chemoCR
chemical compound
reconstruction

From Bitmap Images to Connection Tables

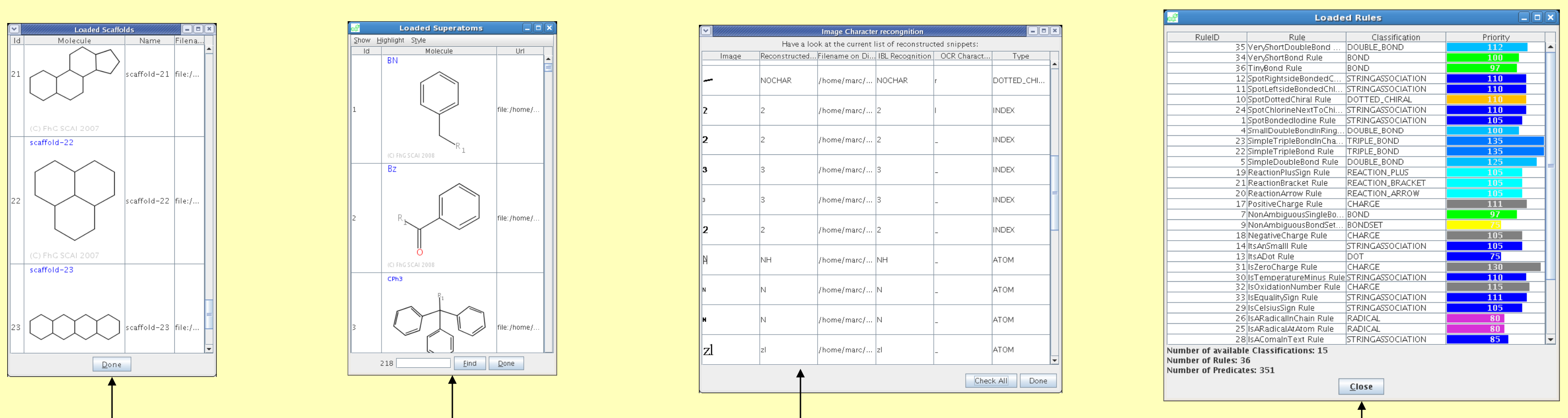
chemoCR™, a system for the reconstruction of chemical information from chemical structure depictions. This tool has been developed to make one of the largest sources of chemical information accessible: the information communicated through images containing chemical structure depictions. Although communication of chemical information through images is quite common amongst chemists, the information contained in images could not be used by machines. chemoCR™ solves this problem by reconstructing chemical information from images.

SCAI
Fraunhofer Institute
Algorithms and
Scientific Computing

chemoCR Workflow:

- I Check Configuration
- II Load Input
- III Start Reconstruction Workflow
- IV Post-process Molecule

I



Known organic scaffolds

Known super atoms and aliases

Trained characters of the OCR

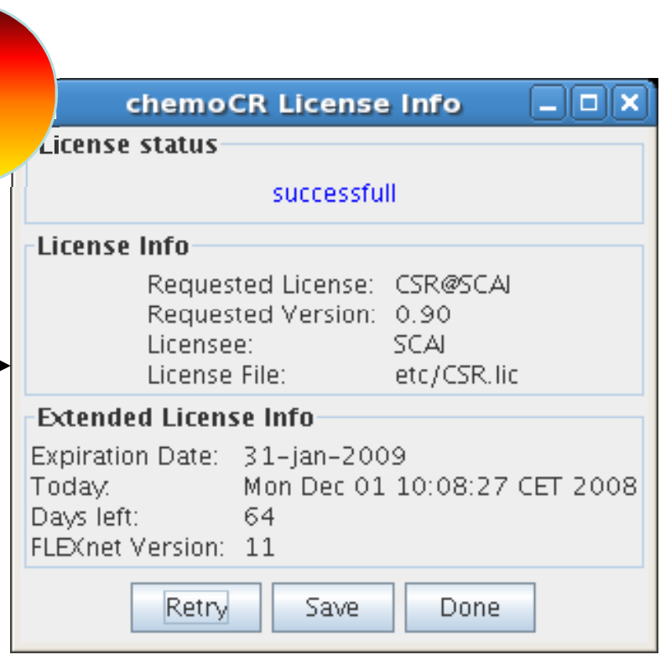
Loaded Expert Rules

Number of available Classifications: 15
Number of Rules: 36
Number of Predicates: 351

Machine learning:

- **instance based learning**
- **knowledge bases**
- **chemical expert rules**

Check the validity of the licenses for a compute cluster



chemoCR License Info

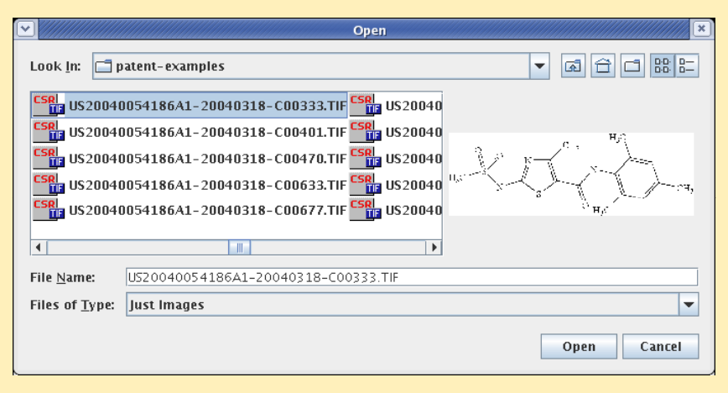
License status
successfull

License Info
Requested License: CSR@SCAI
Requested Version: 0.90
Licensee: SCAI
License File: etc/CSR.lic

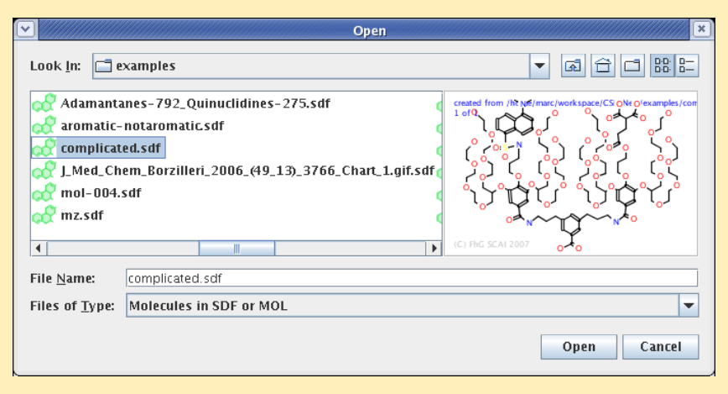
Extended License Info
Expiration Date: 31-jan-2009
Today: Mon Dec 01 10:08:27 CET 2008
Days left: 64
FLEXnet Version: 11

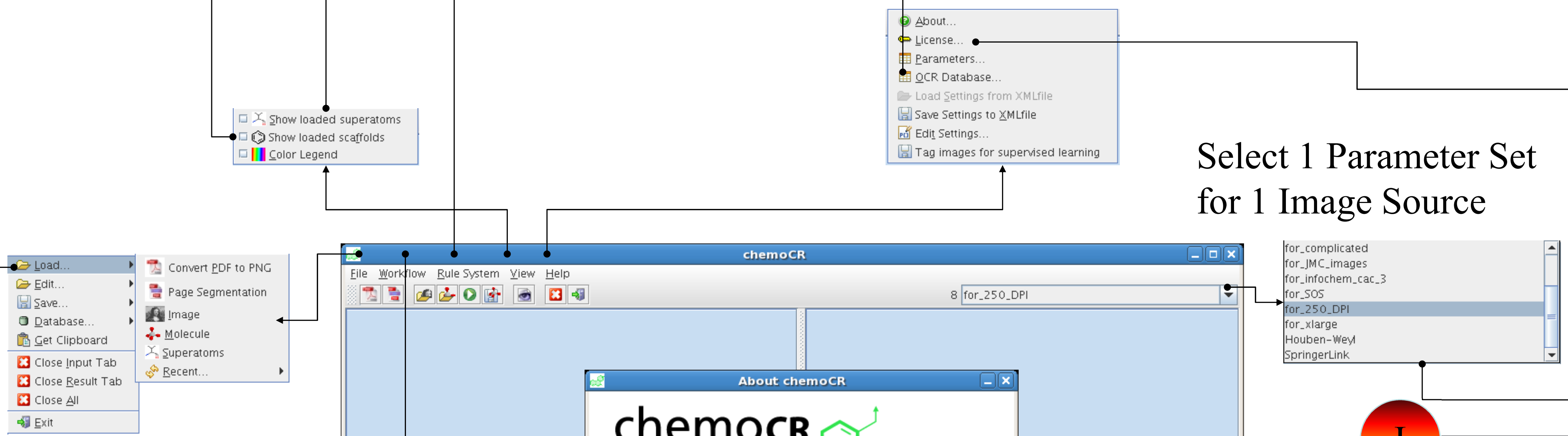
II Input and Output

All bitmap formats (GIF, PNG, BMP, ...)

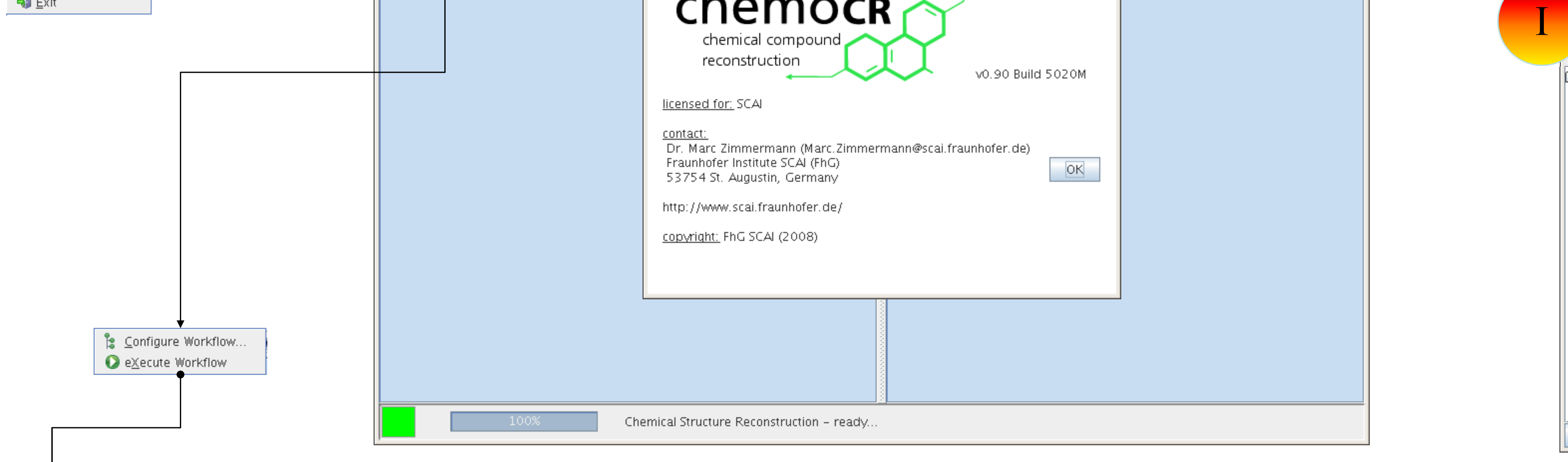


Connection tables (MOL, SDF)





Select 1 Parameter Set for 1 Image Source



About chemoCR

chemoCR
chemical compound
reconstruction

v0.90 Build 5020M

licensed for: SCAI

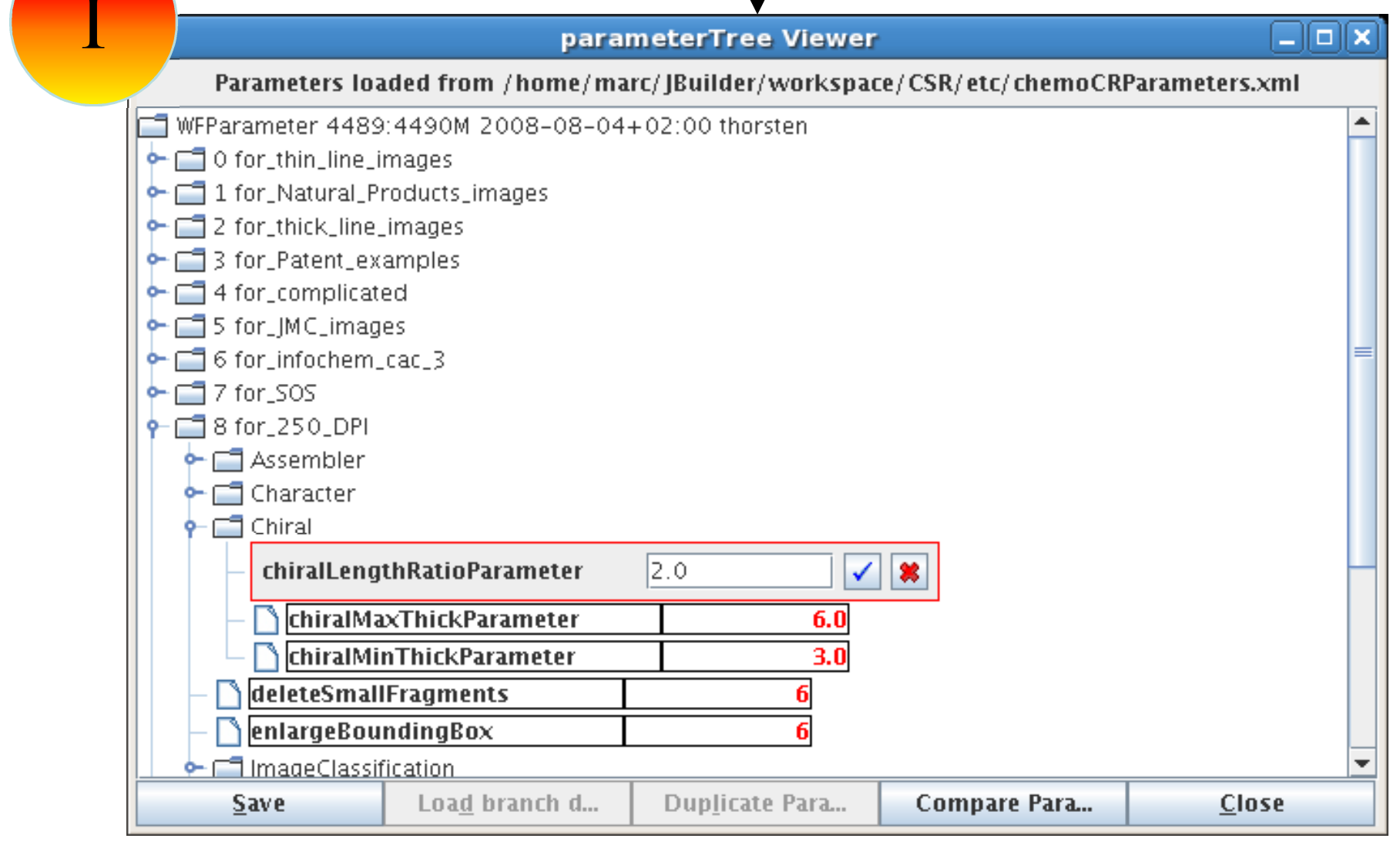
contact:
Dr. Marc Zimmermann (Marc.Zimmermann@scai.fraunhofer.de)
Fraunhofer Institute SCAI (FhG)
53754 St. Augustin, Germany

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

copyright: FhG SCAI (2008)

Chemical Structure Reconstruction – ready...

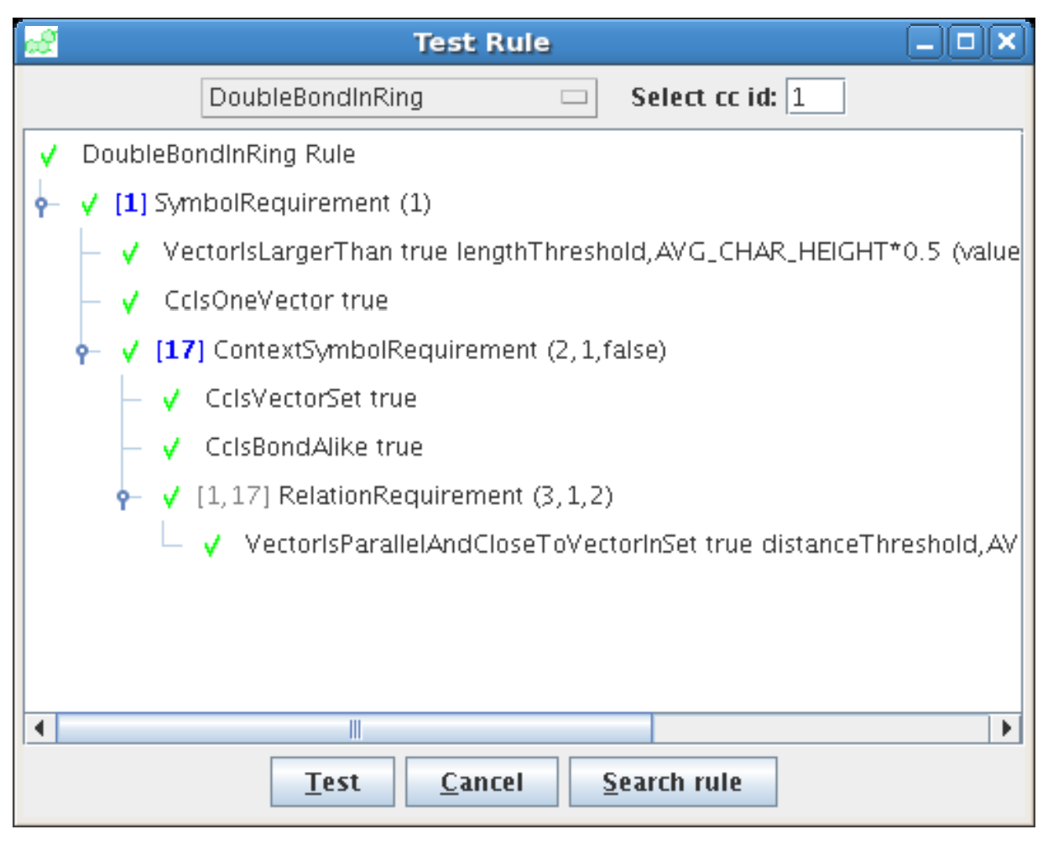
Parameters for the Extractors



parameterTree Viewer

Parameters loaded from /home/marc/JBuilder/workspace/CSR/etc/chemoCRParameters.xml

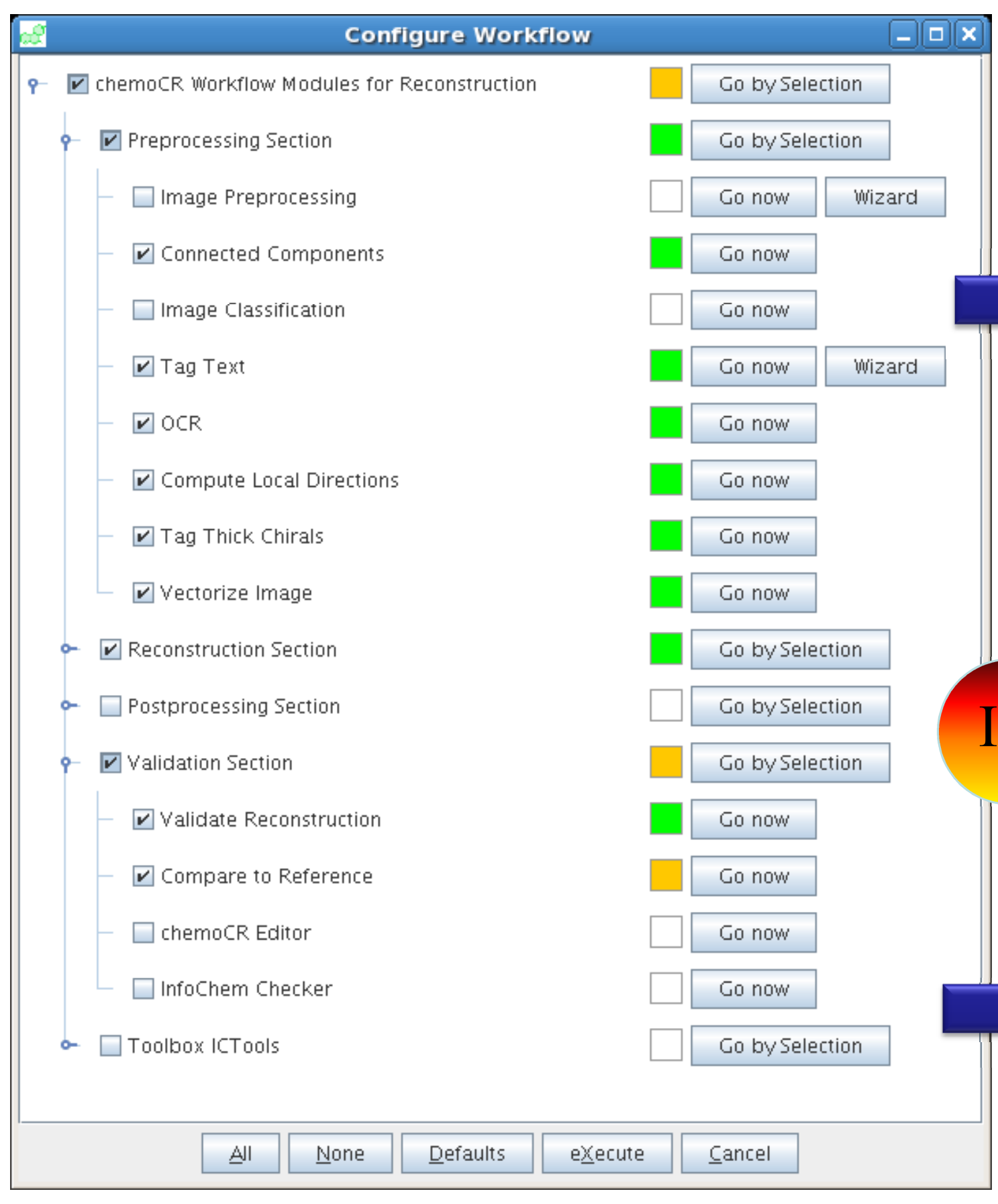
```
WFParameter 4489:4490M 2008-08-04+02:00 thorsten
  0 for_thin_line_images
  1 for_Natural_Products_images
  2 for_thick_line_images
  3 for_Patent_examples
  4 for_complicated
  5 for_JMC_images
  6 for_infochem_cac_3
  7 for_SOS
  8 for_250_DPI
    Assembler
    Character
    Chiral
      chiralLengthRatioParameter   2.0
      chiralMaxThickParameter      6.0
      chiralMinThickParameter      3.0
    deleteSmallFragments           6
    enlargeBoundingBox             6
    ImageClassification
```



Test Rule

DoubleBondInRing   Select cc id: 1

```
✓ DoubleBondInRing Rule
  ✓ [1] SymbolRequirement (1)
    ✓ VectorIsLargerThan true lengthThreshold,AVG_CHAR_HEIGHT*0.5 (value
    ✓ CcIsOneVector true
    ✓ [17] ContextSymbolRequirement (2,1,false)
      ✓ CcIsVectorSet true
      ✓ CcIsBondAlike true
      ✓ [1,17] RelationRequirement (3,1,2)
        ✓ VectorIsParallelAndCloseToVectorInSet true distanceThreshold,AV
```



Configure Workflow

- ☑ chemoCR Workflow Modules for Reconstruction 🟨 Go by Selection
  - ☑ Preprocessing Section 🟩 Go by Selection
    - ☐ Image Preprocessing ⬜ Go now Wizard
    - ☑ Connected Components 🟩 Go now
    - ☐ Image Classification ⬜ Go now
    - ☑ Tag Text 🟩 Go now Wizard
    - ☑ OCR 🟩 Go now
    - ☑ Compute Local Directions 🟩 Go now
    - ☑ Tag Thick Chirals 🟩 Go now
    - ☑ Vectorize Image 🟩 Go now
  - ☑ Reconstruction Section 🟩 Go by Selection
  - ☐ Postprocessing Section ⬜ Go by Selection
  - ☑ Validation Section 🟨 Go by Selection
    - ☑ Validate Reconstruction 🟩 Go now
    - ☑ Compare to Reference 🟨 Go now
    - ☐ chemoCR Editor ⬜ Go now
    - ☐ InfoChem Checker ⬜ Go now
  - ☐ Toolbox ICTools ⬜ Go by Selection

Select the appropriate Modules into a Workflow

III Reconstruction

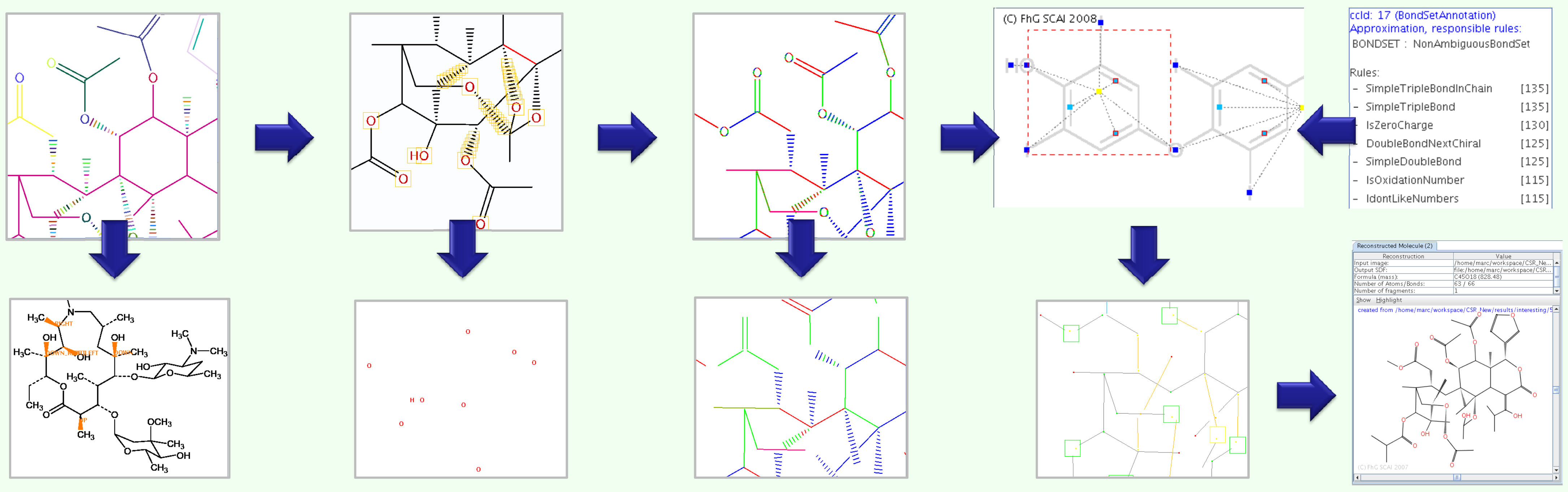
Connected Components

Character Extraction

Bitmap Textures

Chemical Expert System

Choose Rule



(C) FhG SCAI 2008

ccId: 17 (BondSetAnnotation)
Approximation, responsible rules:
BONDSET : NonAmbiguousBondSet

Rules:
- SimpleTripleBondInChain [135]
- SimpleTripleBond [135]
- IsZeroCharge [130]
- DoubleBondNextChiral [125]
- SimpleDoubleBond [125]
- IsOxidationNumber [115]
- IdontLikeNumbers [115]

Reconstructed Molecule (2)

Reconstruction	Value
Input image:	/home/marc/workspace/CSR_Ne...
Output SDF:	file:/home/marc/workspace/CSR...
Formula (mass):	C45O18 (828.48)
Number of Atoms/Bonds:	63 / 66
Number of fragments:	1

created from /home/marc/workspace/CSR_New/results/interesting/5

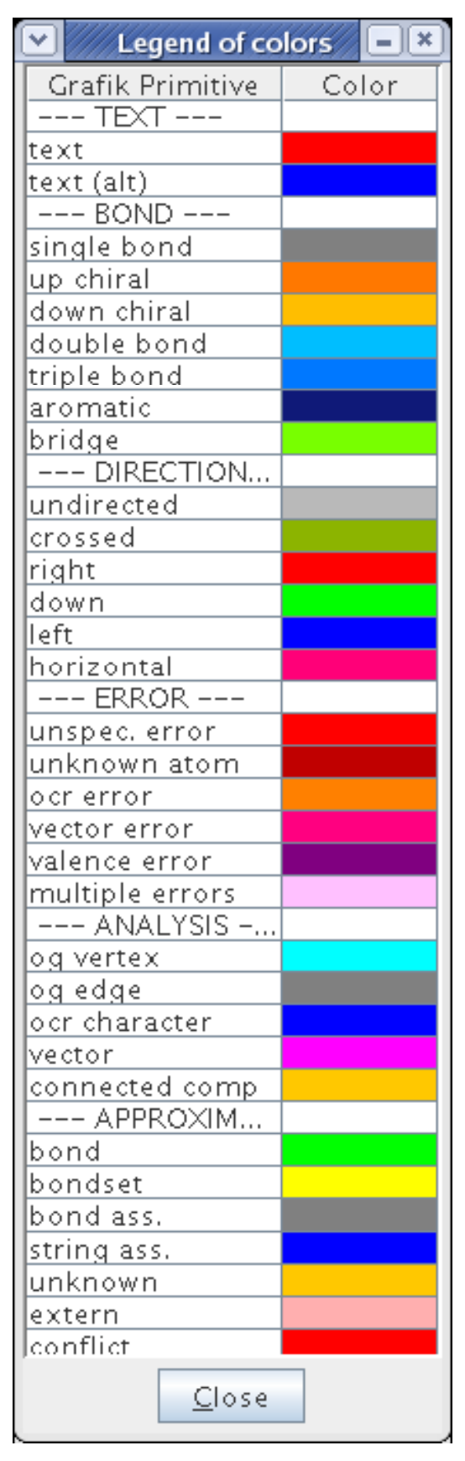
Thick Chirals

OCR

Vectorization

Chemical Graph

Create Molecule

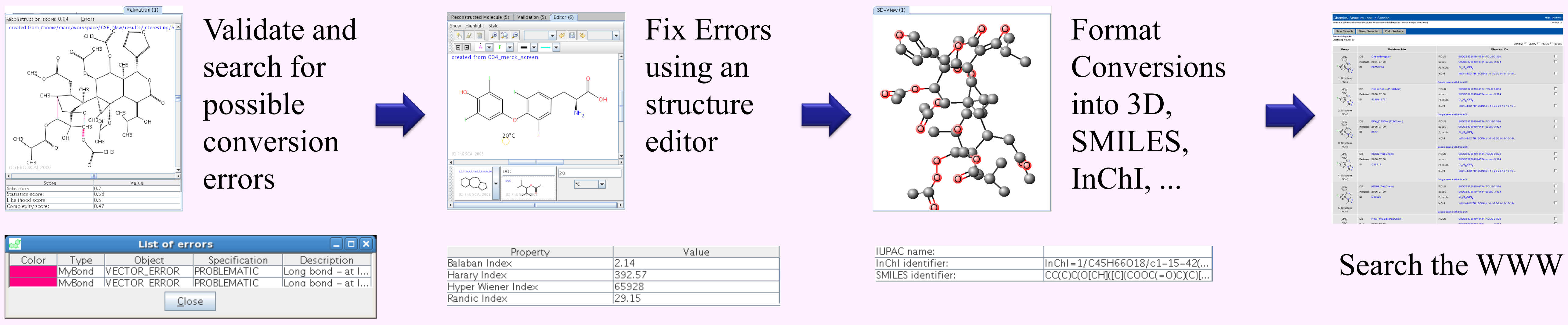


Legend of colors

Grafik Primitive	Color
--- TEXT ---	⬜
text	🟥
text (alt)	🟦
--- BOND ---	⬜
single bond	⬛
up chiral	🟧
down chiral	🟨
double bond	🟦
triple bond	🟦
aromatic	🟦
bridge	🟩
--- DIRECTION...	⬜
undirected	⬛
crossed	🟩
right	🟥
down	🟩
left	🟦
horizontal	🟥
--- ERROR ---	⬜
unspec. error	🟥
unknown atom	🟥
ocr error	🟧
vector error	🟥
valence error	🟪
multiple errors	🟪
--- ANALYSIS -...	⬜
og vertex	🟦
og edge	⬛
ocr character	🟦
vector	🟪
connected comp	🟨
--- APPROXIM...	⬜
bond	🟩
bondset	🟨
bond ass.	⬛
string ass.	🟦
unknown	🟨
extern	🟥
conflict	🟥

Color Coding

IV Validation



Validate and search for possible conversion errors

Score	Value
Subscore:	0.7
Statistics score:	0.58
Likelihood score:	0.5
Complexity score:	0.47

List of errors

Color	Type	Object	Specification	Description
🟥	MyBond	VECTOR_ERROR	PROBLEMATIC	Long bond – at l...
🟥	MyBond	VECTOR_ERROR	PROBLEMATIC	Long bond – at l...

Fix Errors using an structure editor

Property	Value
Balaban Index	2.14
Harary Index	392.57
Hyper Wiener Index	65928
Randic Index	29.15

Format Conversions into 3D, SMILES, InChI, ...

IUPAC name:	
InChI identifier:	InChI=1/C45H66O18/c1-15-42(...
SMILES identifier:	CC(C)C(O[CH]([C](COOC(=O)C)C)[...

Search the WWW

Technology

In order to solve the problem of recognizing and translating chemical structures in image documents, our chemoCR™ system combines pattern recognition techniques with supervised machine-learning concepts. The method is based on the idea of identifying from depictions the most significant semantic entities (e.g. chiral bonds, super atoms, reaction arrows…). The workflow consists of three phases: image pre-processing, semantic entity recognition, and molecule reconstruction plus validation of the result. All steps of the reconstruction process make use of chemical knowledge. The following validation process is for detection and fixing of possible reconstruction errors. The system can be adapted to different sets of input images.

Technical Specification

The chemoCR™ core functionality is based on platform independent JAVA libraries. It has been tested on UNIX™ operating systems (Fedora Linux, Sun Solaris) and on Windows XP. External tools can be easily included in the workflow. Our software can be used interactively by a graphical user interface or it can be run distributed in batch processing mode. Our benchmark test set consisting of 8000 images of natural products is processed in about 2 hours.

For contact and further details: http://www.scai.fraunhofer.de/chemocr.html