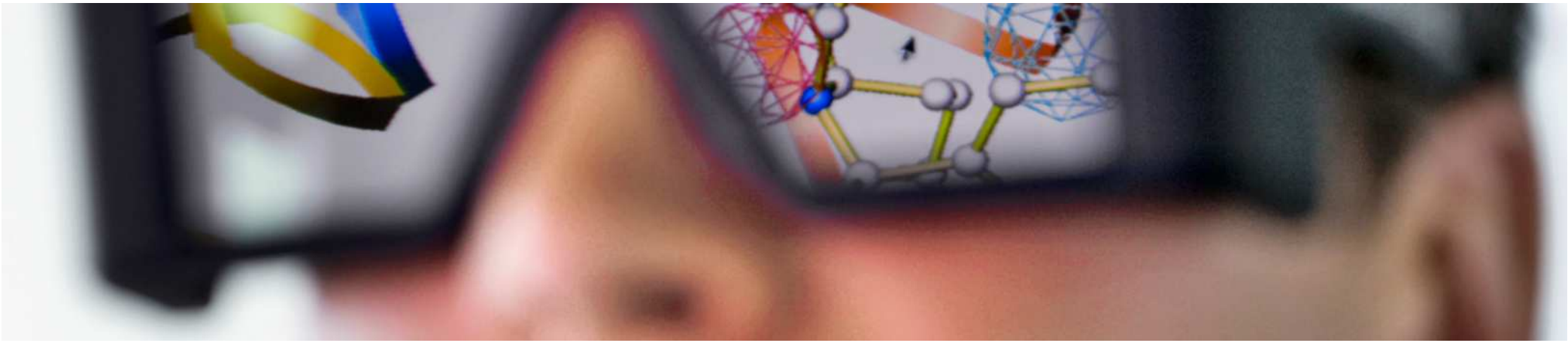
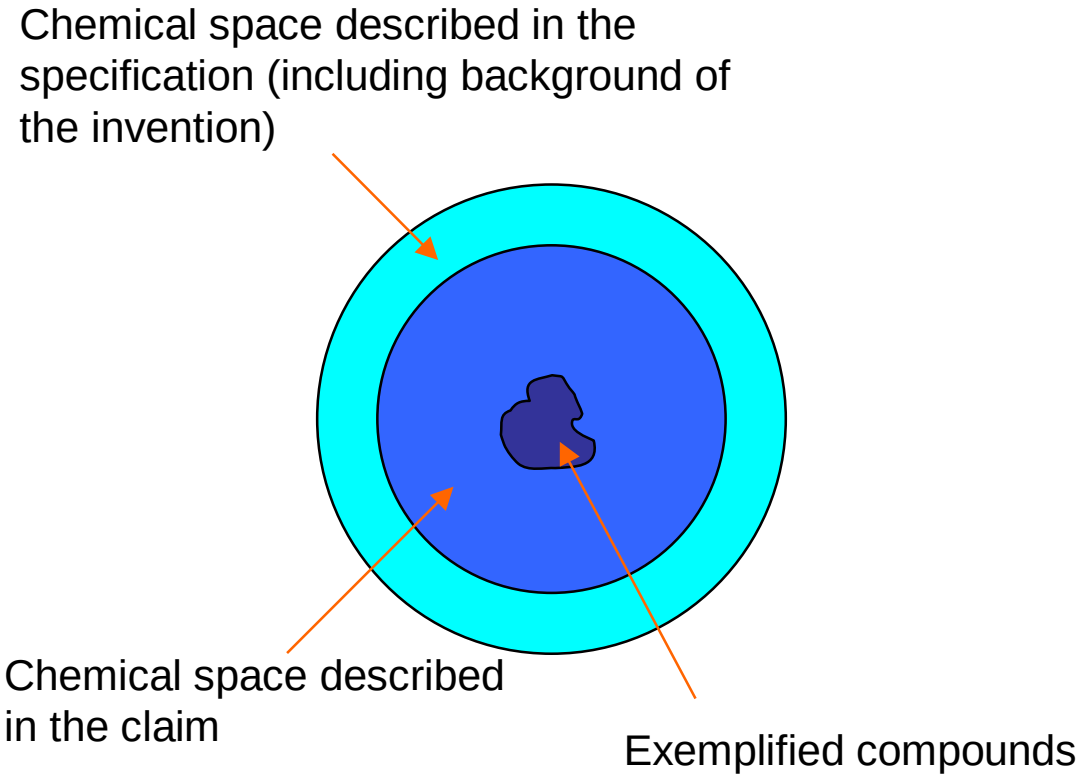
MarVis: A Visualization Tool for Patent Markush Structure Analysis

Wei Deng (David), Steven J Berthel, W. Venus So
Pharma Research & Early Development Informatics, Roche, Nutley, USA
Oct 25th, 2010, ICIC Meeting



Schematic Representation of the Chemical Space in a Patent

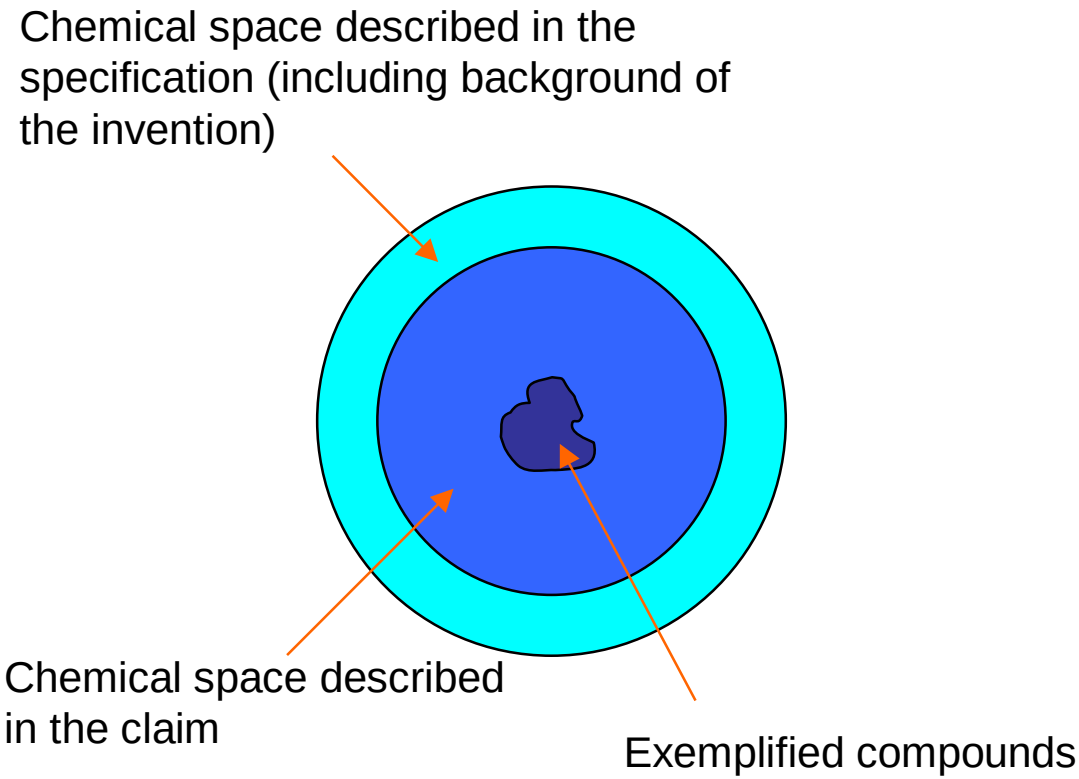
Exemplified and Markush Structures



Chemical Space in Patents

Exemplified and Markush Structures

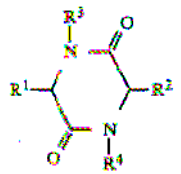
- Exemplified structures
  - Chemical space: 10^0 - 10^3
  - Search: SciFinder
  - OCR and OSR
- Markush structures
  - Chemical space: >> 10^3
  - Search systems:
    - MARPAT (STN)
    - Markush DARC (MMS)
  - Search results: difficult to read



Flexible Markush Structures

In Chemical Patent Documents

The piperazine compounds of this invention can be represented by the following formula:



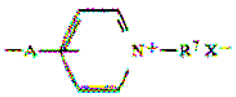
(I)

wherein

R^1 is aryl;

ar(lower)alkyl which may have lower alkoxy; heterocyclic(lower)alkyl which may have substituent(s) selected from the group of lower alkyl and aryl on the heterocyclic ring; or

a group of the formula:



wherein

A is lower alkylene,

R^7 is lower alkyl and

X is acid residue;

R^2 is

lower alkyl;

higher alkyl;

lower alkylthio(lower)alkyl;

lower alkenylthio(lower)alkyl;

hydroxy(lower)alkyl;

protected hydroxy(lower)alkyl;

amino(lower)alkyl;

protected amino(lower)alkyl;

carboxy(lower)alkyl;

protected carboxy(lower)alkyl;

arylthio(lower)alkyl;

ar(lower)alkylthio(lower)alkyl;

heterocyclic(lower)alkyl which may have substituent(s) selected from the group of lower alkyl, halogen, lower alkoxy, aryl and ar(lower)alkoxy on the heterocyclic ring; or

heterocyclic-substituted ar(lower)alkyl which may have lower alkyl on the herocyclic ring; and

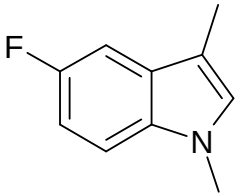
R^3 and R^4 are each hydrogen or lower alkyl.

The piperazine compounds (I) include known and novel compounds.

copied from part of US4806538

Patent Markush Database Search

Current Workflow



query substructure

→

gr
6-1:6-9-8
3-10
5-7-8
7-11
9-12

query language

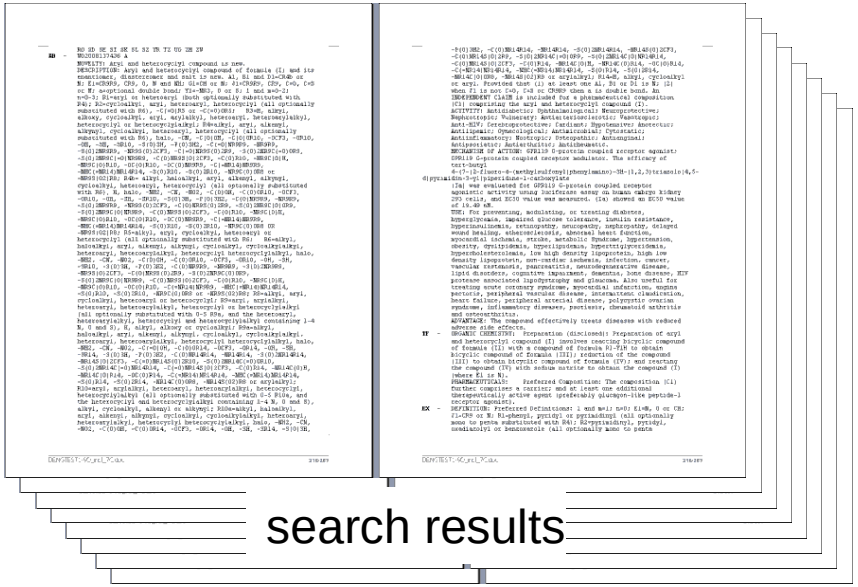
at
N 9
F 10

bo
no 1-2,1-6,2-3,3-4,4-5,5-6
do 7-8
……

↓

MMS or MARPAT

←



search results

Update: SciFinder on the web; ChemAxon

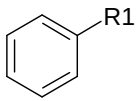
Visualize Markush Structure

Available Tools and Formats

- Convert from patent documents to structure data
  - Already done in both systems
- Visualize Markush Structure
  - Questel
  - ChemAxon
- Markush Exchange Format
  - Proprietary
    - MARPAT
    - MDARC
      - VMN (Thomson Reuter's): binary
      - XML (Questel): chosen as raw input for this study
  - Symyx RgFile
  - Extension to CML
  - Extension to InChi
  - Extension to SLN (Sybyl Line Notation)

Complicated Patent Markush Structure

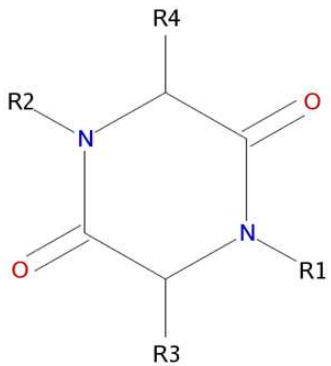
Single connection vs. multiple connections



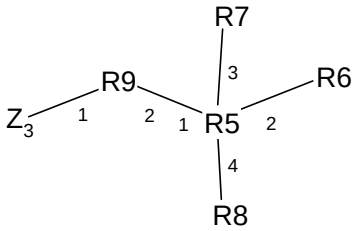
R1 = F, Cl, Br

Complicate Patent Markush Structure

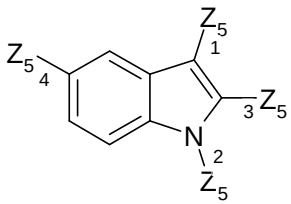
Nested R groups



R3 =



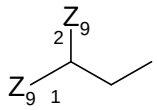
R5 =



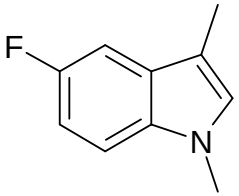
R6 =

R8 =

R9 =



MDARC: up to 50 R groups and 4 levels

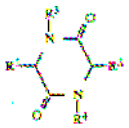


Quick Facts of MarVis

- What does MarVis stand for?
  - **Mar**kush **Vis**ualization
- MarVis is a Pipeline Pilot protocol that
  - Converts MDARC Markush structures to non-proprietary format
  - Generates an R-table report of the chemical space described by Markush structures in a patent
  - Reads the substructure search results from MMS and expand the core
  - Enumerates R groups in a Markush structure
- What is iMarVis
  - The interactive interface of MarVis
  - Allows users to explore patent chemical space

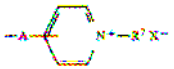
Example MarVis Report

The piperazine compounds of this invention can be represented by the following formula:



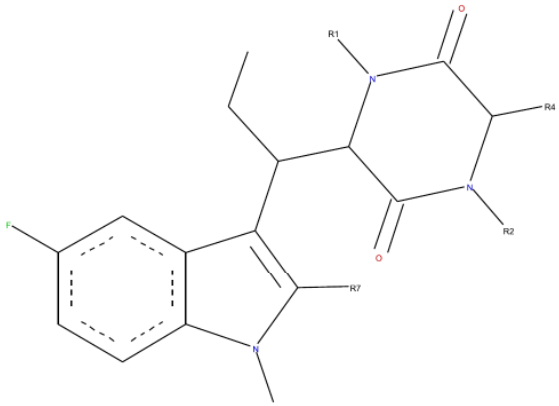
(I)

wherein
R^1 is aryl;
ar(lower)alkyl which may have lower alkoxy; heterocyclic(lower)alkyl which may have substituent(s) selected from the group of lower alkyl and aryl on the heterocyclic ring; or
a group of the formula:



wherein
A is lower alkylene,
R^7 is lower alkyl and
X is acid residue;

Patent Number is US4806538 Markush ID is 8743-08701



R1					
1.1					
CHK—Z					

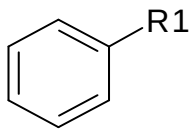
R2	
2.1	
CHK—Z	

R4					
4.1	4.2	4.3	4.4	4.5	4.6
					

R14 Parent = R4.5			
14.1	14.2	14.3	14.4
HET—Z	HEF—Z		

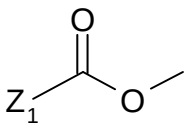
“Open Format” Used in This Study

- SMILES (Simplified molecular input line entry specification)
- Advantage
  - Widely adapted
  - Flexible for manipulation
  - Already has Markush adaption
- Extension needed
- Inspired by work from Barnard et. al.



c1ccccc1[R1]

R1 =



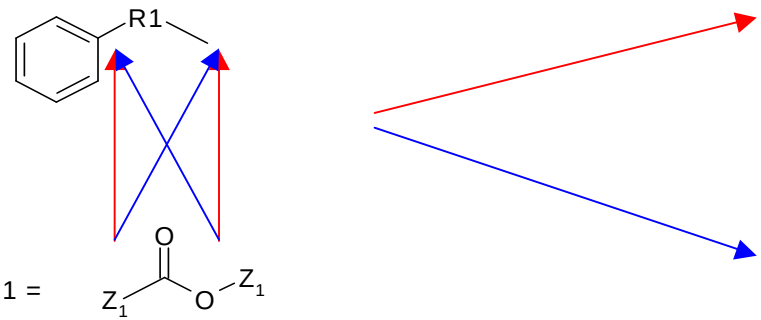
COC(=O)[Z1]

Barnard, J. M.; Downs, G. M.; von Scholley-Pfab, A.; Brown, R. D. Use of Markush Structure Analysis Techniques for Descriptor Generation and Clustering of Large Combinatorial Libraries. *J. Mol. Graphics Modell.* **2000**, *18*, 452-63.

R Group with Multiple Connections

The Challenge and Its Solution

c1ccccc1[R1]C



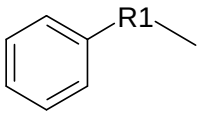
[Z1]OC(=O)[Z1]

R1 =

R Group with Multiple Connections

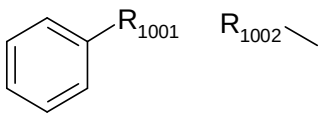
Extended SMILES to Ensure Correction Connection

c1ccccc1[R1]C



C:%10:%11.C:%11:%12.C:%12:%13.C:%13:%14.C:%14:%15.C:%10:%15%16.[R1]%16%17.C%17

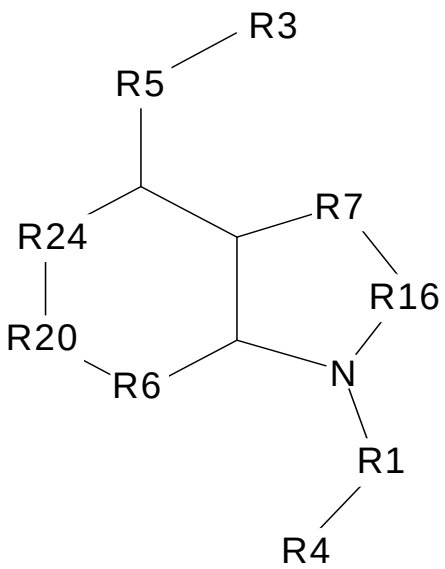
c1ccccc1[R1001]
C[R1002]



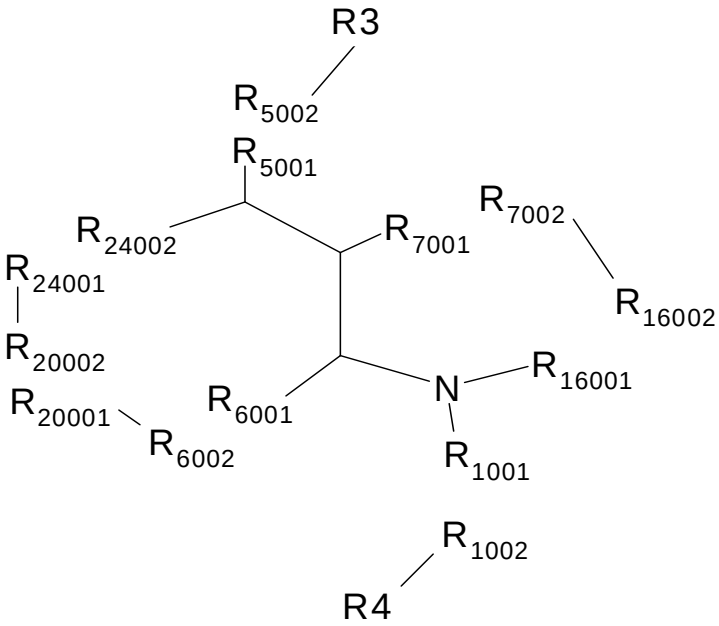
C:%10:%11.C:%11:%12.C:%12:%13.C:%13:%14.C:%14:%15.C:%10:%15%16.[R1001]%16.[R1002]%17.C%17

Applying the “Break R groups” technique

A More Complicated Example



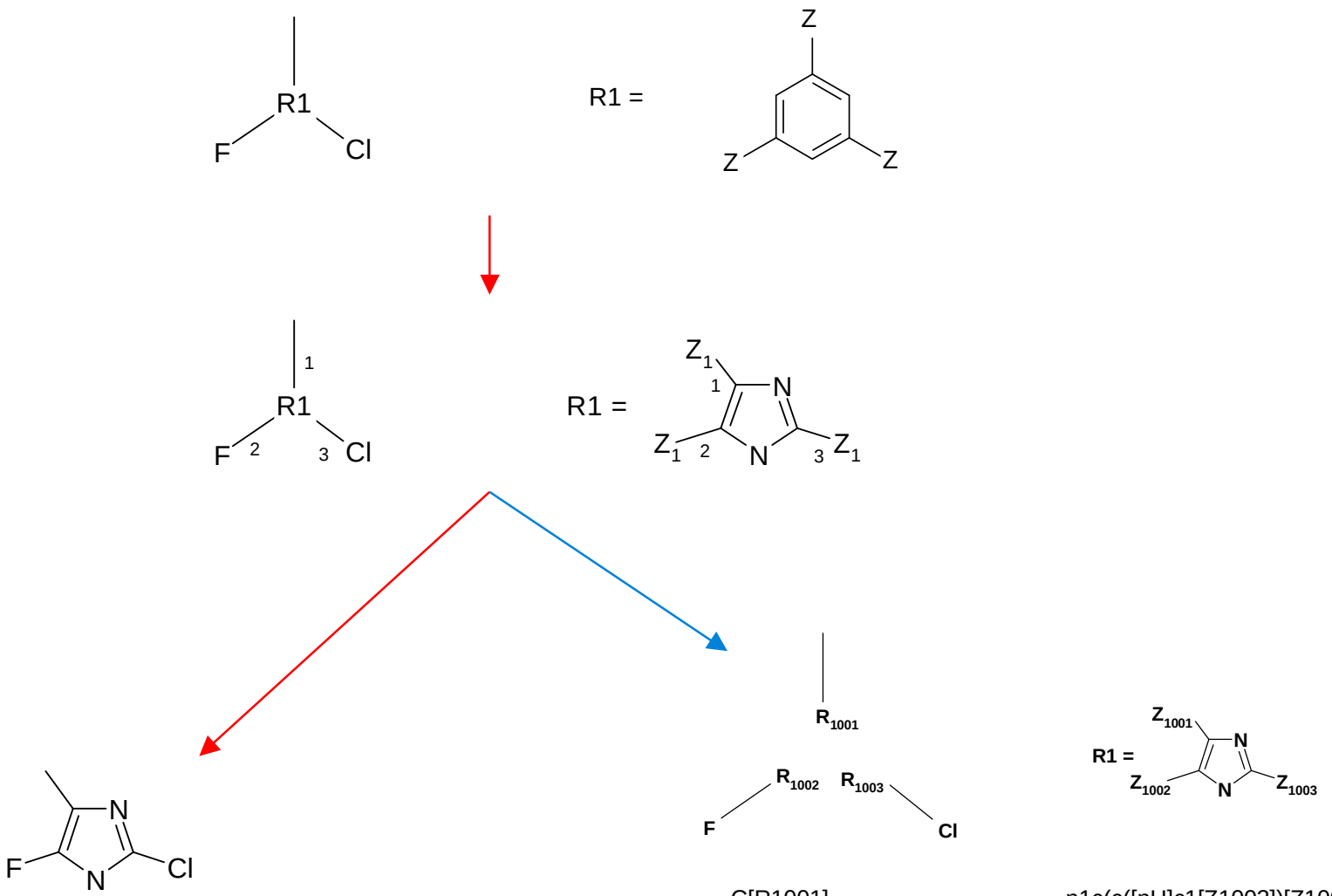
C12C(C([R5][R3])[R24][R20][R6]1)[R7][R16]N2[R1][R4]



C(C(C([R5001])[R24002])[R7001])(N([R16001])[R1001])[R6001]
[R3][R5002]
[R4][R1002]
[R20001][R6002]
[R20002][R24001]
[R7002][R16002]

Nested R groups

Maintaining the Parent-Child Relationship of R Groups

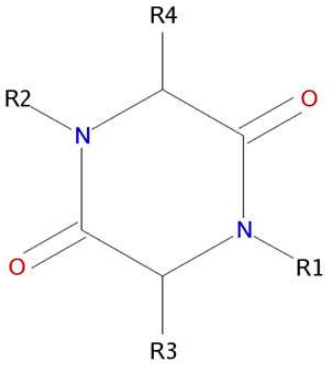


C[R1001]
F[R1002]
Cl[R1003]

n1c(c([nH]c1[Z1003])[Z1002])[Z1001]

Displaying Markush Structures

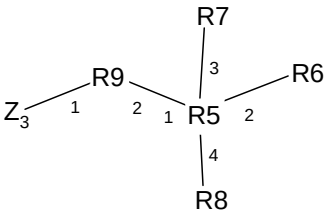
Patching R groups together



[N]%10%11%12.[C]%10%13=%14.[C]%11%15%16.[R1]%12.[C]%13%17%18.[O]=%14.[C]%15%19=%20.[R3]%16.[N]%17%19%21.[R4]%18.[O]=%20.[R2]%21

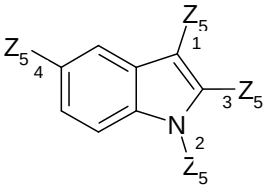
Markush structure of a hit patent

R3 =



[R9002]%10.[R9001]([Z3]).[R5001]%10.[R5002]%11.[R5003]%12.[R5004]%13.[R6]%11.[R7]%12.[R8]%13

R5 =



[C]=%20%21([Z5001]).[C]=%20%22([Z5003]).[C]%21:%23:%24.[N]%22%25([Z5002]).[C]:%23%25:%26.[C]:%24:%27.[C]:%26:%28.[C]:%27:%29([Z5004]).[C]:%28:%29

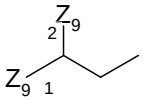
R6 =

[C]([Z6])

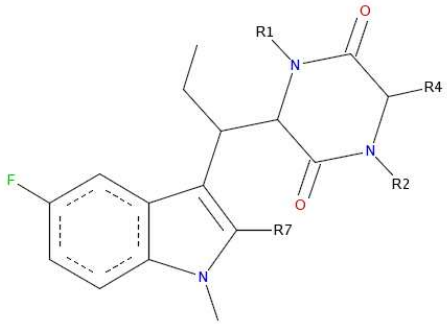
R8 =

[F]([Z8])

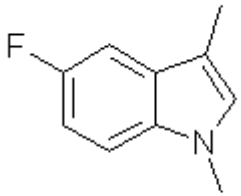
R9 =



[C]%10([Z9001])([Z9002]).[C]%10%11.[C]%1



displaying query substructure in the hit result



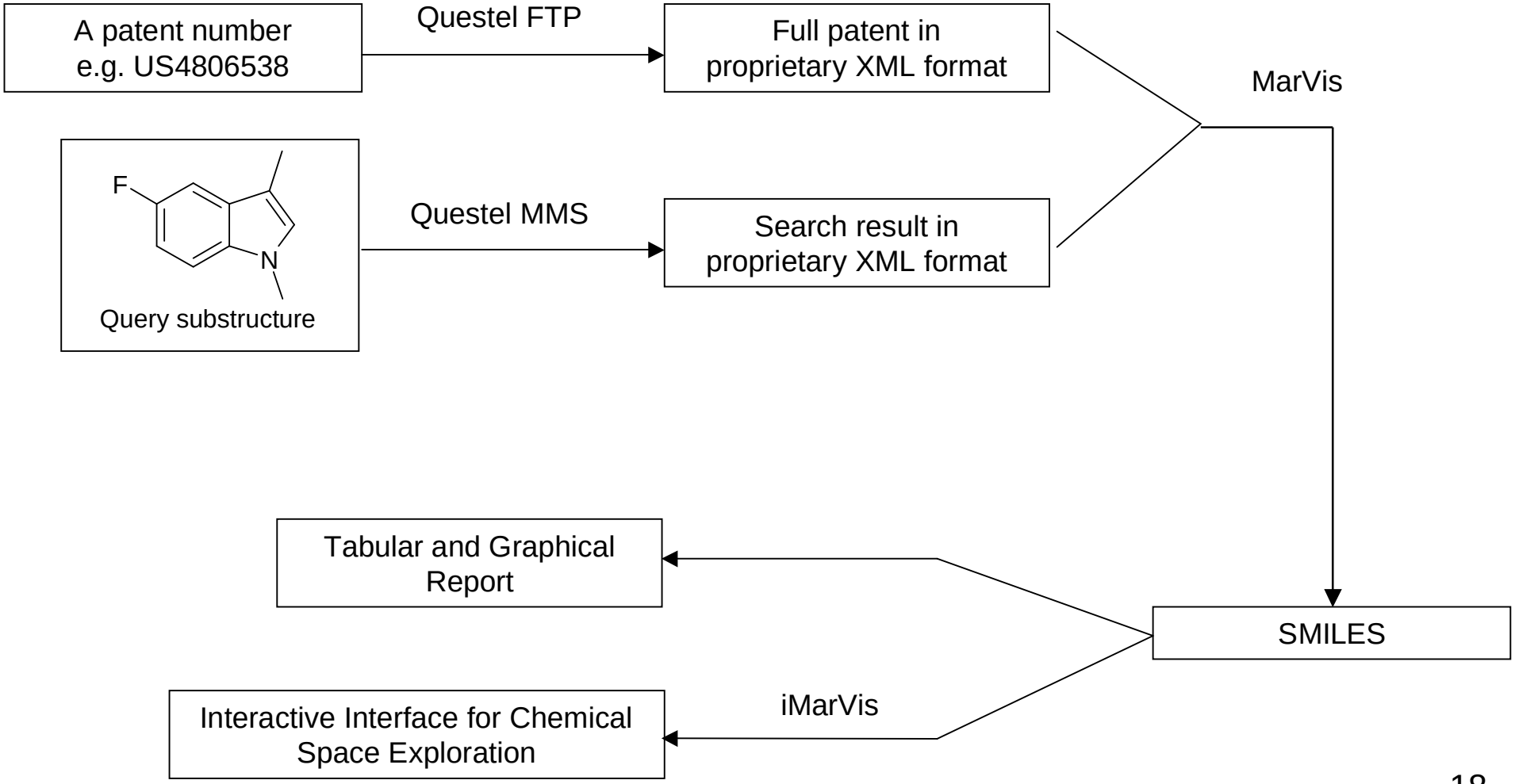
query substructure

Superatoms

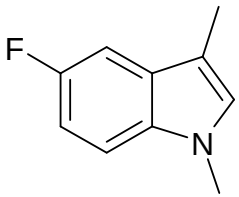
Questel Superatoms	Groups
CHK	Alkyl or alkylene
CHE	Alkenyl or alkenylene
CHY	Alkynyl or alkynylene
ARY	Carbocyclic system, optionally fused, containing at least one benzene (aryl)
CYC	Cycloaliphatic carbocyclic, optionally fused
HEA	Monocyclic, aromatic heterocycle (heteroaryl)
HET	Non-aromatic monocyclic heterocycle
HEF	Fused heterocycle
HAL	Halogen
UNK	Undefined group
…	…

Barnard, J. M.; Downs, G. M.; von Scholley-Pfab, A.; Brown, R. D. Use of Markush Structure Analysis Techniques for Descriptor Generation and Clustering of Large Combinatorial Libraries. *J. Mol. Graphics Modell.* **2000**, *18*, 452-63.

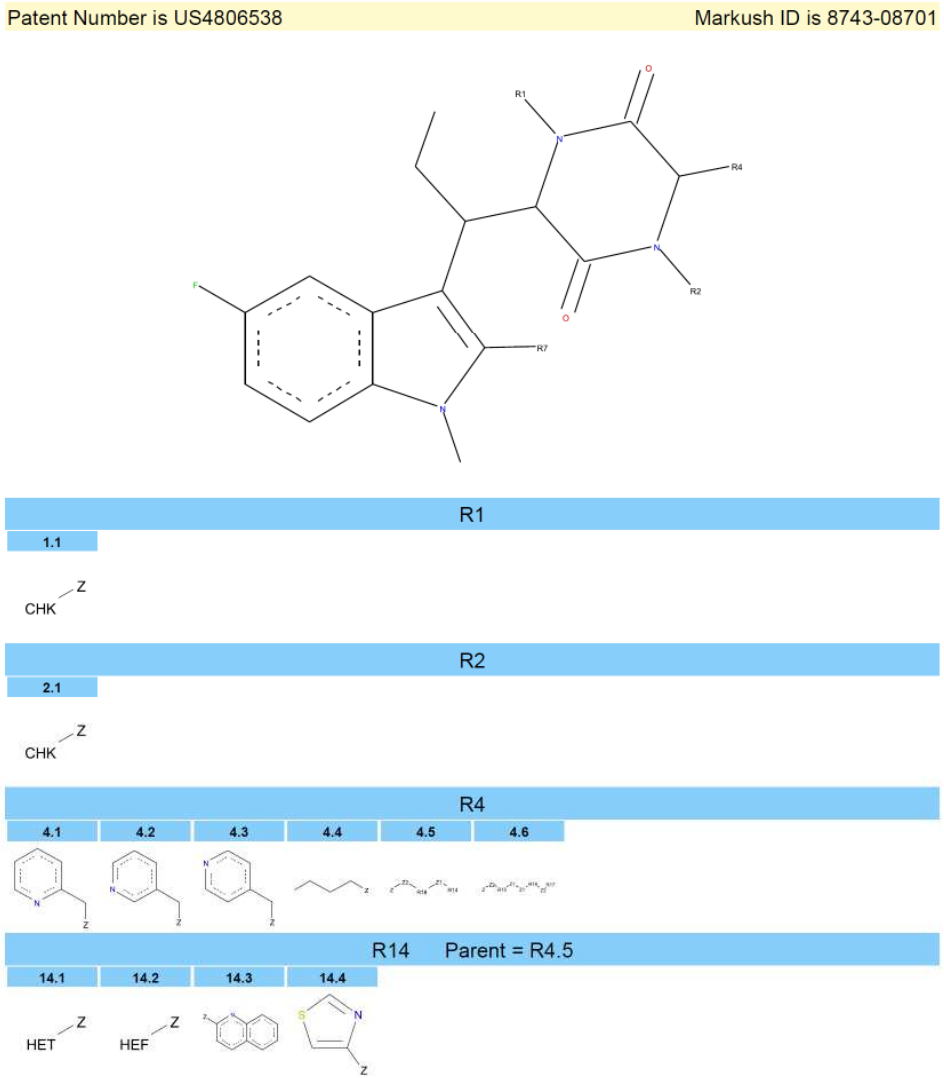
New Work Flow



Example MarVis Report



query substructure

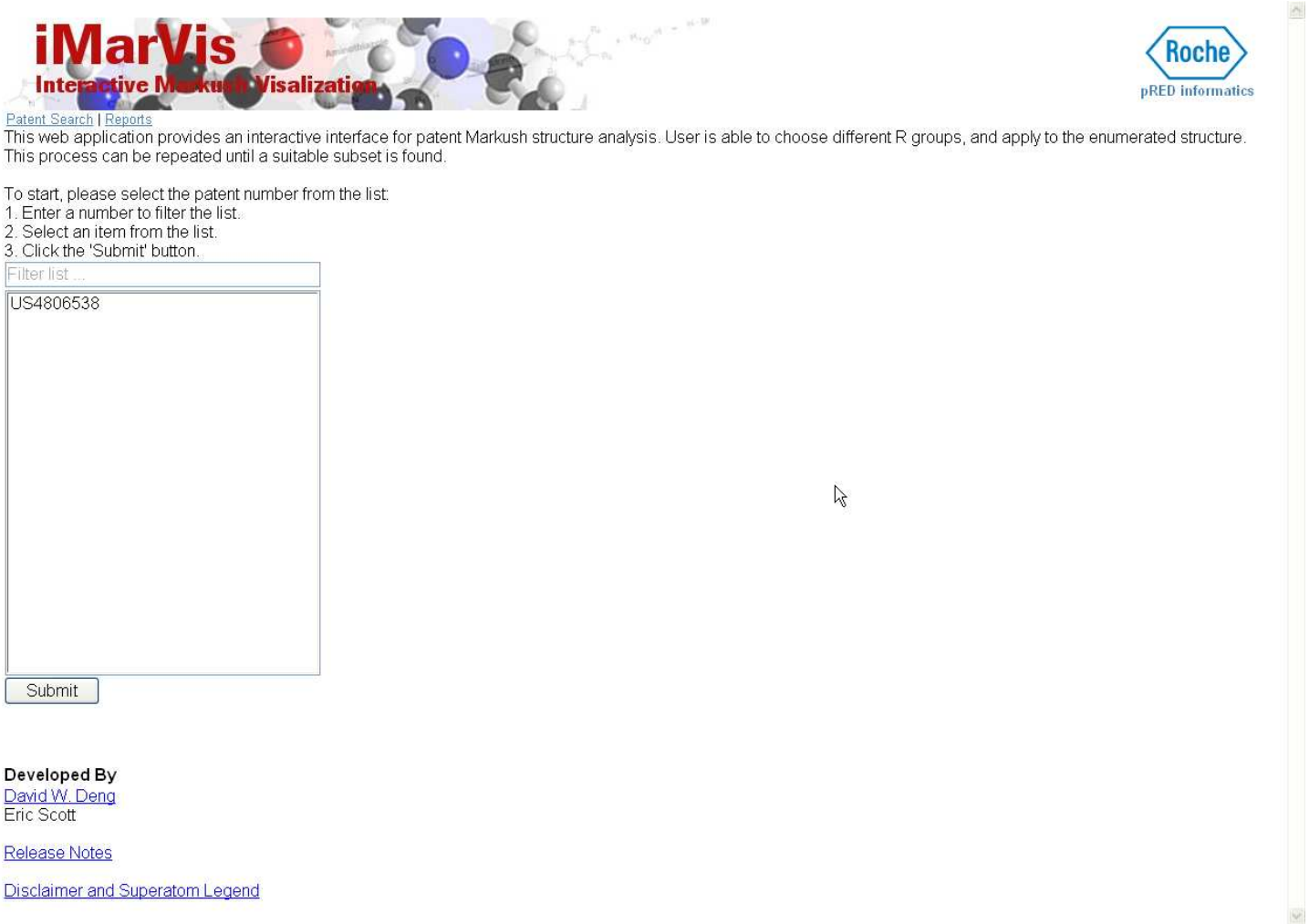


iMarVis
interactive MarVis Interface

- Online interface
- User can zoom in and out of the chemical space of a patent Markush structure
- R group hierarchy
  - Can only choose an R group if its parent is selected
  - Deselect R group will deselect all its children R groups
- Generate report to share with others
- URL link for easy access for project teams

iMarVis Interface

Select Patent Numbers



iMarVis Interface

Select Markush Structure

iMarVis
Interactive Markush Visualization

Roche
pRED informatics

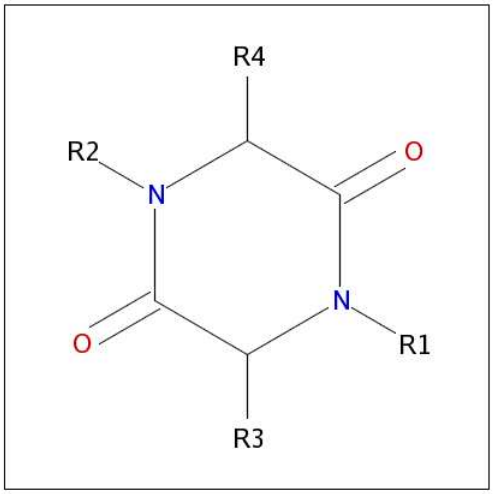
Patent Search | Reports

Select Markush Structure in Patent US4806538
There are 1 Markush structures in this patent

Disclaimer and Superatom Legend
User: dengw1
Date: 08/03/10
XML Location: \\rnumsriggap25\web_apps\iMarVis\Patent_XMLs\US4806538_1_1.xml

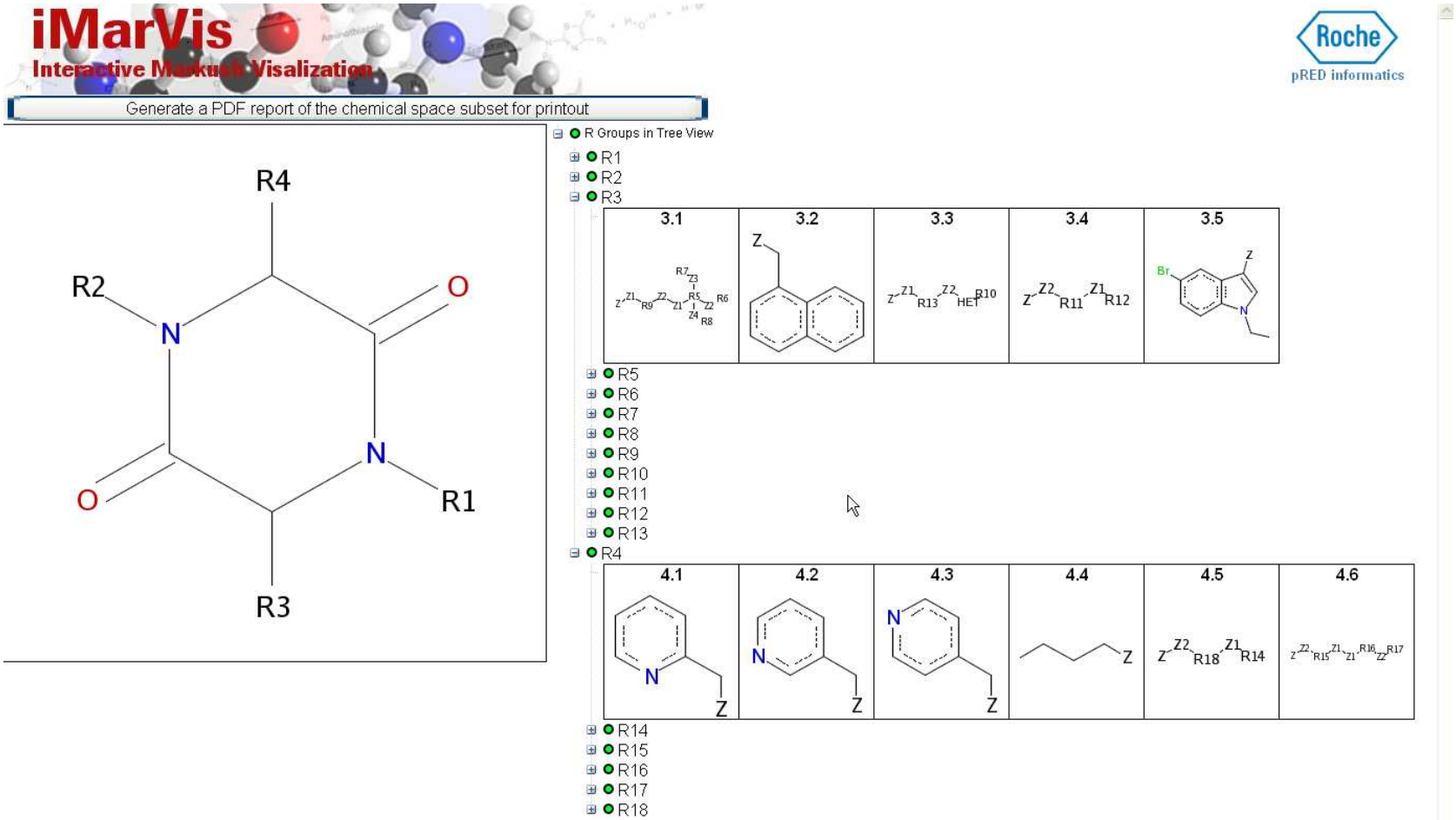
All 1 Markush Core(s)

Click on the Markush core you are interested in. You will be able to see the R groups in the next page, choose R groups and add to the core structure.



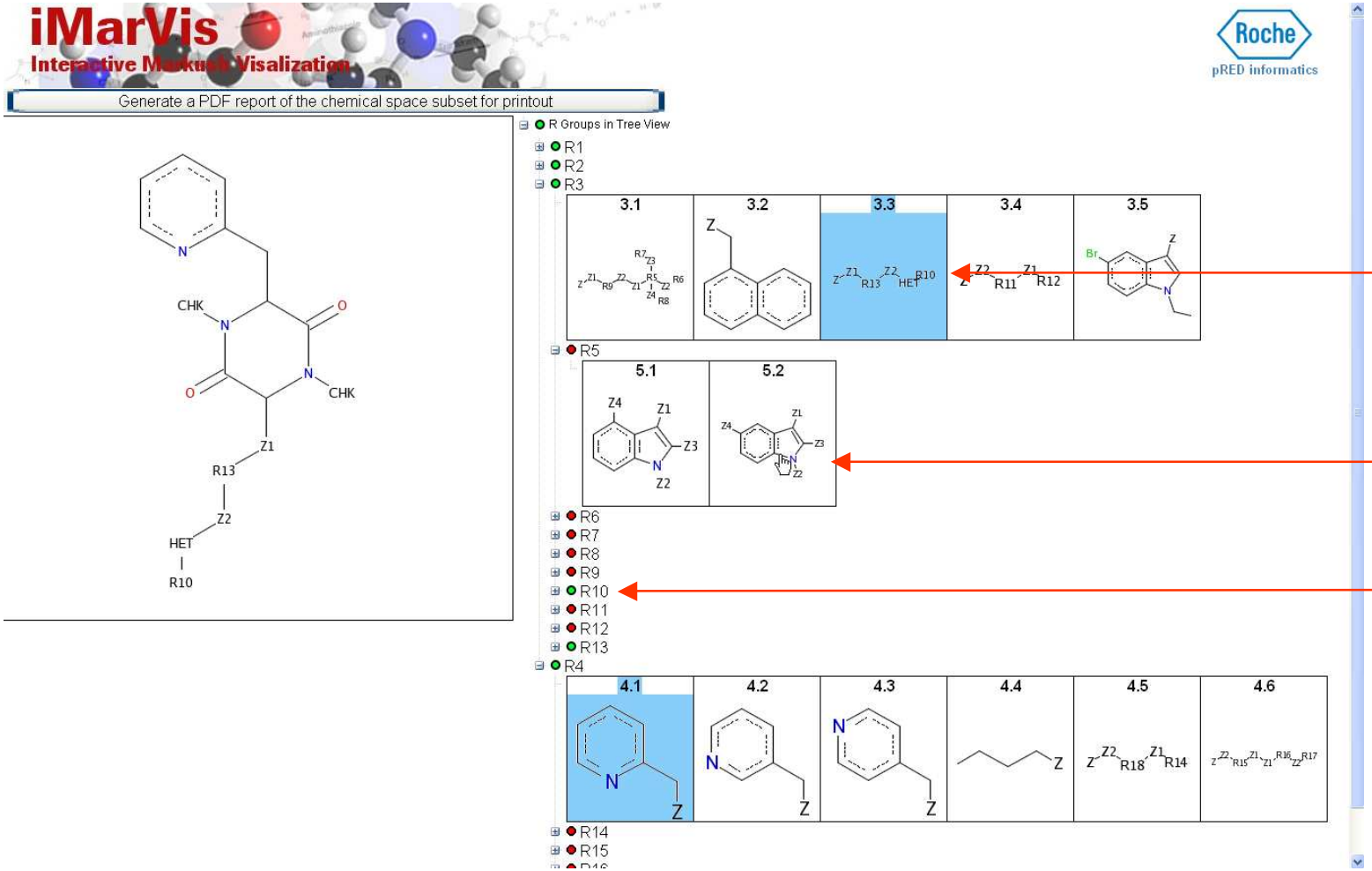
iMarVis

R Group Tree



iMarVis Interface

Select and De-select R Groups



Add to core (<1s)

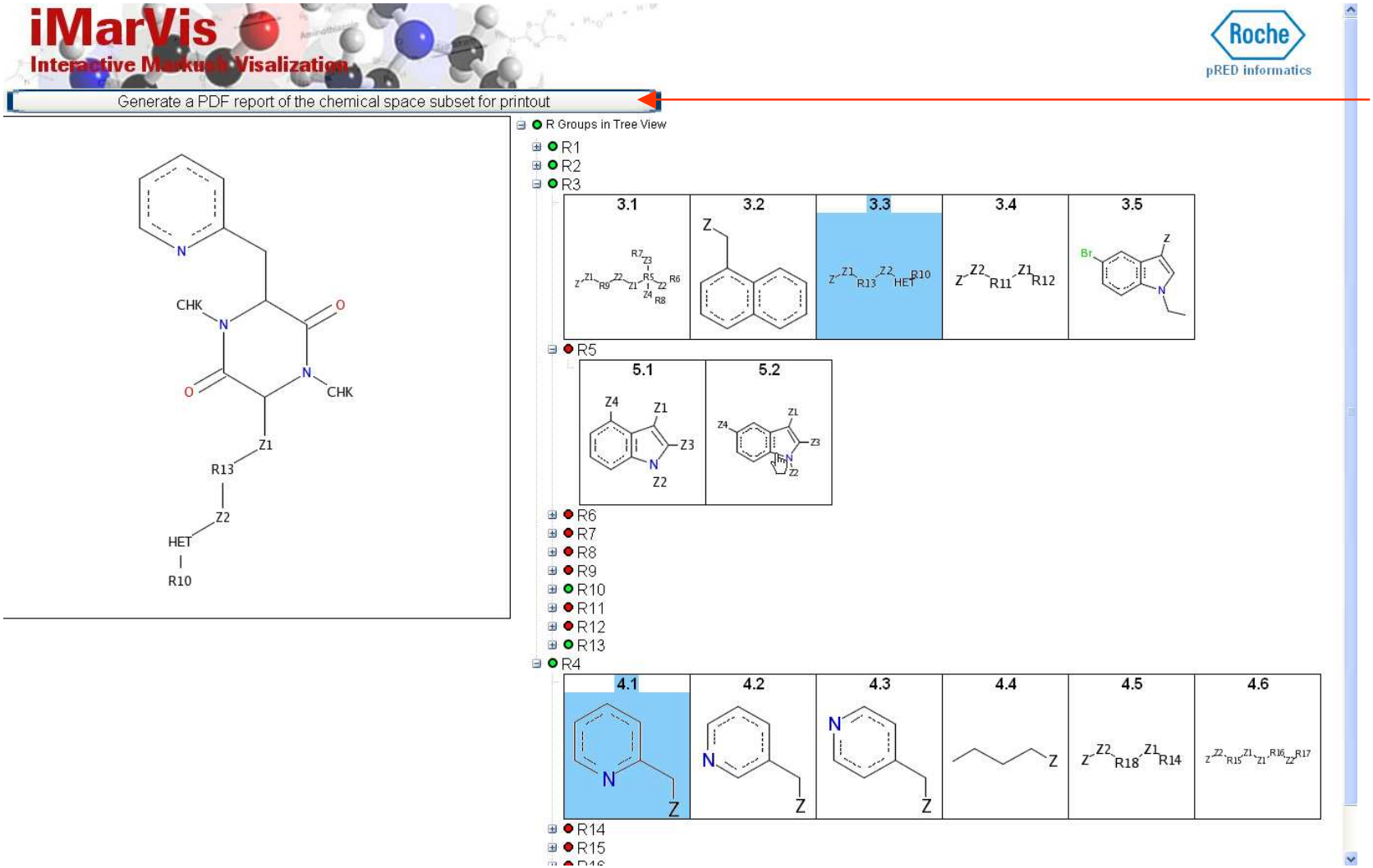
Cannot select if parent not selected

Child R groups can be selected

Pre-selected R groups are highlighted and added to core after the page loads

iMarVis Interface

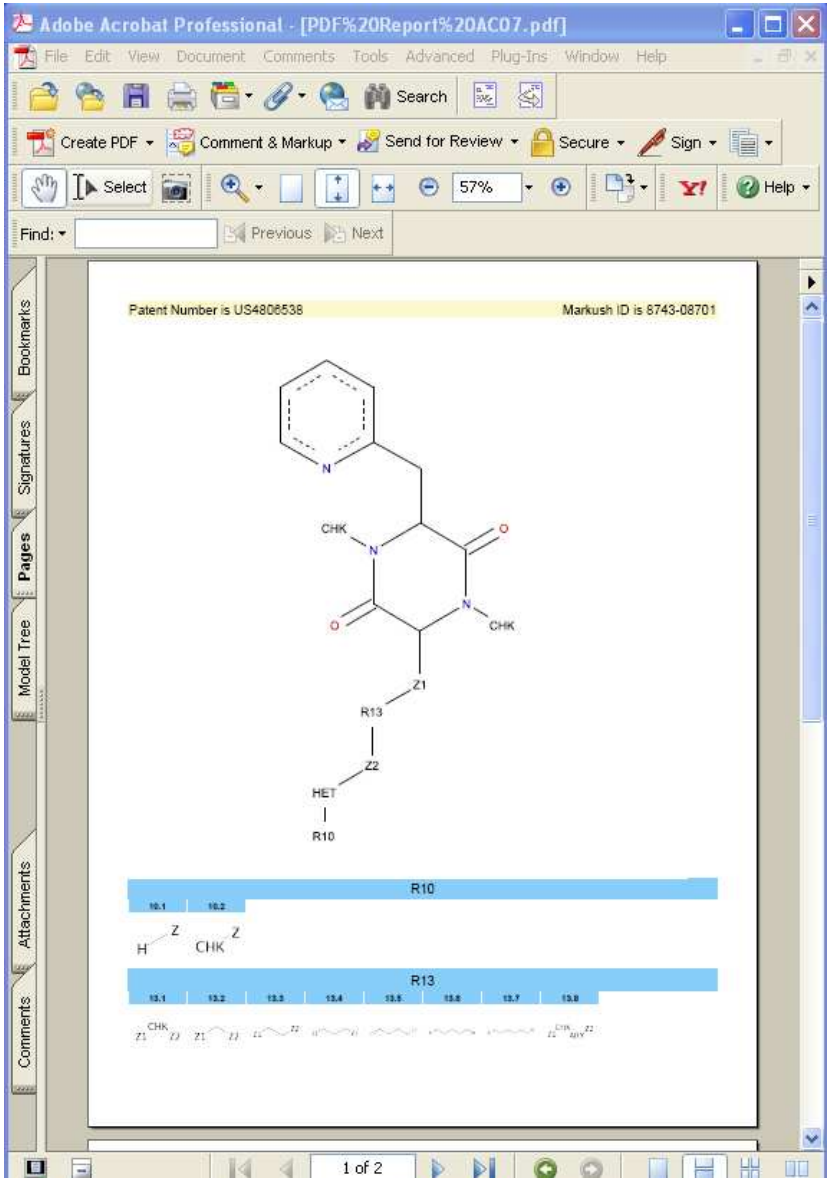
Generate PDF Report



Generate a PDF Report

iMarVis Interface

The PDF Report

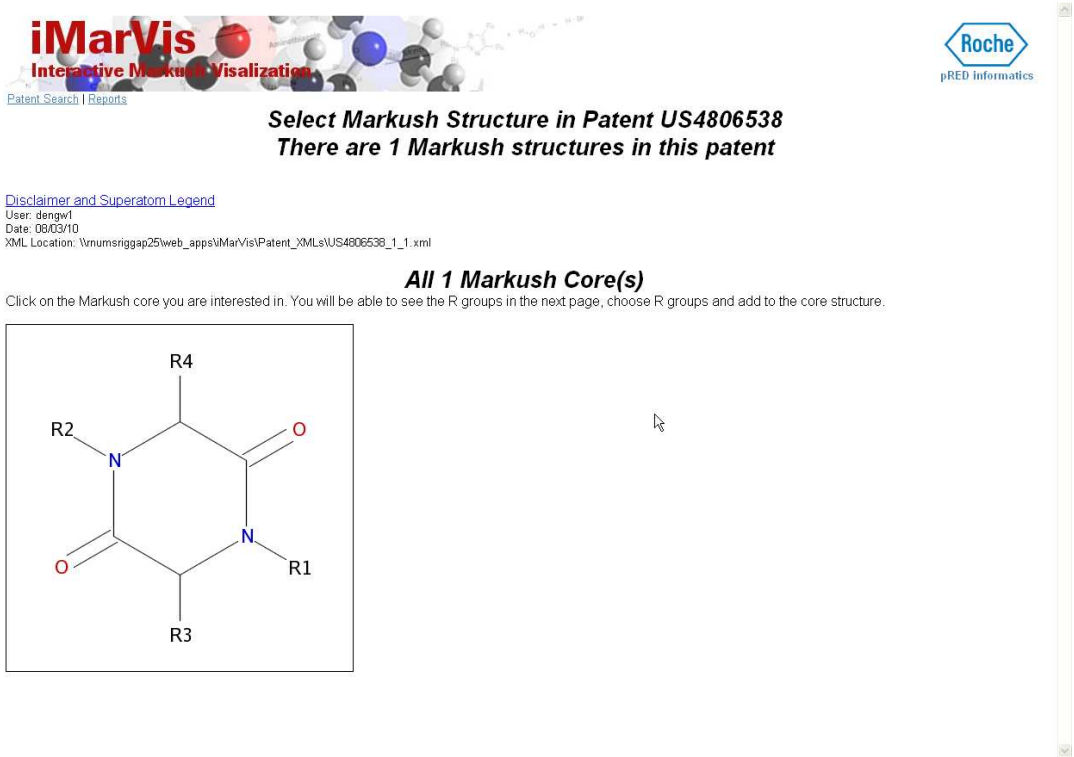


iMarVis Interface

Hyperlink for Easy Access

http://rnumsriggap25:9944/perlbin/runjob.pl?_protocol=%7B758A2270-4214-41CF-A476-B3FCCCF42649%7D&patent_number=US4806538&__QuickRun=true

US4806538



Future Study

- Current Limitations:
  - No stereochemistry
  - No notes
  - Inherit indexing problem
- Other areas for future study

- Manuscript
  - Submitted to Journal of Chemical Information and Modeling

Acknowledgements

- pRED Informatics
  - Venus So
  - Eric Scott
  - Joan Skinner
- Medicinal Chemistry
  - Steve Berthel
  - Paul Gillespie
  - Jeff Tilley
  - Robert Kester
  - Peter Wovkulich
- Molecular Modeling
  - Sung-Sau So
- SCRUM team
  - Li Zhang
- Patent Law
  - Samuel Megerditchian
- Roche Post-doc Fellow Program

We Innovate Healthcare