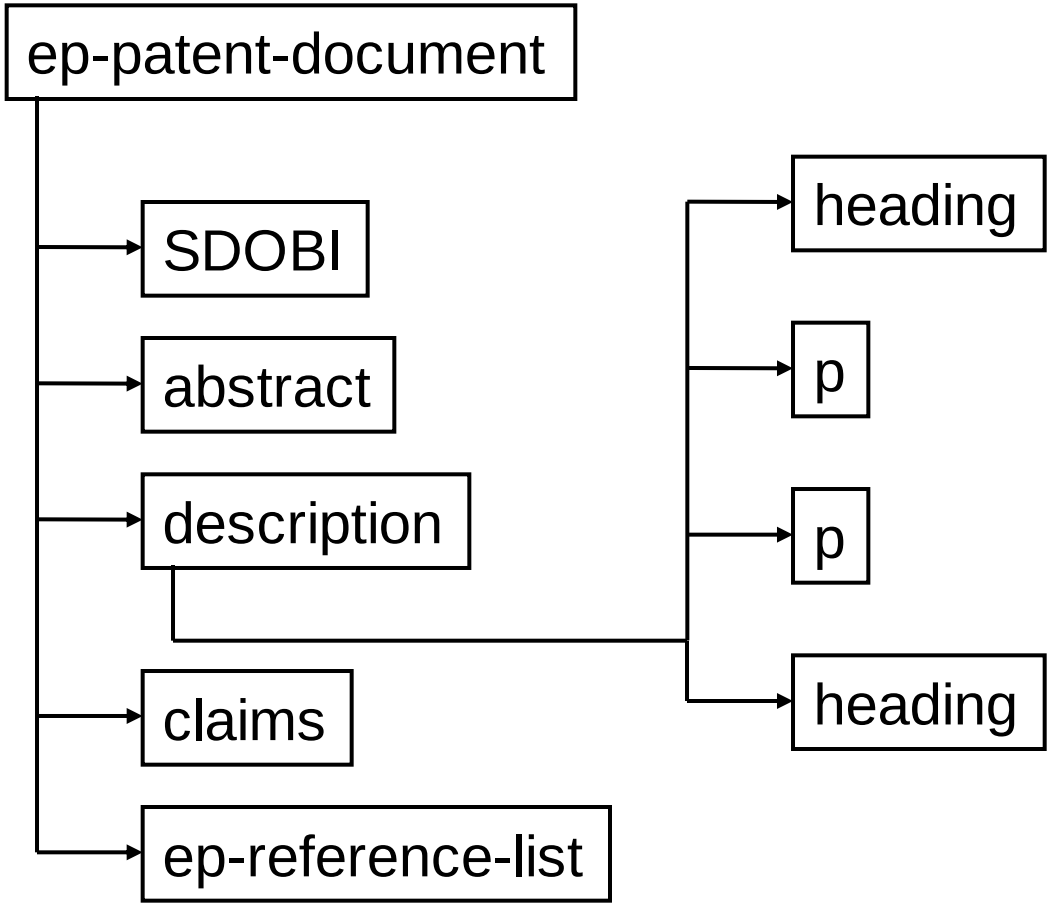
[0136] Isopropyl 3-(hydroxymethyl)pyridine-2-carbox
Chem. 1989, 32, 827), methyl N-[(4-methylphenyl)sulf
1.5 mol) were dissolved in dry THF (3000mls) and coole
(267.6 g, 1.5 mol) was dissolved in dry THF (250 mls)

```
<reaction>
  <productList>
    <product>
      <molecule title='3-([Methoxycarb
    </product>
  </productList>
  <reactantList>
    <reactant>
      <molecule title='Isopropyl 3-(hyd
      <amount units='cml:g'>200</am
      <amount units='cml:mol'>1.0</a
    </reactant>
```

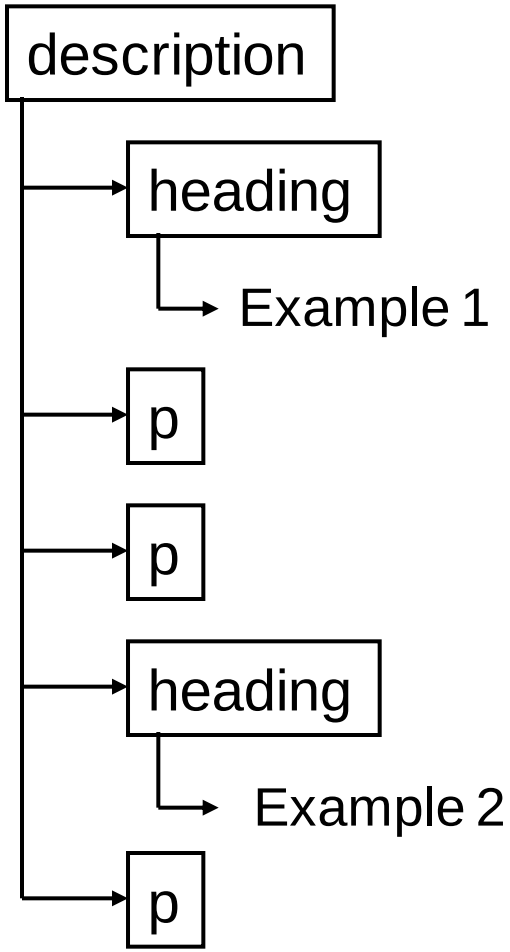
PatentEye Overview

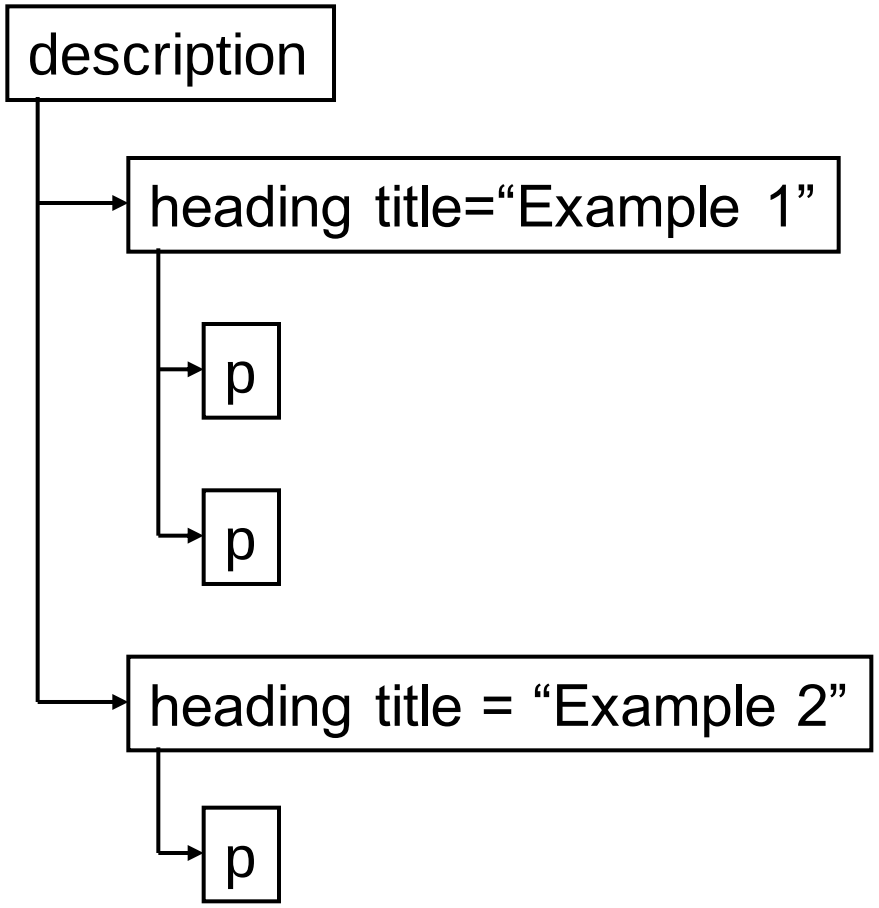


EPO XML Structure



Paragraph Deflattening





Document Segmentation

Description

Field of the Invention

[0001] The i
facture of medi

Background

[0002] The h
presynaptic he
receptor is alsc
[0003] Propc
(Panula et al. A
129), sleep/wal

Description

TECHNICAL FIELD

[0001] The present i
[0002] More specific

(1) piperidine deriv

(wherein all symbc
(2) a process for p
(3) an agent comp

Description

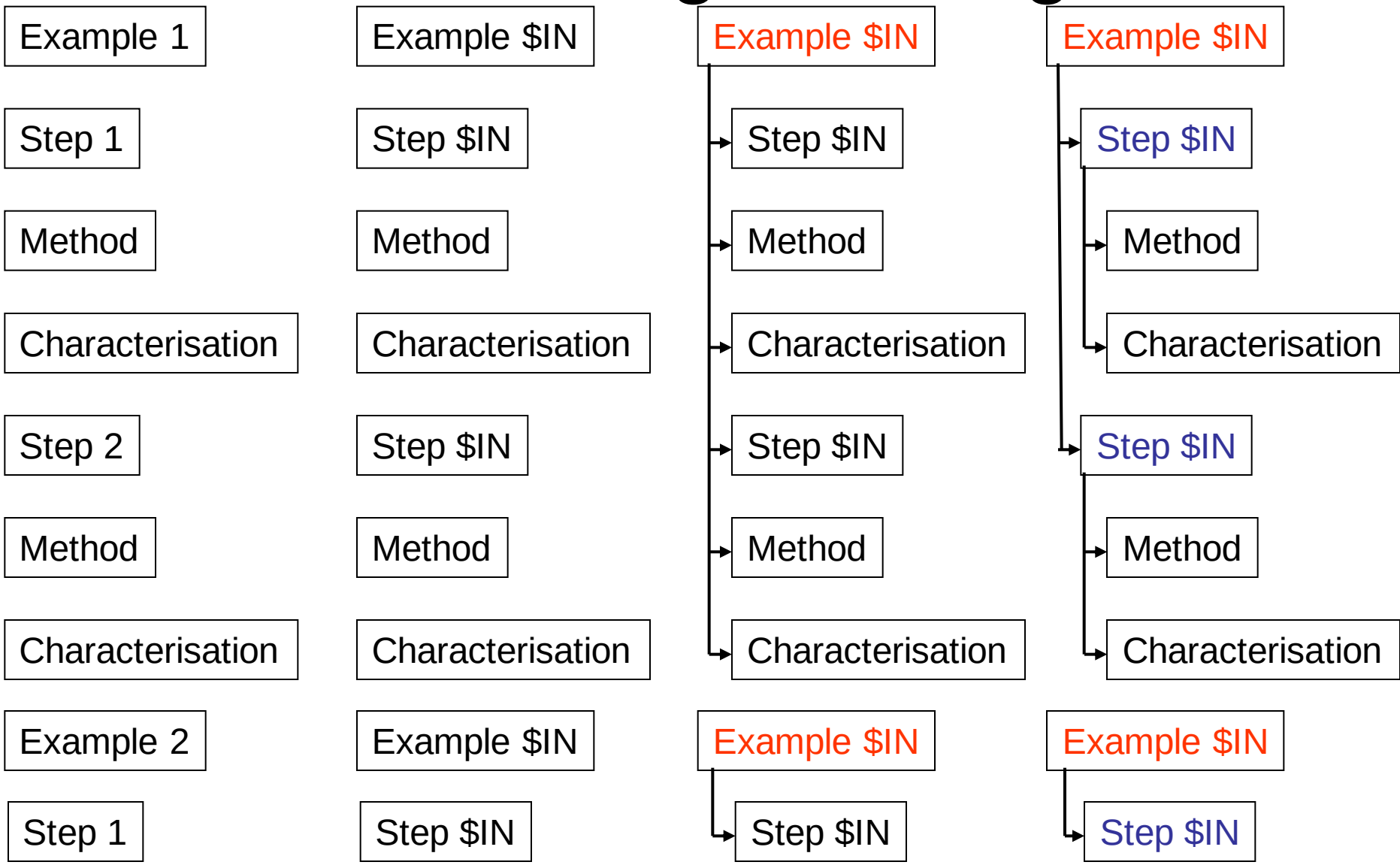
FIELD OF THE INVENTION

[0001] The present invention is directed to aza- a
ceptable salts thereof, their synthesis, and their use
present invention include 7-(N-substituted carboxa
pounds and pharmaceutically acceptable salts ther
infection by HIV and for treating AIDS.

[0002] References are made throughout this applic
the state of the art to which this invention pertains.
reference in their entireties.

BACKGROUND OF THE INVENTION

Unflattening Headings



Classification of Experimental Paragraphs

3-{[Methoxycarbonylmethyl-(toluene-4-sulfonyl)-amino]-methyl}-pyridine-2-carboxylic acid isopropyl ester (1.02 mol) was dissolved in dry methanol (4000ml) and cooled to zero degrees under nitrogen. Then via addition funnel, sodium methoxide (137.8g, 2.5 mol) was added slowly to avoid any exotherm.

The compounds of the present invention can be readily prepared according to the following reaction schemes and examples, or modifications thereof, using readily available starting materials, reagents and conventional synthesis procedures.

Can a machine tell them apart?

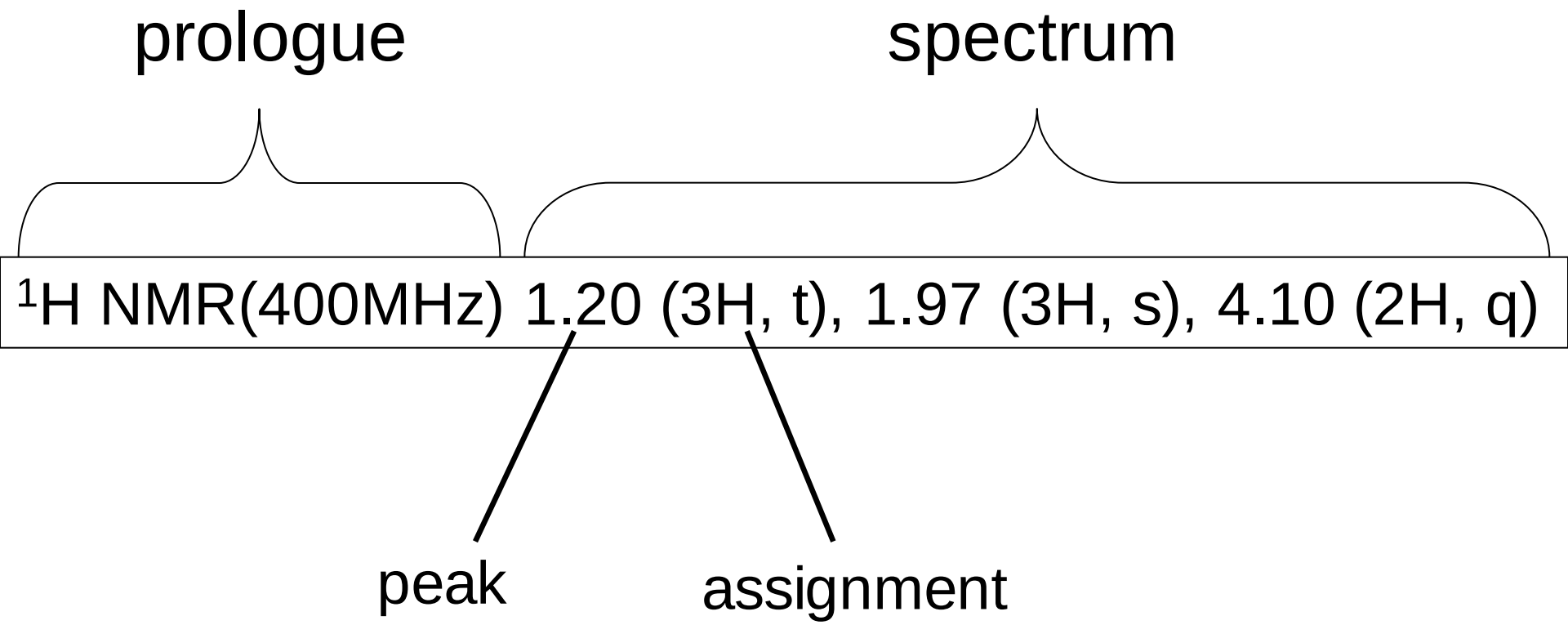
Classification Results

Experimental	
Probability	Frequency
0.99	115
$0.98 \geq p > 0.95$	1
$0.05 \geq p > 0$	3

Non-experimental	
Probability	Frequency
0.99	12
$0.06 < p < 0.5$	2
$0.01 < p \leq 0.06$	3
0.01	102

- Naïve Bayesian Classifier
- Experimental paragraphs: 96.6%
- Non-experimental paragraphs: 89.9%

Data Annotation



Annotation Performance

Spectrum type	# in corpus	OSCAR 3 current performance	
		Precision (%)	Recall (%)
MassSpec	199	70%	61%
HNMR	202	82%	89%
CNMR	24	85%	92%

$$\text{Precision} = \frac{TP}{TP + FP}$$

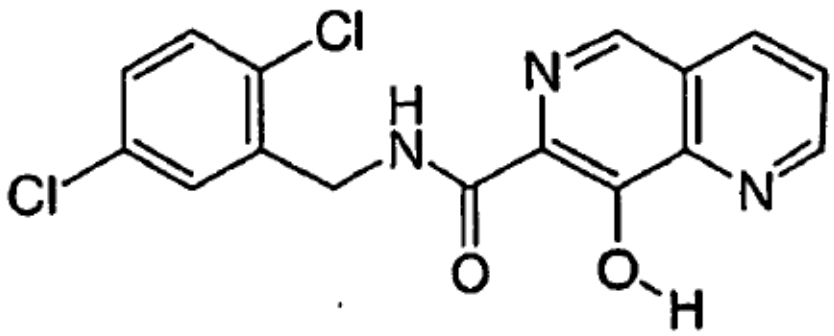
$$\text{Recall} = \frac{TP}{TP + FN}$$

Image Interpretation - OSRA

EXAMPLE 2

N-(2,5-dichlorobenzyl)-8-hydroxy-1,6-naphthyridine-7-carboxamide

[0139]



SMILES

Step 1: Preparation of 8-hydroxy-1,6-naphthyridine-7-carboxylic acid

[0140] To a slurry of methyl 8-hydroxy-1,6-naphthyridine-7-carboxyla
methanol. (45ml) was added lithium hydroxide (22.0ml of a 1M aq. solut
100°C for 7 hrs. Upon cooling to room temperature, hydrochloric acid

Image Interpretation Results

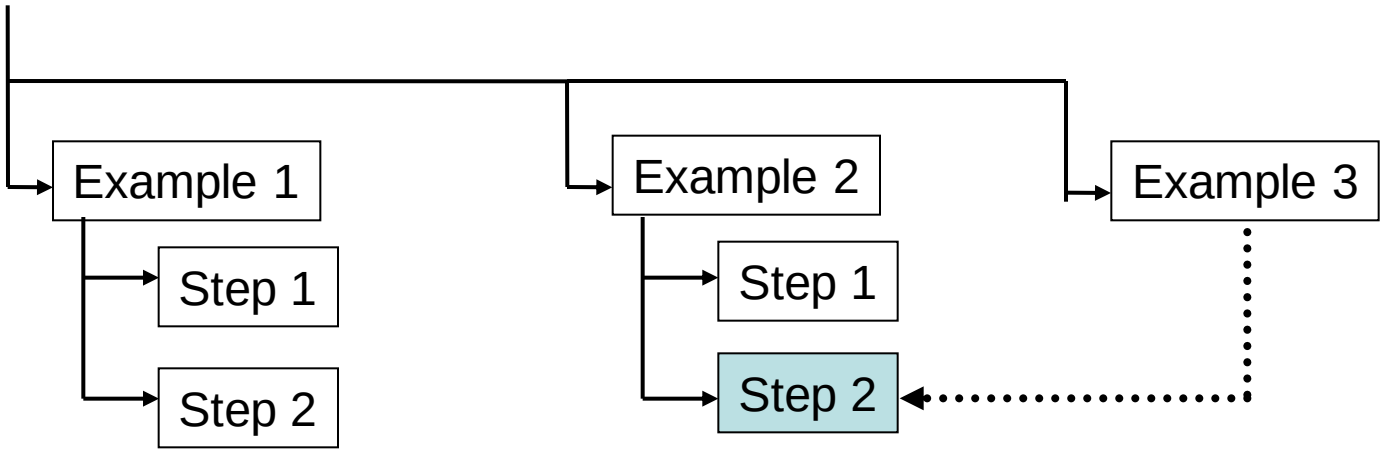
- By comparison to manually-redrawn structures

	%
Correct by InChI	34
Incorrect by InChI	40
Partial SMILES	26
Invalid SMILES	1

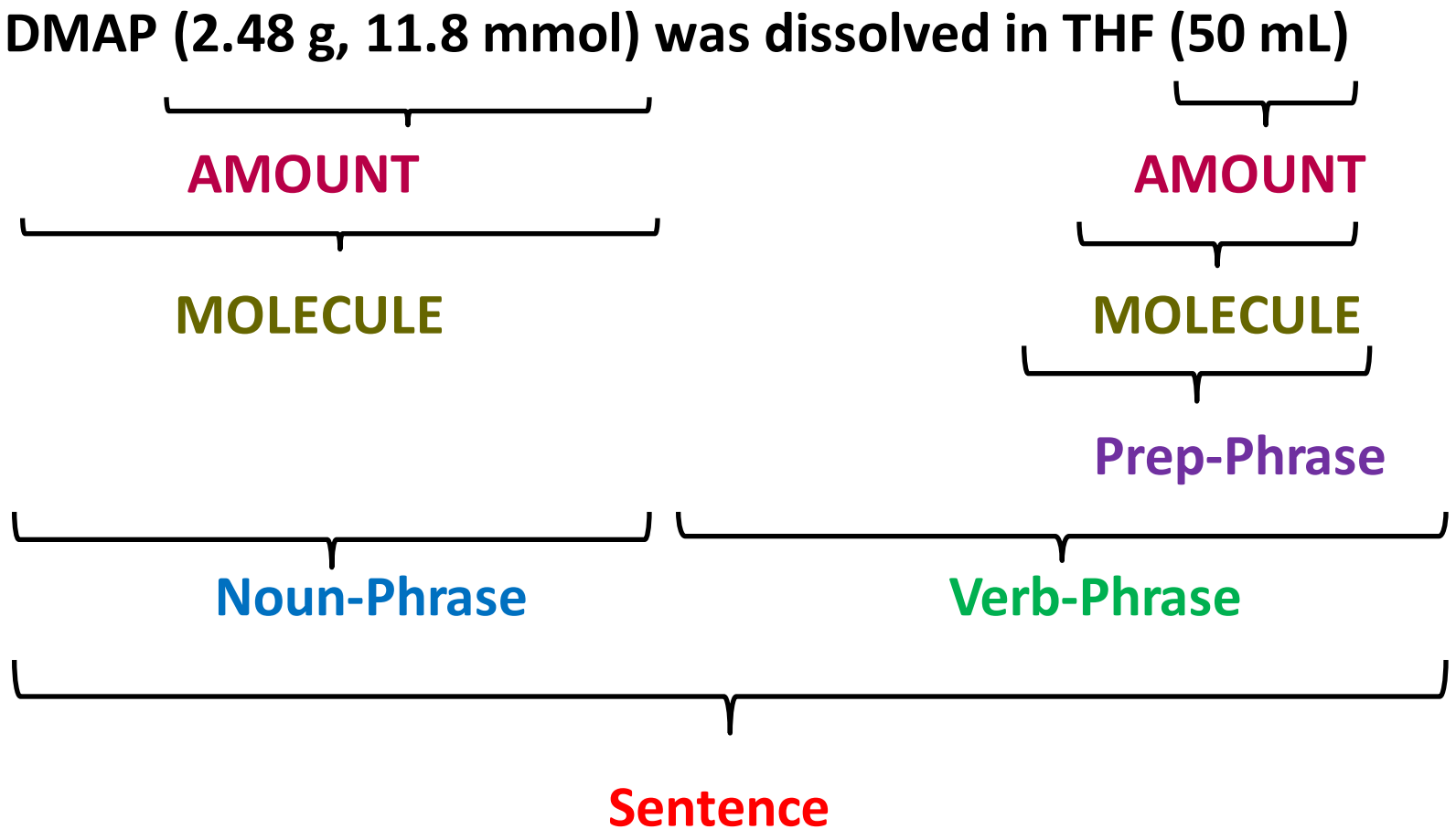
Back Reference Annotation

Example 3:

The title compound was prepared using the procedure described in Example 2, Step 2 from 8-hydroxy-1,6-naphthyridine-7-carboxylic acid and 1(R,S) aminoindane.



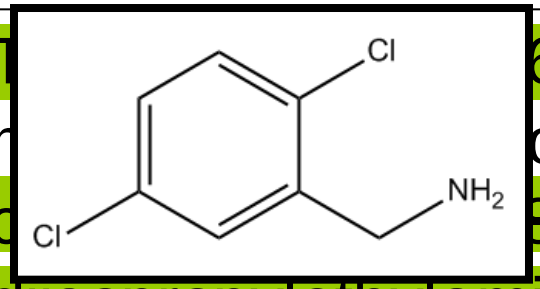
Identification of Reagents with ChemicalTagger

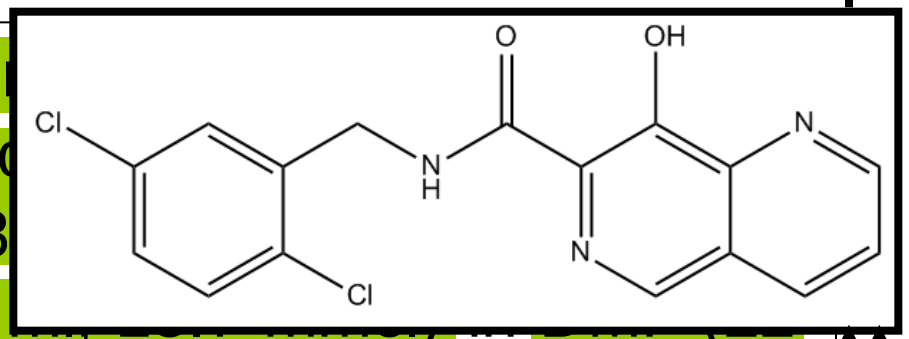


Preparation of N-(2,5-dichlorobenzyl)-8-hydroxy-1,6-naphthyridine-7-carboxamide

Triphosgene (0.556g, 1.87 mmol) was added over 20 mins to a solution of 8-hydroxy-1,6-naphthyridine-7-carboxylic acid (0.89g, 4.68 mmol) and diisopropylethylamine 3.26 ml, 18.7 mmol) in DMF (22 ml) at 0°C. 2,5-dichlorobenzylamine (0.142 ml, 1.05 mmol) was treated with a portion of the above solution (0.58ml, 0.07 mmol) and the resulting mixture was stirred at room temperature for 16 hrs. 1H NMR (d6DMSO, 400MHz) d 9.90 (1H, br t, J=5.0 Hz)), 9.20 (1H, d, J=4.0 Hz), 8.95 (1H, s), 8.65 (1H, d, J=8.0Hz),7.85 (1H, dd, J=8.0 and 4.0 Hz), 7.54 (1H, d, J=8.0Hz), 7.50-7.30 (2H, m), 4.64 (2H, d, J=5.0 Hz) ppm. FAB MS calcd for C16H11N3O2Cl2 348 (MH+), found 348.

Preparation of N-(2,5-dichlorobenzyl)-8-hydroxy-1,6-naphthyridine-7-carboxamide





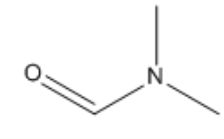
T 6g, 1.87 r of 8-hydr 89g, 4.68 diisopropylethylamine 3.26 ml) at 0°C. 2,5-dichlorobenzylamine (0.142 ml, 1.05 mmol) was treated with a portion of the above solution (0.58ml, 0.07 mmol) and the resulting mixture was stirred at room temperature for 16 hrs. 1H NMR (d6DMSO, 400MHz) d 9.90 (1H, br t, J=5.0 Hz)), 9.20 (1H, d, J=4.0 Hz), 8.95 (1H, s), 8.65 (1H, d, J=8.0Hz),7.85 (1H, dd, J=8.0 and 4.0 Hz), 7.54 (1H, d, J=8.0Hz), 7.50-7.30 (2H, m), 4.64 (2H, d, J=5.0 Hz) ppm. FAB MS calcd for C16H11N3O2Cl2 348 (MH+), found 348.

Resolution of Analogous Reactions

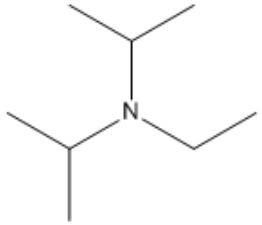
Example 3: N-[(1R,S)-2,3-dihydro-1H-inden-1-yl]-8-hydroxy-1,6-naphthyridine-7-carboxamide

The title compound was prepared using the procedure described in Example 2, Step 2 from 8-hydroxy-1,6-naphthyridine-7-carboxylic acid and 1(R,S) aminoindane.

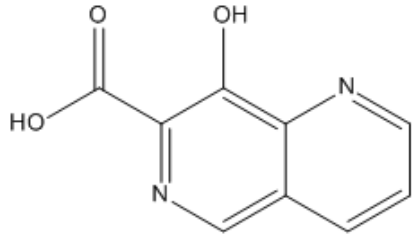
Molecules in Reference Reaction



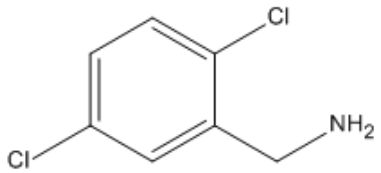
dimethylformamide



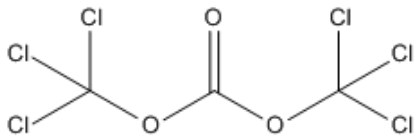
diisopropylethylamine



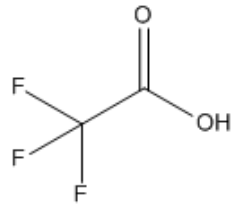
8-hydroxy-1,6-naphthyridine-7-carboxylic acid



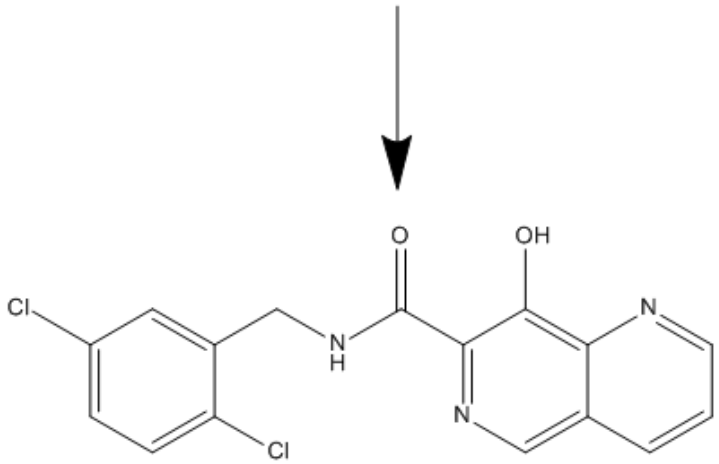
2,5-dichlorobenzylamine



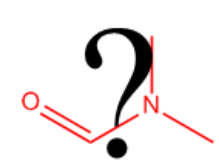
triphosgene



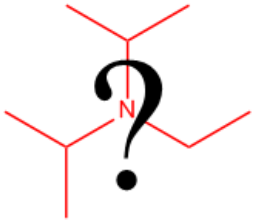
trifluoroacetic acid



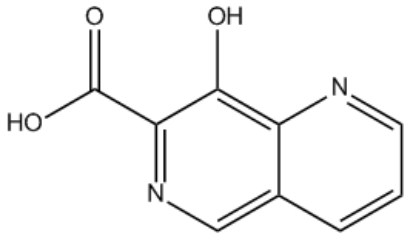
N-(2,5-dichlorobenzyl)-8-hydroxy-1,6-naphthyridine-7-carboxamide



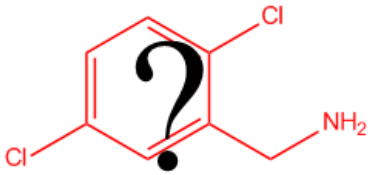
dimethylformamide



diisopropylethylamine

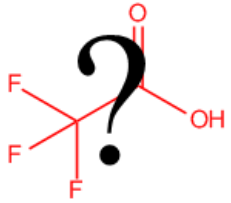


8-hydroxy-1,6-naphthyridine-7-carboxylic acid

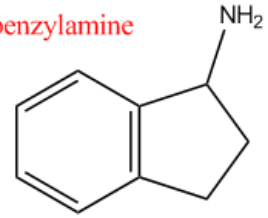


2,5-dichlorobenzylamine

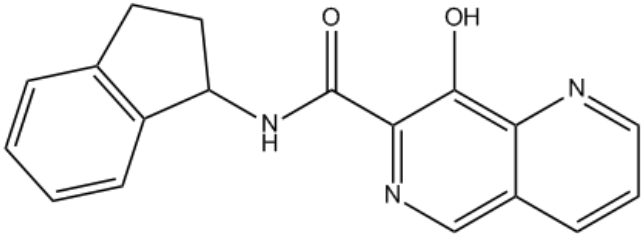
triphosgene



trifluoroacetic acid



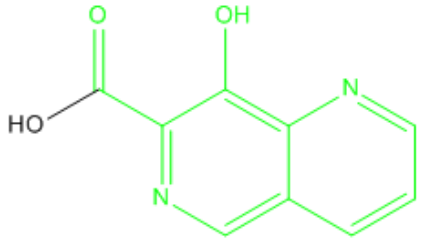
1(R,S) aminoindane



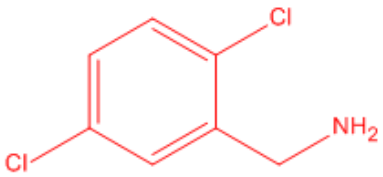
n-[(1r,sr)-2,3-dihydro-1h-inden-1-yl]-8-hydroxy-1,6-naphthyridine-7-carboxamide

dimethylformamide

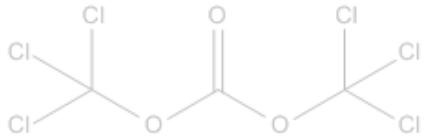
diisopropylethylamine



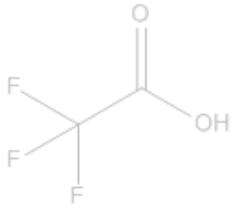
8-hydroxy-1,6-naphthyridine-7-carboxylic acid



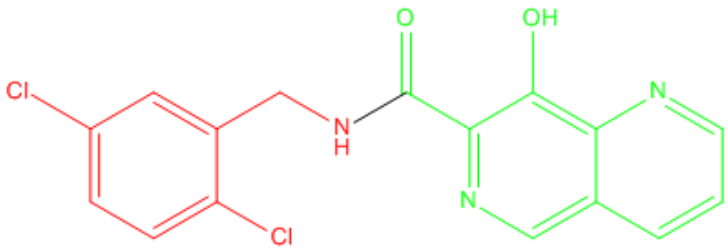
2,5-dichlorobenzylamine



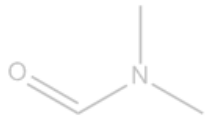
triphosgene



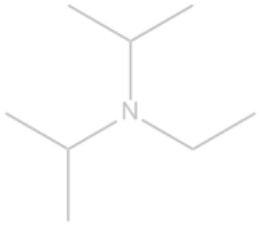
trifluoroacetic acid



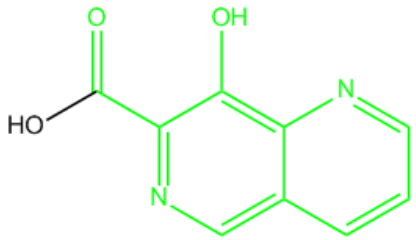
N-(2,5-dichlorobenzyl)-8-hydroxy-1,6-naphthyridine-7-carboxamide



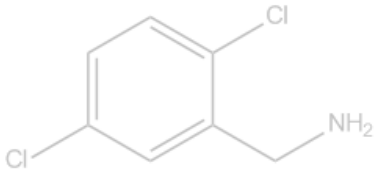
dimethylformamide



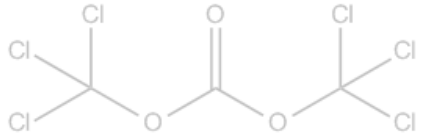
diisopropylethylamine



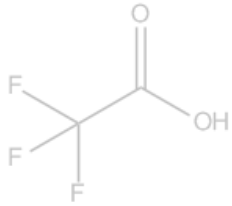
8-hydroxy-1,6-naphthyridine-7-carboxylic acid



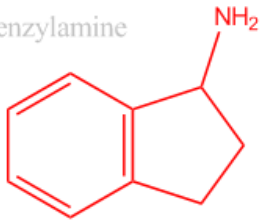
2,5-dichlorobenzylamine



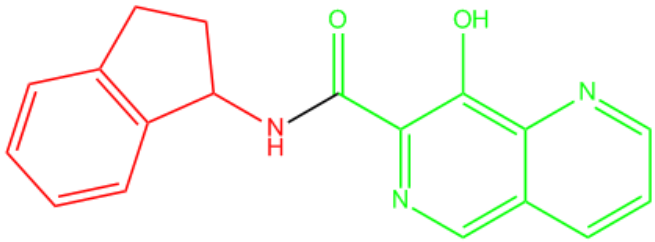
triphosgene



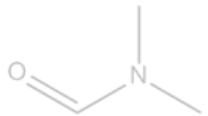
trifluoroacetic acid



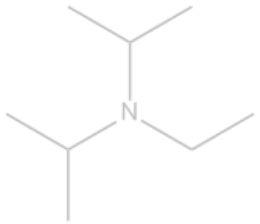
1(R,S) aminoindane



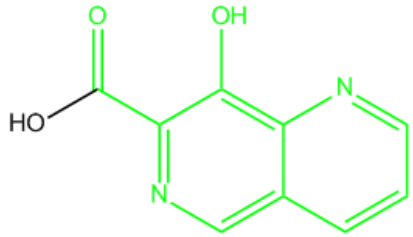
n-[(1r,sr)-2,3-dihydro-1h-inden-1-yl]-8-hydroxy-1,6-naphthyridine-7-carboxamide



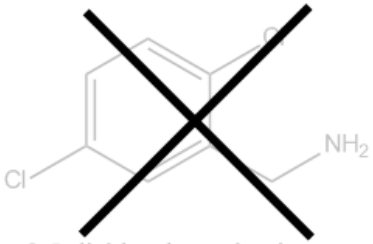
dimethylformamide



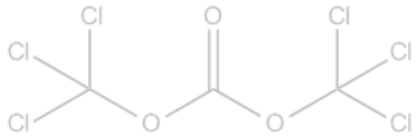
diisopropylethylamine



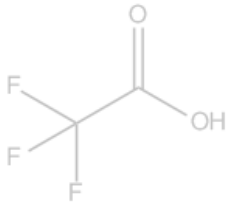
8-hydroxy-1,6-naphthyridine-7-carboxylic acid



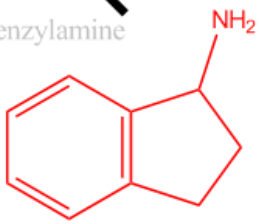
2,5-dichlorobenzylamine



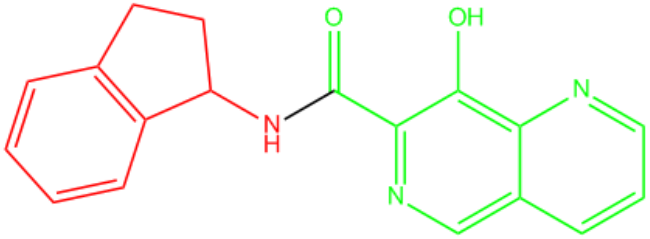
triphosgene



trifluoroacetic acid

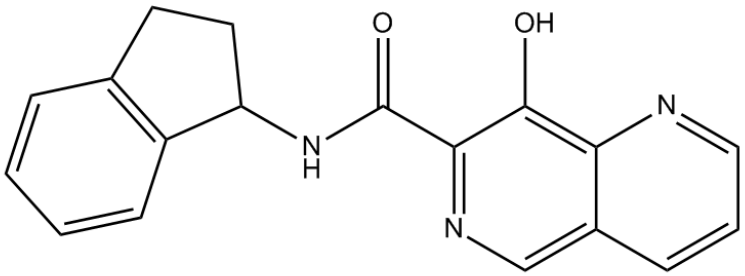


1(R,S) aminoindane



n-[(1r,sr)-2,3-dihydro-1h-inden-1-yl]-8-hydroxy-1,6-naphthyridine-7-carboxamide

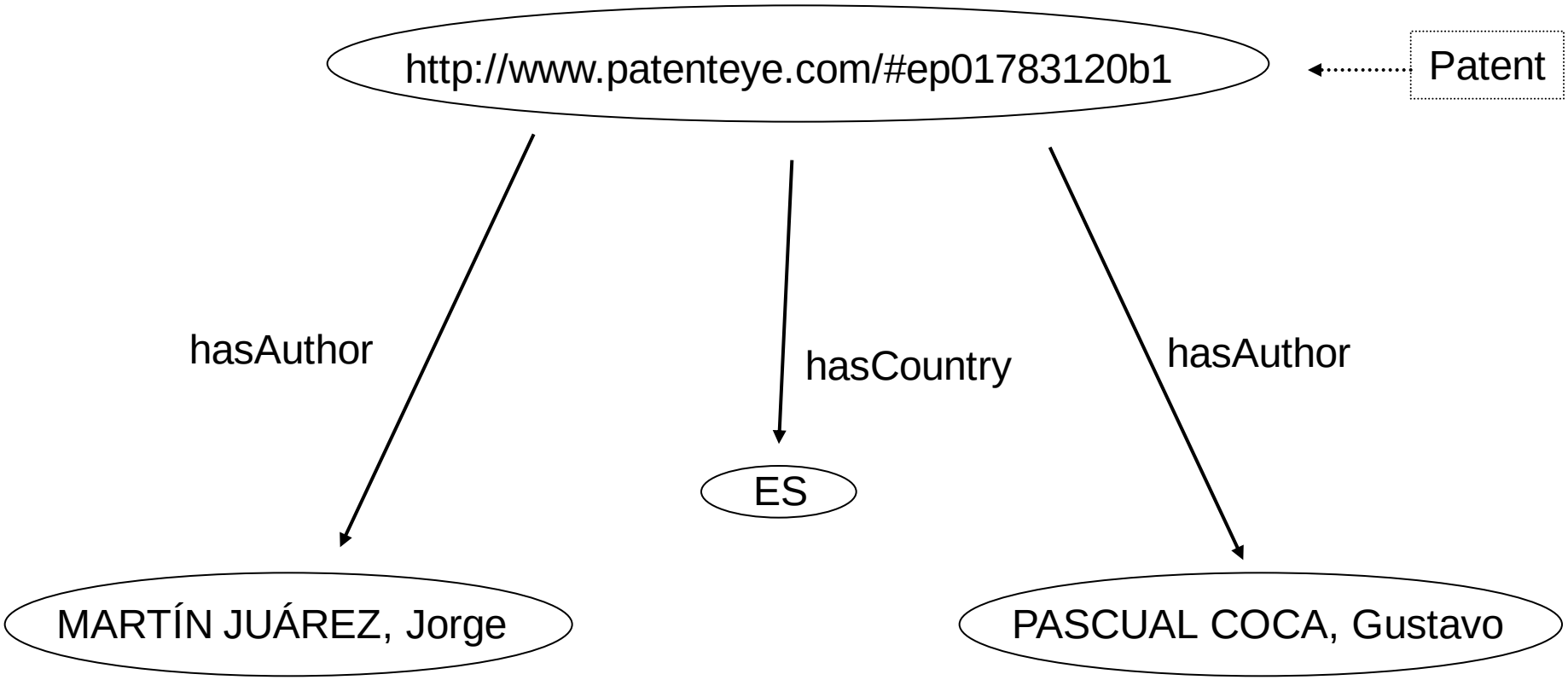
Checking the Product



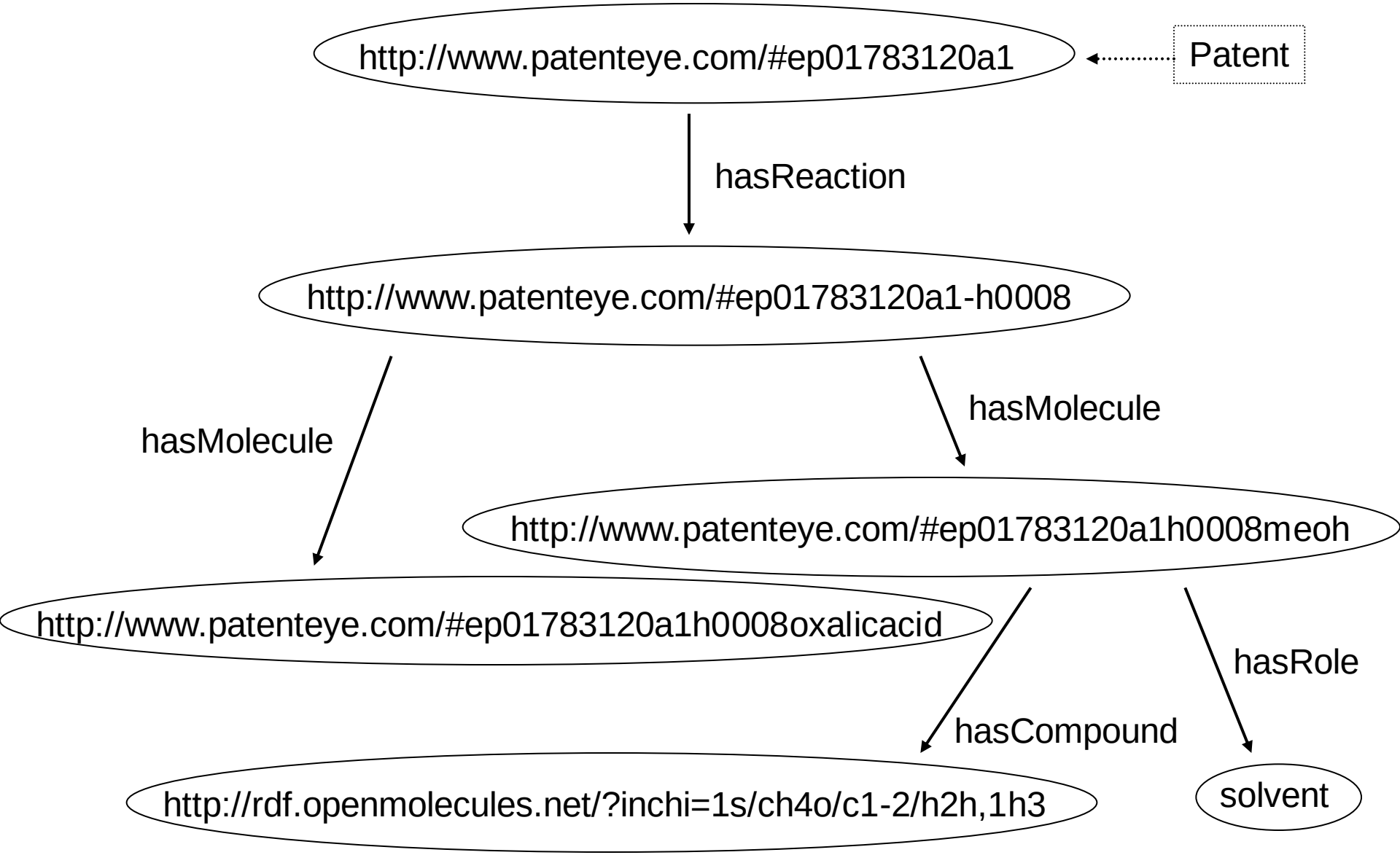

```
<product>
  <molecule title='foo'
      matchesNmr='?'
      matchesMassSpec='?'
      matchesImage='?'
    <atomArray>…</atomArray>
    <bondArray>…</bondArray>
  </molecule>
</product>
```

- ^{1}H NMR – expected proton count
- Mass spectrum – expected mass
- OSRA – expected connection table

Conversion to RDF



Conversion to RDF



Conclusions

- Chemical reactions and data are automatically abstracted from the literature
- Data is semantically encoded to be machine-readable & reusable

- Technology continues to be under development

Acknowledgements

- Prof. Robert Glen
- Prof. Peter Murray-Rust
- Dr Lezan Hawizy

- Unilever

UNIVERSITY OF CAMBRIDGE

Unilever Cambridge
Centre for Molecular Informatics

Any Questions?