Next Generation Chemical Information Mining in Images

Discovery Knowledge & Informatics 2007, Amsterdam

Fraunhofer Institute Algorithms and Scientific Computing

Dr. Marc Zimmermann

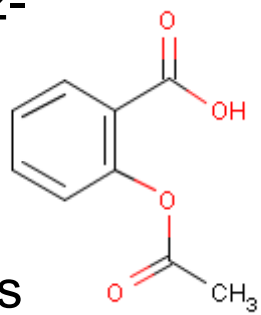
Structure of this Presentation

- Definition of the scientific challenge
- Analysis of existing approaches
- New ideas – the tool of Fraunhofer
  - A chemical vectorizer and OCR
  - CSR through a expert system
  - Error tracking and prediction
  - Current status
- Perspectives for the future: multi-modal information extraction

Identification and Representation in Chemistry

- **Trivial names:** Aspirin, Acetylsalicyclic acid, ….

- **Systematic nomenclatures:**
  - Mass formula: C9H8O4
  - SMILES: OC(=O)C1=C(C=CC=C1)OC(=O)C
  - InChI: 1/C9H8O4/c1-6(10)13-8-5-3-2-4-7(8)9(11)12/h2-5H,1H3,(H,11,12)



- **Structural formula:** universal language between chemists

Chemical properties ~ atom composition + spatial arrangement

Chemical Structure Reconstruction (CSR) Problem

- **Publication process**

  Molecule is published as image

  embedded in

  books, patents, papers, journals,

  websites, internal reports, PhD theses

  ⇒ the machine readable format is lost

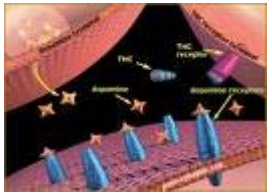
- **CSR Problem definition:**

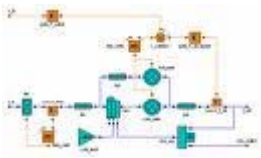
  Molecule image ⇒ Molecule computer representation

Searching for Structural Information in Images

Google™ THC Search Images

Results **1** - **20** of about **113,000** for **THC**. (**0.40** seconds)





New! Want to improve Google Image Search? Try Google Image Labeler.

Welcome to **Google Image Labeler**, a new feature of Google Image Search that allows you to label random images to help improve the quality of Google's image search results.

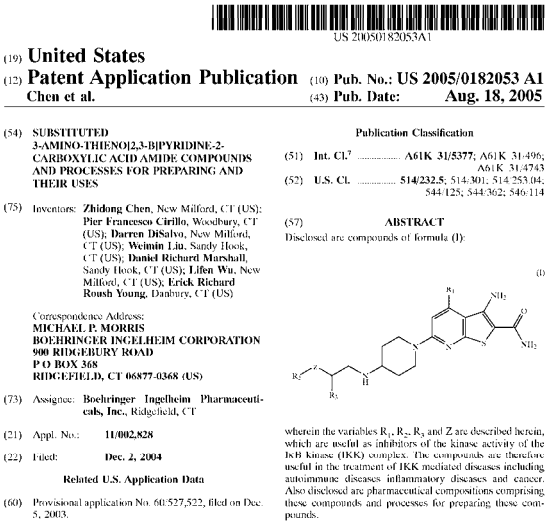
From the Perspective of a Chemist

But what is really interesting is – find me documents containing...

- similar structures
- structures containing a benzene ring
- structures fulfilling the pharmacophore
- the patented core structure
- a synthesis protocol
- …

Chemical Structure Reconstruction – an Overview

1 Document



US 20050182053A1

(19) United States
(12) Patent Application Publication
Chen et al.
(10) Pub. No.: US 2005/0182053 A1
(43) Pub. Date: Aug. 18, 2005

(54) SUBSTITUTED 3-AMINO-THIENO[2,3-B]PYRIDINE-2-CARBOXYLIC ACID AMIDE COMPOUNDS AND PROCESSES FOR PREPARING AND THEIR USES

(75) Inventors: Zhidong Chen, New Milford, CT (US); Pier Francesco Cirillo, Woodbury, CT (US); Darren DiSalvo, New Milford, CT (US); Weimin Liu, Sandy Hook, CT (US); Daniel Richard Marshall, Sandy Hook, CT (US); Lifen Wu, New Milford, CT (US); Erick Richard Roush Young, Danbury, CT (US)

Correspondence Address:
MICHAEL P. MORRIS
BOEHRINGER INGELHEIM CORPORATION
900 RIDGEBURY ROAD
P O BOX 368
RIDGEFIELD, CT 06877-0368 (US)

(73) Assignee: Boehringer Ingelheim Pharmaceuticals, Inc., Ridgefield, CT

(21) Appl. No.: 11/002,828

(22) Filed: Dec. 2, 2004

Related U.S. Application Data

(60) Provisional application No. 60/527,522, filed on Dec. 5, 2003.

Publication Classification

(51) Int. Cl.[7] A61K 31/5377; A61K 31/496; A61K 31/4743

(52) U.S. Cl. 514/232.5; 514/301; 514/253.04; 544/125; 544/362; 546/114

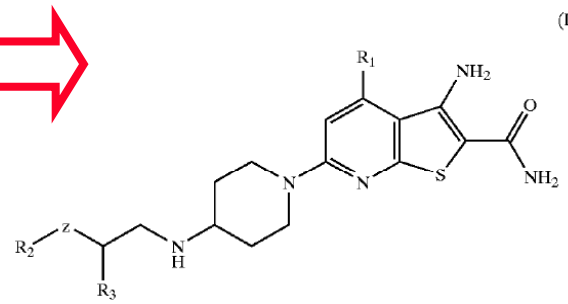
(57) ABSTRACT

Disclosed are compounds of formula (I):

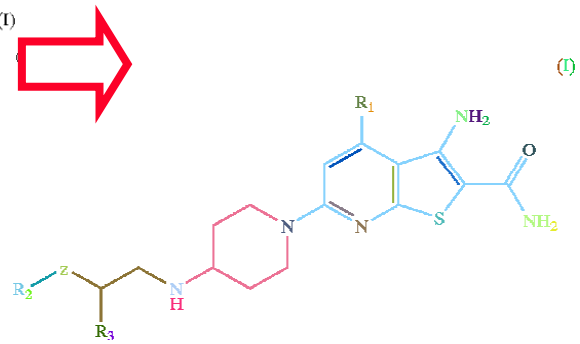
(I)

wherein the variables R_1, R_2, R_3 and Z are described herein, which are useful as inhibitors of the kinase activity of the IκB kinase (IKK) complex. The compounds are therefore useful in the treatment of IKK mediated diseases including autoimmune diseases inflammatory diseases and cancer. Also disclosed are pharmaceutical compositions comprising these compounds and processes for preparing these compounds.

2 Depiction



3 Reconstruction

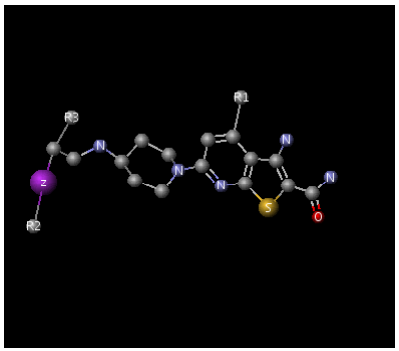


4 SDF file

created from
/home/marc/workspace/CSR/results/CSR/examples/US2005182053/US2005182053_result.pnm
MZCSRv0.5010050621162D 0.00000 0.00000 0

```
  26 28  0     1  0  0  0  0 0999 V2000
204.0000  102.0000    0.0000 C  0  0  0  0  0  0  0  0  0  0  0  0
275.0000   61.0000    0.0000 C  0  0  0  0  0  0  0  0  0  0  0  0
201.0000   59.0000    0.0000 C  0  0  0  0  0  0  0  0  0  0  0  0
422.0000  178.0000    0.0000 C  0  0  0  0  0  0  0  0  0  0  0  0
311.0000  164.0000    0.0000 C  0  0  0  0  0  0  0  0  0  0  0  0
384.0000  165.0000    0.0000 C  0  0  0  0  0  0  0  0  0  0  0  0
447.0000  144.0000    0.0000 C  0  0  0  0  0  0  0  0  0  0  0  0
383.0000  123.0000    0.0000 C  0  0  0  0  0  0  0  0  0  0  0  0
131.0000   60.0000    0.0000 C  0  0  0  0  0  0  0  0  0  0  0  0
239.0000  123.0000    0.0000 C  0  0  0  0  0  0  0  0  0  0  0  0
349.0000  218.0000    0.0000 R#  0  0  0  0  0  0  0  0  0  0  0  0
447.0000  207.0000    0.0000 R#  0  0  0  0  0  0  0  0  0  0  0  0
```

5 *in silico* Chemistry

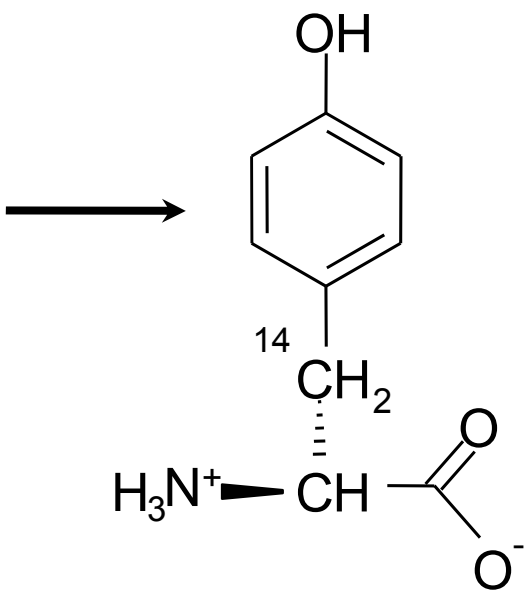


CSR is a famous Problem

- **Kekulé: Ocr-optical chemical (structure) recognition,** R. McDaniel and Jason R. Balmuth. J. Chem. Inf. Comput. Sci., 32(4):373–378, 1992.
- **Chemical Literature Data Extraction: The CLiDE Project,** P. Ibison, M. Jacquot, F. Kam, A. G. Neville, R.W. Simpson, C. Tonnelier, T. Venczel and A.P. Johnson, J. Chem. Inf. Comput. Sci., vol. 33(3): 338-344, 1993.
- **Optical recognition of chemical graphics,** S. Boyer, Document Analysis and Recognition, Proceedings of the Second International Conference on Publication, 627–631, 20-22 Oct 1993.

But what happened to them?

The General Idea is simple



(II)

- use OCR ⇒ atoms, indices, charges
- use vectorizer ⇒ bonds
- use clustering ⇒ superatoms, double bonds
- use pattern matching ⇒ stereochemical configuration
- convert into graph ⇒ molecule

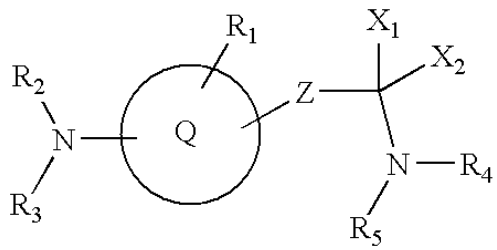
The CSR Process…

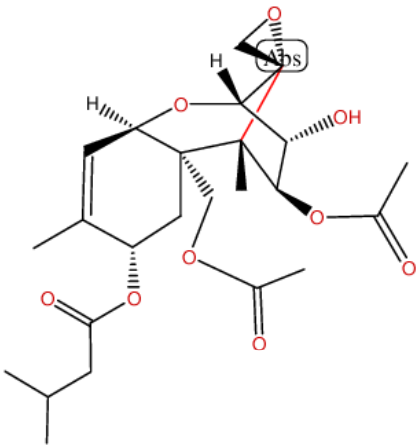
- is a multidisciplinary approach:
  - Image Processing
  - Pattern Recognition
  - Maschine Learning
  - Algorithms
  - Chemoinformatics
  - Organic Chemistry

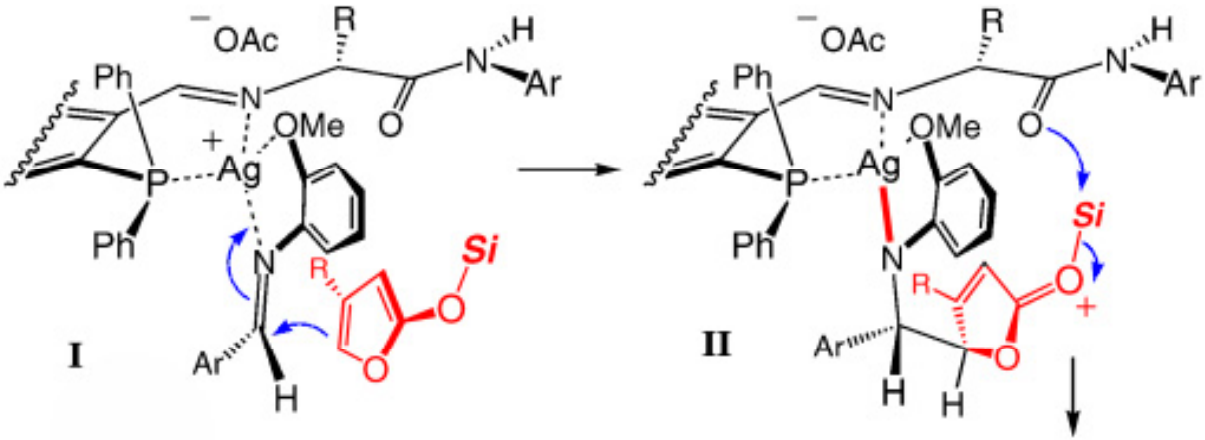
- is a multi step process:
  1. image preprocessing
  2. image conversion
  3. semantic entity recognition
  4. chemical structure assembly
  5. reconstruction validation
  6. post processing
- for each step a specific module has should implemented
- modules should be assembled into workflows

Main Roadblocks in the CSR Process

- OCR is not adapted to chemical depictions
- Vectorization is not chemical aware
- The system cannot be improved by the chemist
- A critical validation is needed
- Chemists like complicated diagrams

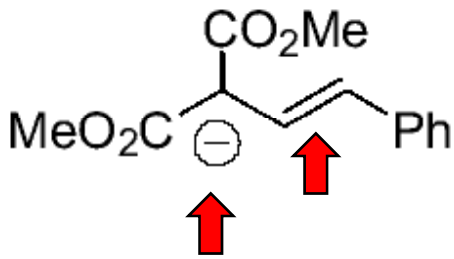






Traditional OCR For Chemical Depictions

Yes! OCR was solved in the 70's, but…



- Existing OCR systems have trouble with
  - non character symbols, e.g: the bonds
  - scattered symbols (looking for a common baseline)
  - have a fixed symbol space (not trainable)
  - don‘t provide certainty values
  - errors are consistently made between classes like “H” and ”N” , “O” and “Q”.

CsrOcr Solution Approach

- We are making use of
  - staged approach for very small characters
  - feature extraction (Zernike functions and Wavelets) instead of template matching
  - machine learning methods (SVM)
  - contextual information
  - a confidence measure

- Advantages
  - trainable
  - probability estimates can be used for certainty estimation

Requirements for Vectorizing Chemical Depictions

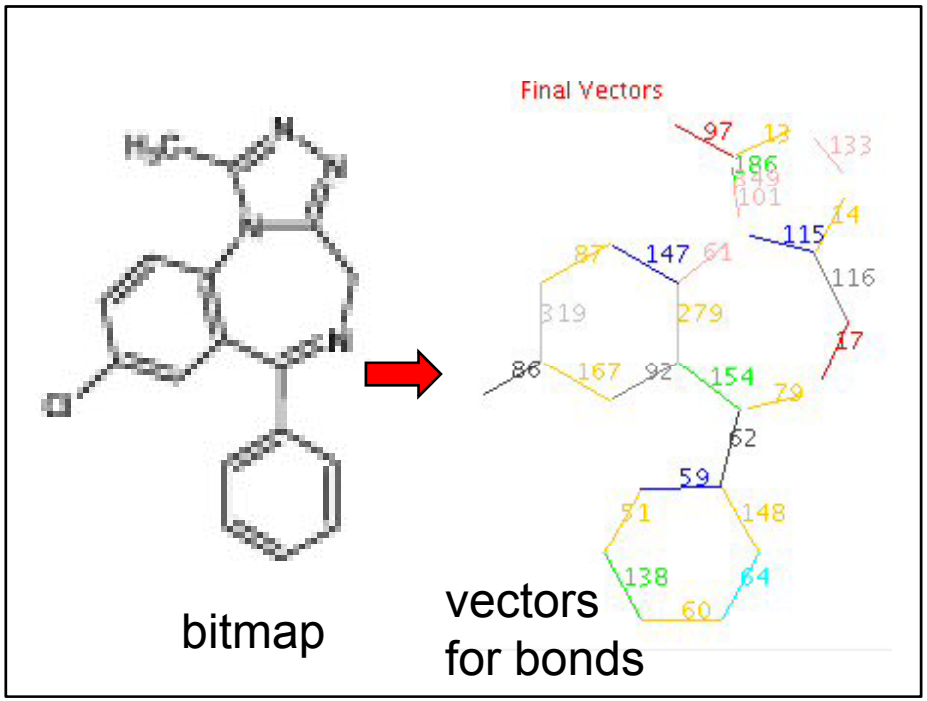
Yes! Vectorization of technical depictions was also solved, but…

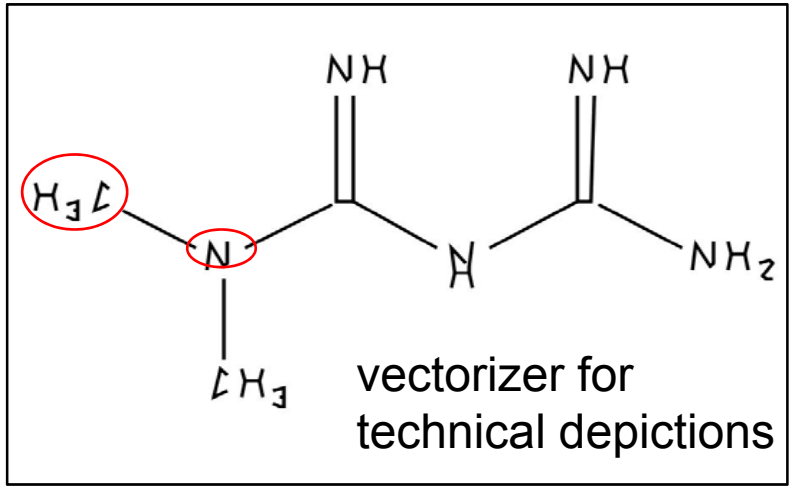
Our vectorizer must preserve the graph characteristics of the image:

- Same number of vectors as bonds are in the image
- Same number of vertices as C-atoms are in the image
- That is: a line cannot be broken into many vectors
- And: a thick joint cannot create small spurious vectors

Vectorizing Chemical Depictions

- we need to reproduce bonds exactly not angles, lenghts, thickness of lines...
- not all objects should be vectorized

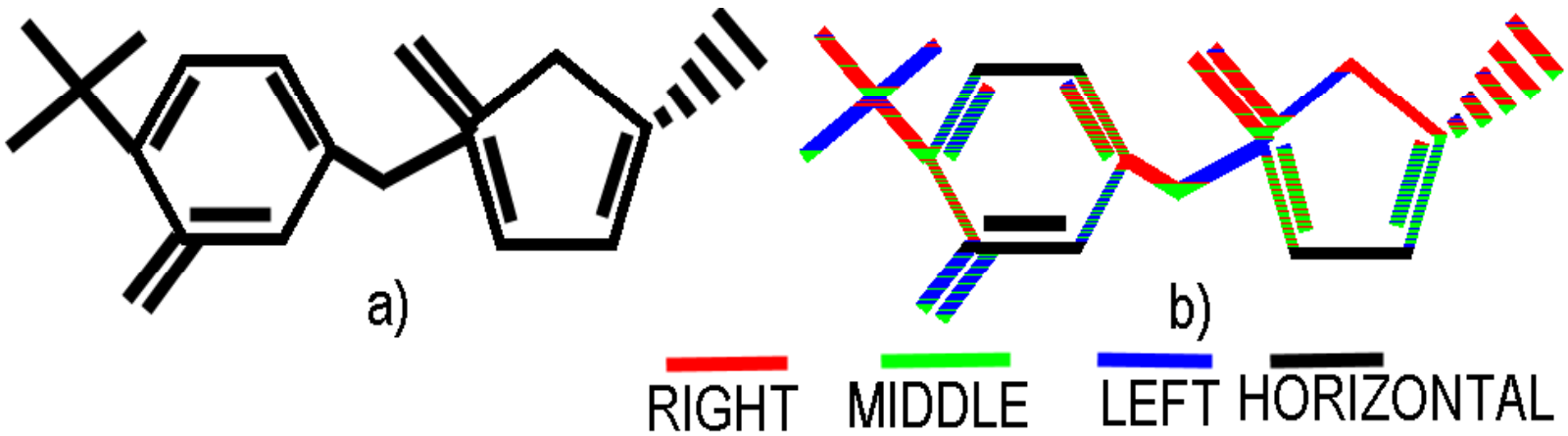




⇒ we need a classifier

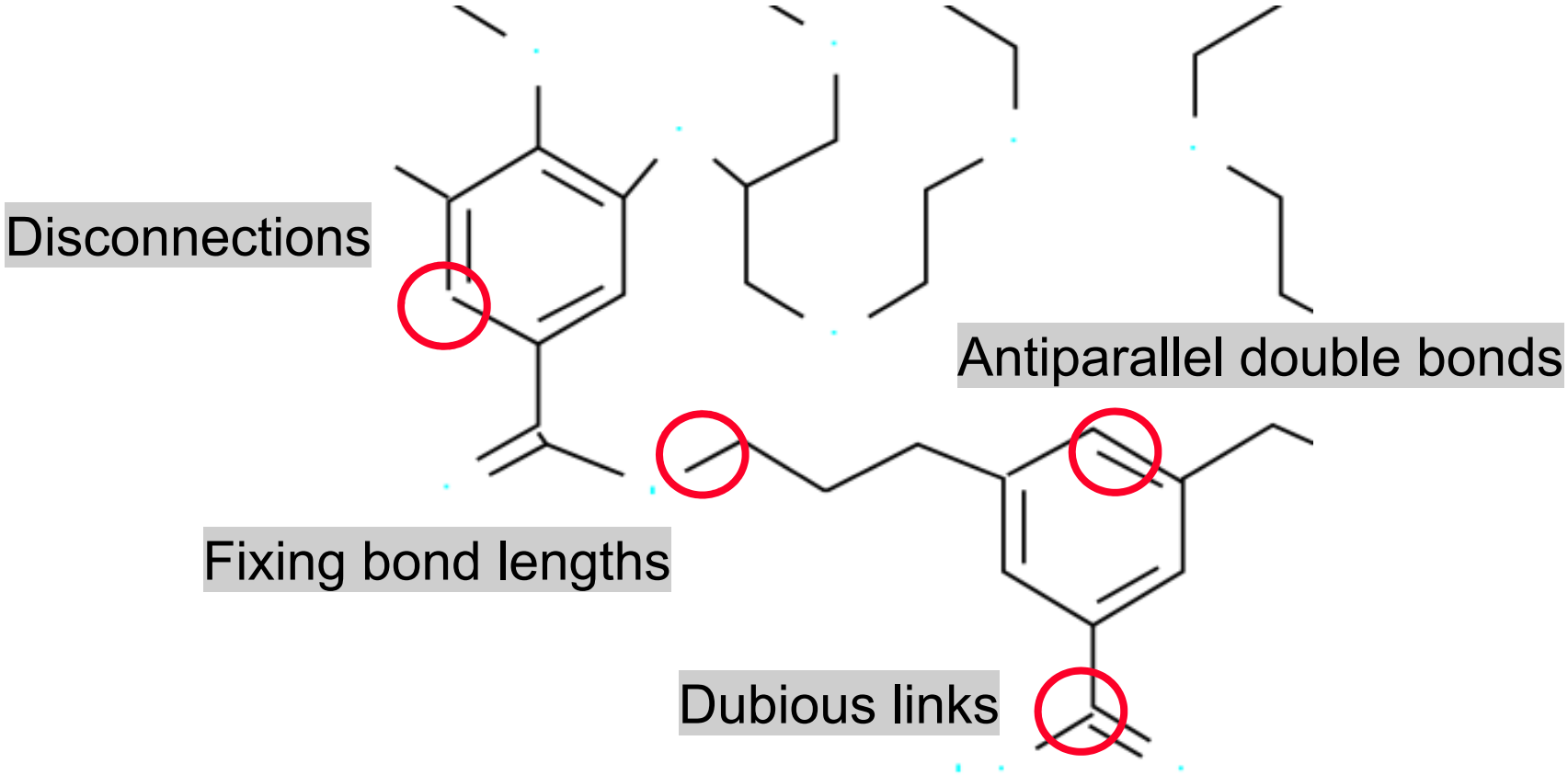
Vectorization by Local Pattern Analysis

- The DIRECTED connected components are examined for patterns of local directions:



- We associate the patterns of local directions with the global directions or vectors present in the image of the molecule.

Vectorization Problems

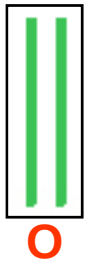


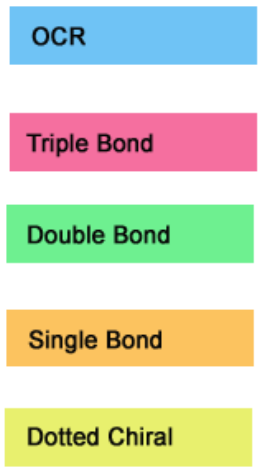
⇒ Fixing vectorization errors using a reconstructor

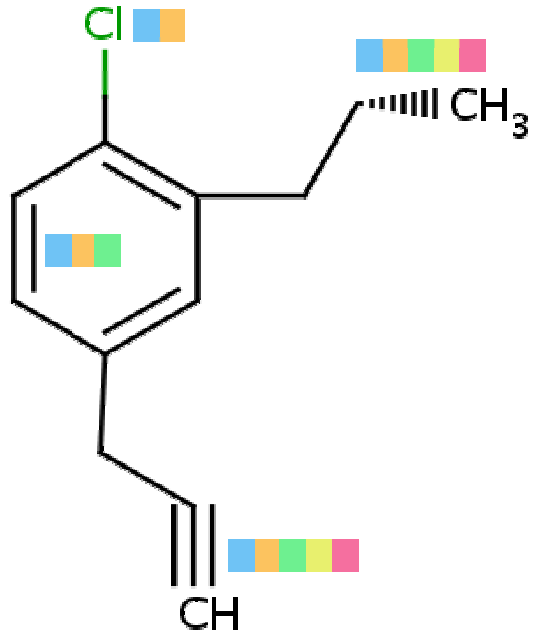
Bottlenecks of CSR

Several problems arise through several conceptual assumptions:

- **Assumptions:**
  - It is possible to develop independent recognizers for each element
  - Recognizers need no context information of the pattern they identify
- **But:** Overloaded symbols exist







New Concept to Address the CSR Problem

CSR through an expert system determined graph exploration

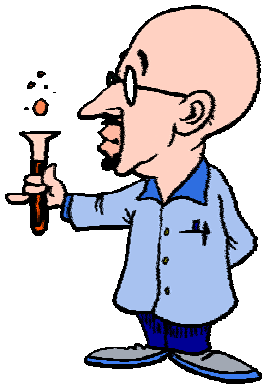
Main ideas:

- Address the CSR problem with an expert system
- Introduce new information level:

  Spatial arrangement approximation through graph approach
- Context specific recognition (determined exploration)

Expert System Description

- Part of artificial intelligence
- Difference to Machine Learning
  - No reduction on a general mathematical formalism
  - No training to model / simulate knowledge
- Special language allows explicit expert knowledge formulation
- Often rule based inference engine
- Explanation component



Address CSR Problem with Expert System

Motivation:

- Separate recognition and extraction and reconstruction
- Introduce context specific chemical knowledge based recognition
- Recognition rules for all elements in a common language
- Rules centralized in a common rule repository
- Expert/chemist can specify new rules without programming (HMI)

The KnowledgeFactBase

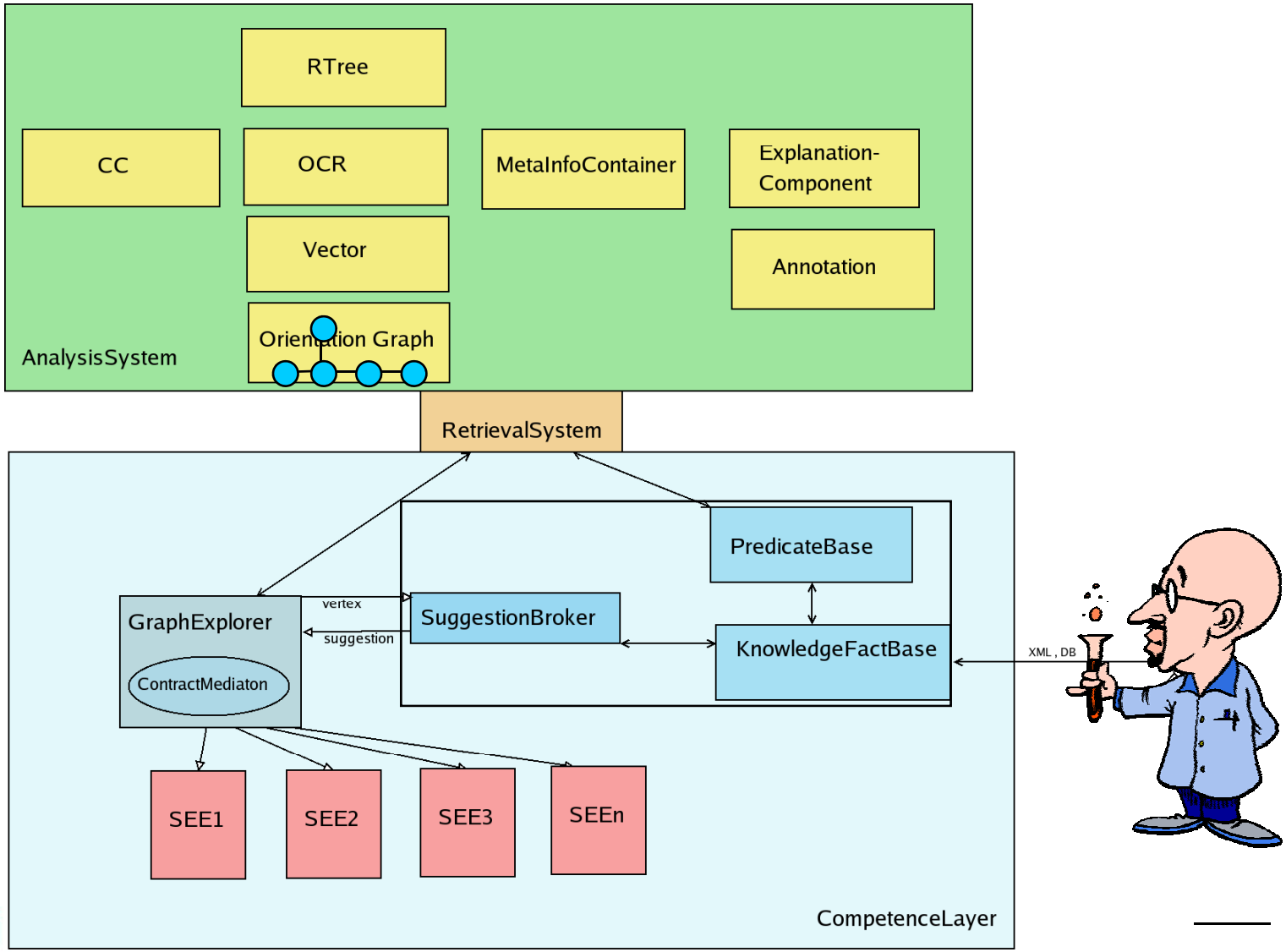
KnowledgeFactBase
$Rule_1$: $pred_1 \wedge pred_2 \wedge pred_5 \rightarrow SE_1$
…

What knowledge can be formulated:

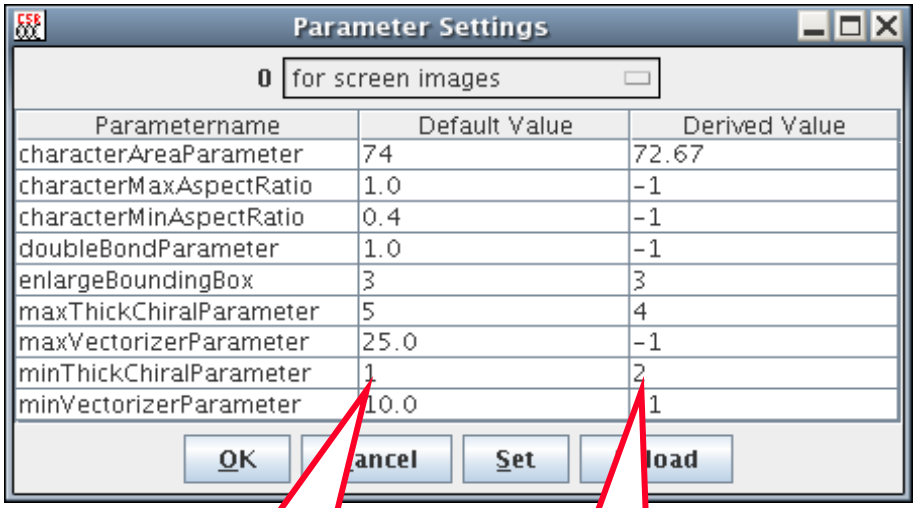
- Structural Formula knowledge: how structural formulas are drawn
- Chemistry knowledge: e.g. Chlorine is a halogen → valence 1
- Example:

a.isOneVector() Λ
a.hasOrientation(12) Λ
b.isCharacter(C) Λ
a.relativePosition(EAST, b) →
CHLORINE

Current Architecture



Many Images ⇨ Many Parameters?



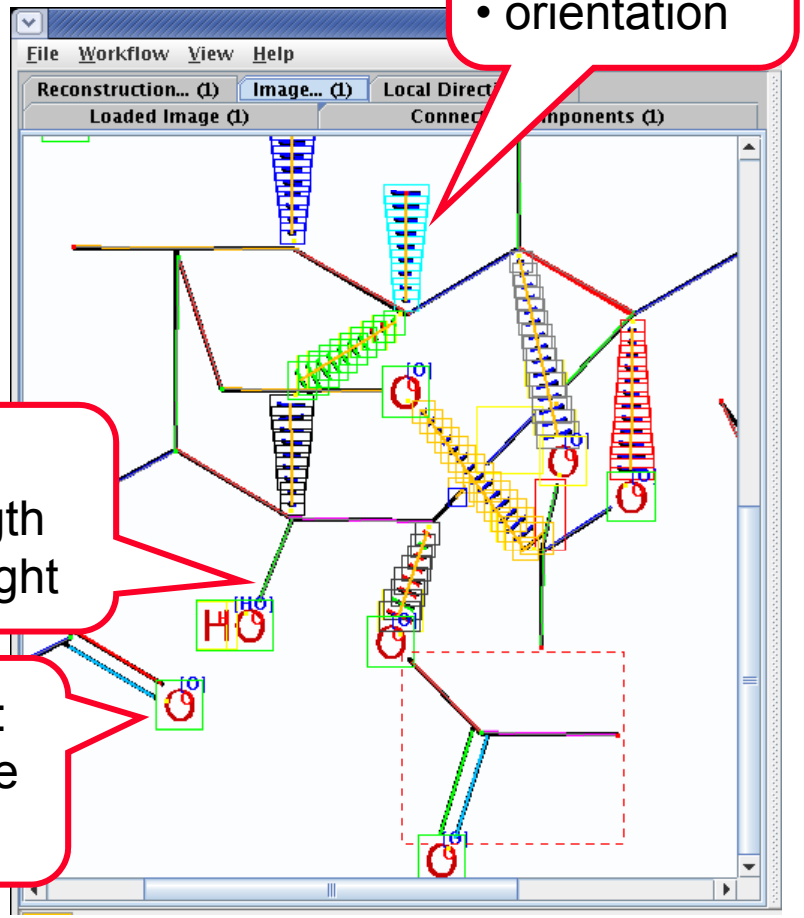
Parametername	Default Value	Derived Value
characterAreaParameter	74	72.67
characterMaxAspectRatio	1.0	-1
characterMinAspectRatio	0.4	-1
doubleBondParameter	1.0	-1
enlargeBoundingBox	3	3
maxThickChiralParameter	5	4
maxVectorizerParameter	25.0	-1
minThickChiralParameter	1	2
minVectorizerParameter	10.0	1

predefined for image sets – currently 4

estimated from the image itself

geometric constraints

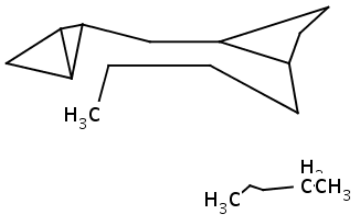
chiral:
- # segments
- orientation

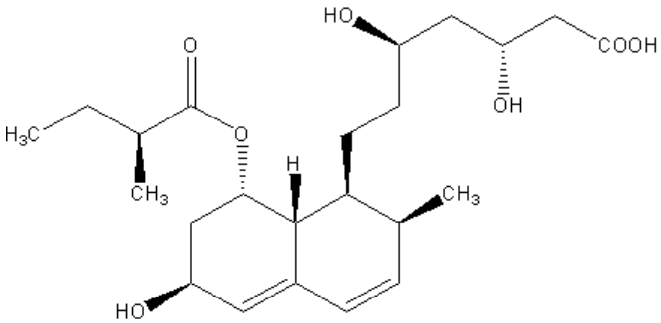


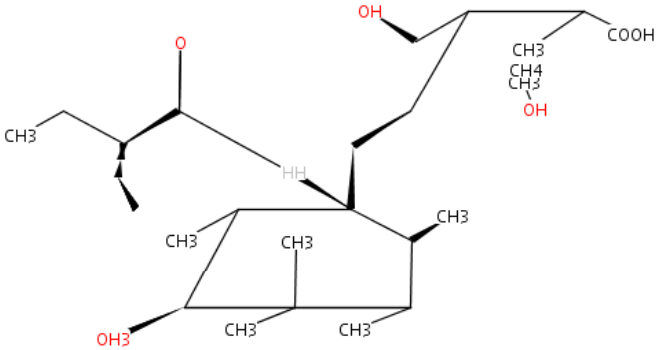
bond:
- min length
- max lenght

character:
- area size
- shape

Beautiful Artwork But Wrong Molecule







Reconstruction Error Prediction As an Alert System

result validation can only be used if the molecule is already known or the expert is checking the result:

- good for bug fixing and training of the process
- can‘t be used for the data generation process

⇨ need a different strategy for the *batch mode*:

- identify and predict reconstruction errors
- alert the user only if interaction is needed
- choose a threshold for the precision

Error Prediction – the Theory

prediction and *recognition* can be based on

- the use of chemical knowledge bases
- image properties, i.e. measure the complexity of the problem
- instance based machine learning, i.e. teach the system

the main goal is to assemble a *reconstruction score* without knowing the correct solution

$$R_{score} = w_1 \cdot \text{complexity} + w_2 \cdot \text{chemical likelihood} + w_3 \cdot \text{known errors} < T_{alert} \ ?$$

weights *w* can be set by regression analysis

Established Error Classes

chemical knowledge bases

- OCR errors and unknown super atoms
- valence checking
- known scaffolds

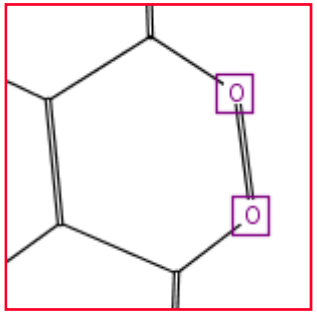
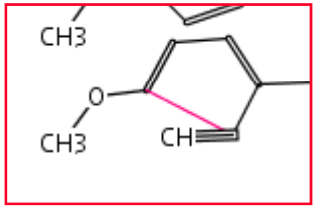


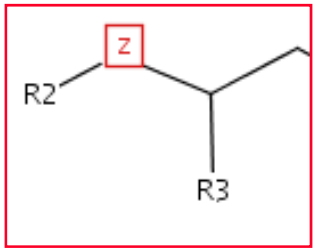
image properties

- strange bond drawings (size, angles, ...)
- pixel density, size of connected components
- complexity



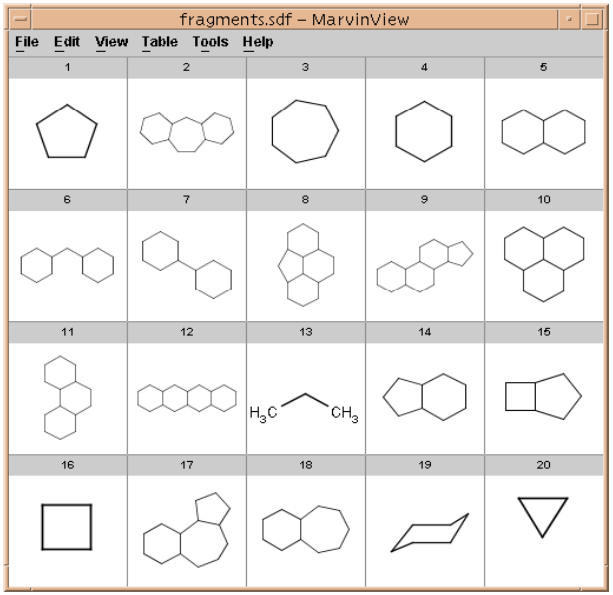
instance based machine learning (IBL)

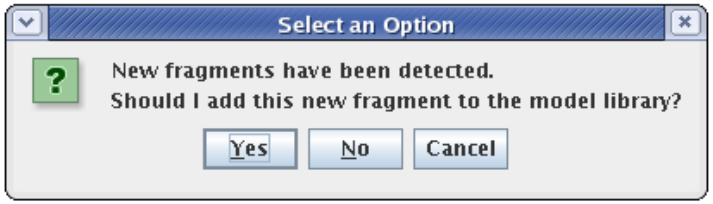
- atom and bond distributions
- Lipinski score (i.e. drug like)



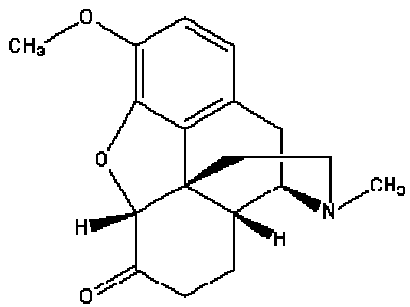
Graph Matching of Known Scaffolds

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

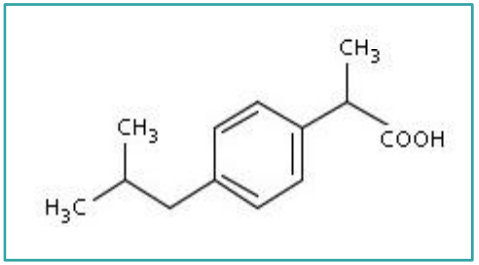




- Still needed: mapping bridged ring systems



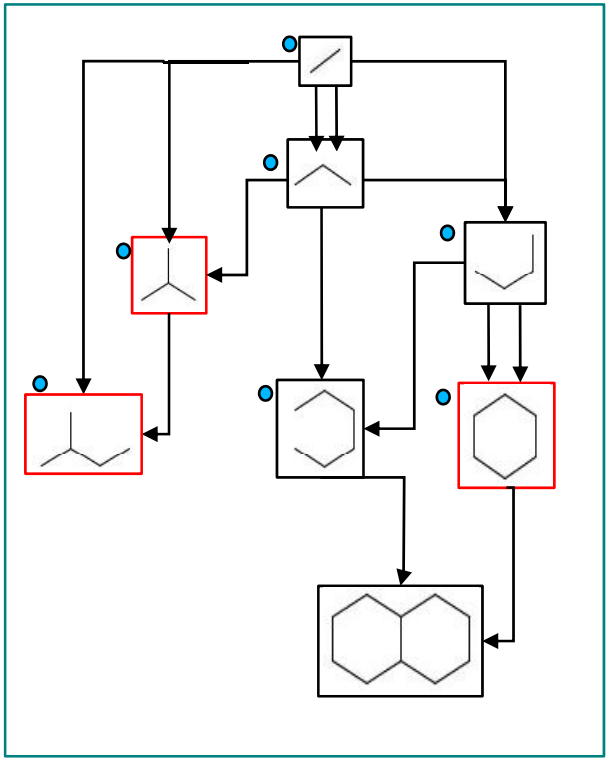
Combing several Graphs into a Decomposition Network



Input-Graph

found subgraph isomorphismn

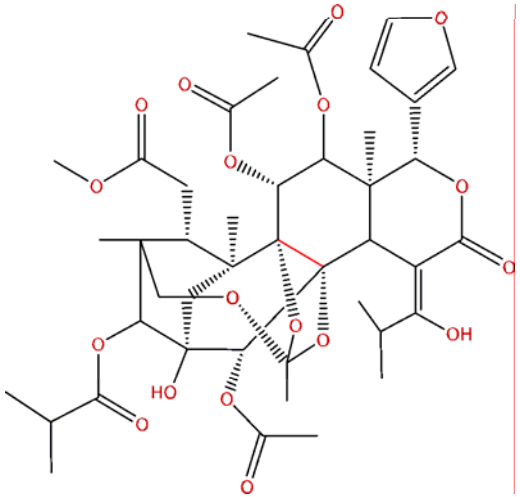
chosen fragments for the reconstruction



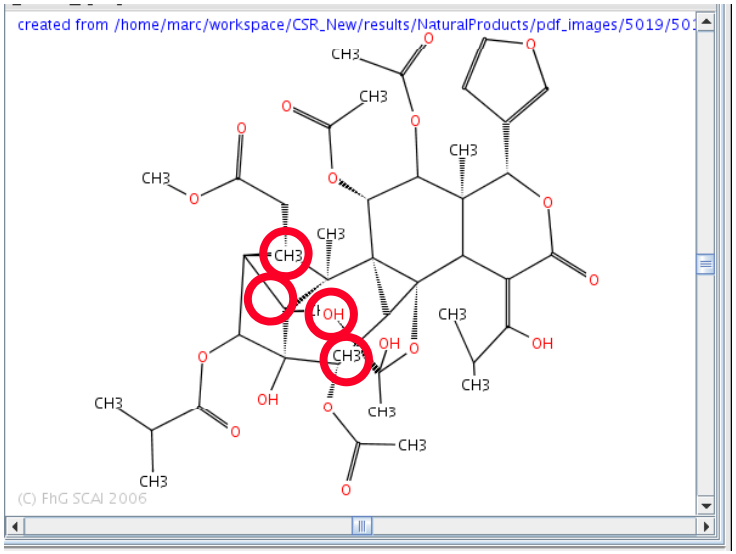
decomposition network

So We Got an Error Reported

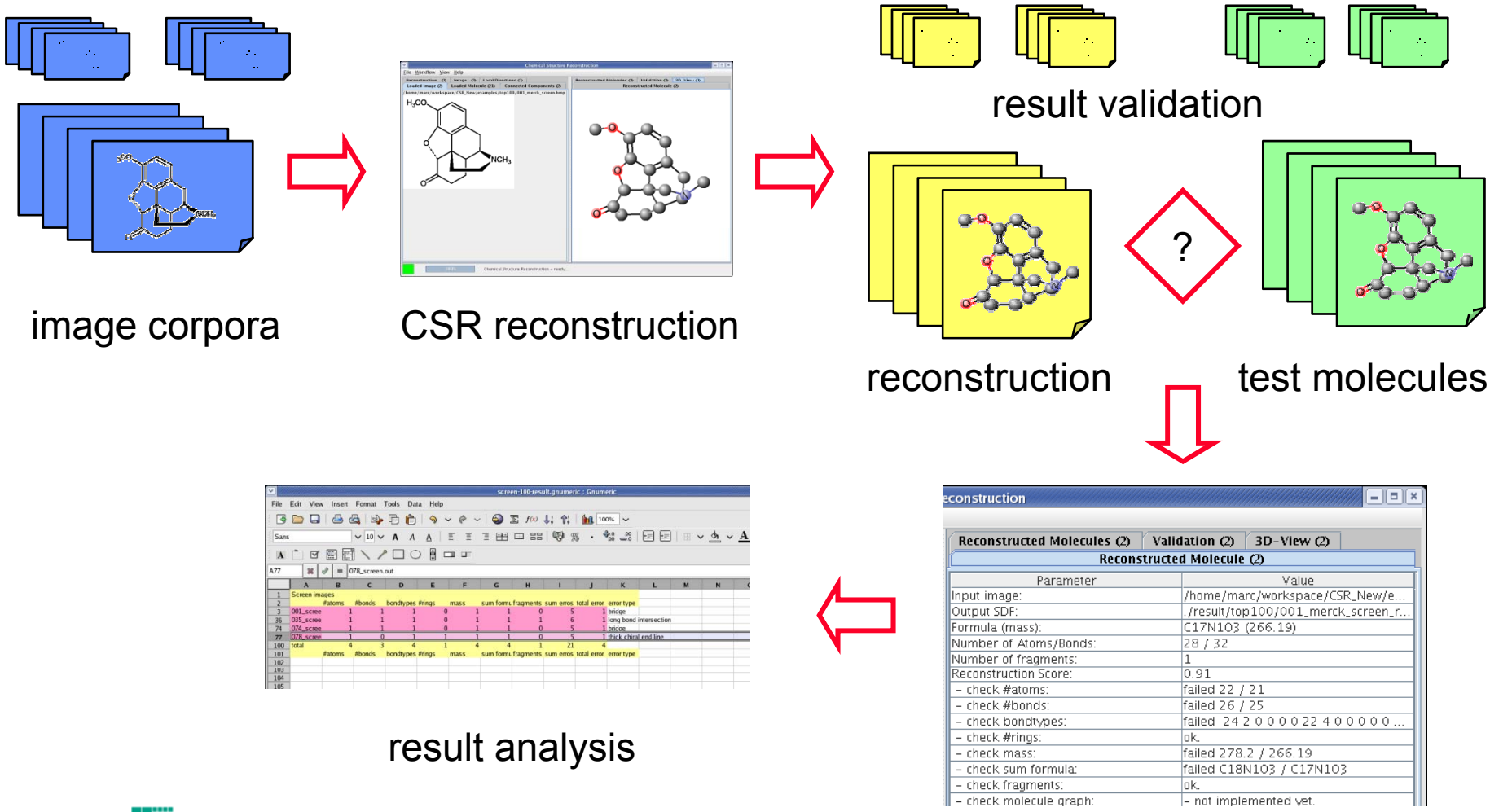
- need perfect reconstruction ⇨ start molecule editor
- need for indexing and retrieval
  - ⇨ use similarity and substructure searches
  - ⇨ specify reporting threshold



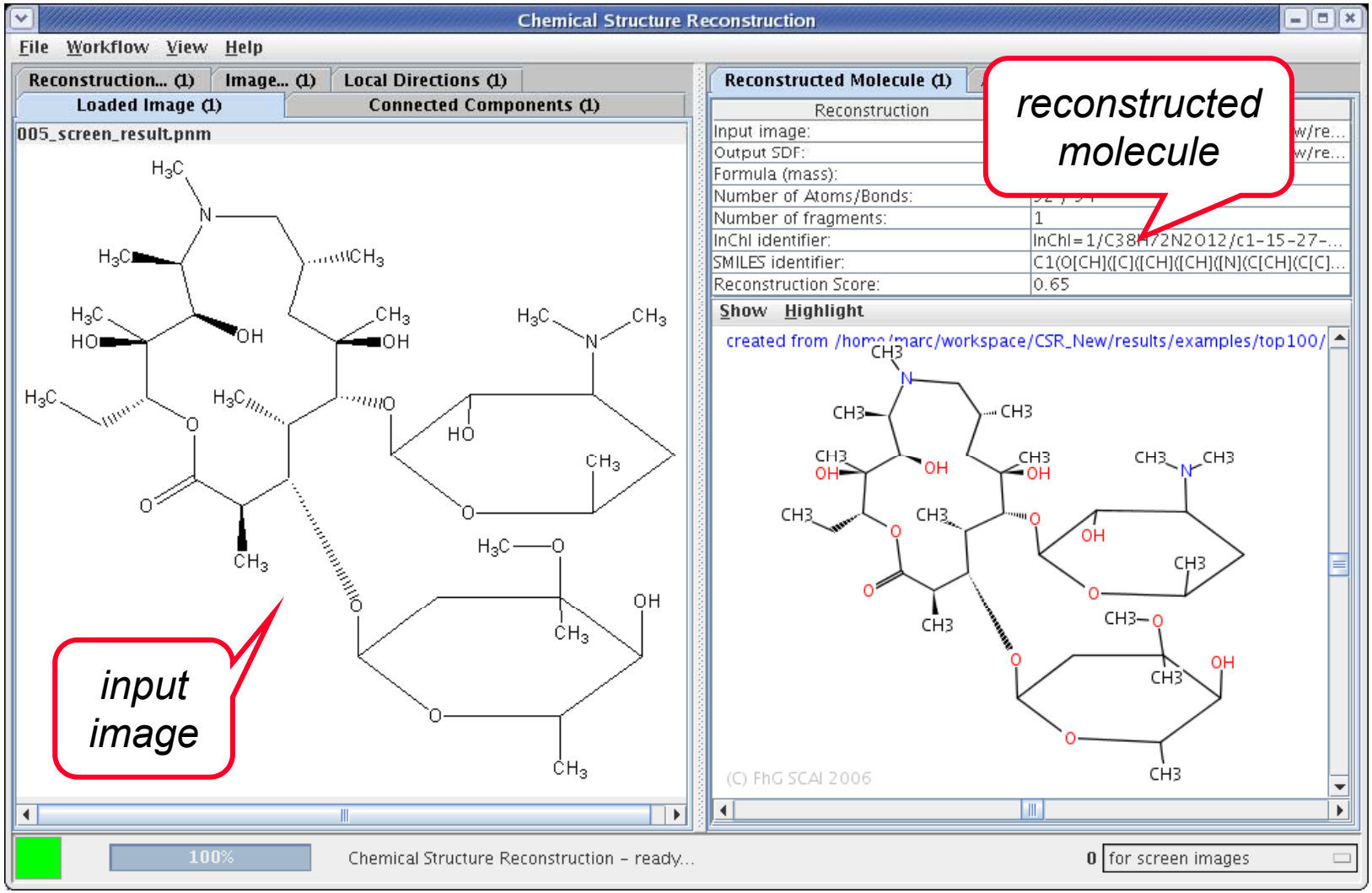
to be tolerated?



Result Validation Using Training and Test Data

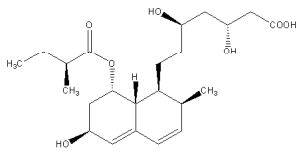


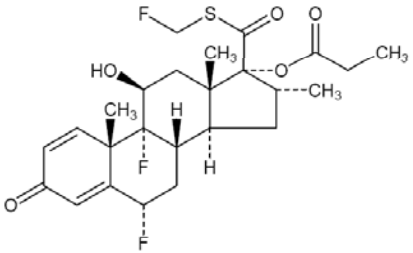
Look And Feel of chemoCR

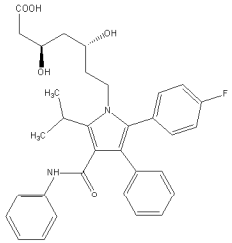


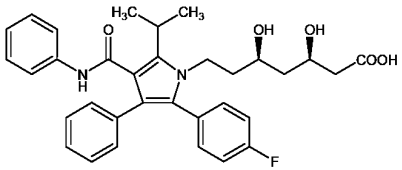
The Results – Current Status

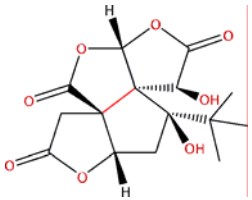
test sets:	2005	2006	2007
100x	100x	7600x	750,000x

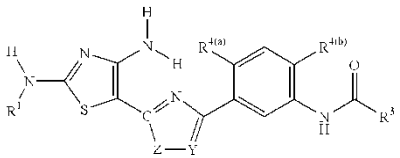








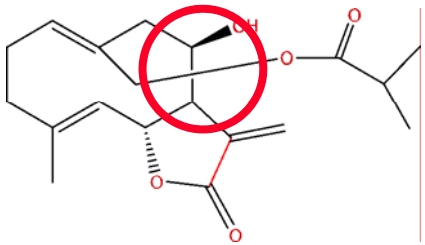




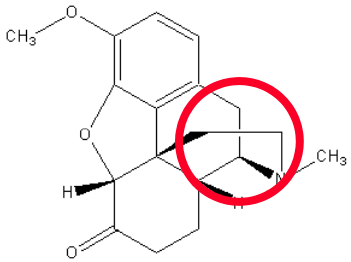
<u>measurement:</u>
correct: input identical to output
incorrect: more than one error

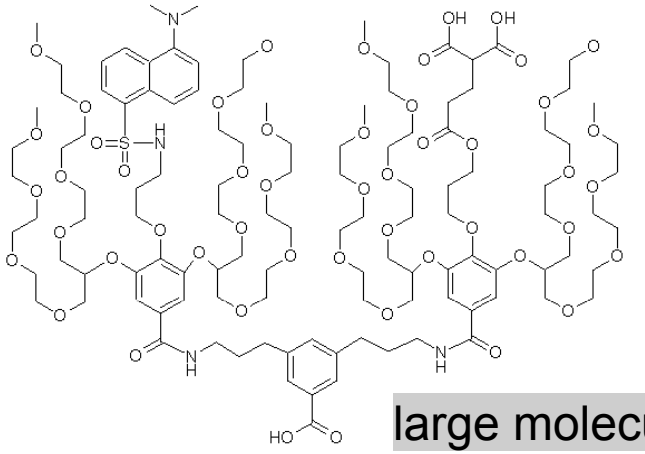
thin & clean > 95%	thick & clean > 75%	challenging mixture > 55%	patents > ??%

Not Too Bad...

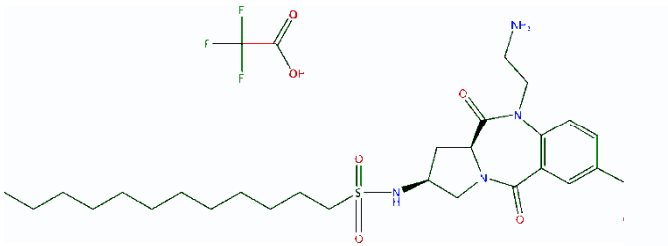


simple bridges

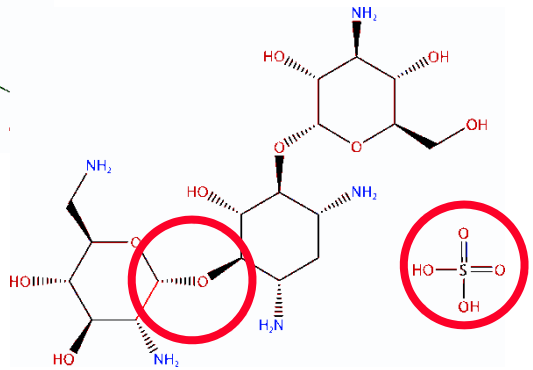




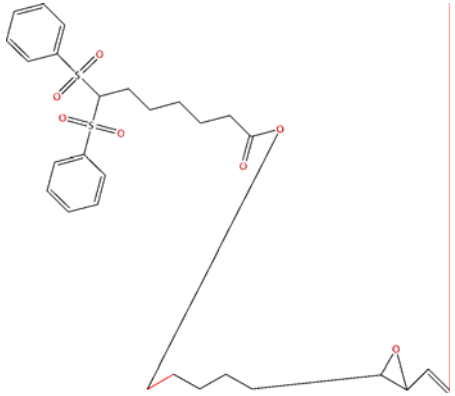
large molecules

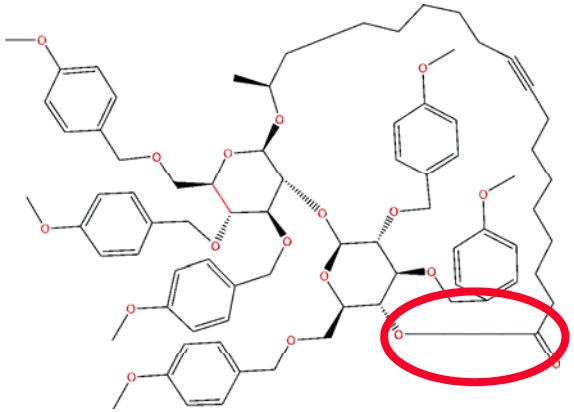


salts and chirals

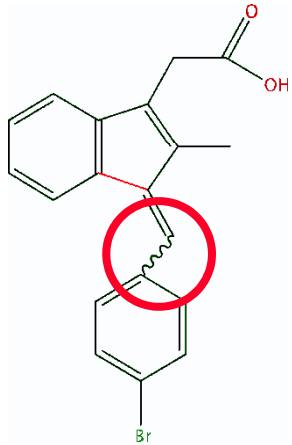


Questionable…

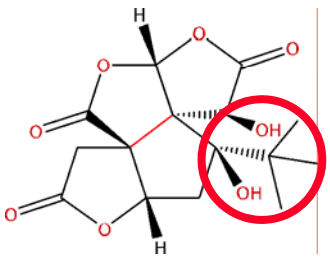


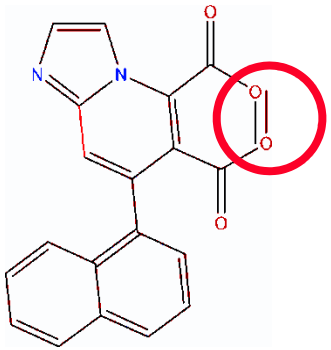


extreme long bonds

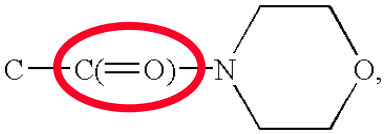


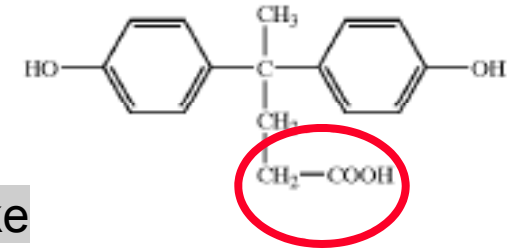
new bond types





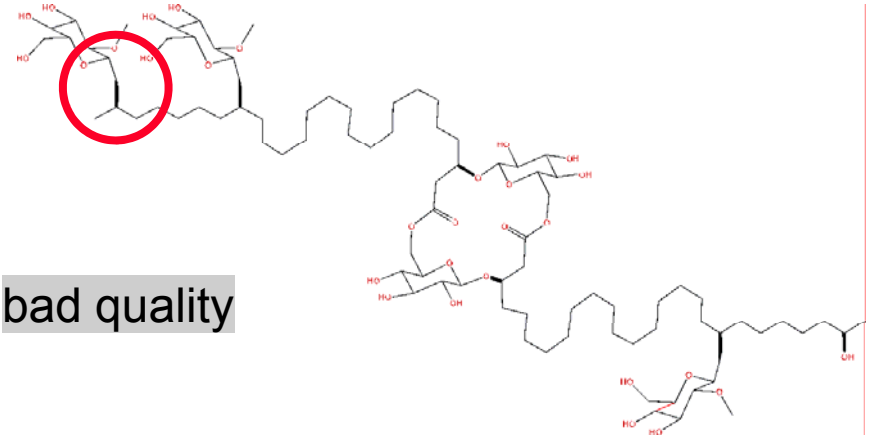
close connections



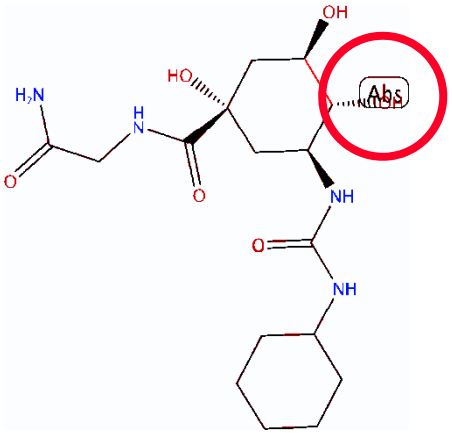


SMILES and alike

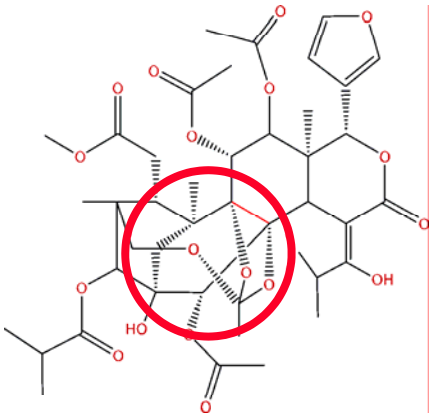
Really Bad...



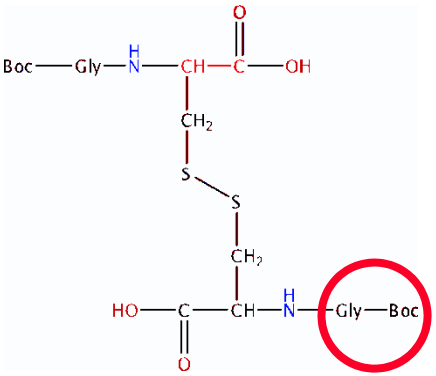
bad quality



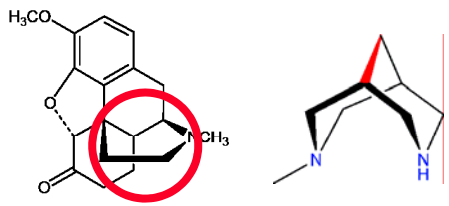
overlaps



complicated bridged rings

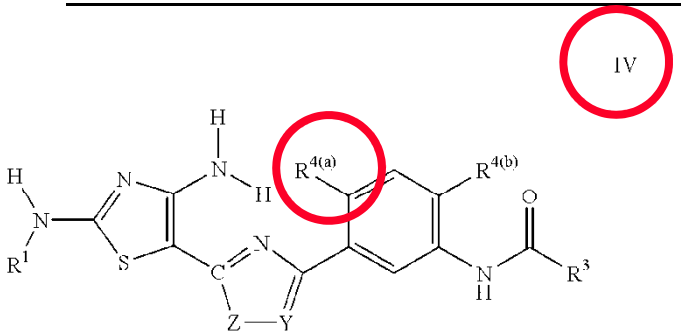


amino acid orientation

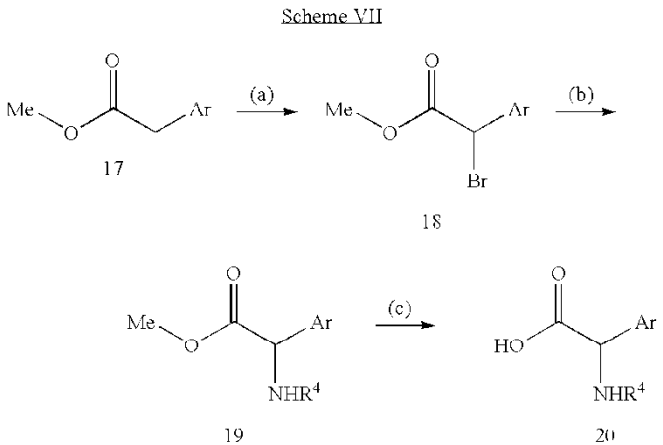


thick bridges

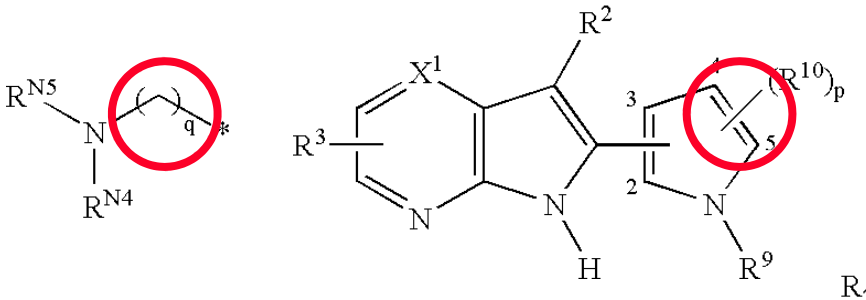
Patent Images...



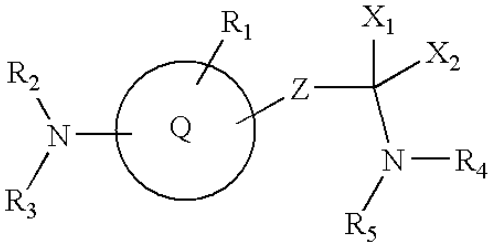
R-groups and captions [✓] SDF



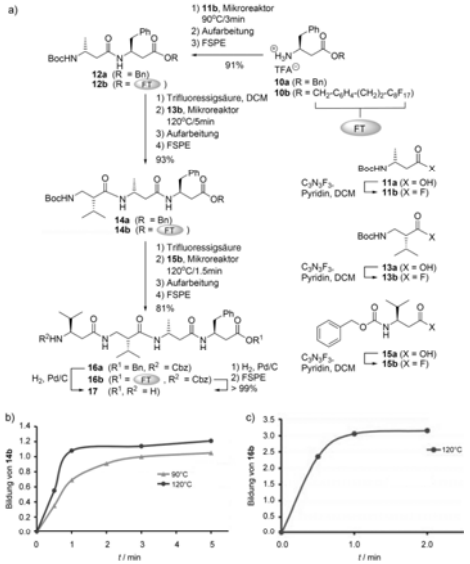
reaction schema [?] RDF



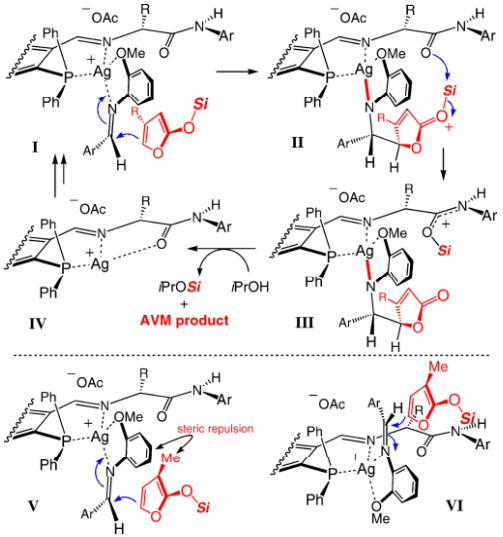
Markush structures [×] file format



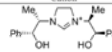
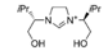
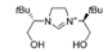
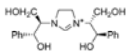
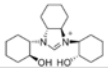
Most wanted...



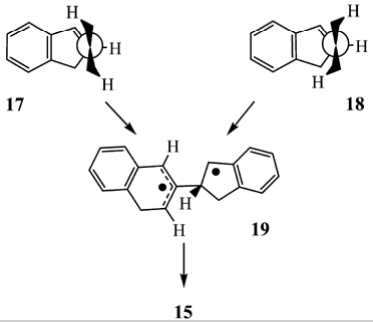
multimodal information



3D depictions

Entry	Diamine	Cation	Anion	Salt	Yield (%)
1	5		BF_4^-	10A	87
2	5		NTf_2^-	10B	93
3	5		$B[C_6H_3(CF_3)_2]_4^-$	10C	85
4	6		BF_4^-	11A	94
5	6		NTf_2^-	11B	52
6	7		BF_4^-	12A	88
7	7		NTf_2^-	12B	59
8	8		BF_4^-	13A	98
9	8		$B[C_6H_3(CF_3)_2]_4^-$	13C	80
10	9		BF_4^-	14A	93

QSAR tables

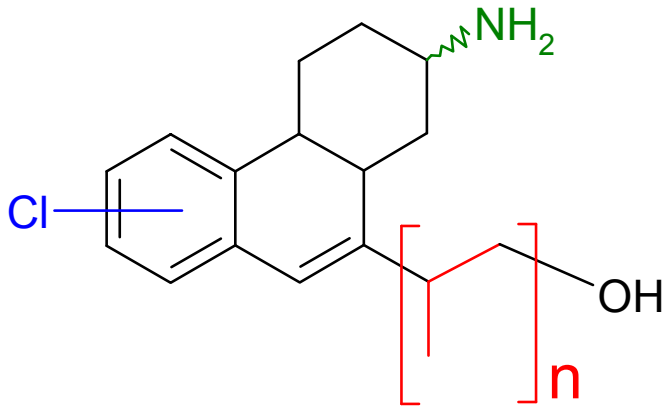


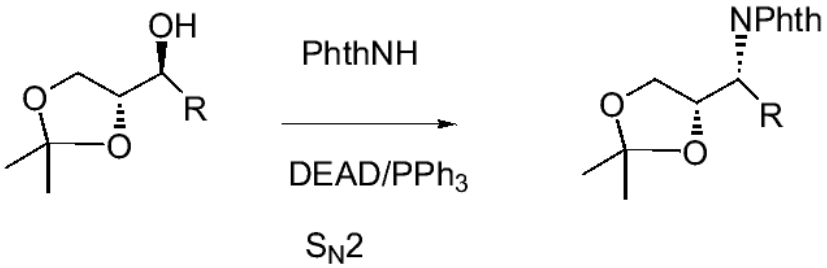
new graphic primitives

Future Work

Incompletely-defined substances and reaction schemes:

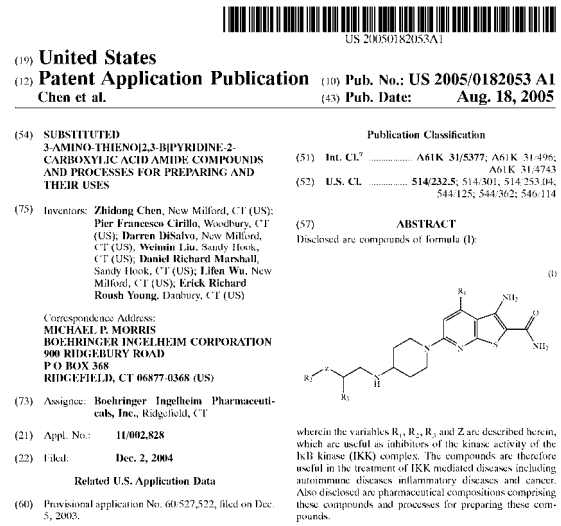
- unknown stereochemistry
- unknown attachment position
- unknown repetition
- sets of structures with common features
- structures with R-groups
- reaction syntax





(1)

A Glimpse at the Future: Multi Modal Extraction From Patents



US 20050182053A1

(19) United States
(12) Patent Application Publication
Chen et al.
(10) Pub. No.: US 2005/0182053 A1
(43) Pub. Date: Aug. 18, 2005

(54) SUBSTITUTED 3-AMINO-THIENO[2,3-B]PYRIDINE-2-CARBOXYLIC ACID AMIDE COMPOUNDS AND PROCESSES FOR PREPARING AND THEIR USES

(75) Inventors: Zhidong Chen, New Milford, CT (US); Pier Francesco Cirillo, Woodbury, CT (US); Darren DiSalvo, New Milford, CT (US); Weinin Liu, Sandy Hook, CT (US); Daniel Richard Marshall, Sandy Hook, CT (US); Lifen Wu, New Milford, CT (US); Erick Richard Roush Young, Danbury, CT (US)

Correspondence Address:
MICHAEL P. MORRIS
BOEHRINGER INGELHEIM CORPORATION
900 RIDGEBURY ROAD
P O BOX 368
RIDGEFIELD, CT 06877-0368 (US)

(73) Assignee: Boehringer Ingelheim Pharmaceuticals, Inc., Ridgefield, CT

(21) Appl. No.: 11/002,828

(22) Filed: Dec. 2, 2004

Related U.S. Application Data

(60) Provisional application No. 60/527,522, filed on Dec. 5, 2003.

Publication Classification

(51) Int. Cl.[7] A61K 31/5377; A61K 31/496; A61K 31/4743

(52) U.S. Cl. 514/232.5; 514/301; 514/253.04; 544/125; 544/362; 546/114

(57) ABSTRACT

Disclosed are compounds of formula (I):

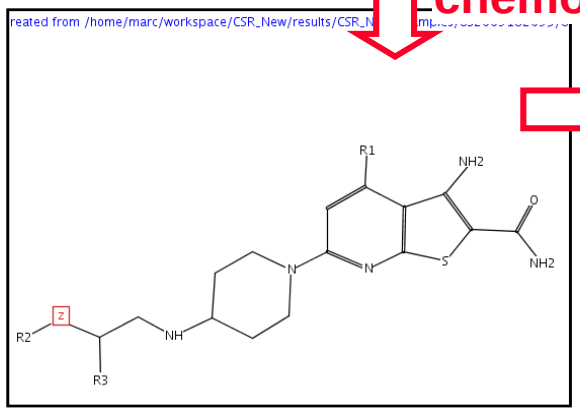
(I)

wherein the variables R_1, R_2, R_3 and Z are described herein, which are useful as inhibitors of the kinase activity of the IkB kinase (IKK) complex. The compounds are therefore useful in the treatment of IKK mediated diseases including autoimmune diseases, inflammatory diseases and cancer. Also disclosed are pharmaceutical compositions comprising these compounds and processes for preparing these compounds.

ProMiner (NER)

PatentMine by SCAI

chemoCR



PatentSpace

graphical mining

3-AMINO-THIENO[2,3-BIPYRIDINE-2-CAR
PROCESSES FOR PREPARING AND TH

RELATED APPLICATIONS

[0001] This application claims priority to U.
application No. 601527,522 filed on Dec. 5
application is also related to U.S. patent ap
Nos. 101453,175 and 101730,172.

TECHNICAL FIELD OF THE INVENTION

[0002] This invention relates to substituted 3-amino-thieno[2,3-b]pyridine-2-carboxylic acid amide compounds useful as inhibitors of the kinase activity of the IKB kinase (IKK) complex. The compounds are therefore useful in the treatment of IKK-mediated diseases including autoimmune diseases, inflammatory diseases and cancer. The invention also relates to processes for preparing such compounds and pharmaceutical compositions comprising them.

BACKGROUND OF THE INVENTION

[003] NF-KB or nuclear factor KB is a transcription factor that induces the expression of a large number of pro-inflammatory and anti-apoptotic genes. These include cytokines such as IL-1, IL-2, TNF-a and IL-6, chemokines including IL-8 and RANTES, as well as other pro-inflammatory molecules including COX-2 and cell adhesion molecules such as ICAM-I, VCAM-I, and E-selectin. The NF-KB family includes homo- and heterodimeric transcription factors composed of members of the Re1 family (see for example P. A. Baeurle and D. Baltimore, Cell, 1996,87, 13). Under resting conditions, NF-KB is present in the cytosol of cells as a complex with IKB. The IKB family of proteins

proteins
protein families
protein complex
compound
process
drug class
disease
pathways

Lessons Learned

- a generic chemoCR framework has been established
- there is and there will not be a “one-fits-all” solution
- chemoCR can be adapted and optimized (parameters, error models, image preprocessing, …)
- although we have looked into many examples, we have not seen so far all sorts of image sources (e.g. legacy of old documents)
- we will continuously improve our methods as new challenges come along

- *You* can get hands on experience on chemoCR in an *evaluation project*
- *SCAI* provides: training, installation support, bug fixing, fitting chemoCR to the data, *long term research agenda*

chemoCR
chemical compound reconstruction

http://www.scai.fraunhofer.de/chemocr

Fachhochschule Bonn-Rhein-Sieg

Tanja Fey
Carina Haupt
Prof. Dr. Ralf Thiele

Julius-Maximilians-UNIVERSITÄT WÜRZBURG

Le Thuy Bui Thi
Prof. Dr. Noltemeier

LMU Ludwig-Maximilians-Universität München

Peter Kral
Karsten Borgwardt
Prof. Dr. Hans-Peter Kriegel

ITAM INSTITUTO TECNOLÓGICO AUTÓNOMO DE MÉXICO

Miguel Alvarez
Santiago Akle Serrano
Prof. Dr. Maria Elena Algorri

b-it Bonn-Aachen International Center for Information Technology

Yuan Wang
Wei Wang
Albert Ou
Prof. Dr. M. Hofmann-Apitius

SCAI Fraunhofer Institut Algorithmen und Wissenschaftliches Rechnen

Dr. Marc Zimmermann
Dr. Christoph Friedrich
Angelika Weihermüller

InfoChem

Hans Kraut
Dr. Peter Löw

Thank you for your attention