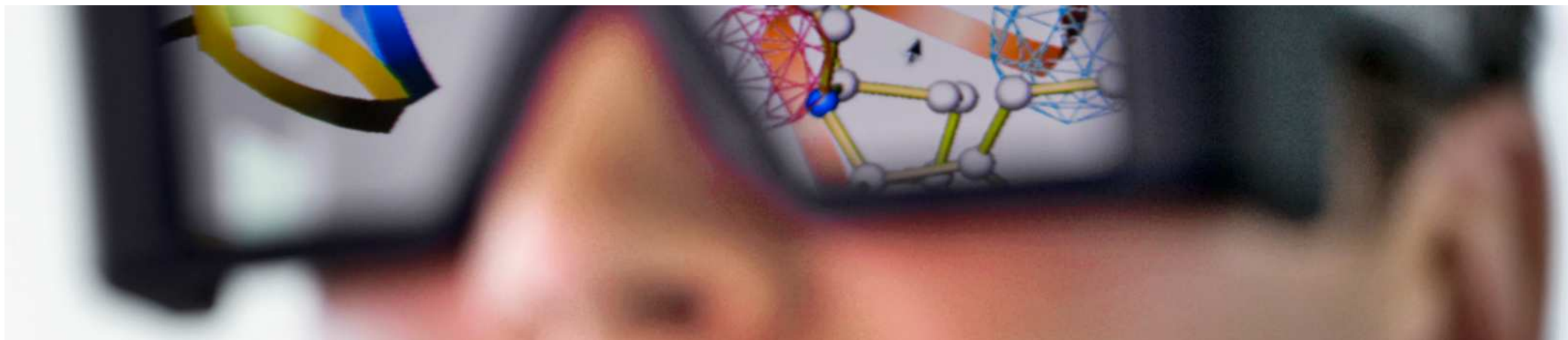

MarVis: A Visualization Tool for Patent Markush Structure Analysis

Wei Deng (David), Steven J Berthel, W. Venus So

Pharma Research & Early Development Informatics, Roche, Nutley, USA

Oct 25th, 2010, ICIC Meeting

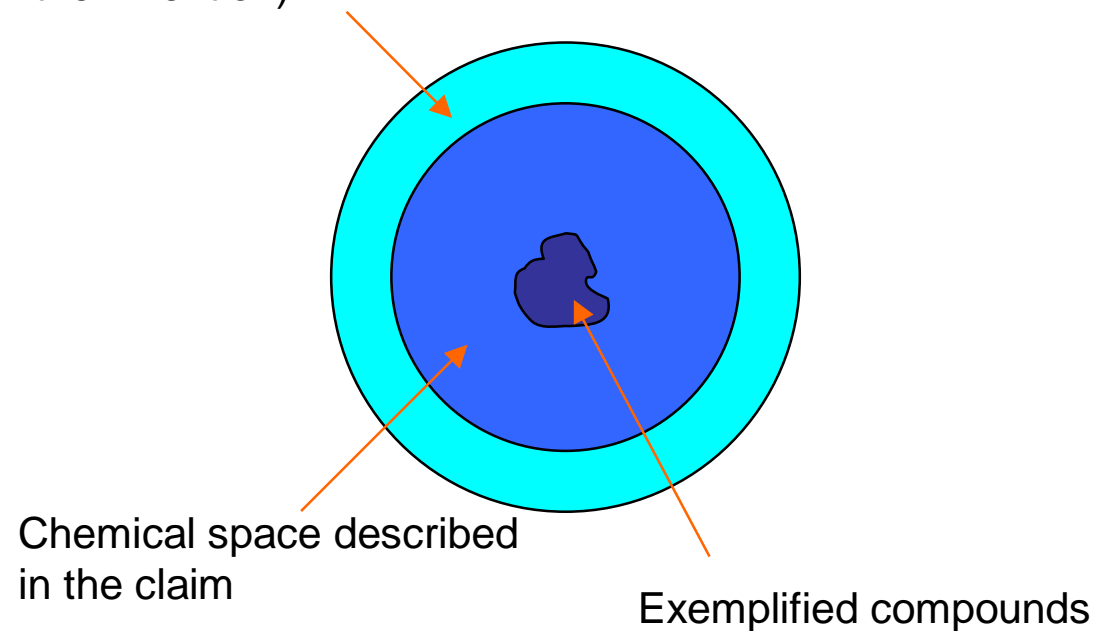


Schematic Representation of the Chemical Space in a Patent

Exemplified and Markush Structures



Chemical space described in the specification (including background of the invention)

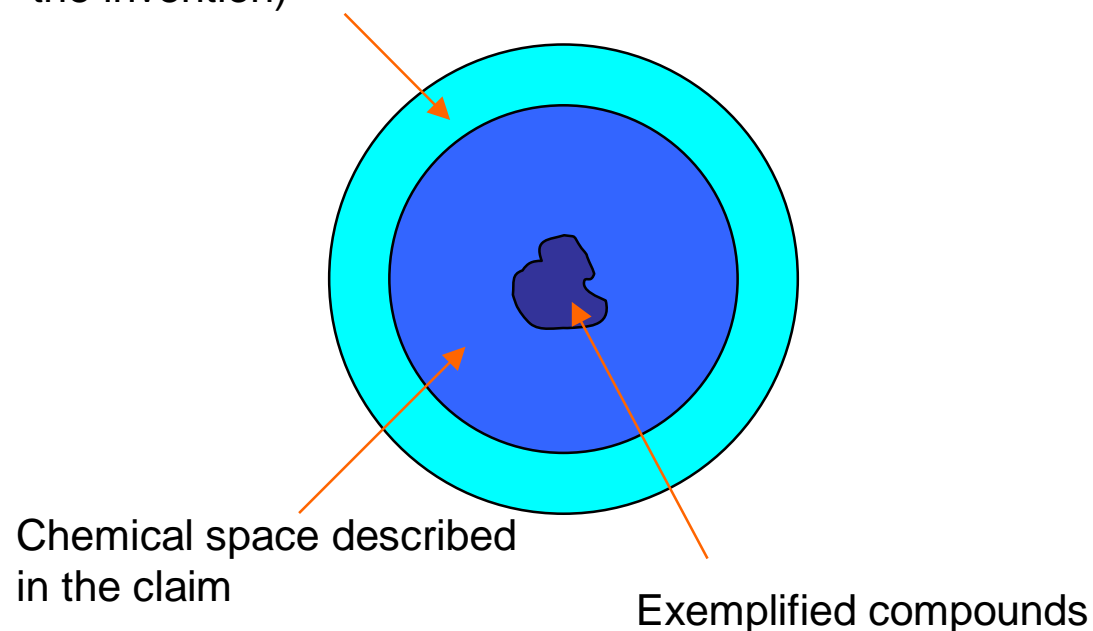


Chemical Space in Patents

Exemplified and Markush Structures

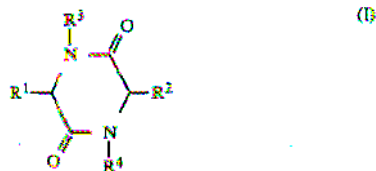
- Exemplified structures
 - Chemical space: $10^0 - 10^3$
 - Search: SciFinder
 - OCR and OSR
- Markush structures
 - Chemical space: $\gg 10^3$
 - Search systems:
 - MARPAT (STN)
 - Markush DARC (MMS)
 - Search results: difficult to read

Chemical space described in the specification (including background of the invention)



Flexible Markush Structures In Chemical Patent Documents

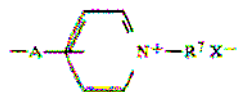
The piperazine compounds of this invention can be represented by the following formula:



wherein

R¹ 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,

R⁷ is lower alkyl and

X is acid residue;

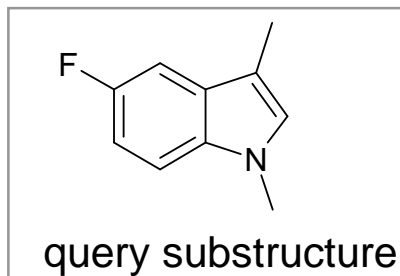
R² is

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;
arythio(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
R³ and R⁴ are each hydrogen or lower alkyl.

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

Patent Markush Database Search

Current Workflow



```

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
.....
  
```

search results



Update: SciFinder on the web; ChemAxon

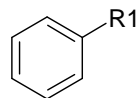
Visualize Markush Structure

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

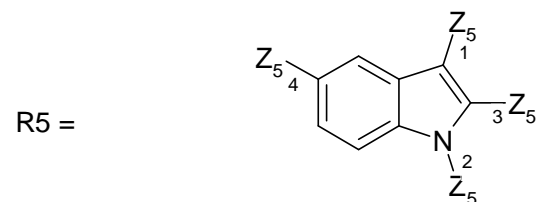
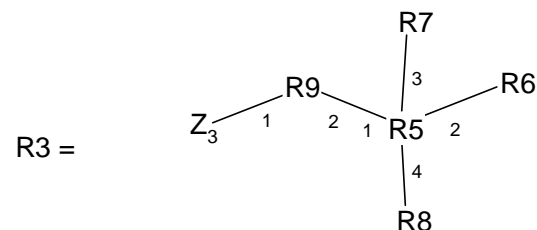
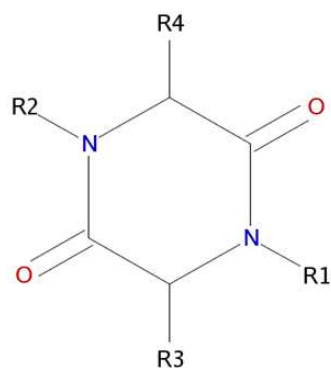
Single connection vs. multiple connections



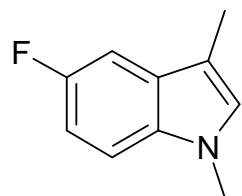
R1 = F, Cl, Br

Complicate Patent Markush Structure

Nested R groups



MDARC: up to 50 R groups and 4 levels

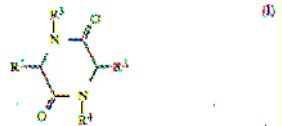


Quick Facts of MarVis

- What does MarVis stand for?
 - **Markush Visualization**
- 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:



wherein
 R^1 is aryl;
 R^2 is (lower)alkyl which may have lower alkoxy;
 R^3 is heterocycle(lower)alkyl which may have substituent(s) selected from the group of lower alkyl and aryl on the heterocyclic ring; or
 R^4 is a group of the formula:

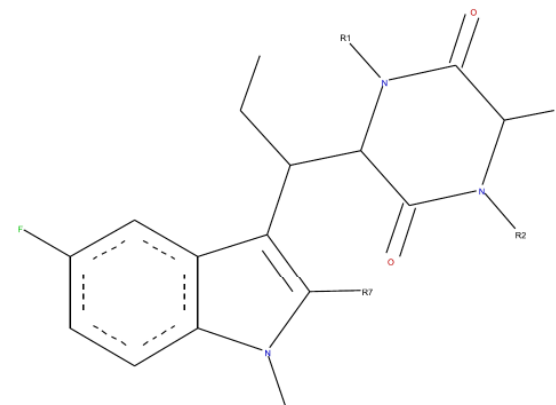


wherein
 A is lower alkylene,
 R^2 is lower alkyl and
 X is acid residue;



Patent Number is US4806538

Markush ID is 8743-08701



R1

1.1



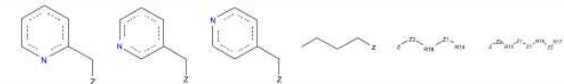
R2

2.1



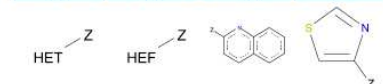
R4

4.1 4.2 4.3 4.4 4.5 4.6



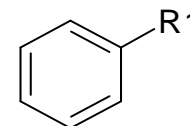
R14 Parent = R4.5

14.1 14.2 14.3 14.4

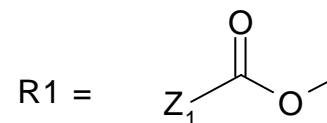


“Open Format” Used in This Study

- 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]

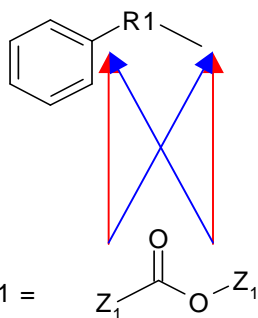


COC(=O)[Z1]

R Group with Multiple Connections

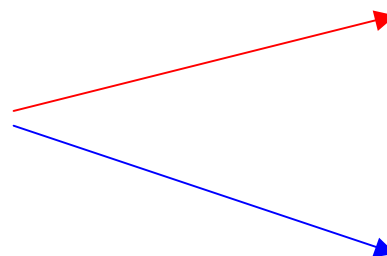
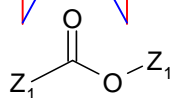
The Challenge and Its Solution

c1ccccc1[R1]C



[Z1]OC(=O)[Z1]

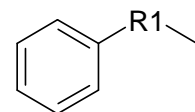
R1 =



R Group with Multiple Connections

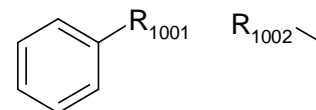
Extended SMILES to Ensure Correction Connection

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

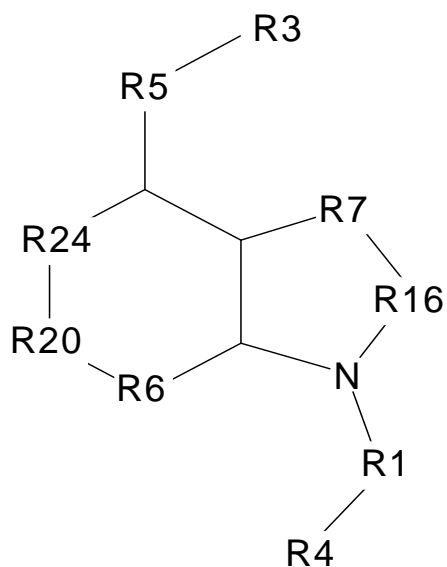
c1ccccc1[R1001]
C[R1002]



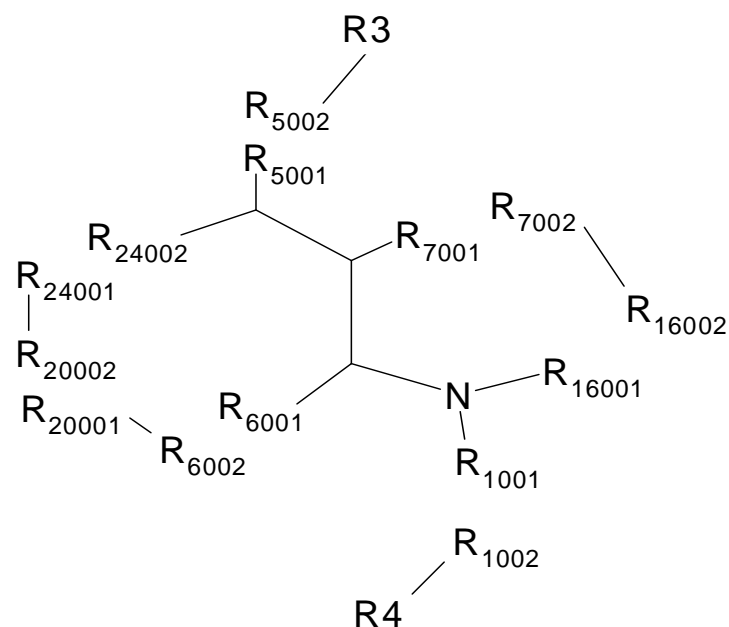
C:%10:%11.C:%11:%12.C:%12:%13.C:%13:%14.C:%14:%15.C:%10:%15%16.[R1001]%16.[R1002]%17.C%17

Applying the “Break R groups” technique

A More Complicated Example



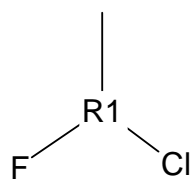
C12C(C([R5][R3])[R24][R20][R6]1)[R7][R16]N2[R1][R4]



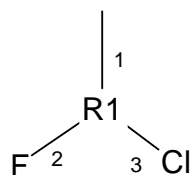
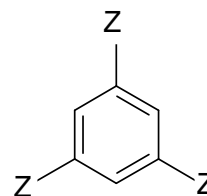
C(C(C([R5001])[R24002])[R7001])(N([R16001])[R1001])[R6001]
 [R3][R5002]
 [R4][R1002]
 [R20001][R6002]
 [R20002][R24001]
 [R7002][R16002]

Nested R groups

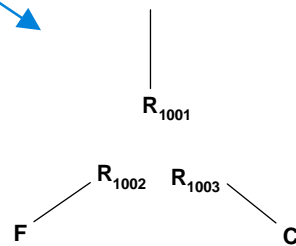
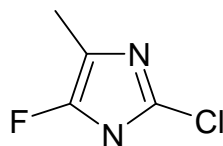
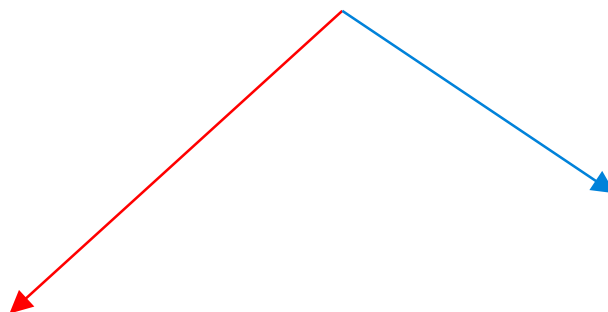
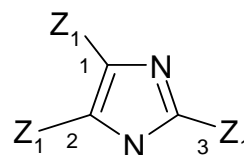
Maintaining the Parent-Child Relationship of R Groups



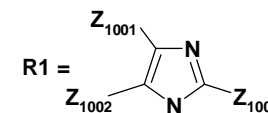
R1 =



R1 =



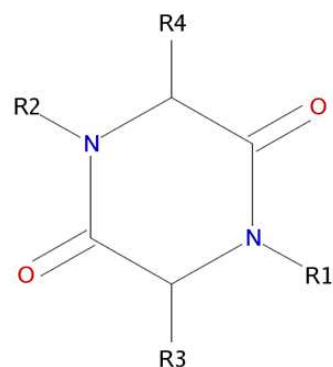
C[R1001]
F[R1002]
Cl[R1003]



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

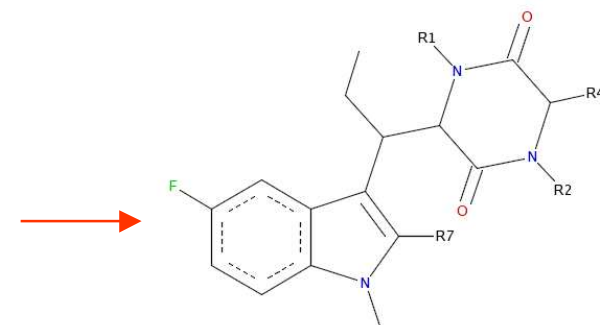
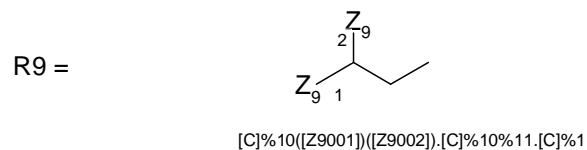
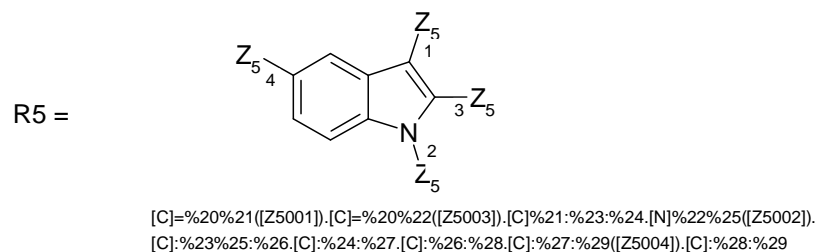
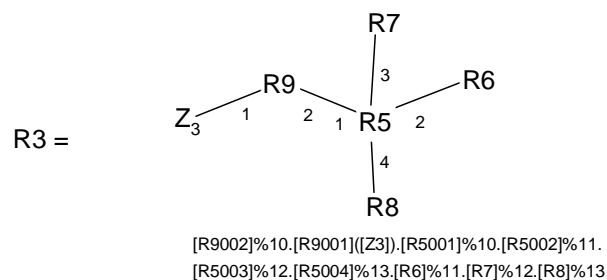
Displaying Markush Structures

Patching *R* groups together

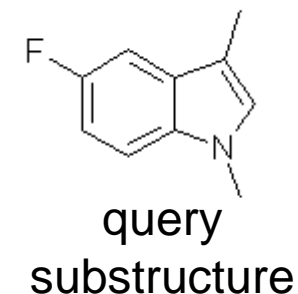


[N]%10%11%12.[C]%10%13=%14.[C]%11%15%16.[R1]%12.[C]%13%17%18.[O]=%14.[C]%15%19=%20.[R3]%16.[N]%17%19%21.[R4]%18.[O]=%20.[R2]%21

Markush structure of a hit patent



displaying query substructure in the hit result

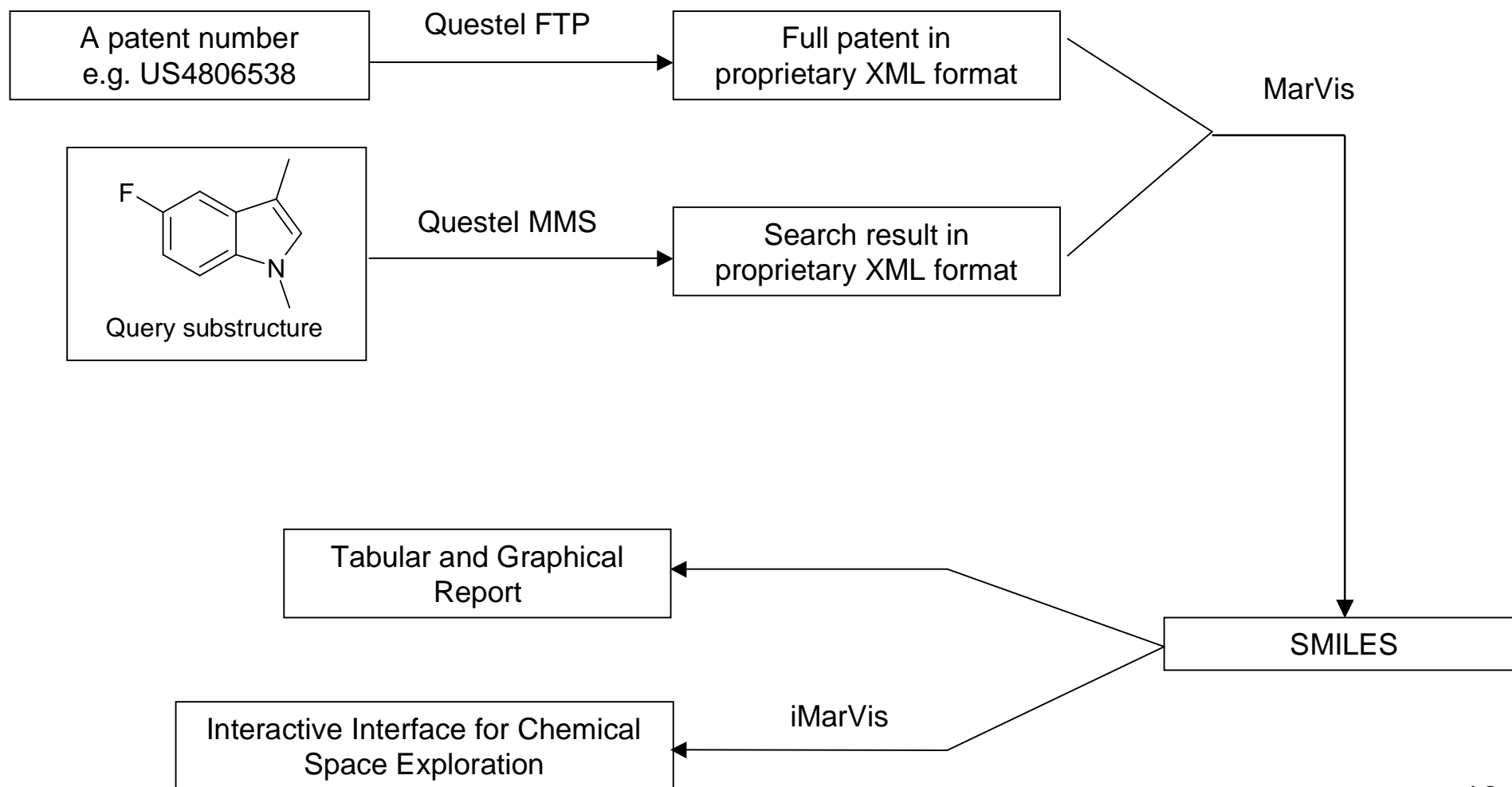


Superatoms

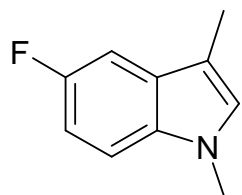
Questel Superatoms	Groups
CHK	Alkyl or alkylene
CHE	Alkenyl or alkenylene
CHY	Alkynyl or alkynylene
ARY	Carbocyclic system, optionally fused, containing at least one benzene (aryl)
CYC	Cycloaliphatic carbocyclic, optionally fused
HEA	Monocyclic, aromatic heterocycle (heteroaryl)
HET	Non-aromatic monocyclic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

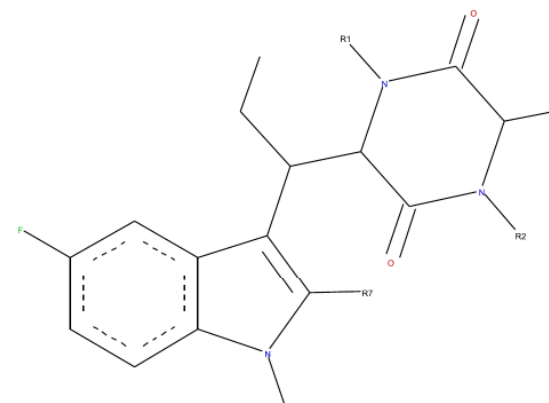


query
substructure



Patent Number is US4806538

Markush ID is 8743-08701



R1

1.1



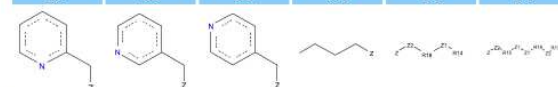
R2

2.1



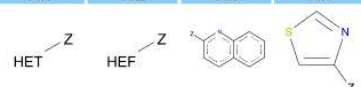
R4

4.1 4.2 4.3 4.4 4.5 4.6



R14 Parent = R4.5

14.1 14.2 14.3 14.4



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



[Patent Search](#) | [Reports](#)

This web application provides an interactive interface for patent Markush structure analysis. User is able to choose different R groups, and apply to the enumerated structure. This process can be repeated until a suitable subset is found.

To start, please select the patent number from the list:

1. Enter a number to filter the list.
2. Select an item from the list.
3. Click the 'Submit' button.

US4806538

Developed By

[David W. Deng](#)

[Eric Scott](#)

[Release Notes](#)

[Disclaimer and Superatom Legend](#)

iMarVis Interface

Select Markush Structure



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

[Disclaimer and Superatom Legend](#)

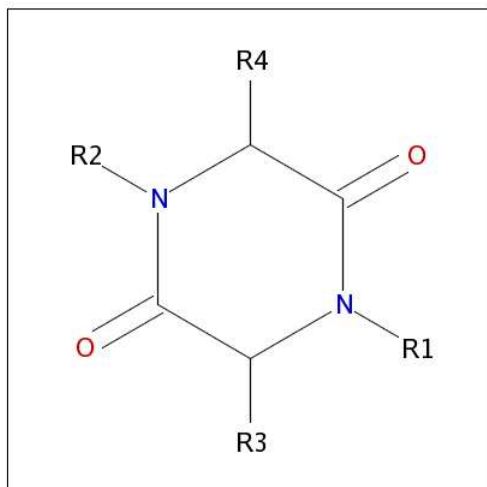
User: dengw1

Date: 08/03/10

XML Location: \\numrigger25\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
Interactive Molecular Visualization

Generate a PDF report of the chemical space subset for printout

R Groups in Tree View

- R1
- R2
- R3
- R4
- R5
- R6
- R7
- R8
- R9
- R10
- R11
- R12
- R13
- R14
- R15
- R16
- R17
- R18

3.1 	3.2 	3.3 	3.4 	3.5 	
4.1 	4.2 	4.3 	4.4 	4.5 	4.6

iMarVis Interface

Select and De-select R Groups

The screenshot displays the iMarVis interface with a chemical structure on the left and a tree view of R groups on the right. The tree view is organized into levels (R1 to R4) and contains various chemical fragments. Red arrows point from text annotations to specific elements in the interface:

- An arrow points from the text "Add to core (<1s)" to the highlighted R3.3 fragment in the tree view.
- An arrow points from the text "Cannot select if parent not selected" to the R5.2 fragment, which is not highlighted.
- An arrow points from the text "Child R groups can be selected" to the R10 fragment, which is highlighted.

The interface also includes a header with the iMarVis logo and the text "Interactive Molecular Visualization", a button to "Generate a PDF report of the chemical space subset for printout", and the Roche pRED informatics logo.

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

iMarVis Interface

Generate PDF Report

iMarVis
Interactive Markup Visualization

Generate a PDF report of the chemical space subset for printout

R Groups in Tree View

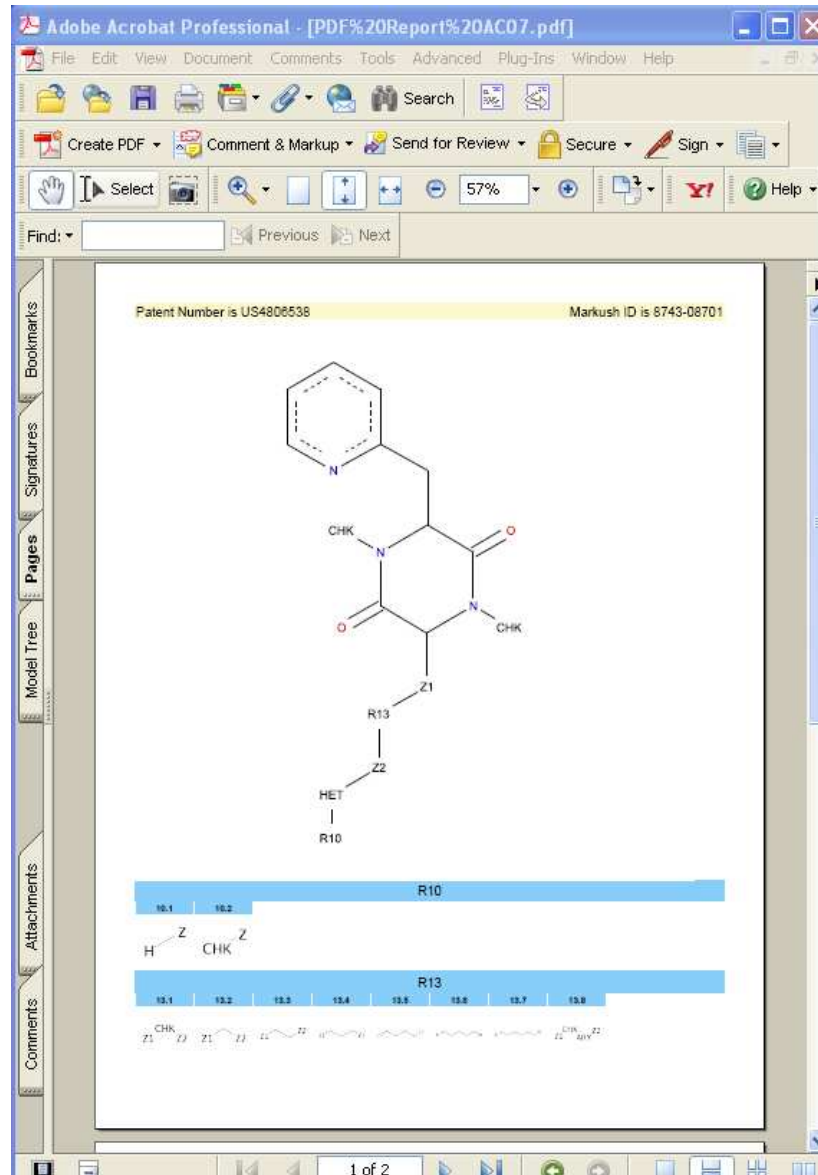
- R1
- R2
- R3
- R5
 - 5.1
 - 5.2
- R6
- R7
- R8
- R9
- R10
- R11
- R12
- R13
- R4
 - 4.1
 - 4.2
 - 4.3
 - 4.4
 - 4.5
 - 4.6
- R14
- R15
- R16

Roche
pRED informatics

Generate a PDF Report

iMarVis Interface

The PDF Report



Adobe Acrobat Professional - [PDF%20Report%20AC07.pdf]

File Edit View Document Comments Tools Advanced Plug-Ins Window Help

Create PDF Comment & Markup Send for Review Secure Sign

Find: Previous Next

Patent Number is US4806538 Markush ID is 8743-08701

Chemical structure diagram showing a central pyridine ring connected to a six-membered ring containing two nitrogen atoms (N) and two carbonyl groups (C=O). The nitrogen atoms are labeled with 'CHK'. The central ring is connected to a side chain containing a nitrogen atom (N) labeled 'Z1', which is further connected to a carbon atom (C) labeled 'R13', which is connected to a carbon atom (C) labeled 'Z2', which is connected to a hydrogen atom (H) labeled 'HET' and a carbon atom (C) labeled 'R10'.

10.1 10.2 R10

H Z CHK Z

10.1 10.2 10.3 10.4 10.5 10.6 10.7 10.8 R13

CHK Z1 Z2

1 of 2



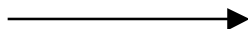
iMarVis Interface

Hyperlink for Easy Access

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



[US4806538](#)



[Patent Search | Reports](#)

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

[Disclaimer and Superatom Legend](#)

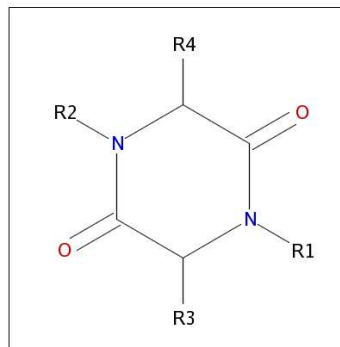
User: dengwl

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.



pRED Informatics



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