Large Scale Evaluation of Chemical Structure Recognition

4th Text Mining Symposium in Life Sciences

October 10, 2006

SCAI

Fraunhofer Institut
Algorithmen und Wissenschaftliches Rechnen

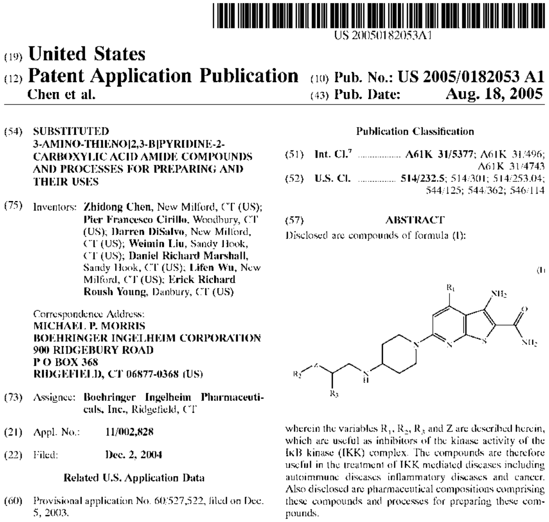
Dr. Marc Zimmermann

Overview

- Brief introduction Chemical Structure Recognition (chemOCR)
- Manual conversion of images
- Up scaling and automatisation
- Protocol database and parameter evaluation
- 2 methods of validation
- Test and benchmark data sets
- Examples, results and lessons learned

Chemical Structure Recognition – 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

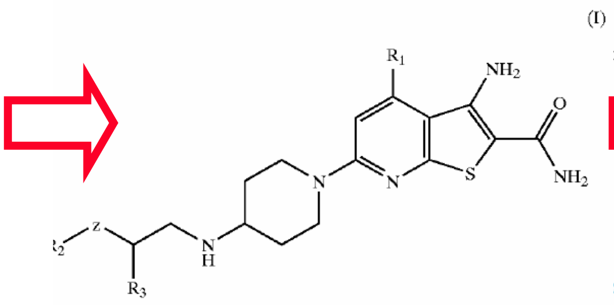
(52) **U.S. Cl.** **514/232.5**; 514/301; 514/253.04; 544/125; 544/362; 546/114

(57) **ABSTRACT**

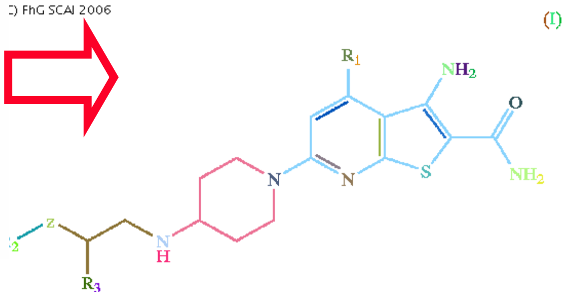
Disclosed are compounds of formula (I):

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

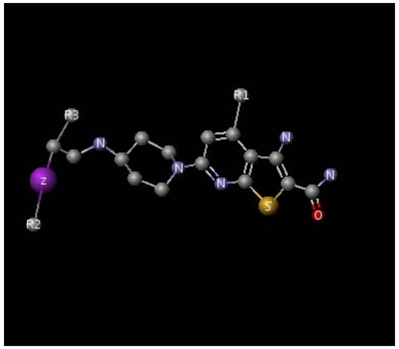


4 SDF file

```
created from
/home/marc/workspace/CSR/results/CSR/examples/US20051820
53/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
```

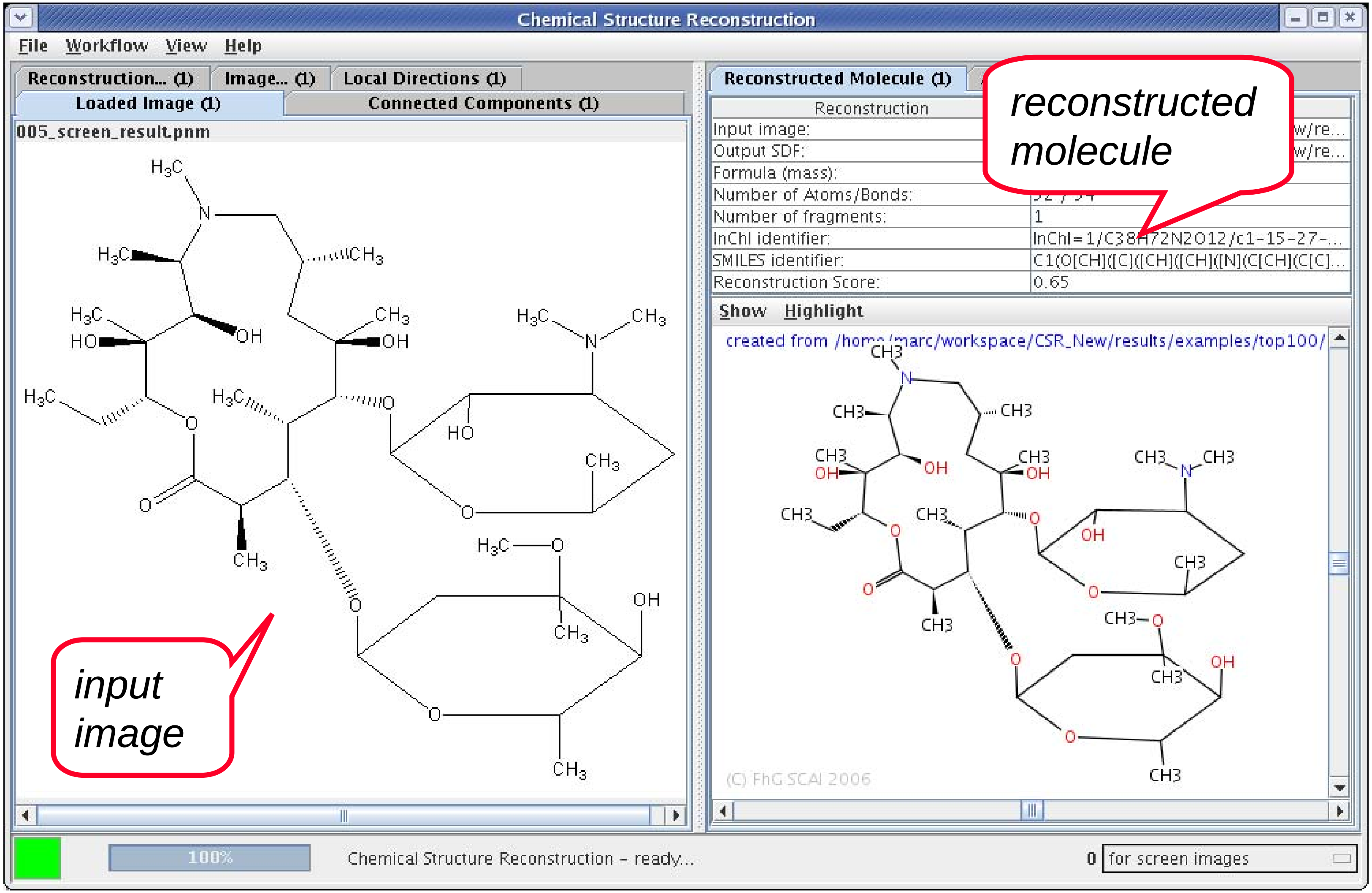
5 *in silico* Chemistry



The chemOCR Process

- 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 been implemented
- modules can be assembled into workflows

Look And Feel Of CSR


Chemical Structure Reconstruction
File Workflow View Help
Reconstruction... (1) Image... (1) Local Directions (1)
Loaded Image (1) Connected Components (1)
005_screen_result.pnm
Reconstructed Molecule (1)
Show Highlight
(C) FhG SCAI 2006
100%
0

Reconstruction	
Input image:	w/re...
Output SDF:	w/re...
Formula (mass):	
Number of Atoms/Bonds:	52 / 54
Number of fragments:	1
InChI identifier:	InChI=1/C38H72N2O12/c1-15-27-...
SMILES identifier:	C1(O[CH]([C]([CH]([CH]([N](C[CH](C[C]...
Reconstruction Score:	0.65

created from /home/marc/workspace/CSR_New/results/examples/top100/

Chemical Structure Reconstruction – ready...

for screen images

reconstructed molecule

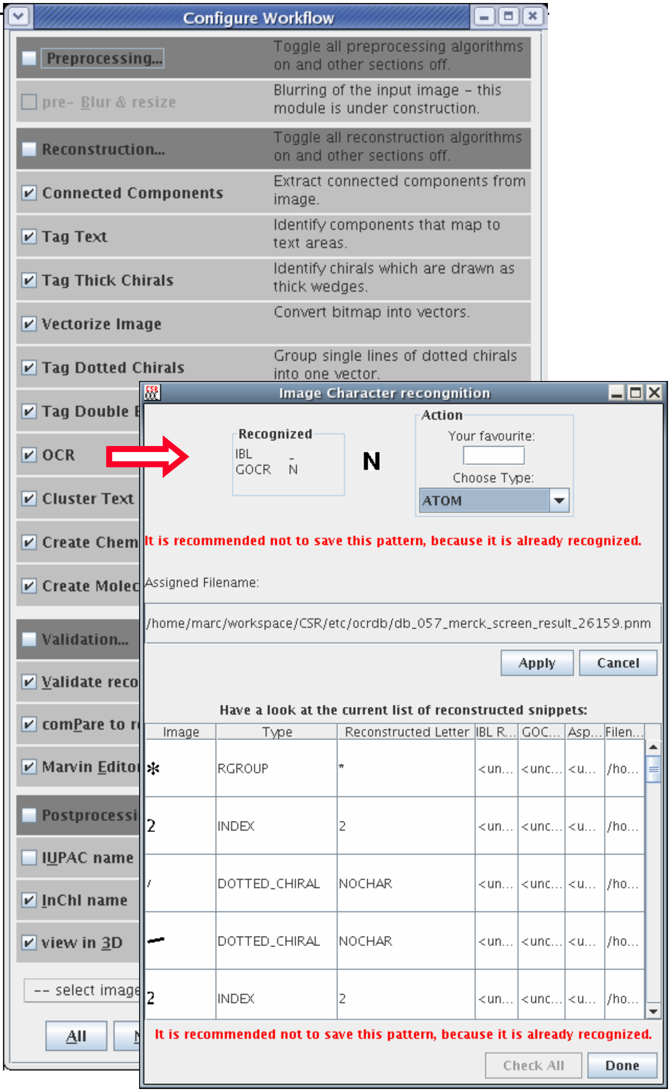
input image

The Interactive User Mode

- a graphical user interface has been developed
- the user can trigger each module separately
- there are curators and editors to interfere with the process

the main advantages are:

- full control of the process
- easier than redrawing of the image
- teaching and learning capabilities of the system



Configure Workflow

- □ Preprocessing... — Toggle all preprocessing algorithms on and other sections off.
- □ pre- Blur & resize — Blurring of the input image – this module is under construction.
- □ Reconstruction... — Toggle all reconstruction algorithms on and other sections off.
- ☑ Connected Components — Extract connected components from image.
- ☑ Tag Text — Identify components that map to text areas.
- ☑ Tag Thick Chirals — Identify chirals which are drawn as thick wedges.
- ☑ Vectorize Image — Convert bitmap into vectors.
- ☑ Tag Dotted Chirals — Group single lines of dotted chirals into one vector.
- ☑ Tag Double E
- ☑ OCR
- ☑ Cluster Text
- ☑ Create Chem
- ☑ Create Molec
- □ Validation...
- ☑ Validate reco
- ☑ comPare to r
- ☑ Marvin Editor
- □ Postprocessi
- □ IUPAC name
- ☑ InChI name
- ☑ view in 3D

-- select image

Image Character recongnition

Recognized
IBL -
GOCR N

N

Action
Your favourite:
Choose Type: ATOM

It is recommended not to save this pattern, because it is already recognized.

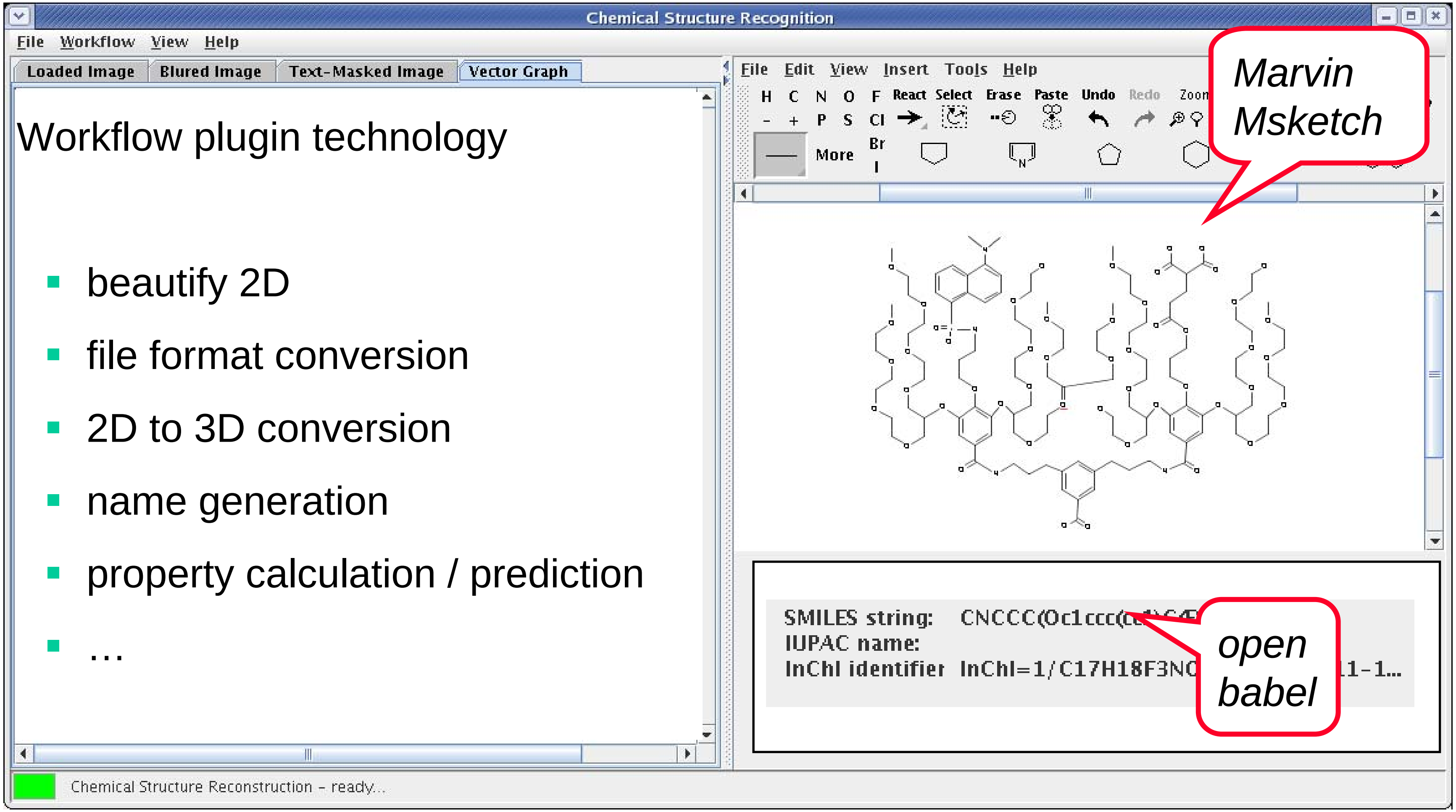
Assigned Filename:
/home/marc/workspace/CSR/etc/ocrdb/db_057_merck_screen_result_26159.pnm

Have a look at the current list of reconstructed snippets:

Image	Type	Reconstructed Letter	IBL R...	GOC...	Asp...	Filen...
✻	RGROUP	*	<un...	<unc...	<u...	/ho...
2	INDEX	2	<un...	<unc...	<u...	/ho...
/	DOTTED_CHIRAL	NOCHAR	<un...	<unc...	<u...	/ho...
–	DOTTED_CHIRAL	NOCHAR	<un...	<unc...	<u...	/ho...
2	INDEX	2	<un...	<unc...	<u...	/ho...

It is recommended not to save this pattern, because it is already recognized.

Adding New Modules – Using JAVA APIs and RPCs


Chemical Structure Recognition
File Workflow View Help
Loaded Image
Blured Image
Text-Masked Image
Vector Graph
File Edit View Insert Tools Help
H C N O F React Select Erase Paste Undo Redo Zoom
- + P S Cl
Br
More
I

Workflow plugin technology

- beautify 2D
- file format conversion
- 2D to 3D conversion
- name generation
- property calculation / prediction
- …

SMILES string: CNCCC(Oc1ccc(c

IUPAC name:

InChI identifier InChI=1/C17H18F3NO 1-1...

Chemical Structure Reconstruction - ready...

Marvin Msketch

open babel

The Distributed Batch Mode: Scaling Up the Process

setting up the batch mode:

- a specific workflow is predefined
- a suitable parameter set is chosen
- each image becomes one job which is send to one computer
- all results are assembled

advantages:

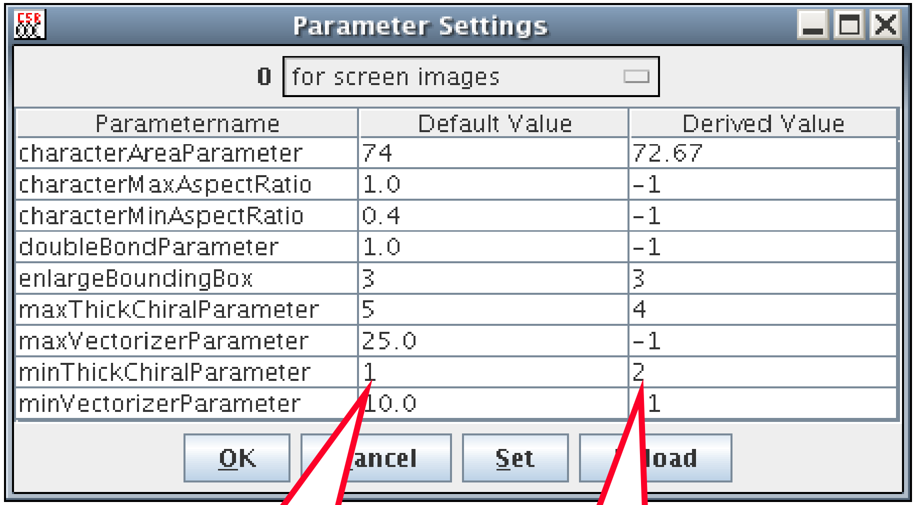
- large speed up
- less human resources
- vast number of results

⇨

disadvantages:

- no control
- errors occur
- checking the results is time consuming

Many Images ⇨ Many Parameters?



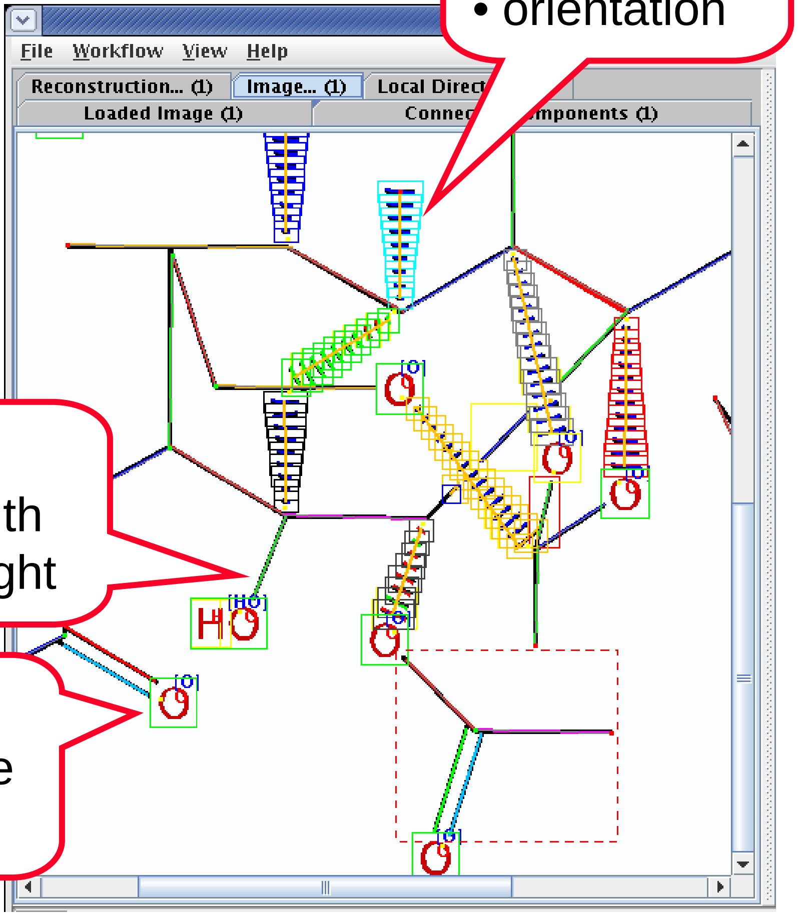
0 for screen images

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

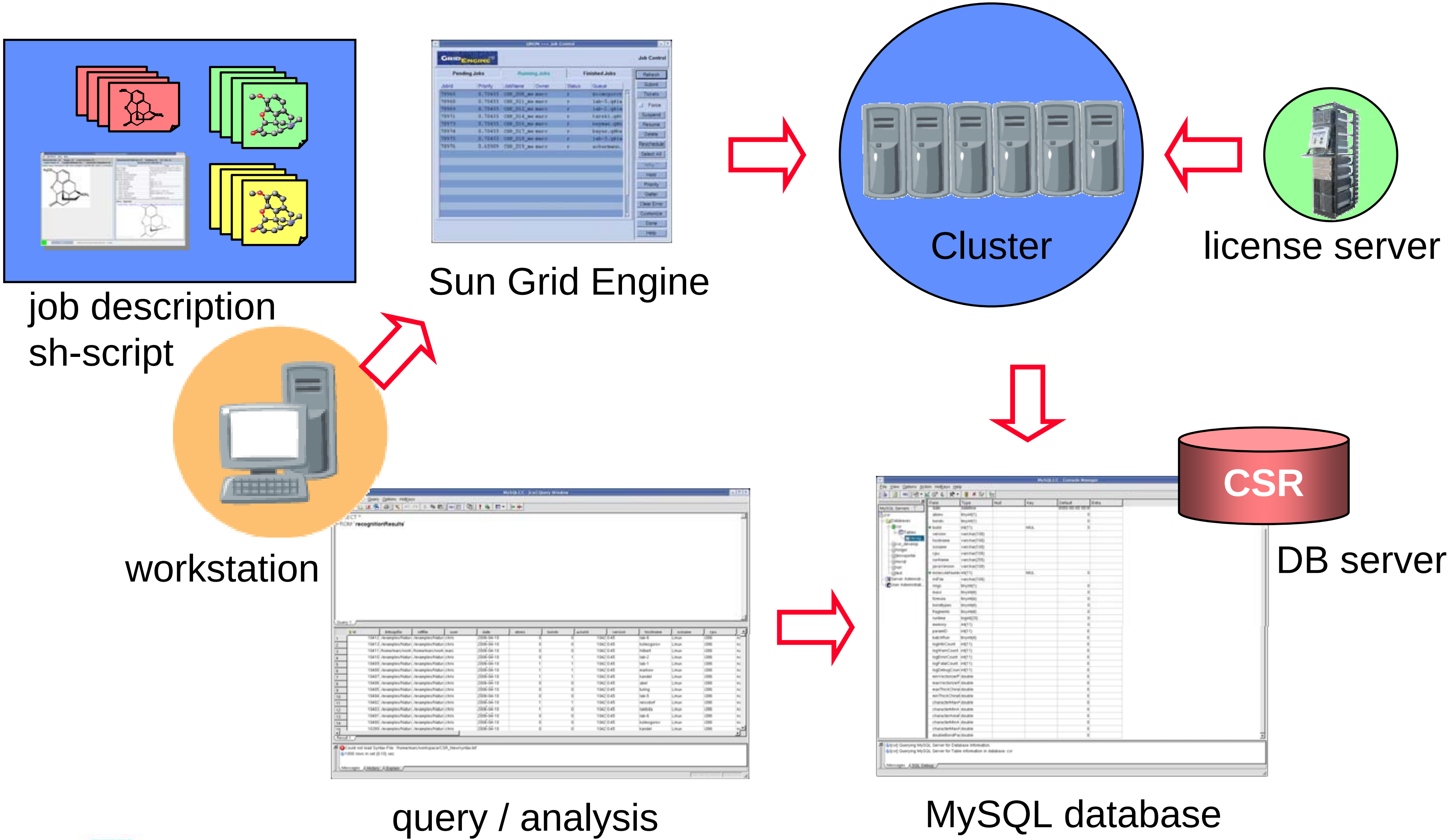


chiral:
- # segments
- orientation

bond:
- min length
- max lenght

character:
- area size
- shape

Technical Solution For Up Scaling



job description
sh-script

Sun Grid Engine

Cluster

license server

workstation

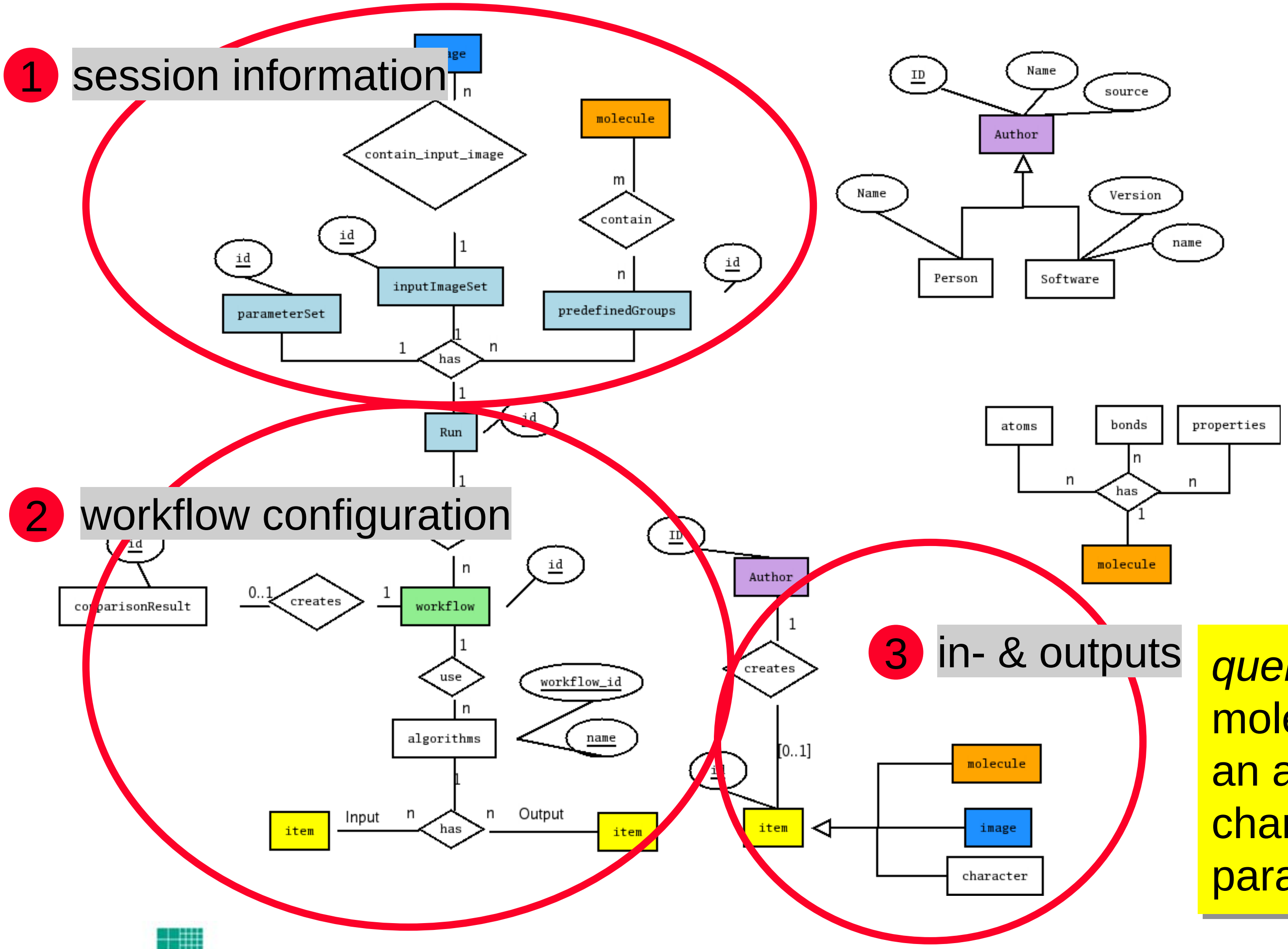
CSR

DB server

query / analysis

MySQL database

Protocol Database for the Reconstruction Process



1 session information

2 workflow configuration

3 in- & outputs

ER diagram

query: show me all molecules which have an atom error after changing the text_area parameter

Result Validation Using Training and Test Data

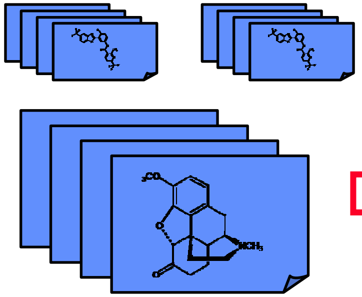
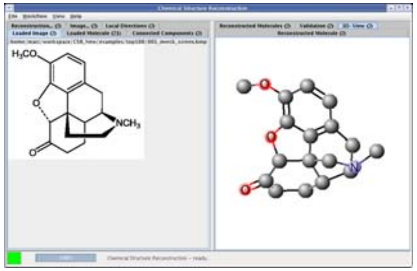
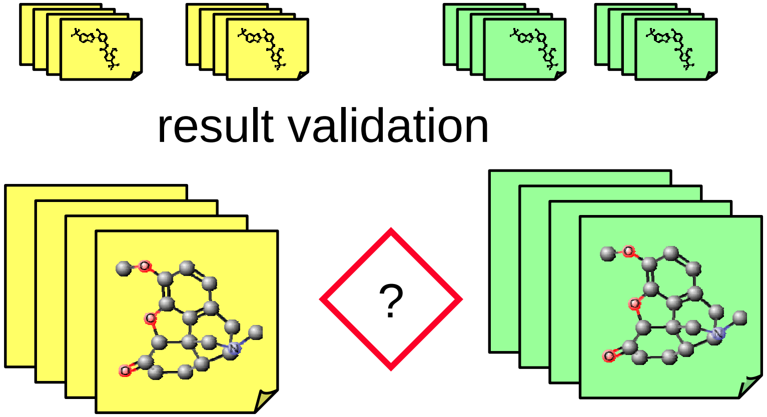


image corpora


Chemical Structure Reconstruction
File Workflow View Help
Reconstruction... (2) Image... (2) Local Directions (2)
Loaded Image (2) Loaded Molecule (21) Connected Components (2)
Reconstructed Molecules (2) Validation (2) 3D-View (2)
Reconstructed Molecule (2)
H3CO
O
NCH3
O

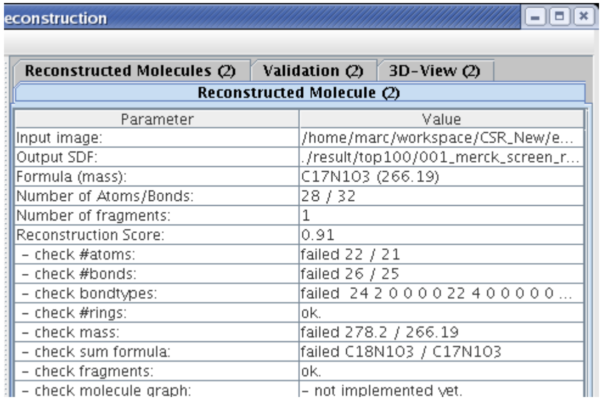
CSR reconstruction



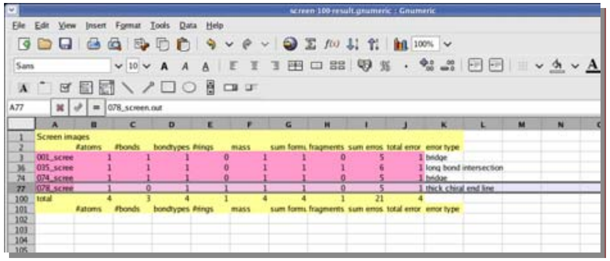
result validation

reconstruction

test molecules


econstruction
Reconstructed Molecules (2) Validation (2) 3D-View (2)
Reconstructed Molecule (2)

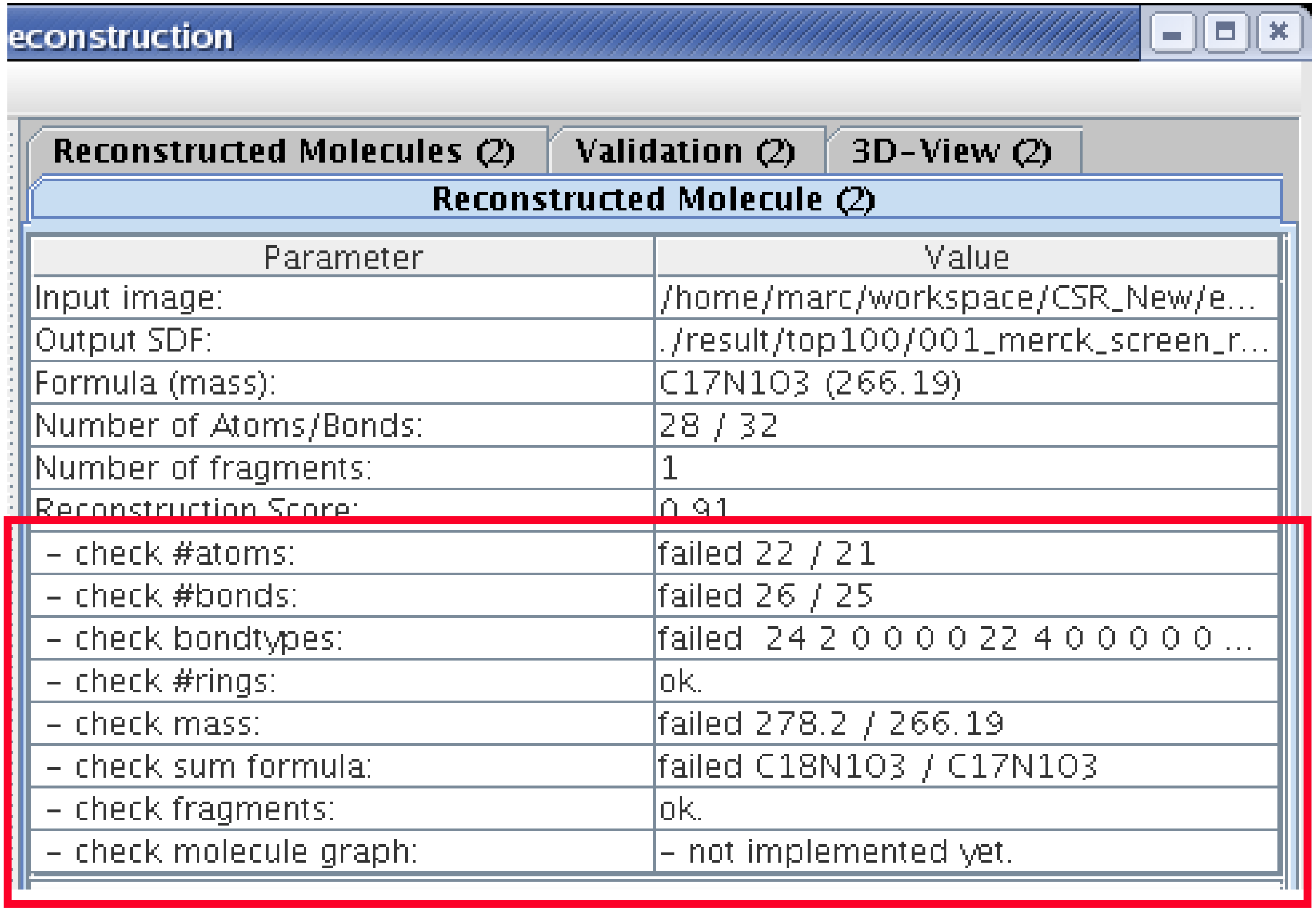
Parameter	Value
Input image:	/home/marc/workspace/CSR_New/e...
Output SDF:	./result/top100/001_merck_screen_r...
Formula (mass):	C17N1O3 (266.19)
Number of Atoms/Bonds:	28 / 32
Number of fragments:	1
Reconstruction Score:	0.91
- check #atoms:	failed 22 / 21
- check #bonds:	failed 26 / 25
- check bondtypes:	failed 24 2 0 0 0 0 22 4 0 0 0 0 0 ...
- check #rings:	ok.
- check mass:	failed 278.2 / 266.19
- check sum formula:	failed C18N1O3 / C17N1O3
- check fragments:	ok.
- check molecule graph:	- not implemented yet.


screen 100 result.gnumeric : Gnumeric
File Edit View Insert Format Tools Data Help
Sans 10
A77 078_screen.out

	A	B	C	D	E	F	G	H	I	J	K
1	Screen images										
2		#atoms	#bonds	bondtypes	#rings	mass	sum form.	fragments	sum erros	total error	error type
3	001_scree	1	1	1	0	1	1	0	5	1	bridge
36	035_scree	1	1	1	0	1	1	1	6	1	long bond intersection
74	074_scree	1	1	1	0	1	1	0	5	1	bridge
77	078_scree	1	0	1	1	1	1	0	5	1	thick chiral end line
100	total	4	3	4	1	4	4	1	21	4	
101		#atoms	#bonds	bondtypes	#rings	mass	sum form.	fragments	sum erros	total error	error type

result analysis

Validation Classes – A Closer Look



Reconstructed Molecule (2)

Parameter	Value
Input image:	/home/marc/workspace/CSR_New/e...
Output SDF:	./result/top100/001_merck_screen_r...
Formula (mass):	C17N1O3 (266.19)
Number of Atoms/Bonds:	28 / 32
Number of fragments:	1
Reconstruction Score:	0.91
- check #atoms:	failed 22 / 21
- check #bonds:	failed 26 / 25
- check bondtypes:	failed 24 2 0 0 0 0 22 4 0 0 0 0 0 ...
- check #rings:	ok.
- check mass:	failed 278.2 / 266.19
- check sum formula:	failed C18N1O3 / C17N1O3
- check fragments:	ok.
- check molecule graph:	- not implemented yet.

validation test

reconstruction / test 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

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

weights *w* can be set by regression analysis

Fraunhofer Institut Algorithmen und Wissenschaftliches Rechnen SCAI

Established Error Classes

chemical knowledge bases

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

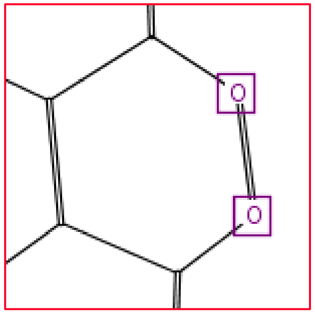
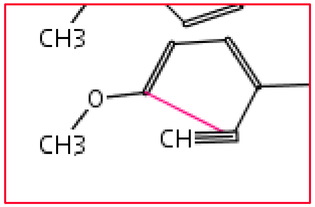


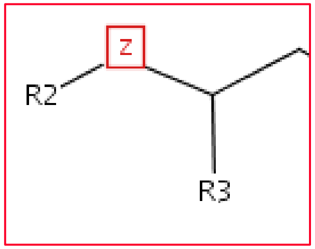
image properties

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



instance based machine learning (IBL)

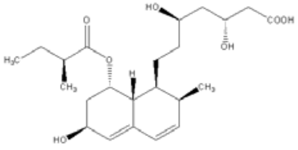
- atom and bond distributions
- Lipinski score (i.e. drug like)

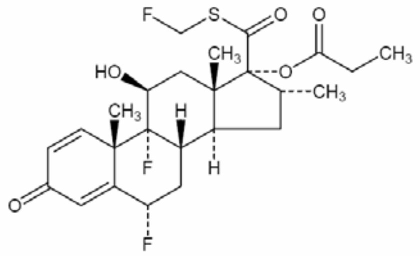


The Results – Current Status

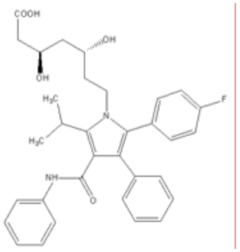
test sets: 2005 2006 2007

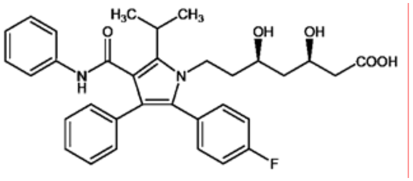
100x 100x 7600x 750,000x

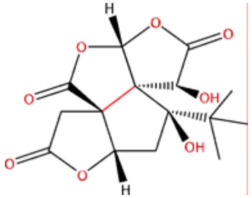




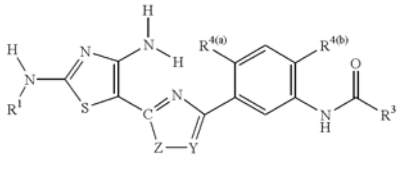
<u>measurement:</u>
correct: input identical to output
incorrect: more than one error







IV



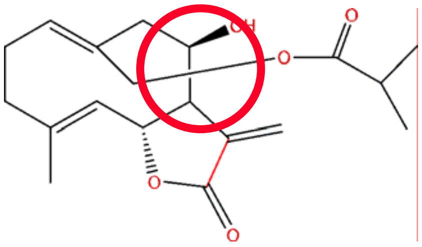
thin & clean > 95%

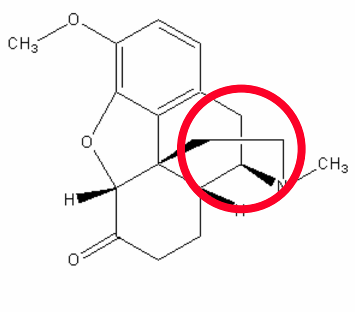
thick & clean > 75%

challenging mixture > 55%

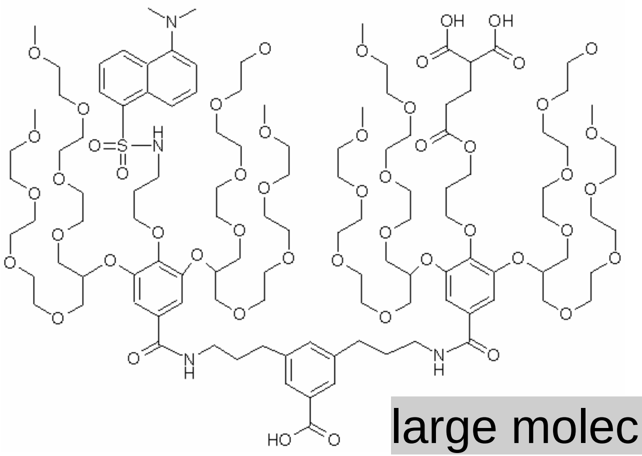
patents > ??%

Not Too Bad...

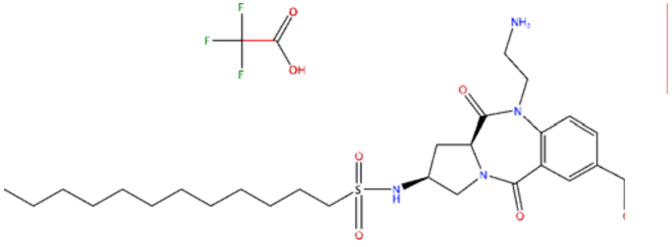


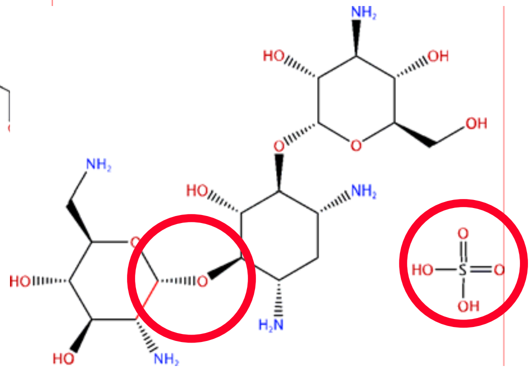


simple bridges



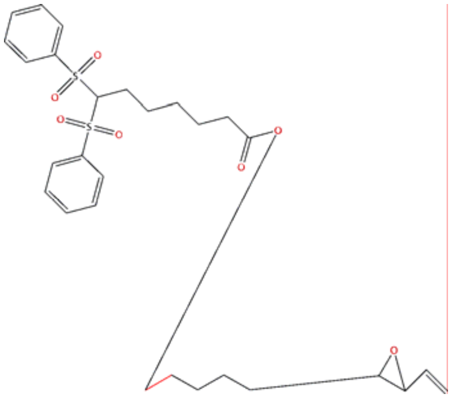
large molecules

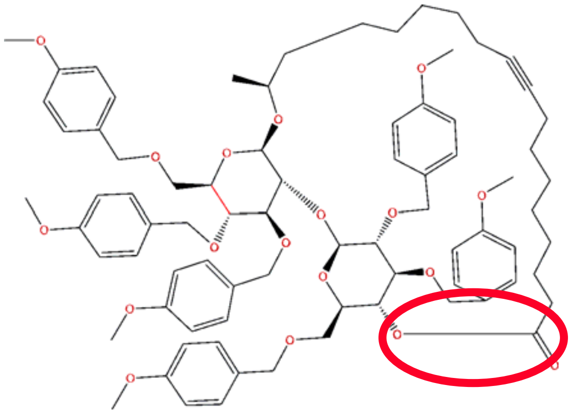




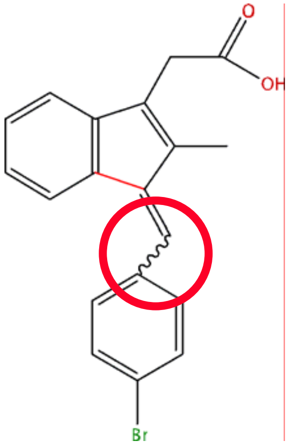
salts and chirals

Questionable…

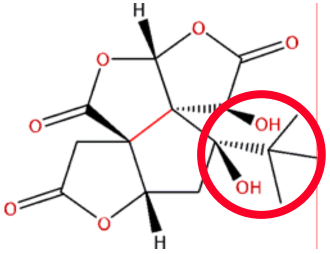


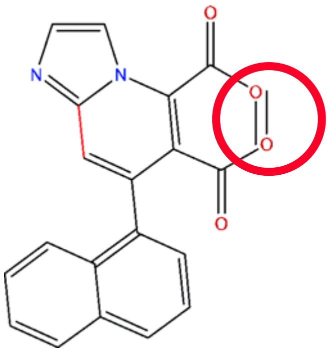


extreme long bonds

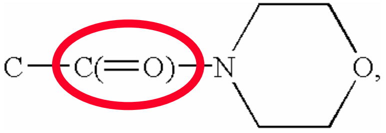


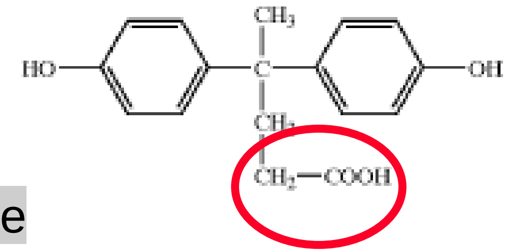
new bond types





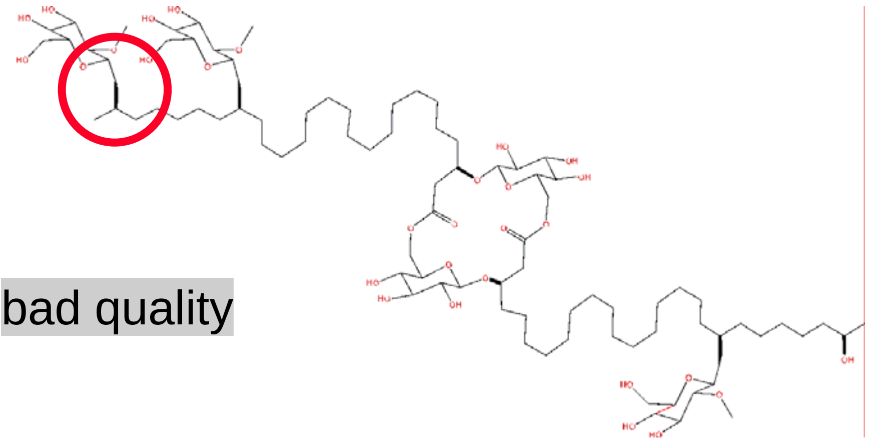
close connections



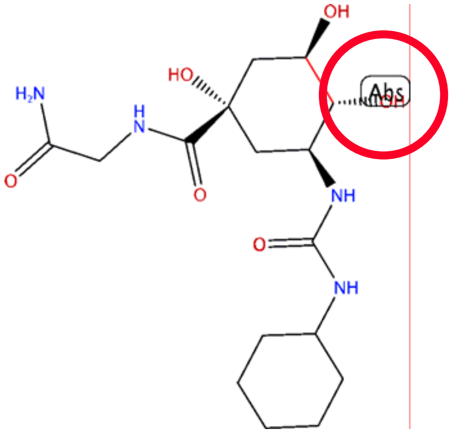


SMILES and alike

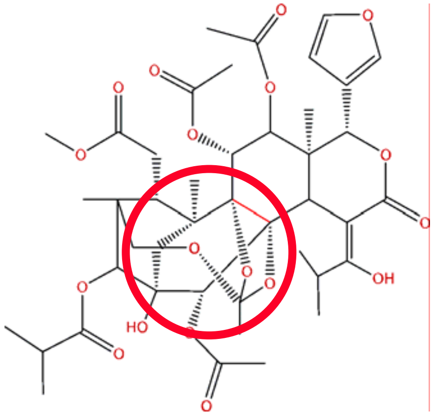
Really Bad...



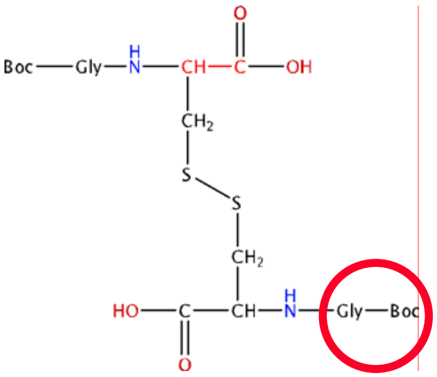
bad quality



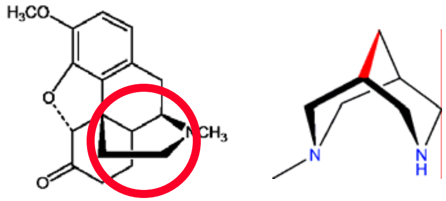
overlaps



complicated bridged rings

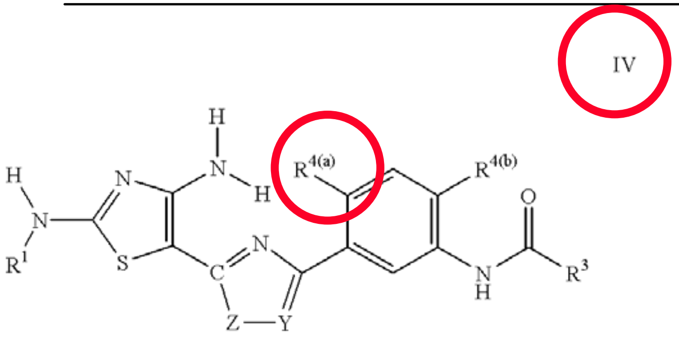


amino acid orientation

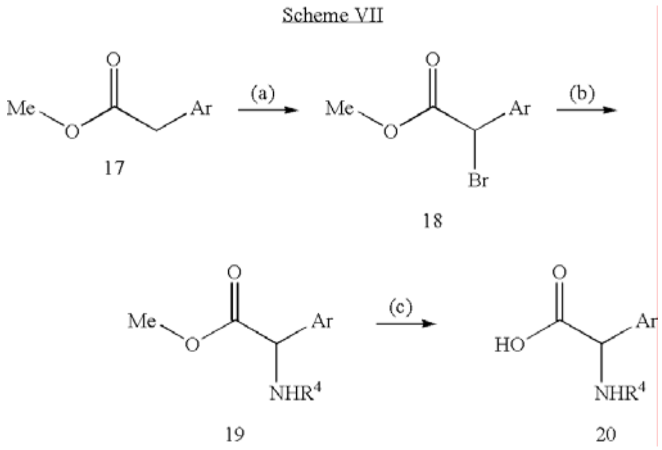


thick bridges

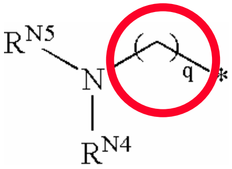
Patent Images...

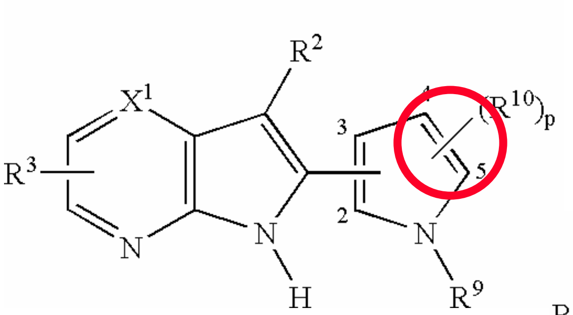


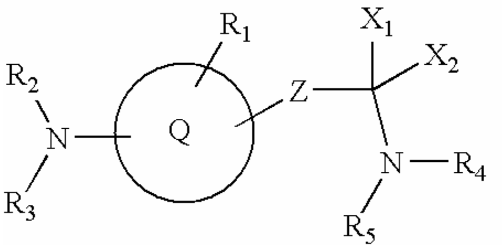
R-groups and captions [✓] SDF



reaction schema [?] RDF



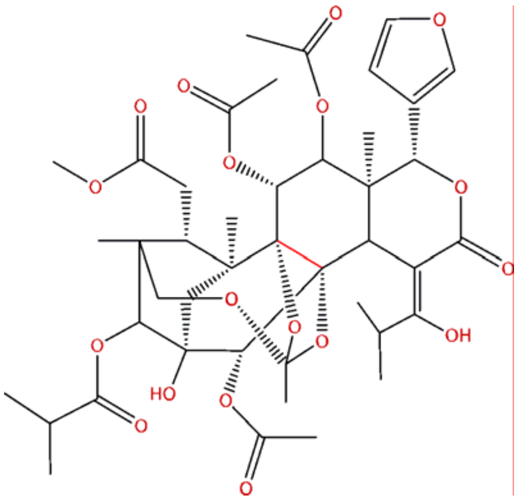




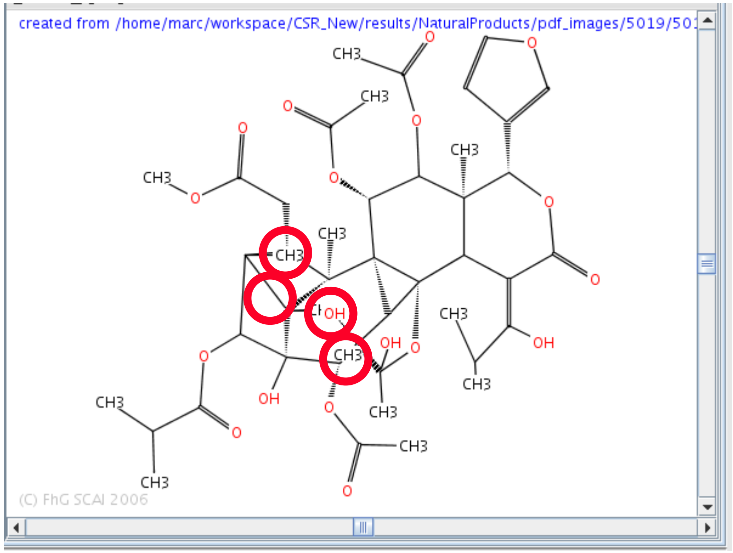
Markush structures [×] file format

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



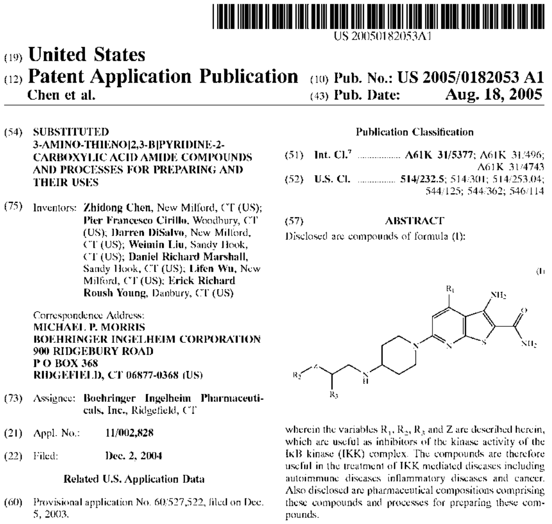
to be tolerated?



created from /home/marc/workspace/CSR_New/results/NaturalProducts/pdf_images/5019/501

(C) FhG SCAI 2006

A Glimpse at the Future: Multi Modal Extraction From Patents



US 20050182053A1

(19) **United States**
(12) **Patent Application Publication** (10) **Pub. No.: US 2005/0182053 A1**
Chen et al. (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

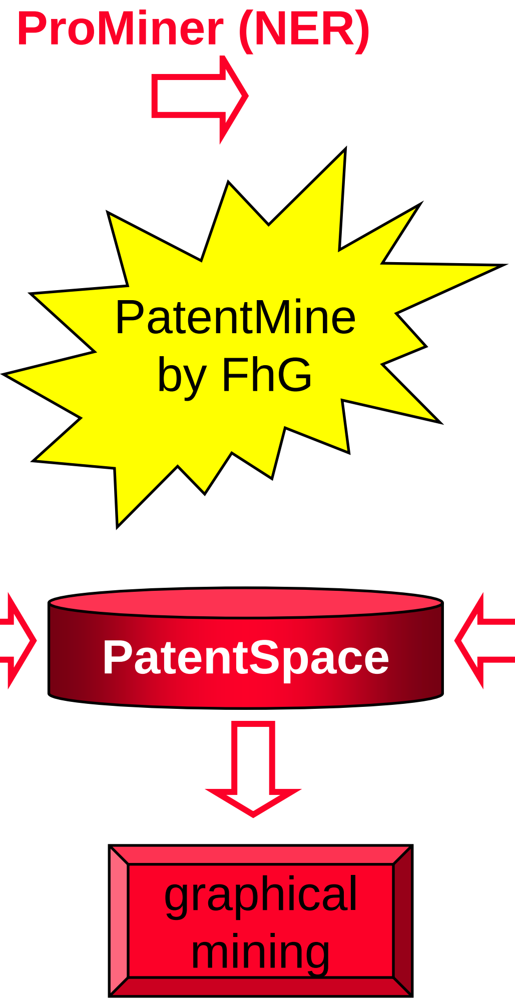
(51) **Int. Cl.**^{7} **A61K 31/5377**; A61K 31/496; A61K 31/4743
(52) **U.S. Cl.** **514/232.5**; 514/301; 514/253.04; 544/125; 544/362; 546/114

(57) **ABSTRACT**

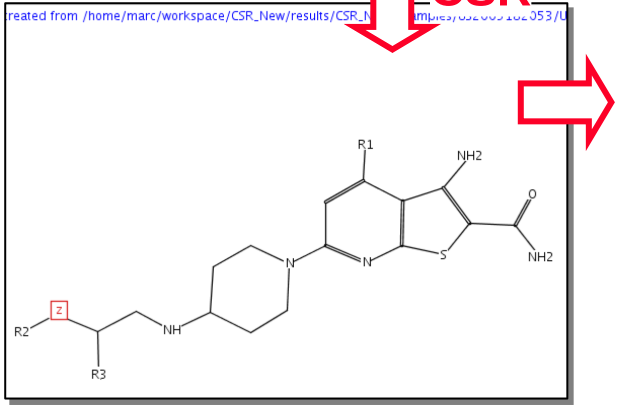
Disclosed are compounds of formula (I):

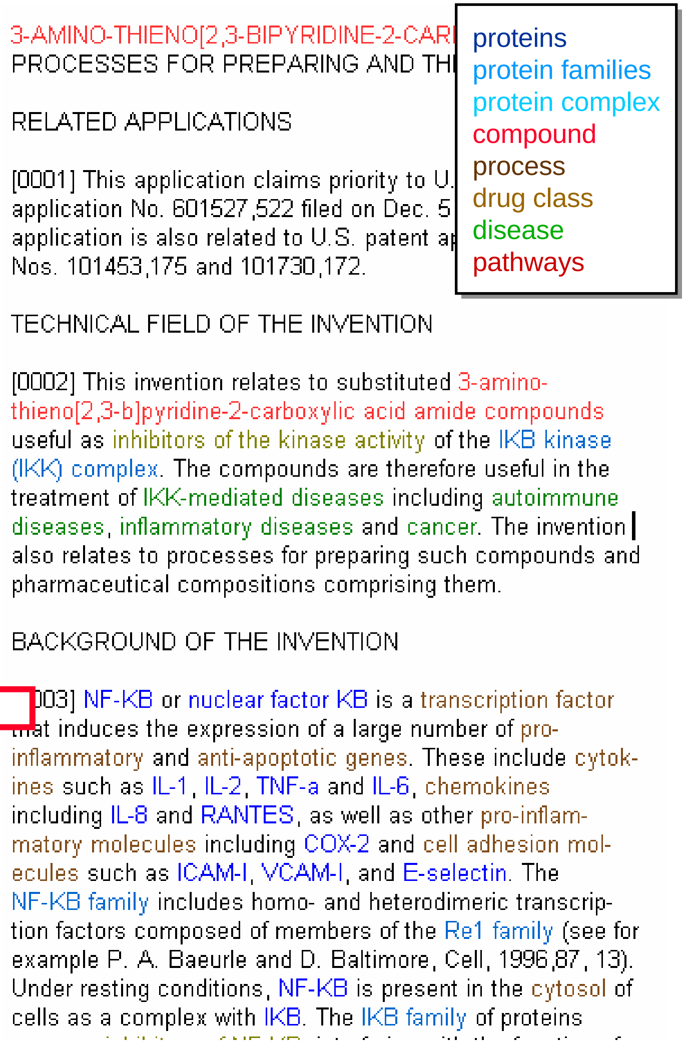
(I)

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.



CSR


reated from /home/marc/workspace/CSR_New/results/CSR_N... mples/US20050182053/U



3-AMINO-THIENO[2,3-B]PYRIDINE-2-CARI
PROCESSES FOR PREPARING AND THI

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

[0003] 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

Lessons Learned

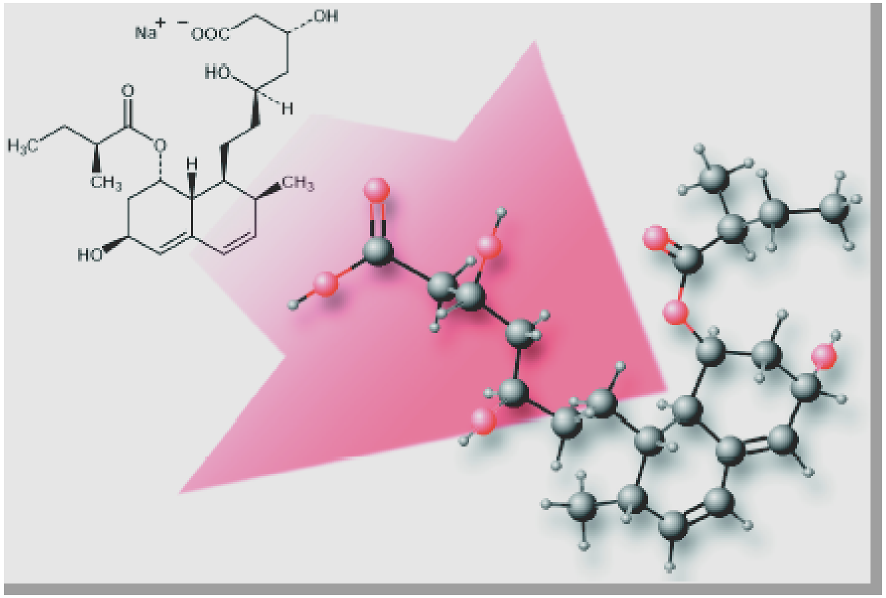
- a generic chemOCR framework has been established
- there is and there will not be a “one-fits-all” solution
- CSR can be adapted and optimized (parameters, error models, image preprocessing, …)
- although we have looked into many examples, we have not seen so far all sorts of image sources (e.g. legacy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

⇨

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

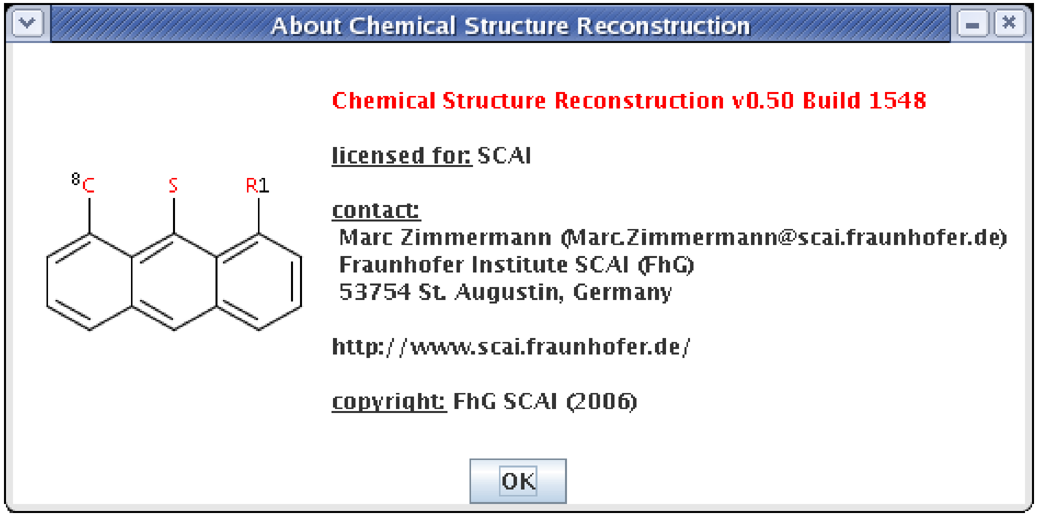
The Team (in the order of appearance)

- Marc Zimmermann*
- Tanja Fey
- Le Thuy Bui Thi
- Christoph Friedrich*
- Yuan Wang
- Maria-Elena Algorri*
- Miguel Alvarez
- Angelika Weihermüller*
- Wei Wang
- Peter Kral*
- Carina Haupt*



*) currently improving CSR

CSR Online Demo Available During The Break



Chemical Structure Reconstruction v0.50 Build 1548

licensed for: SCAI

contact:
Marc Zimmermann (Marc.Zimmermann@scai.fraunhofer.de)
Fraunhofer Institute SCAI (FhG)
53754 St. Augustin, Germany

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

copyright: FhG SCAI (2006)

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.