Mining for Chemistry in Text and Images. A Real-World Example revealing the Challenge, Scope, Limitation and Usability of the current Technology

V. Eigner-Pitto, J. Eiblmaier, U. Frieske, L. Isenko, H. Kraut, H. Saller and P. Loew.
InfoChem GmbH, Landsberger Strasse 408, Munich, 81241, Germany

Paracetamol Connection Table

Paracetamol

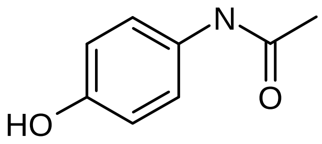
Acetaminophen

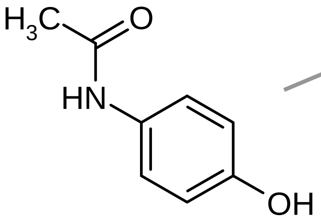
Panadol

Paralief

N-acetyl-p-aminophenol

....





CAS Nr.: 103-90-2

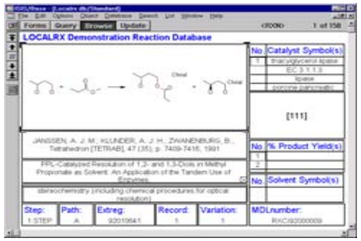
```
11 11  0  0  0  0  0  0  0  0999 V2000
   1.4944   -0.4750    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   1.4932   -1.3023    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   2.2081   -1.7152    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   2.9245   -1.3019    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   2.9216   -0.4713    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   2.2063   -0.0622    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   0.7784   -1.7143    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
   3.6346   -0.0561    0.0000 N   0  0  0  0  0  0  0  0  0  0  0  0
   4.3458   -0.4667    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   4.3459   -1.2917    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
   5.0602   -0.0541    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
  5  6  2  0  0  0  0
  6  1  1  0  0  0  0
  1  2  2  0  0  0  0
  2  7  1  0  0  0  0
  3  4  2  0  0  0  0
  5  8  1  0  0  0  0
  8  9  1  0  0  0  0
  4  5  1  0  0  0  0
  9 10  2  0  0  0  0
  2  3  1  0  0  0  0
  9 11  1  0  0  0  0
M  END
```

Conventional Database Building: Concept

Print Version

Manual Abstraction →

Connection Tables


File Edit Options Object Database Search List Window Help
Forms Query Browse Update
1 of 158
Chiral
Chiral

LOCALRX Demonstration Reaction Database

No.	Catalyst Symbol(s)
1	triacylglycerol lipase
	EC 3.1.1.3
	lipase
	porcine pancreatic

[111]

JANSSEN, A. J. M.; KLUNDER, A. J. H.; ZWANENBURG, B., Tetrahedron [TETRAB], 47 (35), p. 7409-7416, 1991

PPL-Catalyzed Resolution of 1,2- and 1,3-Diols in Methyl Propionate as Solvent. An Application of the Tandem Use of Enzymes

stereochemistry (including chemical procedures for optical resolution)

No.	% Product Yield(s)
1	
2	

No.	Solvent Symbol(s)

Step:	Path:	Extreg:	Record:	Variation:	MDLnumber:
1 STEP	A	92010641	1	1	RXCI92000009

Structure / Reaction Database Building

- Structures/RXN
- Factual Data
- Catalysts
- Solvents

Advantages

✓ High quality

✓ Currently no alternative

Disadvantages

✗ Time-consuming, Time-gap

✗ Not-comprehensive patent extraction

✗ Legacy data extraction not affordable

✗ High costs:

- 5-7 USD per reaction (depending on source)
- 2,5-4 USD per structure (depending on source)

Mining for Chemistry in Text and Images

- Chemical Named Entity Recognition (NER):
  - Automatic identification of chemical names in text documents
  - Generation of connection table
- Image Recognition (IR):
  - Images of chemical structures in text are not computer searchable
  - IR enables generation of computer readable structures
- Mining for Chemistry goal:
  - Automatic generation of structure and reaction databases

Activities in the Area of Chemical Mining

- Chemical Named Entity Recognition:
    - Oscar (Corbett, Murray-Rust, Teufel et al.)
    - Chemical Entity Relationships Skill Cartridge for Luxid (Temis)
    - SureChem (former ReelTwo)
    - ChemBrowser (IBM / InfoChem)
    - IC*ANNOTATOR* (InfoChem)
    - …
- Image Recognition:
    - chemoCR (Fraunhofer SCAI / InfoChem)
    - CLiDE Pro (Keymodule Ltd.)
    - OSRA (NIH)
    - …
- Pilot project InfoChem / SCAI:
    - Status of some of the technologies used in these areas

Our Real-World Example

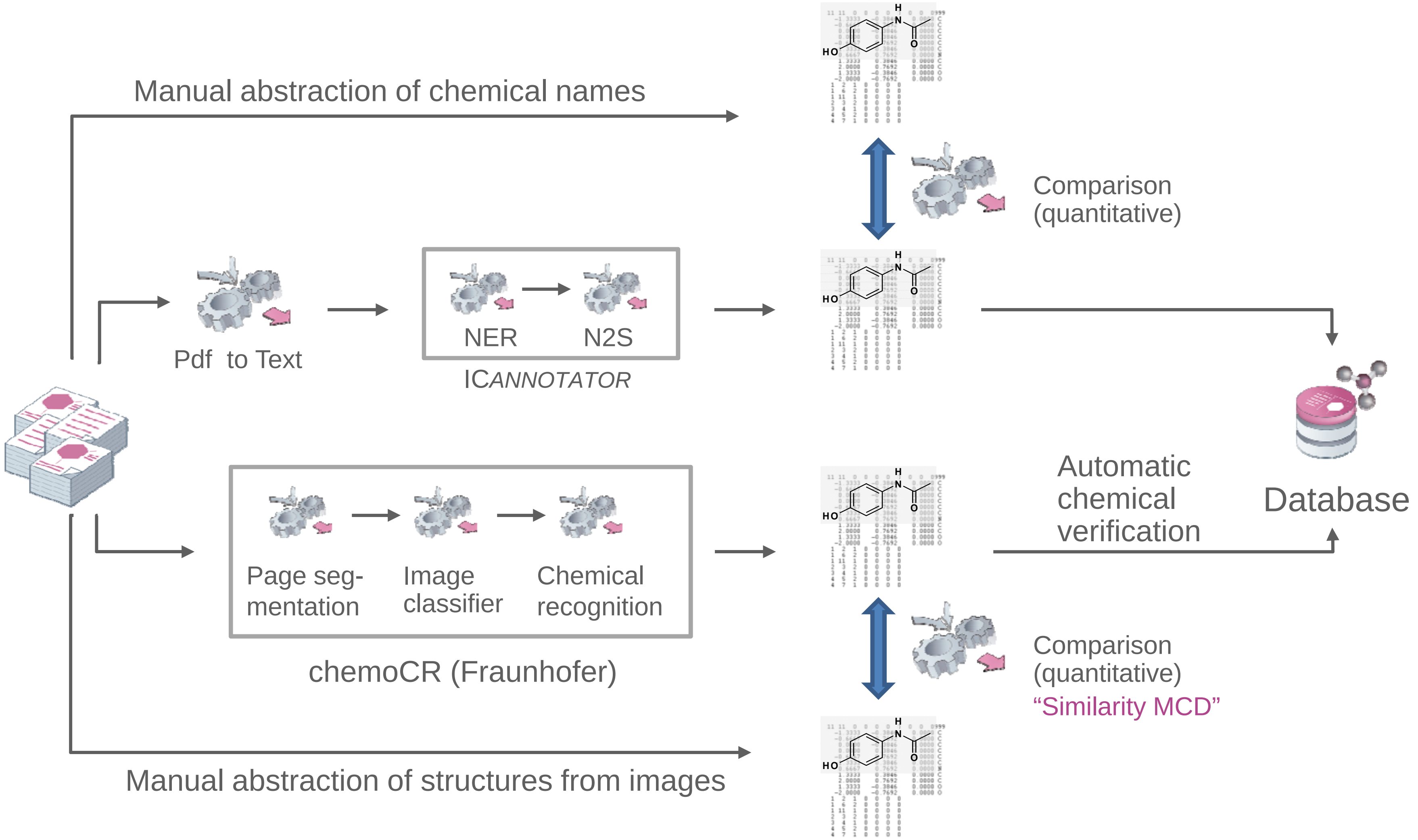
Pilot project with highly heterogeneous source set:

- Sample set: 15 titles (journals, book series)
  - contains chemical compounds in text and images
  - covers different areas of chemistry
  - diverse source formats (scanned, text, pdfs, etc.)

- Aim of project: Proof of concept on sample set

  Assessment of results and quality

Overview of Approach and Applied Technology



Manual abstraction of chemical names

Comparison (quantitative)

Pdf to Text

NER → N2S

ICANNOTATOR

Page seg-mentation → Image classifier → Chemical recognition

chemoCR (Fraunhofer)

Automatic chemical verification

Database

Comparison (quantitative)

“Similarity MCD”

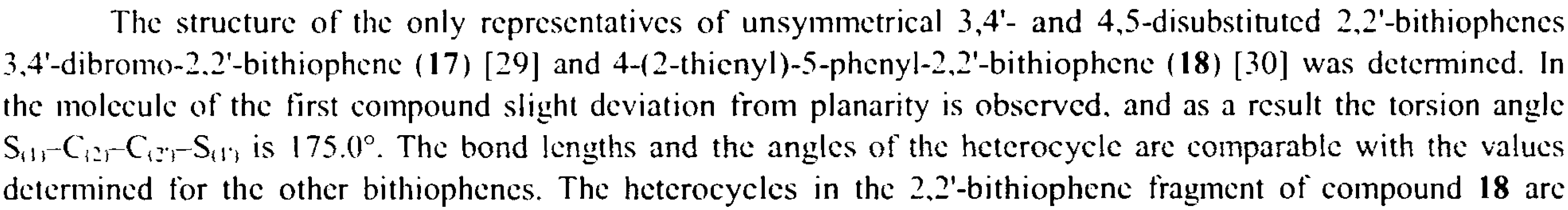
Manual abstraction of structures from images

Selection of Sample Set

- Sources: 15 titles covering time period 2000-2008
  - 10 journals
  - 5 book series
- Definition of random subset:
  - 10 articles of each title equally distributed over the time period
  - 3 pages equally distributed over every article are processed
- Total: 450 pages containing text and images

Challenges NER: Scanned Documents

- Visible and extracted text not identical



The structure of the only representatives of unsymmetrical 3,4'- and 4,5-disubstituted 2,2'-bithiophenes 3,4'-dibromo-2,2'-bithiophene (**17**) [29] and 4-(2-thienyl)-5-phenyl-2,2'-bithiophene (**18**) [30] was determined. In the molecule of the first compound slight deviation from planarity is observed, and as a result the torsion angle $S_{(1)}$–$C_{(2)}$–$C_{(2')}$–$S_{(1')}$ is 175.0°. The bond lengths and the angles of the heterocycle are comparable with the values determined for the other bithiophenes. The heterocycles in the 2,2'-bithiophene fragment of compound **18** are

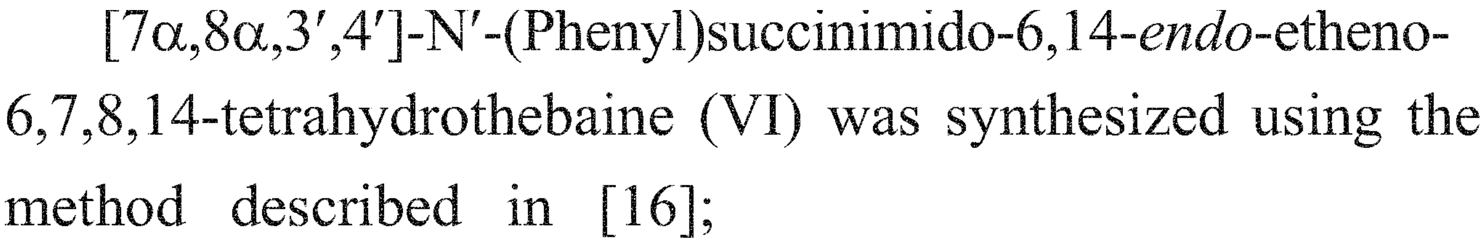
OCR results:

Tile structure of the only reprcsentativcs o f unsyrnrnctrical 3,4'- and 4,5-disubstitutcd 2,2'-bithiophcncs 3,4'-dibromo-2.2'-bithiophcne (17) [29] and 4-(2-thienyl)-5-phcnyl-2,2'-bithiophene (18) [30] was determined. In the moleculc of the first compound slight deviation from planarity is observed, and as a result the torsion angle S,~,-C~:~-Cn.~-S~n is 175.(I~ The bond lengths and the angles of the heterocycle arc comparable with the values determined lbr thc other bithiophenes. The heterocyclcs in thc 2,2'-bithiophene fragmcnt of compound 18 are

Challenges NER: „Native“ pdf

- Visible and extracted text not identical

1. Greek letters and apostrophes:

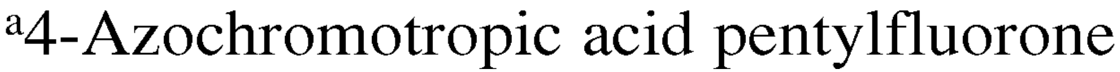


[7α,8α,3′,4′]-N′-(Phenyl)succinimido-6,14-*endo*-etheno-6,7,8,14-tetrahydrothebaine (VI) was synthesized using the method described in [16];

PdfToText results:

[7**a**,8**a**,3**¢**,4**¢**]-N**¢**-(Phenyl)succinimido-6,14-endo-etheno….

2. Notes, indexes:



[a]4-Azochromotropic acid pentylfluorone

PdfToText results:

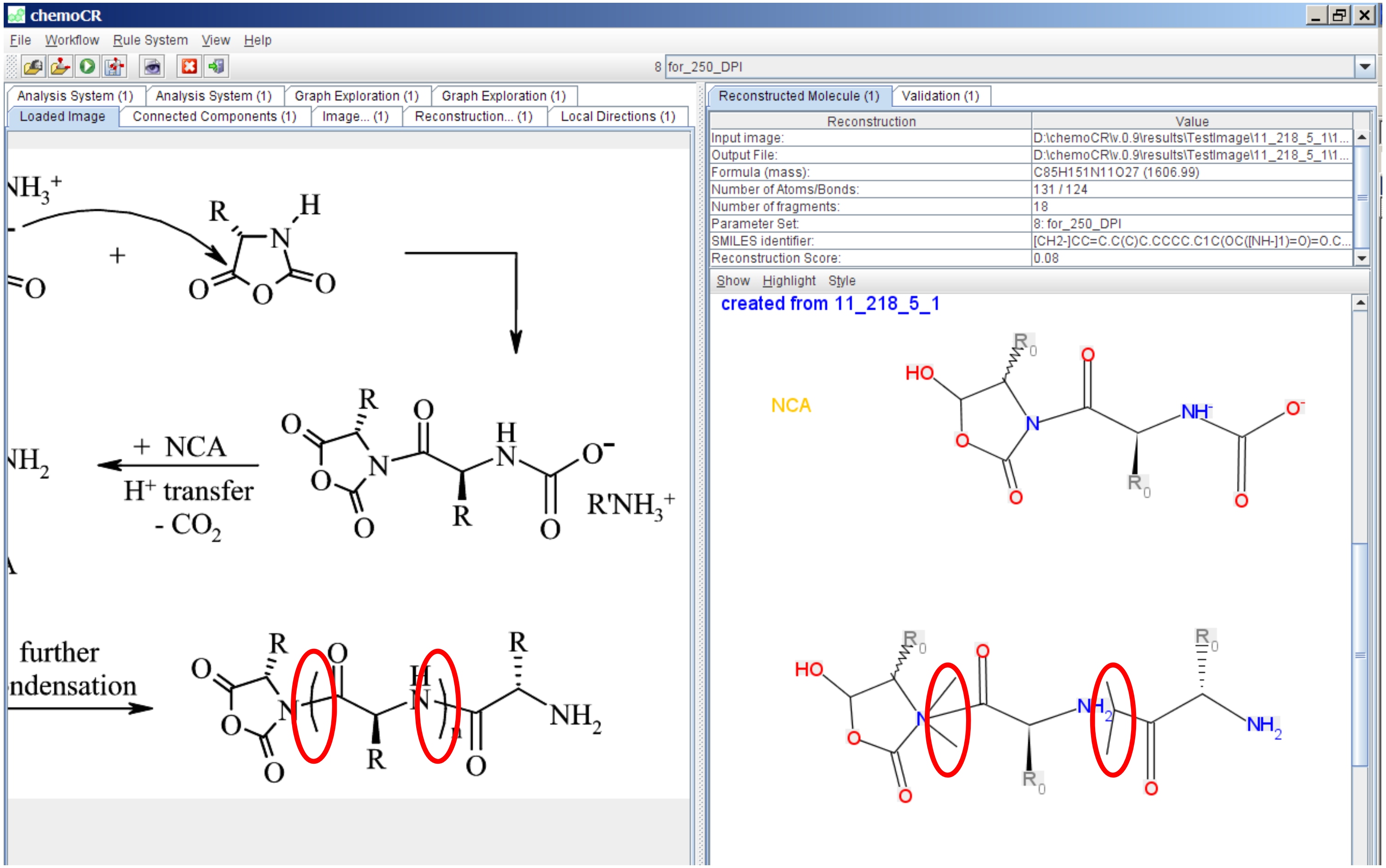
a4-Azochromotropic acid pentylfluorone

More Challenges NER

- Misspelled names (typing errors in original document)
- Errors introduced by N2S conversion tools
- Incomplete and erroneous dictionaries
- …..

Challenges Image Recognition (1)

- Brackets are converted into bonds:

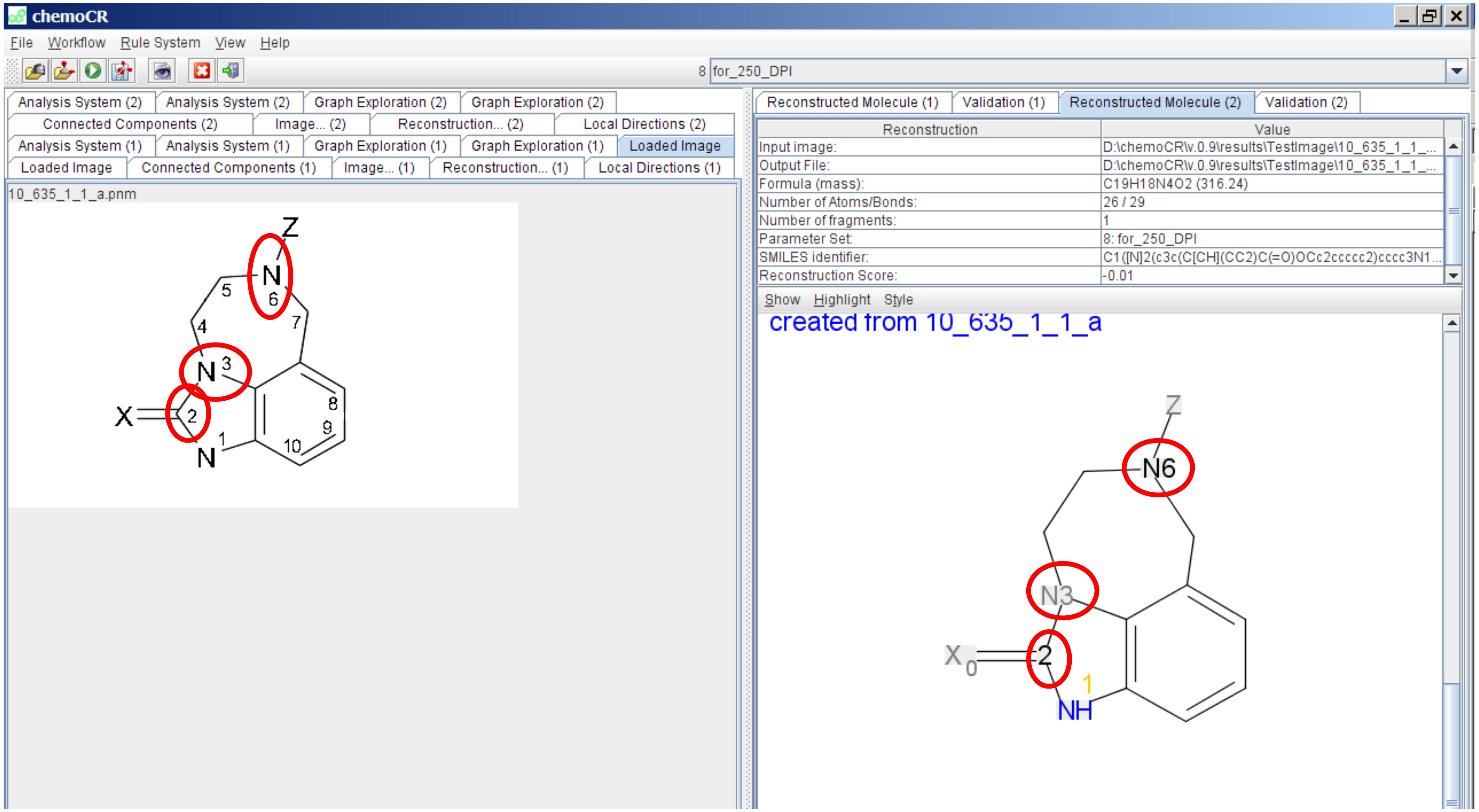


Reconstruction	Value
Input image:	D:\chemoCR\v.0.9\results\TestImage\11_218_5_1\1...
Output File:	D:\chemoCR\v.0.9\results\TestImage\11_218_5_1\1...
Formula (mass):	C85H151N11O27 (1606.99)
Number of Atoms/Bonds:	131 / 124
Number of fragments:	18
Parameter Set:	8: for_250_DPI
SMILES identifier:	[CH2-]CC=C.C(C)C.CCCC.C1C(OC([NH-]1)=O)=O.C...
Reconstruction Score:	0.08

created from 11_218_5_1

Challenges Image Recognition (2)

- Atom numbers in ring are recognized as part of the structure:

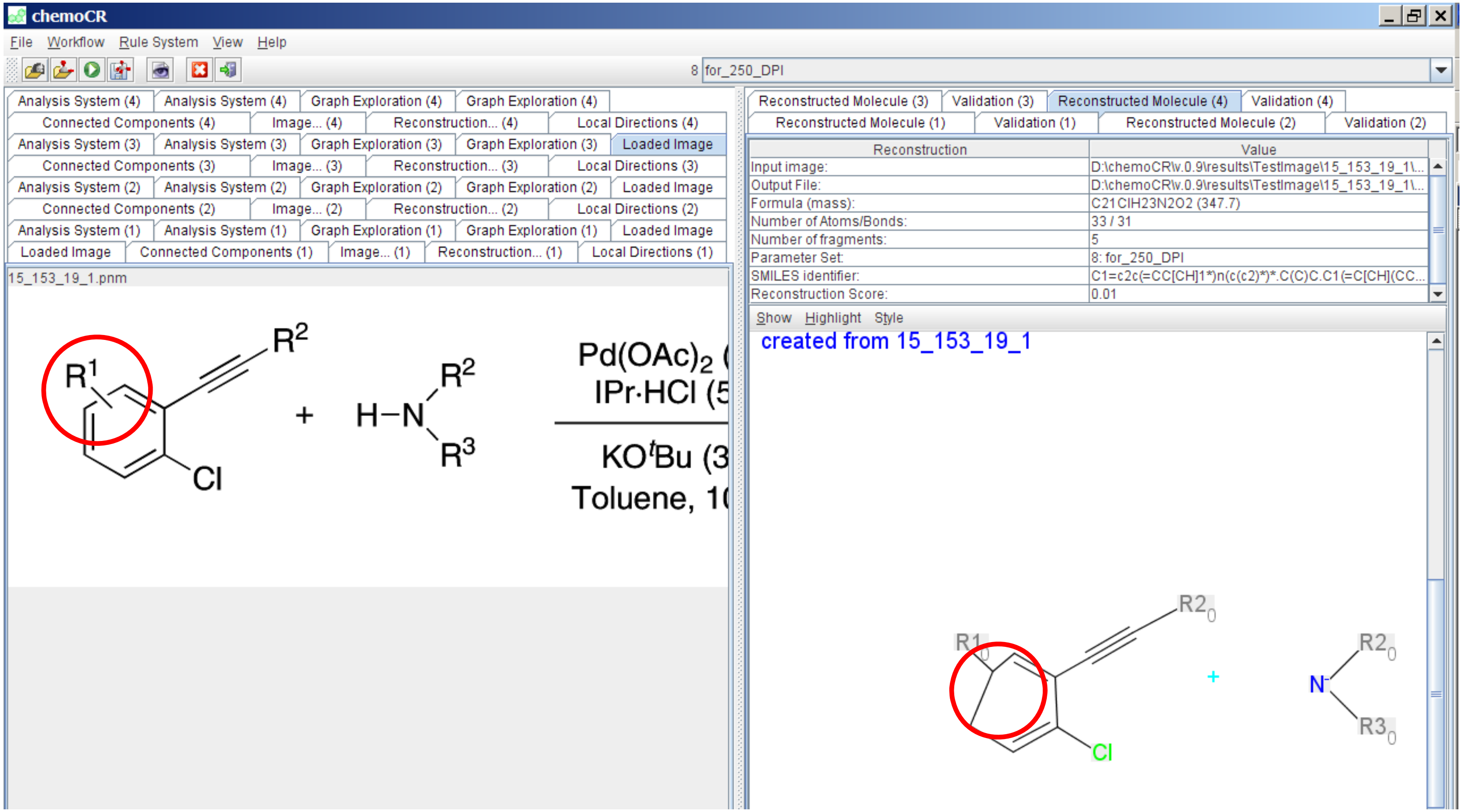

chemoCR
File Workflow Rule System View Help
8 for_250_DPI
Analysis System (2) Analysis System (2) Graph Exploration (2) Graph Exploration (2)
Connected Components (2) Image... (2) Reconstruction... (2) Local Directions (2)
Analysis System (1) Analysis System (1) Graph Exploration (1) Graph Exploration (1) Loaded Image
Loaded Image Connected Components (1) Image... (1) Reconstruction... (1) Local Directions (1)
10_635_1_1_a.pnm
Z
N
5
6
4
7
N
3
X
2
8
9
1
10
N
Reconstructed Molecule (1) Validation (1) Reconstructed Molecule (2) Validation (2)
Show Highlight Style
Z
N6
N3
X
0
2
1
NH

Reconstruction	Value
Input image:	D:\chemoCR\v.0.9\results\TestImage\10_635_1_1_...
Output File:	D:\chemoCR\v.0.9\results\TestImage\10_635_1_1_...
Formula (mass):	C19H18N4O2 (316.24)
Number of Atoms/Bonds:	26 / 29
Number of fragments:	1
Parameter Set:	8: for_250_DPI
SMILES identifier:	C1([N]2(c3c(C[CH](CC2)C(=O)OCc2ccccc2)cccc3N1...
Reconstruction Score:	-0.01

created from 10_635_1_1_a

Challenges Image Recognition (3)

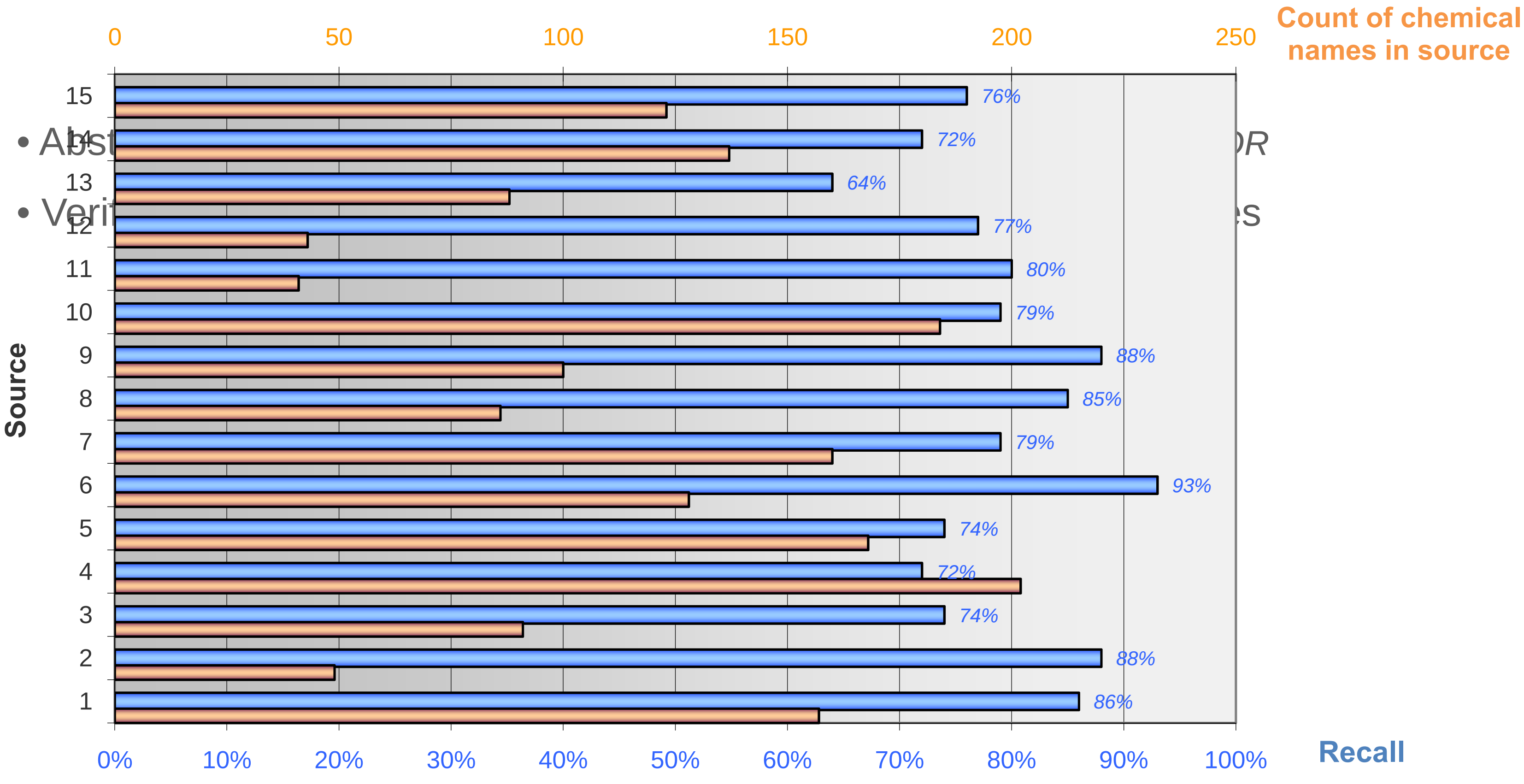
- Variable point of attachment bonds are not recognized:


chemoCR
File Workflow Rule System View Help
8 for_250_DPI
Analysis System (4) Analysis System (4) Graph Exploration (4) Graph Exploration (4)
Connected Components (4) Image... (4) Reconstruction... (4) Local Directions (4)
Analysis System (3) Analysis System (3) Graph Exploration (3) Graph Exploration (3) Loaded Image
Connected Components (3) Image... (3) Reconstruction... (3) Local Directions (3)
Analysis System (2) Analysis System (2) Graph Exploration (2) Graph Exploration (2) Loaded Image
Connected Components (2) Image... (2) Reconstruction... (2) Local Directions (2)
Analysis System (1) Analysis System (1) Graph Exploration (1) Graph Exploration (1) Loaded Image
Loaded Image Connected Components (1) Image... (1) Reconstruction... (1) Local Directions (1)
15_153_19_1.pnm
R1
R2
Cl
+
H−N
R2
R3
Pd(OAc)2 (
IPr·HCl (5
KOtBu (3
Toluene, 10
Reconstructed Molecule (3) Validation (3) Reconstructed Molecule (4) Validation (4)
Reconstructed Molecule (1) Validation (1) Reconstructed Molecule (2) Validation (2)
Show Highlight Style
R1
R2
Cl
+
N
R2
R3

Reconstruction	Value
Input image:	D:\chemoCR\v.0.9\results\TestImage\15_153_19_1\...
Output File:	D:\chemoCR\v.0.9\results\TestImage\15_153_19_1\...
Formula (mass):	C21ClH23N2O2 (347.7)
Number of Atoms/Bonds:	33 / 31
Number of fragments:	5
Parameter Set:	8: for_250_DPI
SMILES identifier:	C1=c2c(=CC[CH]1*)n(c(c2)*)*.C(C)C.C1(=C[CH](CC...
Reconstruction Score:	0.01

created from 15_153_19_1

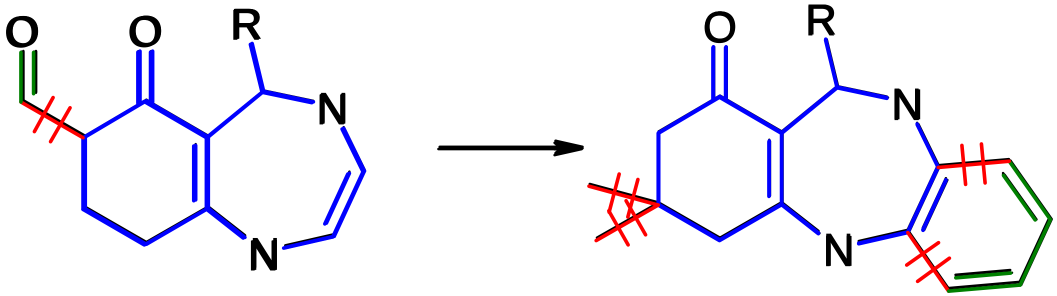
Verification and Results NER



Recall	Precision	F-score
79%	69%	73%

Verification Image Recognition: SimilarityMCD

- Rating method used to compare manually and automatically abstracted chemical structures
- Basic idea is to rate the chemical difference of the two compared molecules



chemoCR result

Original image

MCD = breaking a minimum number of bonds in the reactant and making a minimum number of bonds in the product

- The identical parts and different parts of the two compared structures are rated in a percent value from 0% to 100% based on the Largest Common Subgraph (LCS) and the Minimum Chemical Distance (MCD)

Verification Image Recognition: Rating

- All bonds of the two molecules get different ratings for single, double and triple bonds, for C-C and C-hetero atoms bonds, depending on bond stability

$$\text{SimilarityMCD(Mol1, Mol2)} = \frac{\text{Mapped Value}}{\text{Total Value}} \times 100\ [\ \%\ ]$$

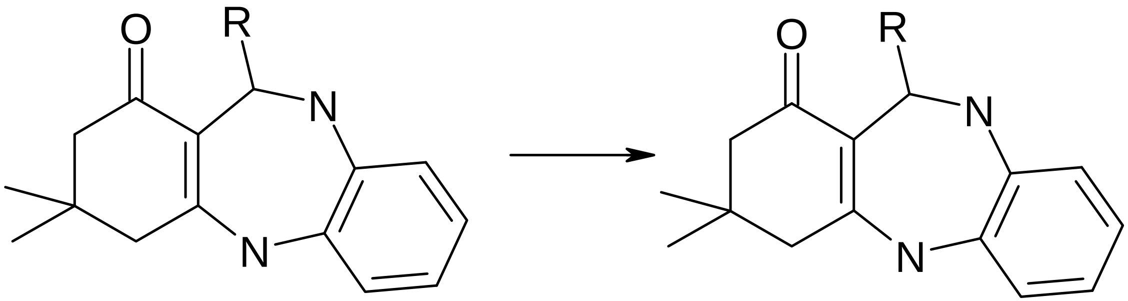
"Mapped Value"

= Σ of ratings for mapped bonds and atoms

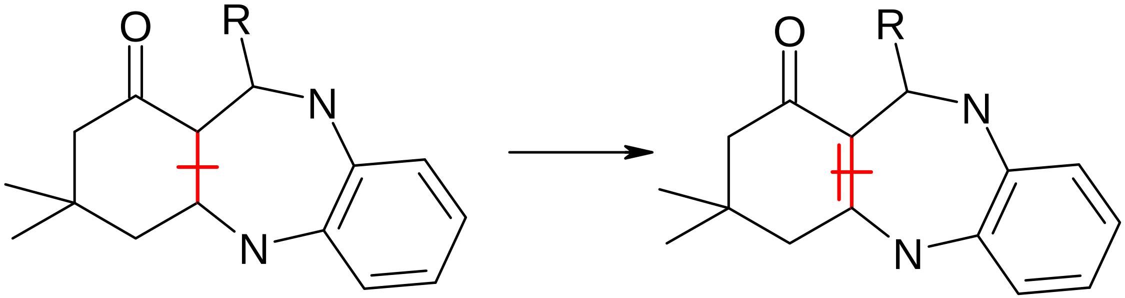
"Total Value" (per definition 100% rating)

= Σ of all bond ratings of the two compared molecules

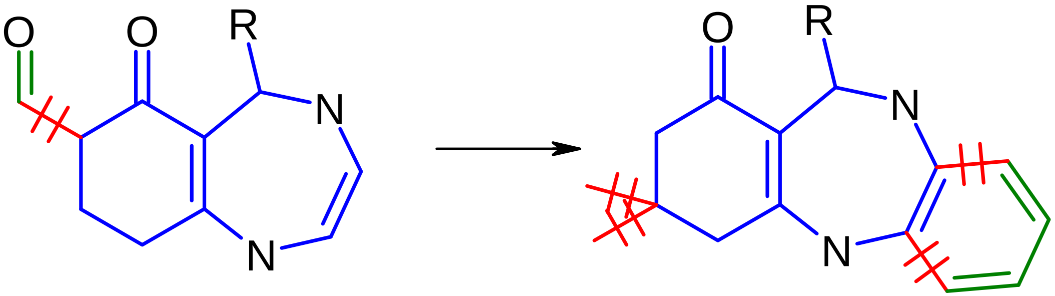
Verification Image Recognition: Examples



100%

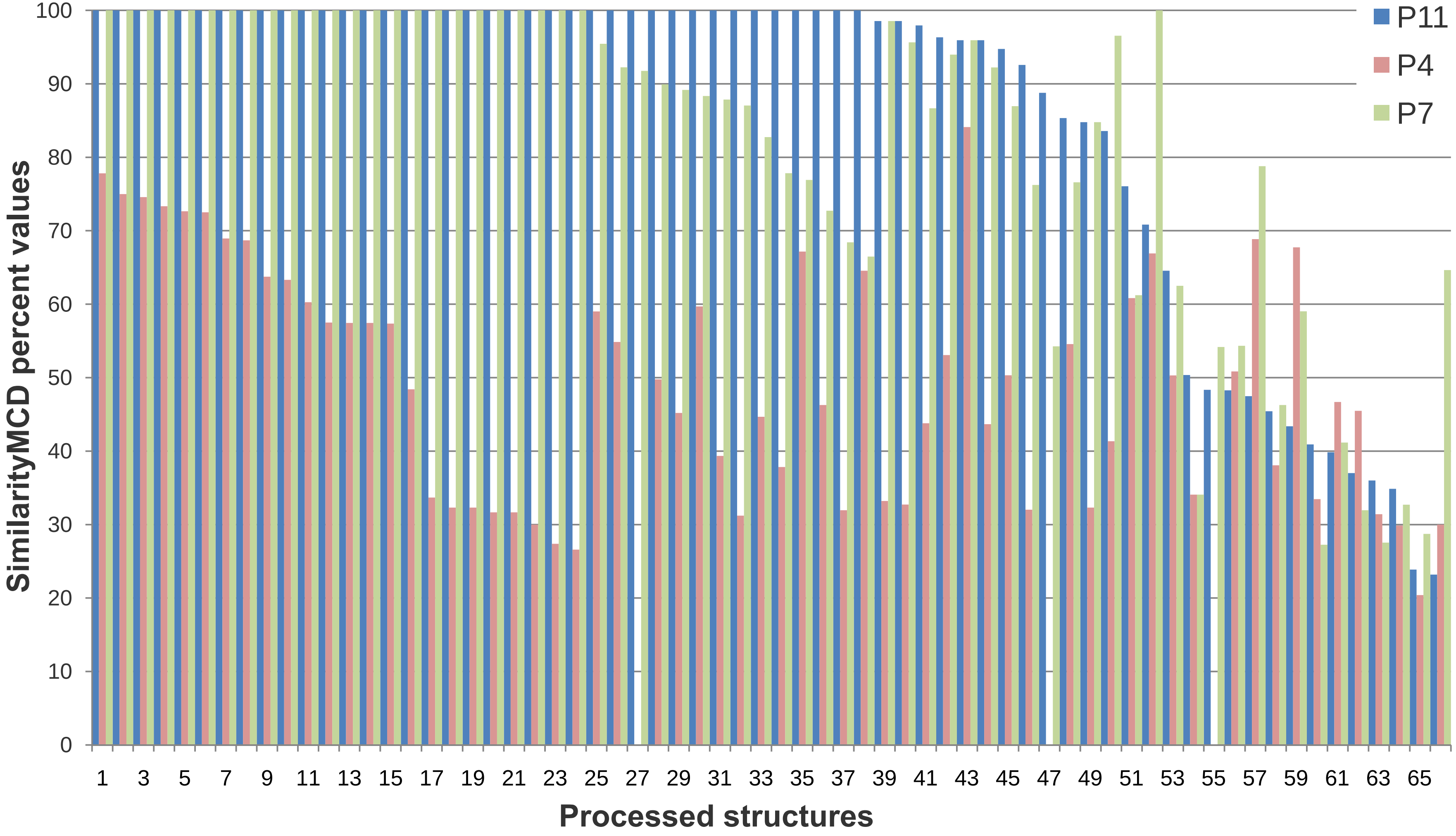


97.73%

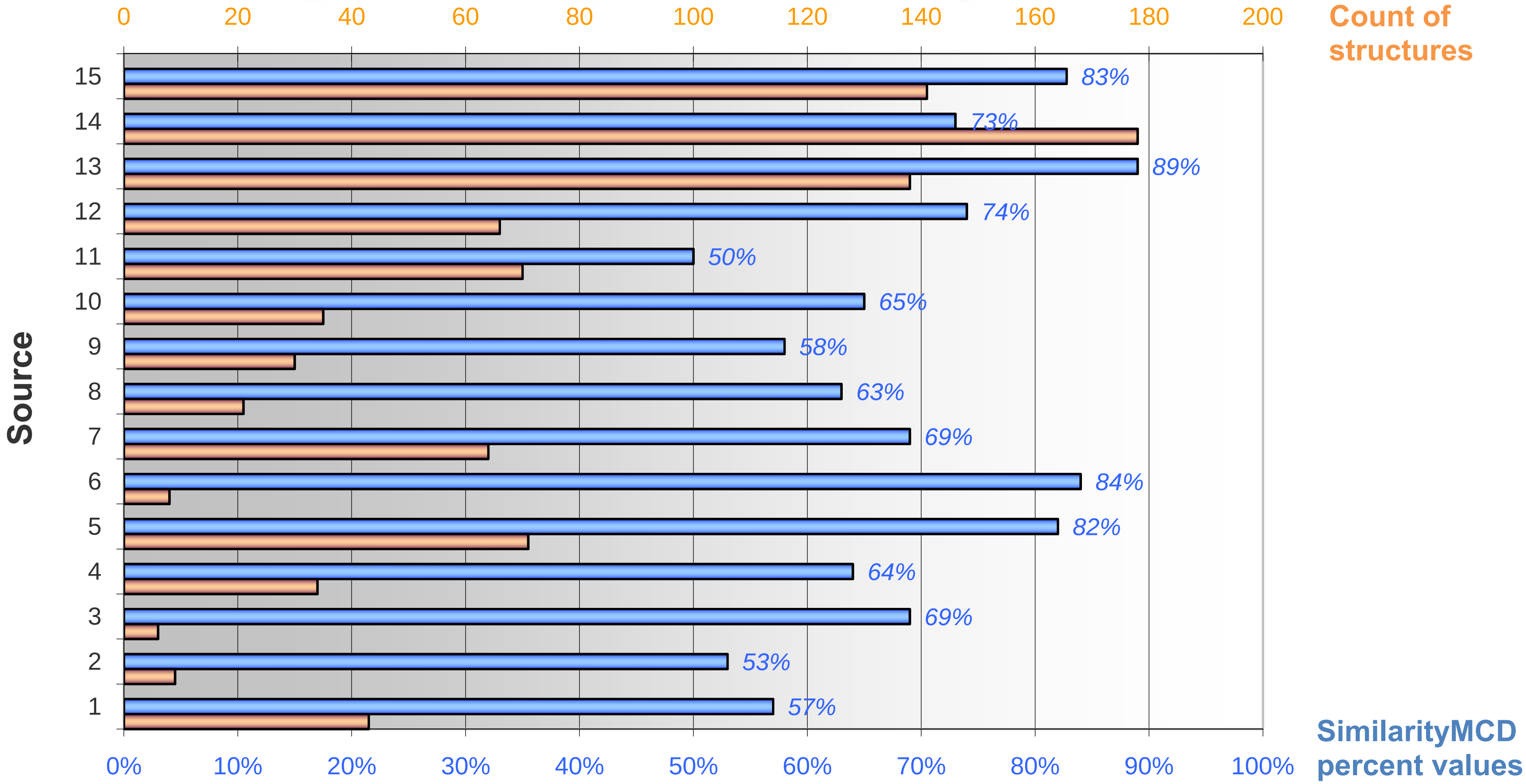


57.97%

Results Image Recognition (Source No. 5)



Results Image Recognition (P11)



Overall SimilarityMCD	All Sources 100% Match	Best Source 100% Match
72%	33%	58%

Conclusions

- Using a highly heterogeneous sample set from a real world example we have:
  - illustrated the challenges of chemical NER and image recognition
  - demonstrated with a quantitative comparison the quality of IC*ANNOTATOR*
  - explained how the SimilarityMCD algorithm enables quantitative comparison and can be used to optimize chemoCR parameter sets
- We have achieved up to 93% recall for text and up to 58% exactly recognized structures
- All components are in the process of further development

Acknowledgements

- Prof. Dr. Martin Hofmann-Apitius, Fraunhofer SCAI
- Dr. Marc Zimmermann and his Team, Fraunhofer SCAI

InfoChem GmbH:

www.infochem.de
www.spresi.com
info@infochem.de