Information Extraction from Chemical Images

3rd Text Mining Symposium in Life Sciences

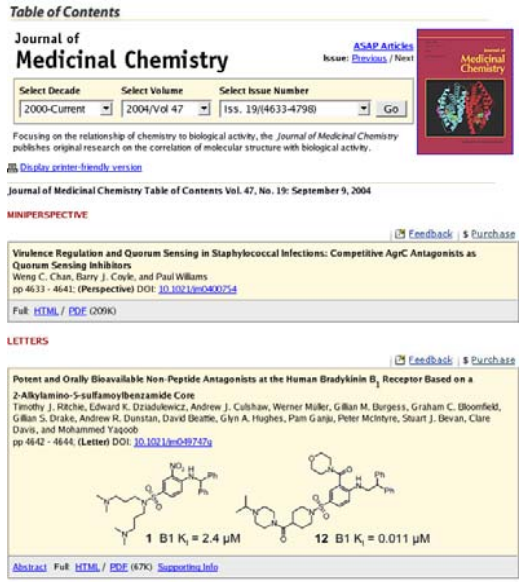
October 13, 2005

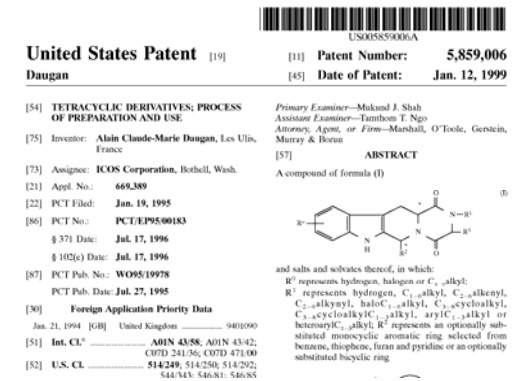
Fraunhofer Institut Algorithmen und Wissenschaftliches Rechnen

Dr. Marc Zimmermann

Available Chemical Information

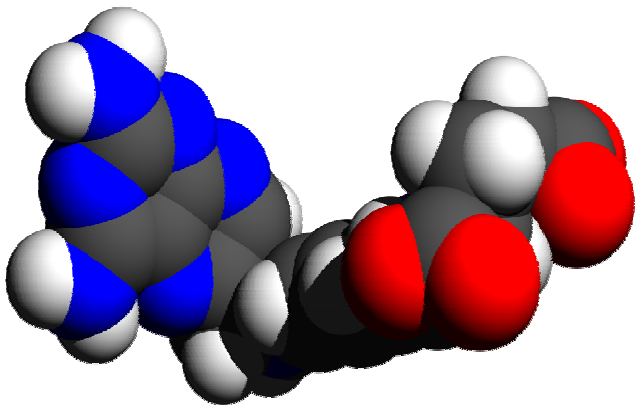
- Textbooks
- Reports
- Patents
- Databases
- Scientific journals and publications
- Websites





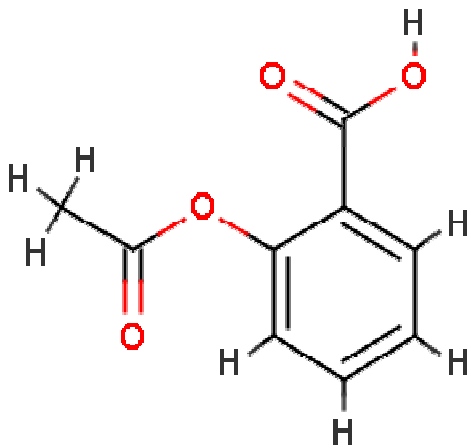
Representations of Chemical Compounds

- Name (trivial, trade, brand, INN, USAN)
- Registration numbers (CAS, NCI, Beilstein)
- Formal description (sum formula, SMILES)
- Chemical nomenclature (IUPAC, CAS, InChI)
- Depictions



Example: Aspirin

- **Name**: Acetylsalicylic acid, Aspirin, Bayer, Colfarit, Dolean PH 8, Duramax, Ecotrin, …
- **CAS**: 50-78-2, **SID**: 35870,
- **Formula**: C9H8O4
- **IUPAC** Name: 2-acetoxybenzoic acid
- **SMILES**: CC(=O)OC1=CC=CC=C1C(=O)O
- **InChI**: 1.12Beta/C9H8O4/c1-6(10)13-8-5-3-2-4-7(8)9(11)12/h1H3,2-5H,(H,11,12)
- **Depiction**:



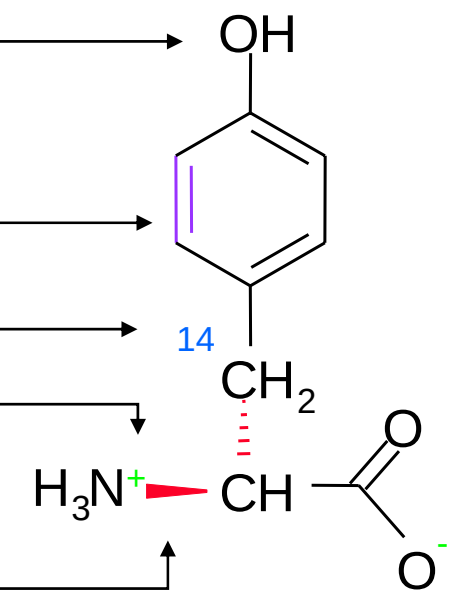
Information Extraction Methods

- Names ⇨ Dictionary based
- Registration numbers ⇨ Databases
- Formal descriptions ⇨ Rule based
- Depictions ⇨ *chemical OCR*

Representing a Chemical Compound

How much information do you want to include?

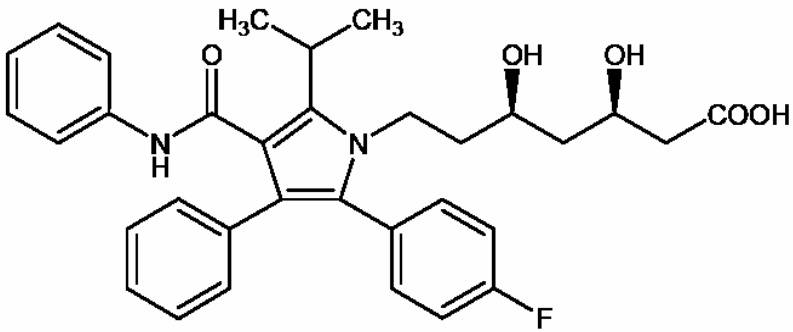
- Atoms present
- Connections between atoms
  - bond types
- Isotopes
- Charges
- Stereochemical configuration



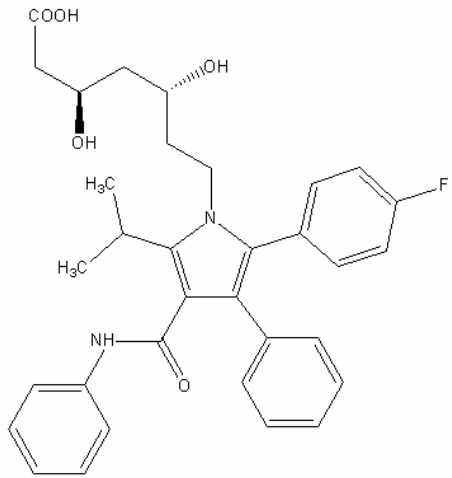
Modeling of Chemicals as Graphs

Why use graph theory?

- Established mathematical field
- Graphs can be easily represented in computers
- Existing algorithms for comparison, searching, etc.
- Unlike humans, computers aren't very good at pattern recognition



Similar or Same?



Computer Representation

A typical example: MDL MOL file (SDF)

```
L-Alanine (13C)
GSMACCS-II10169115362D 1   0.00366     0.00000     0

  6  5  0  0  1  0              3 V2000
   -0.6622    0.5342    0.0000 C   0  0  2  0  0  0
    0.6622   -0.3000    0.0000 C   0  0  0  0  0  0
   -0.7207    2.0817    0.0000 C   1  0  0  0  0  0
   -1.8622   -0.3695    0.0000 N   0  3  0  0  0  0
    0.6220   -1.8037    0.0000 O   0  0  0  0  0  0
    1.9464    0.4244    0.0000 O   0  5  0  0  0  0
  1  2  1  0  0  0
  1  3  1  1  0  0
  1  4  1  0  0  0
  2  5  2  0  0  0
  2  6  1  0  0  0
M  CHG  2   4   1   6  -1
M  ISO  1   3  13
M  END
```

Header block

Counts line

Atom block

Connection table (Ctab)

Bond block

Blocks not used in this Ctab: List block, Stext block

Properties block

For more information on MDL formats, see http://www.mdl.com/downloads/public/ctfile/ctfile.jsp

Disadvantages of Using Graphs

- Many graph algorithms are inherently slow
- Analogy between chemical structures and graphs is not perfect
- Realities of chemical structures cause problems
  - aromaticity
  - stereochemistry
  - tautomerism
  - inorganic compounds
  - macromolecules and polymers
  - incompletely-defined substances

Good News

There is only a limited number of chemical drawing tools

(these are using templates):

- ChemDraw (CambridgeSoft)
- ChemSketch (ACD)
- ISISdraw (MDL)
- JAVA applets (ChemAxon)
- ...

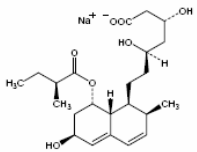
⇨ Reduced complexity

chemOCR: Reconstruction of Chemical Compounds

1 Document

Pravastatin Sodium

Search Monographs



The Merck Index

THE MERCK INDEX® is a trademark of Merck & Company Incorporated, Whitehouse Station, New Jersey, USA and is registered in the United States Patent and Trademark Office.

Published originally as The Merck Index, Thirteenth Edition.

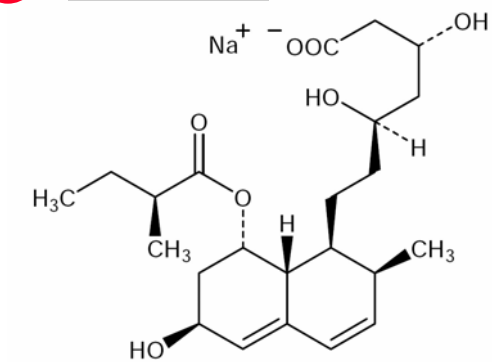
MERCK CambridgeSoft

Monograph Number: 0007800
Title: Pravastatin Sodium
CAS Registry Number: 81131-70-6
CAS Name: (βR,δR,1S,2S,6S,8S,8aR)-1,2,6,7,8,8a-Hexahydro-β,δ,6-trihydroxy-2-methyl-8-[(2S)-2-methyl-1-oxobutoxy]-1-naphthaleneheptanoic acid monosodium salt
Additional Names: sodium (+)-(3R,5R)-3,5-dihydroxy-7-[(1S,2S,6S,8S,8aR)-6-hydroxy-2-methyl-8-[(S)-2-methylbutyryloxy]-1,2,6,7,8,8a-hexahydro-1-naphthyl]heptanoate; eptastatin sodium; 3β-hydroxycompactin sodium salt
Manufacturers' Codes: CS-514; SQ-31000
Trademarks: Elisor (Bristol-Myers Squibb); Lipostat (Bristol-Myers Squibb); Liprevil (Schwarz); Mevalotin (Sankyo); Oilprevin (Bristol-Myers Squibb); Pravachol (Bristol-Myers Squibb); Pravaselect (Menarini); Pravasin (Bristol-Myers Squibb); Selectin (Bristol-Myers Squibb); Selipran (Bristol-Myers Squibb); Vasten (Specia)
Molecular Formula: $C_{23}H_{35}NaO_7$
Molecular Weight: 446.51.
Percent Composition: C 61.87%, H 7.90%, Na 5.15%, O 25.08%
Literature References: HMG-CoA reductase inhibitor; bioactive metabolite of mevastatin, *q.v.* Prepn by microbial hydroxylation: A. Terahara, M. Tanaka, DE 3122499; *eldem*, US 4346227 (1981, 1982 both to Sankyo); N. Sertzawa *et al., J. Antibiot.* **36**, 604 (1983). Structure elucidation: H. Haruyama *et al., Chem. Pharm. Bull.* **34**, 1459 (1986). Effect on serum lipid concentration: N. Nakaya *et al., Atherosclerosis* **61**, 125 (1986); on hepatic metabolism of cholesterol: E. Reihner *et al., N. Engl. J. Med.* **323**, 224 (1990). Clinical comparison with probucol, *q.v.*: G. Yoshino *et al., Lancet* **2**, 740 (1986). Clinical reduction of risk of major cardiovascular events in patients with coronary heart disease: LIPID Study Group, *N. Engl. J. Med.* **339**, 1349 (1998). Clinical effect on risk of stroke: H. D. White *et al., Ibid.* **343**, 317 (2000).
Properties: Odorless, white to off-white, fine or crystalline powder. uv max (methanol): 230, 237, 245 nm. Sol in methanol, water; slightly sol in isopropanol. Practically insol in acetone, acetonitrile, chloroform, ether.
Absorption maximum: uv max (methanol): 230, 237, 245 nm

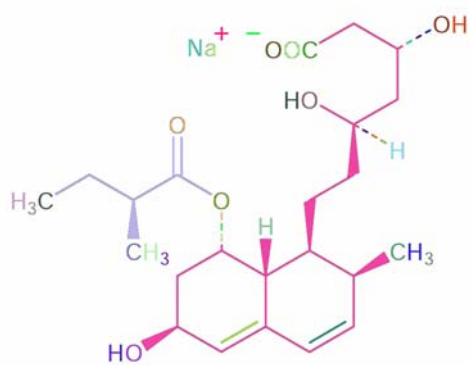
Derivative Type: Lactone
Molecular Formula: $C_{23}H_{34}O_6$
Molecular Weight: 406.51.
Percent Composition: C 67.96%, H 8.43%, O 23.61%
Properties: Colorless plate crystals, mp 138-142°. $[\alpha]_D^{22}$ +194.0° (c = 0.51 in methanol). uv max (methanol): 230, 237, 245 nm.
Melting point: mp 138-142°
Optical Rotation: $[\alpha]_D^{22}$ +194.0° (c = 0.51 in methanol)
Absorption maximum: uv max (methanol): 230, 237, 245 nm

Therap-Cat: Antilipemic.

2 Depiction



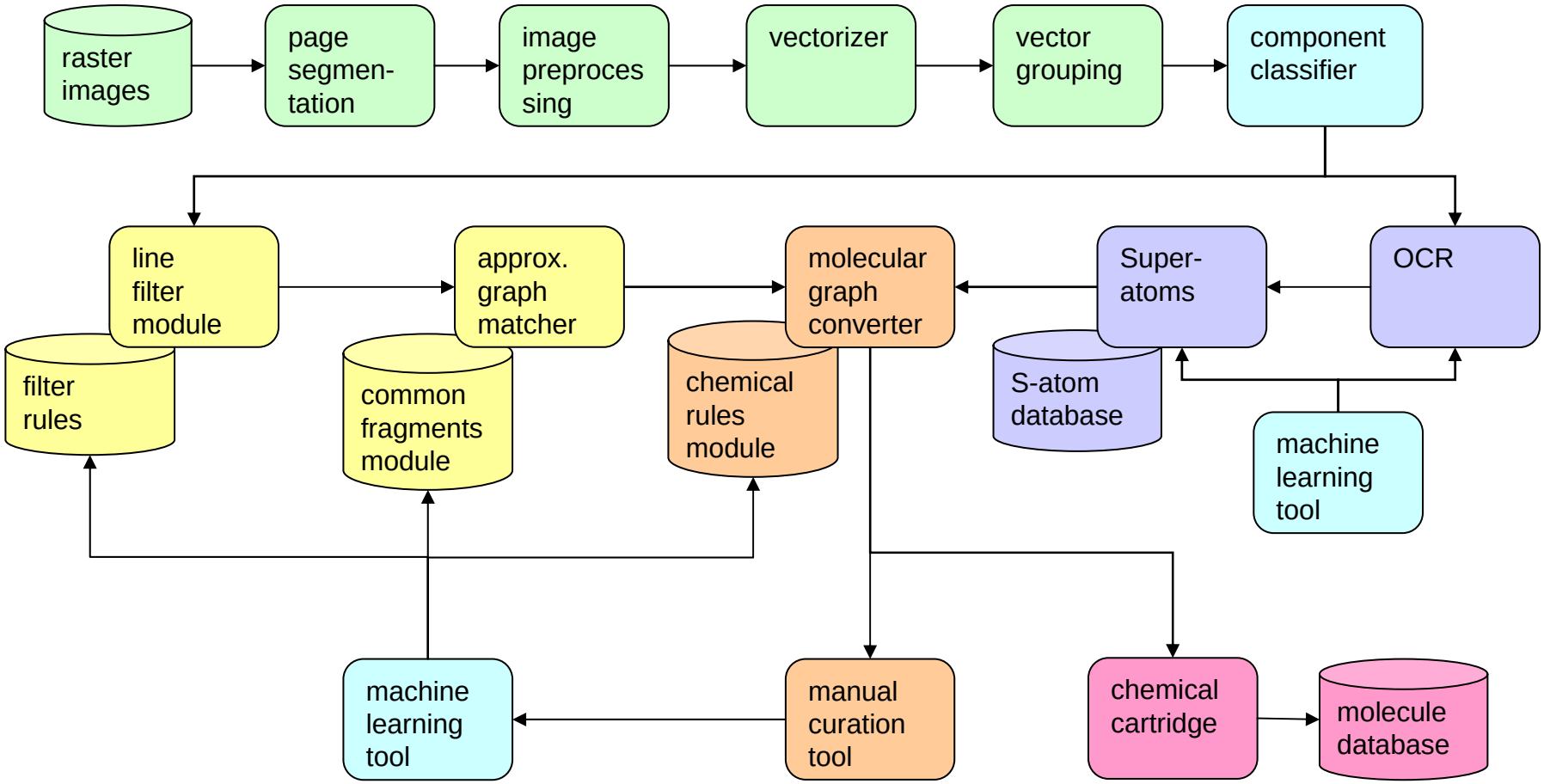
3 Reconstruction



4 SDF file

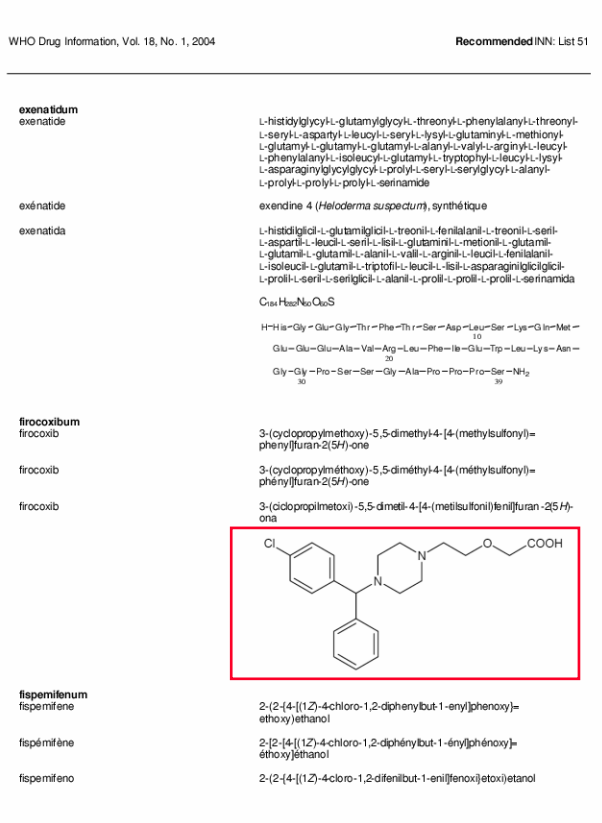
```
-ISIS- 09230315072D

27 29 0 0 0 0 0 0 0 0 0999 V2000
 -0.9348  -0.4000   0.0000 C  0 0 0 0 0 0 0 0 0 0 0 0
 -0.9359  -1.2274   0.0000 C  0 0 0 0 0 0 0 0 0 0 0 0
 -0.2211  -1.6402   0.0000 C  0 0 0 0 0 0 0 0 0 0 0 0
  0.4953  -1.2269   0.0000 C  0 0 0 0 0 0 0 0 0 0 0 0
  0.4925  -0.3964   0.0000 C  0 0 0 0 0 0 0 0 0 0 0 0
 -0.2229   0.0128   0.0000 C  0 0 0 0 0 0 0 0 0 0 0 0
  1.0750  -1.8084   0.0000 C  0 0 0 0 0 0 0 0 0 0 0 0
  1.0708  -2.6334   0.0000 C  0 0 0 0 0 0 0 0 0 0 0 0
  1.7875  -1.3917   0.0000 C  0 0 0 0 0 0 0 0 0 0 0 0
  2.5042  -1.8034   0.0000 C  0 0 0 0 0 0 0 0 0 0 0 0
  3.2162  -1.3874   0.0000 C  0 0 0 0 0 0 0 0 0 0 0 0
  3.2120  -0.5611   0.0000 C  0 0 0 0 0 0 0 0 0 0 0 0
  2.4899  -0.1526   0.0000 C  0 0 0 0 0 0 0 0 0 0 0 0
  1.7808  -0.5709   0.0000 C  0 0 0 0 0 0 0 0 0 0 0 0
  4.0042  -0.3417   0.0000 N  0 0 0 0 0 0 0 0 0 0 0 0
  4.0083  -1.5959   0.0000 N  0 0 3 0 0 0 0 0 0 0 0 0
  4.4125  -1.0542   0.0000 C  0 0 0 0 0 0 0 0 0 0 0 0
  5.2375  -1.0542   0.0000 N  0 0 0 0 0 0 0 0 0 0 0 0
  0.4792  -3.2167   0.0000 C  0 0 0 0 0 0 0 0 0 0 0 0
  4.2167  -2.3917   0.0000 S  0 0 3 0 0 0 0 0 0 0 0 0
  5.0125  -2.1750   0.0000 O  0 0 0 0 0 0 0 0 0 0 0 0
  3.4167  -2.6042   0.0000 O  0 0 0 0 0 0 0 0 0 0 0 0
  4.4292  -3.1875   0.0000 C  0 0 3 0 0 0 0 0 0 0 0 0
  5.2250  -3.4000   0.0000 C  0 0 0 0 0 0 0 0 0 0 0 0
  4.0125  -3.9000   0.0000 C  0 0 0 0 0 0 0 0 0 0 0 0
 -0.3458  -3.2125   0.0000 O  0 0 0 0 0 0 0 0 0 0 0 0
  0.8875  -3.9292   0.0000 N  0 0 0 0 0 0 0 0 0 0 0 0
```

CSR (Compound Structure Reconstruction)



Preprocessing Steps

- Page segmentation
- Image extraction
- Image conversion (image restauration, adaptive binarization ...)



WHO Drug Information, Vol. 18, No. 1, 2004 **Recommended** INN: List 51

exenatidum
exenatide — L-histidylglycyl-L-glutamylglycyl-L-threonyl-L-phenylalanyl-L-threonyl-L-seryl-L-aspartyl-L-leucyl-L-seryl-L-lysyl-L-glutaminyl-L-methionyl-L-glutamyl-L-glutamyl-L-glutamyl-L-alanyl-L-valyl-L-arginyl-L-leucyl-L-phenylalanyl-L-isoleucyl-L-glutamyl-L-tryptophyl-L-leucyl-L-lysyl-L-asparaginylglycylglycyl-L-prolyl-L-seryl-L-serylglycyl-L-alanyl-L-prolyl-L-prolyl-L-prolyl-L-serinamide

exénatide — exendine 4 (*Heloderma suspectum*), synthétique

exenatida — L-histidilglicil-L-glutamilglicil-L-treonil-L-fenilalanil-L-treonil-L-seril-L-aspartil-L-leucil-L-seril-L-lisil-L-glutaminil-L-metionil-L-glutamil-L-glutamil-L-glutamil-L-alanil-L-valil-L-arginil-L-leucil-L-fenilalanil-L-isoleucil-L-glutamil-L-triptofil-L-leucil-L-lisil-L-asparaginilglicilglicil-L-prolil-L-seril-L-serilglicil-L-alanil-L-prolil-L-prolil-L-prolil-L-serinamida

$C_{184}H_{282}N_{50}O_{60}S$

H–His–Gly–Glu–Gly–Thr–Phe–Thr–Ser–Asp–Leu–Ser–Lys–Gln–Met–Glu–Glu–Glu–Ala–Val–Arg–Leu–Phe–Ile–Glu–Trp–Leu–Lys–Asn–Gly–Gly–Pro–Ser–Ser–Gly–Ala–Pro–Pro–Pro–Ser–NH_2

firocoxibum
firocoxib — 3-(cyclopropylmethoxy)-5,5-dimethyl-4-[4-(methylsulfonyl)-phenyl]furan-2(5*H*)-one

firocoxib — 3-(cyclopropylméthoxy)-5,5-diméthyl-4-[4-(méthylsulfonyl)-phényl]furan-2(5*H*)-one

firocoxib — 3-(ciclopropilmetoxi)-5,5-dimetil-4-[4-(metilsulfonil)fenil]furan-2(5*H*)-ona

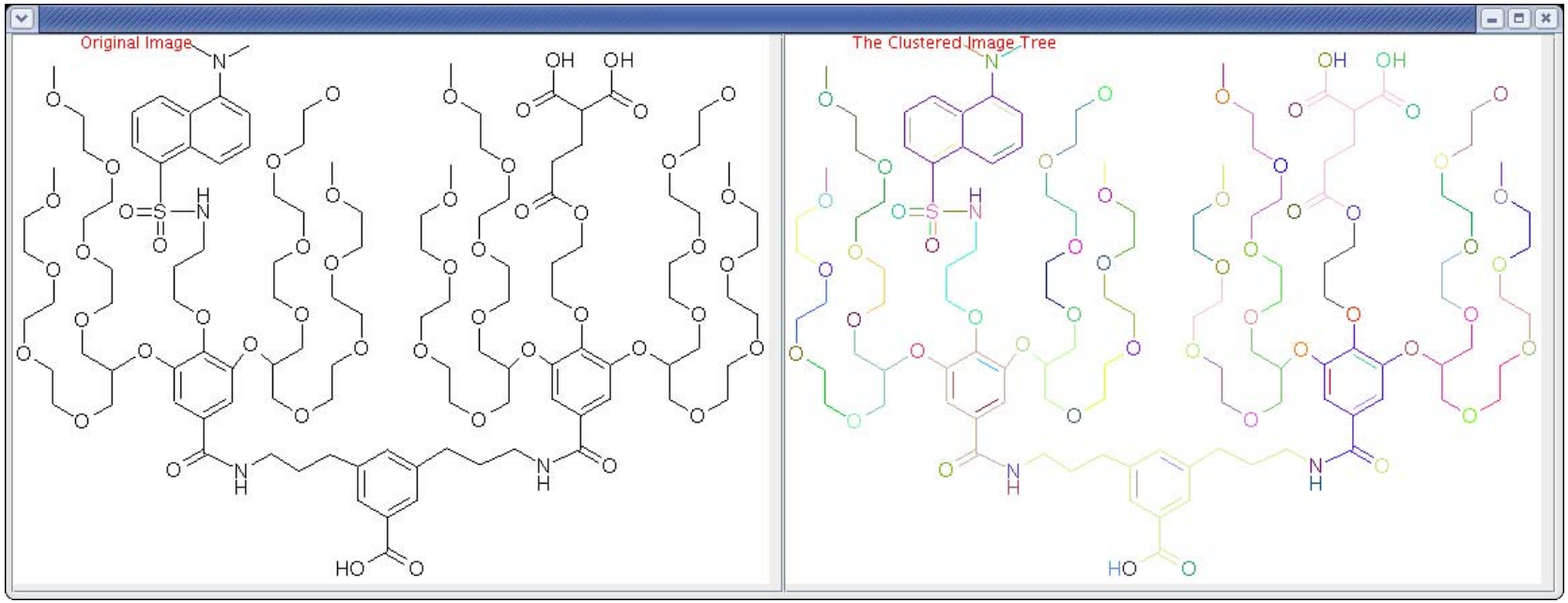
fispemifenum
fispemifene — 2-(2-{4-[(1*Z*)-4-chloro-1,2-diphenylbut-1-enyl]phenoxy}-ethoxy)ethanol

fispémifène — 2-[2-[4-[(1*Z*)-4-chloro-1,2-diphénylbut-1-ényl]phénoxy]-éthoxy]éthanol

fispemifeno — 2-(2-{4-[(1*Z*)-4-cloro-1,2-difenilbut-1-enil]fenoxi}etoxi)etanol

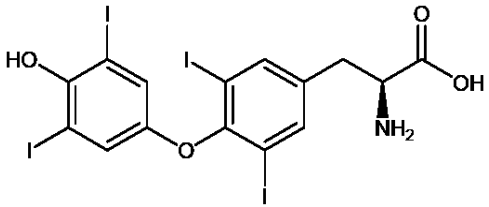
Connected Component Analysis

- Building an image tree
- Using nested TreeMaps

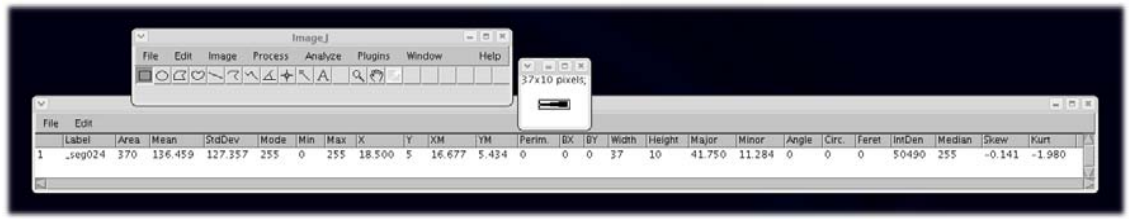


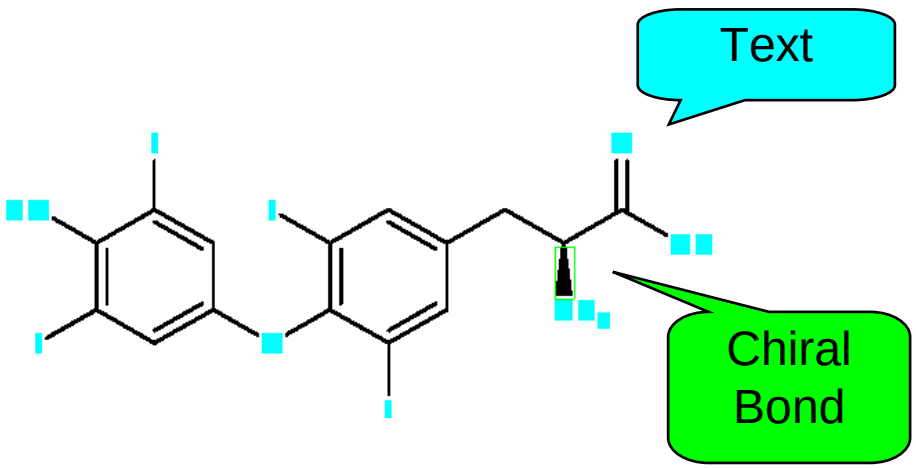
Component Classification

1 Raster image

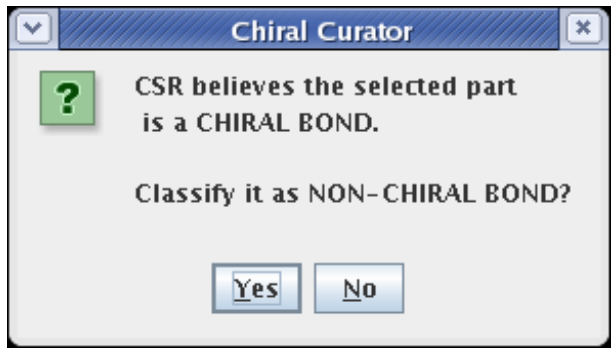


2 Extract features





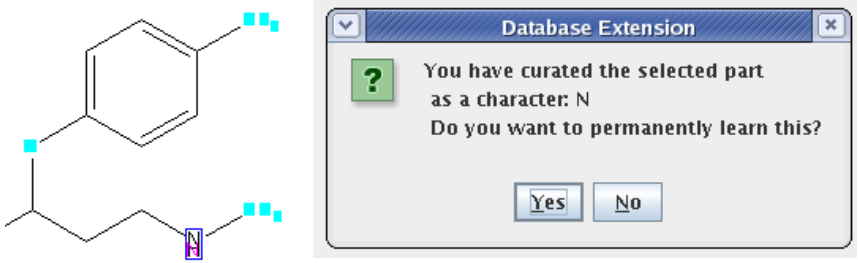
3 Classify



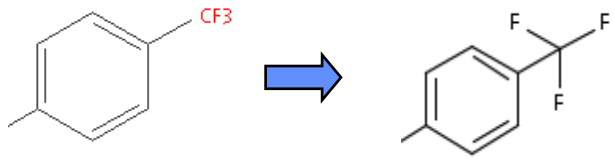
4 Manual curation

Atomtype Reconstruction

1 Train new characters

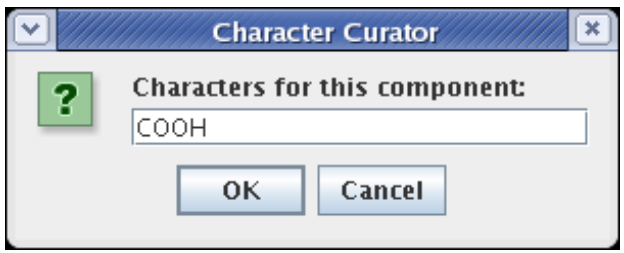


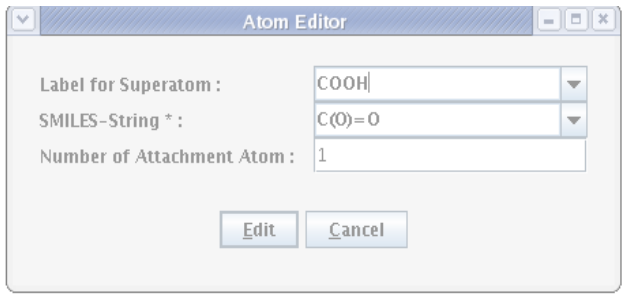
3 Expand superatoms



Need of a chemical intelligent OCR

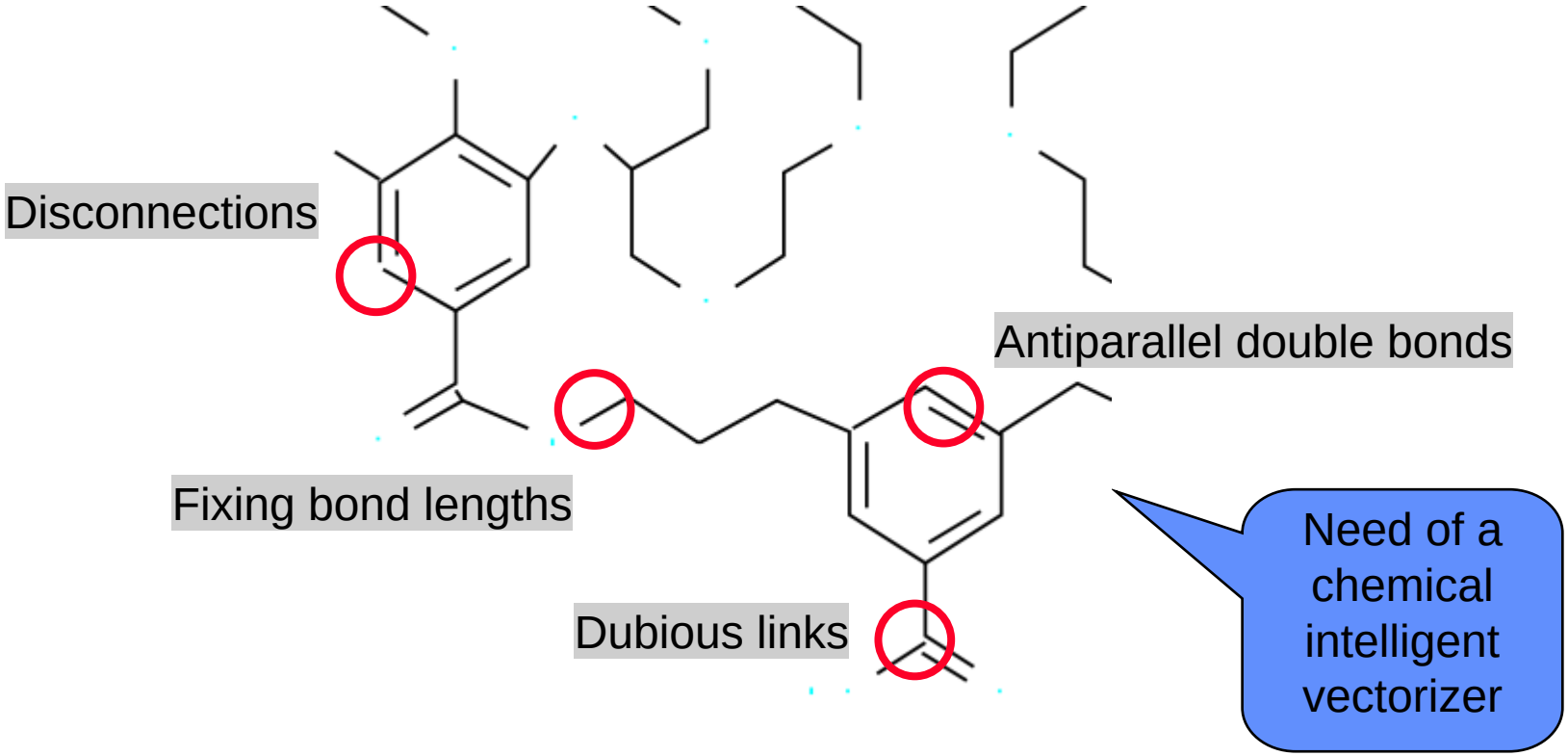
2 Define new superatoms





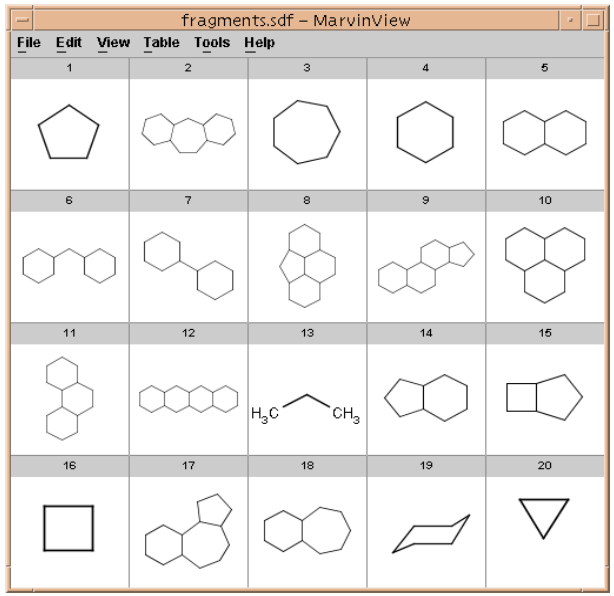
Vectorization

Fixing vectorization errors using relative neighborhood graphs

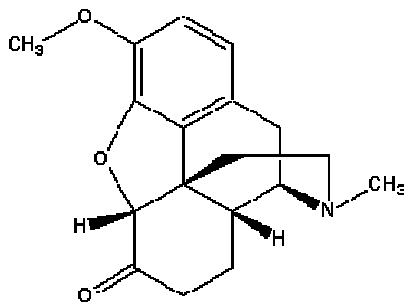


Graph Matching

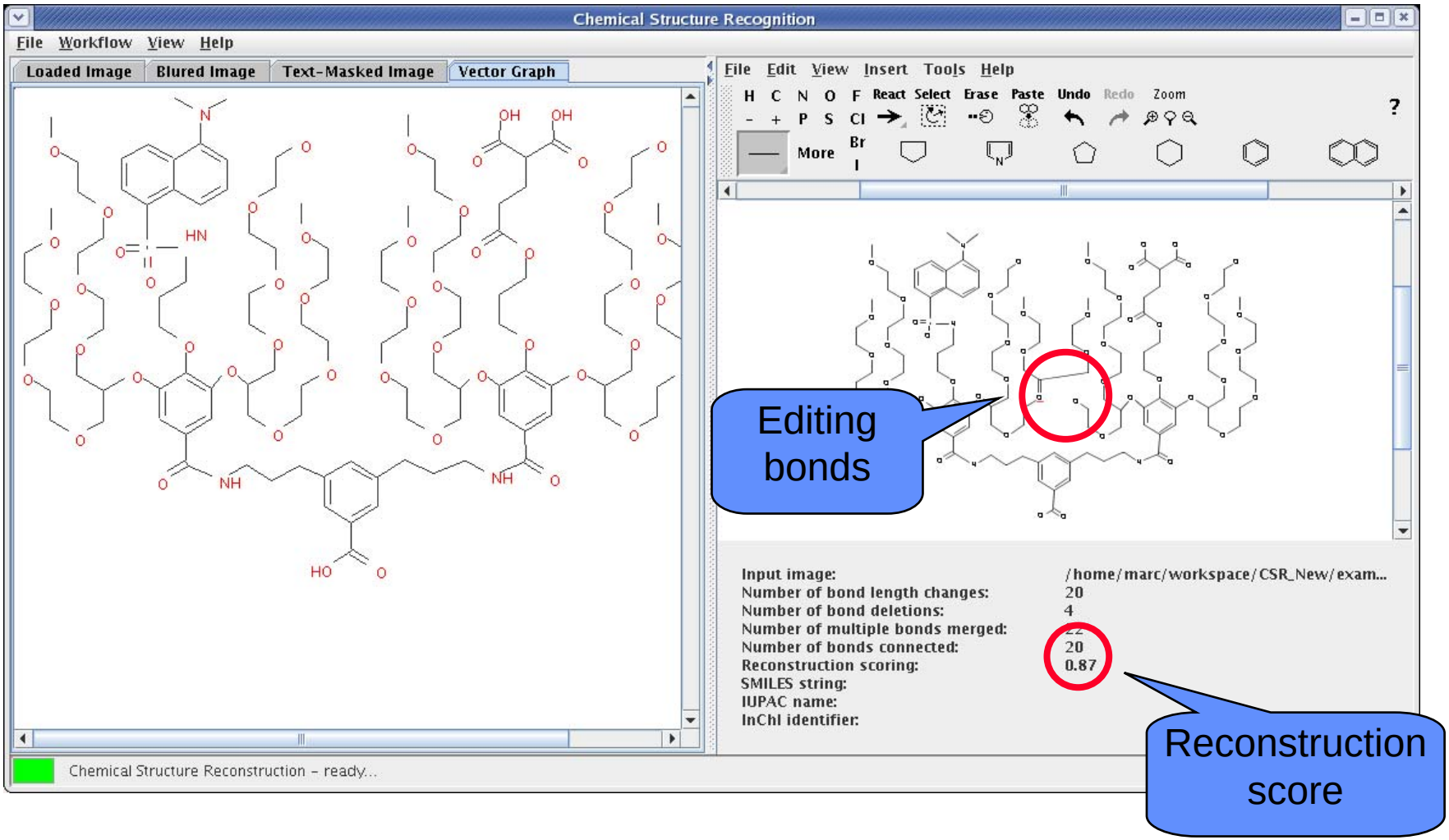
- Using a line graph representation
- Searching for subgraph isomorphism
- Database with common fragments
- Decomposition network for fragments
- Recognizing new fragments



- Still needed: mapping bridged ring systems



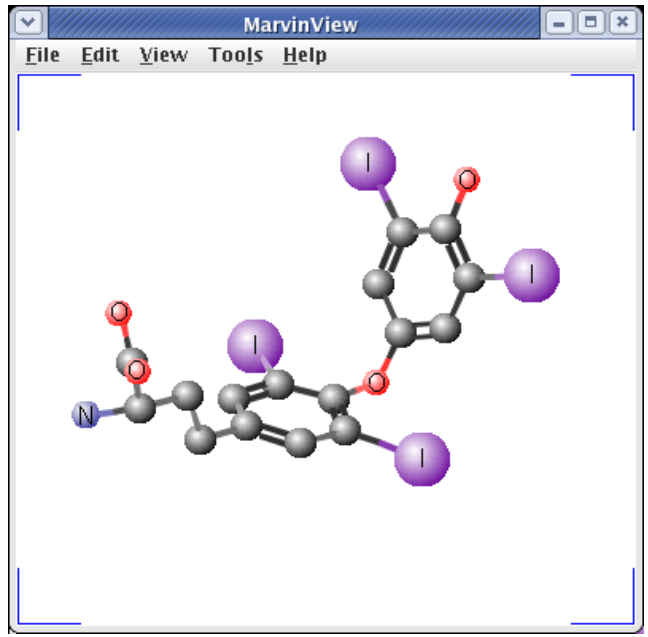
Manual Curation of Errors



Post Processing

Workflow plugin technology

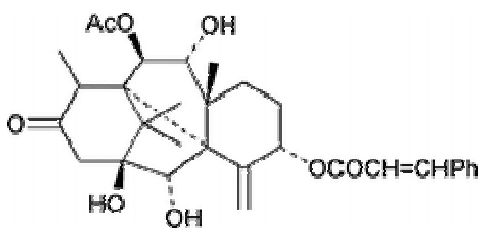
- 2D beautify
- File format conversion
- 2D to 3D conversion
- Name generation
- Property calculation / prediction
- …

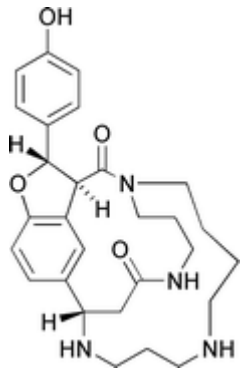


Input image:	/home/marc/workspace/CSR_New/exam...
Number of bond length changes:	2
Number of multiple bonds merged:	6
Reconstruction scoring:	1
SMILES string:	CNCCC(Oc1ccc(cc1)C(F)(F)F)c2ccccc2
IUPAC name:	
InChI identifier:	InChI=1/C17H18F3NO/c1-21-12-11-1...

A Real Challenge (coming soon)

- Data set with ~10.000 depictions of natural products
  - We will train our machine learning methods
  - We will incorporate the CSR workflow into a grid service
  - We will add a database interface





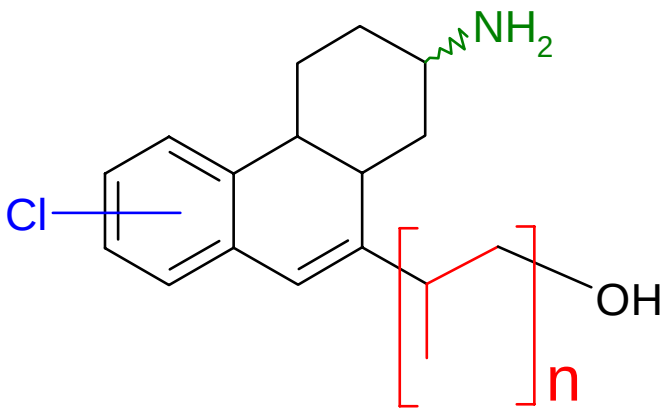
But we need more real training sets…

(i.e. pictures and the solved structure)

Future Works

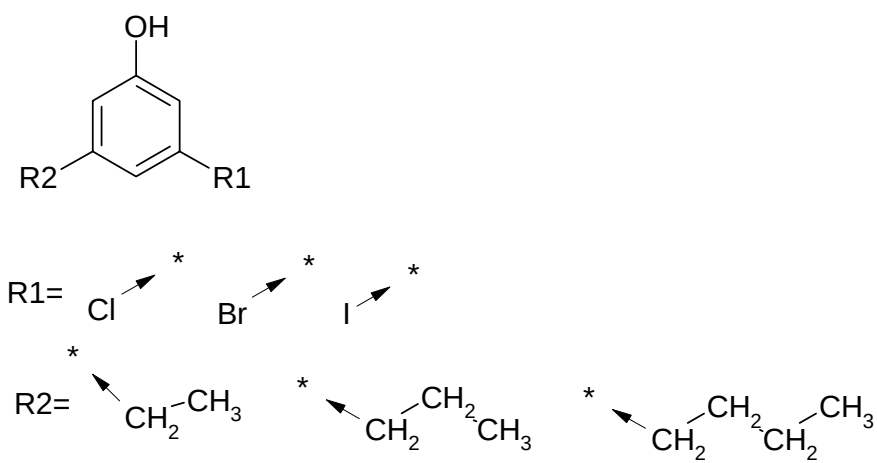
Incompletely-defined substances:

- unknown stereochemistry
- unknown attachment position
- unknown repetition

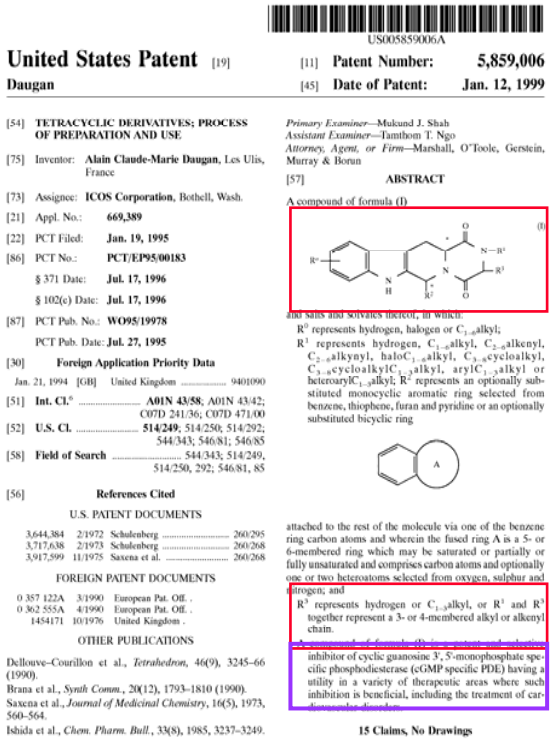


Markush ("Generic") Structures and Reaction Schemes

- shorthand for describing sets of structures with common features
- structures with R-groups
- very important in chemical patents
- can be used to describe combinatorial libraries
- can be used as queries in database searches



The Mission: Combination of CSR and Text Mining



US005859006A

United States Patent [19]
Daugan

[11] **Patent Number: 5,859,006**
[45] **Date of Patent: Jan. 12, 1999**

[54] **TETRACYCLIC DERIVATIVES; PROCESS OF PREPARATION AND USE**

[75] Inventor: **Alain Claude-Marie Daugan**, Les Ulis, France

[73] Assignee: **ICOS Corporation**, Bothell, Wash.

[21] Appl. No.: **669,389**

[22] PCT Filed: **Jan. 19, 1995**

[86] PCT No.: **PCT/EP95/00183**

§ 371 Date: **Jul. 17, 1996**

§ 102(e) Date: **Jul. 17, 1996**

[87] PCT Pub. No.: **WO95/19978**

PCT Pub. Date: **Jul. 27, 1995**

[30] **Foreign Application Priority Data**

Jan. 21, 1994 [GB] United Kingdom 9401090

[51] **Int. Cl.**[6] **A01N 43/58**; A01N 43/42; C07D 241/36; C07D 471/00

[52] **U.S. Cl.** **514/249**; 514/250; 514/292; 544/343; 546/81; 546/85

[58] **Field of Search** 544/343; 514/249, 514/250, 292; 546/81, 85

[56] **References Cited**

U.S. PATENT DOCUMENTS

3,644,384	2/1972	Schulenberg	260/295
3,717,638	2/1973	Schulenberg	260/268
3,917,599	11/1975	Saxena et al.	260/268

FOREIGN PATENT DOCUMENTS

0 357 122A	3/1990	European Pat. Off. .
0 362 555A	4/1990	European Pat. Off. .
1454171	10/1976	United Kingdom .

OTHER PUBLICATIONS

Dellouve–Courillon et al., *Tetrahedron*, 46(9), 3245–66 (1990).
Brana et al., *Synth Comm.*, 20(12), 1793–1810 (1990).
Saxena et al., *Journal of Medicinal Chemistry*, 16(5), 1973, 560–564.
Ishida et al., *Chem. Pharm. Bull.*, 33(8), 1985, 3237–3249.

Primary Examiner—Mukund J. Shah
Assistant Examiner—Tamthom T. Ngo
Attorney, Agent, or Firm—Marshall, O'Toole, Gerstein, Murray & Borun

[57] **ABSTRACT**

A compound of formula (I)

(I)

and salts and solvates thereof, in which:

R^0 represents hydrogen, halogen or C_{1-6}alkyl;

R^1 represents hydrogen, C_{1-6}alkyl, C_{2-6}alkenyl, C_{2-6}alkynyl, haloC_{1-6}alkyl, C_{3-8}cycloalkyl, C_{3-8}cycloalkylC_{1-3}alkyl, arylC_{1-3}alkyl or heteroarylC_{1-3}alkyl; R^1 represents an optionally substituted monocyclic aromatic ring selected from benzene, thiophene, furan and pyridine or an optionally substituted bicyclic ring

A

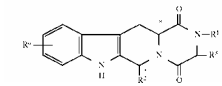
attached to the rest of the molecule via one of the benzene ring carbon atoms and wherein the fused ring A is a 5- or 6-membered ring which may be saturated or partially or fully unsaturated and comprises carbon atoms and optionally one or two heteroatoms selected from oxygen, sulphur and nitrogen; and

R^3 represents hydrogen or C_{1-6}alkyl, or R^1 and R^3 together represent a 3- or 4-membered alkyl or alkenyl chain.

inhibitor of cyclic guanosine 3', 5'-monophosphate specific phosphodiesterase (cGMP specific PDE) having a utility in a variety of therapeutic areas where such inhibition is beneficial, including the treatment of car-

15 Claims, No Drawings

Image Analysis /
Structure Reconstruction



-CH3

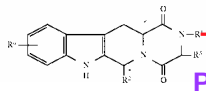
-CH2-CH3

-CH2-CNHS

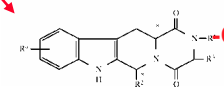
-COOH

Cytochrome inhibition
PPAR activation
Stability in serum
Side effect
Blood-brain-barrier

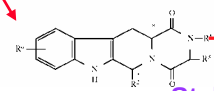
Text Analysis /
Entity Recognition



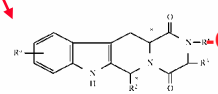
-CH3
PPAR activation



-CH2-CH3
Cytochrome inhibition



-CH2-CNHS
Stability in serum

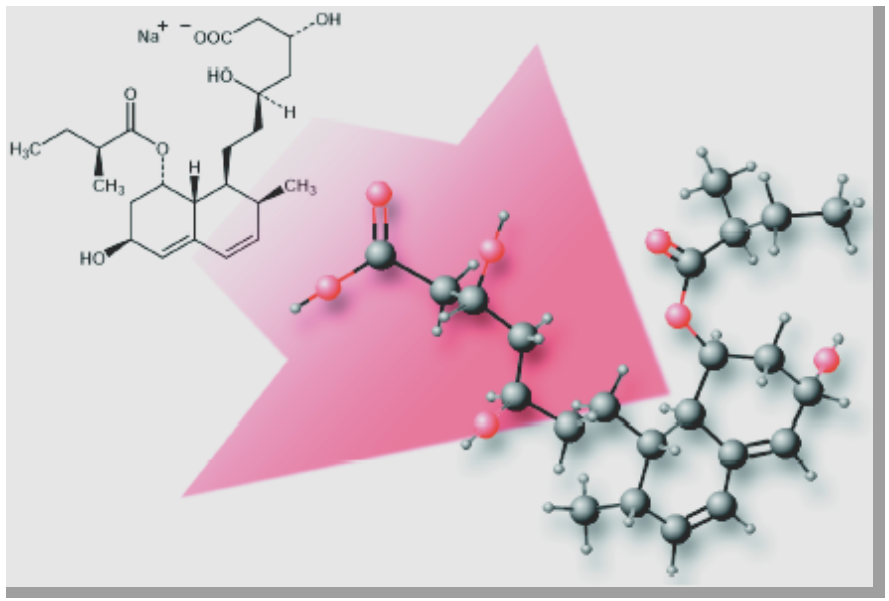


-COOH
Side effect

Reconstruction of
Published Chem-, Pharm-
and PatentSpace

The Team (in the order of appearance)

- Marc Zimmermann
- Tanja Fey
- Le Thuy Bui Thi
- Christoph Friedrich
- Yuan Wang
- Maria-Elena Algorri
- Miguel Alvarez



CSR Software Demo available

CSR can extract chemical depictions from various image sources and convert them into SMILES and SD files, which can be further used in nearly all chemical software; it allows for the modification of reconstructed molecules by a structure editor; it maintains the superatom and bond (single, double, triple, or chiral) information; and it accepts user curation in each stage and scoring schema to improve its performance.