

Markush project – Introducing Markush DARC support

Szabolcs Csepregi

September 2010, Cxn US UGM, Boston

Acknowledgements

ChemAxon

Nóra Máté, Róbert Wagner, Szilárd Dóránt, Tamás Csizmazia, Tim Dudgeon, Erika Bíró, Ali Baharev, Ferenc Csizmadia, et al.

Thomson Reuters

Tim Miller, Steve Hajkowski, Gez Cross and Linda Clark

Many early adopters and colleagues in the field for suggestions and feedback.

Contents

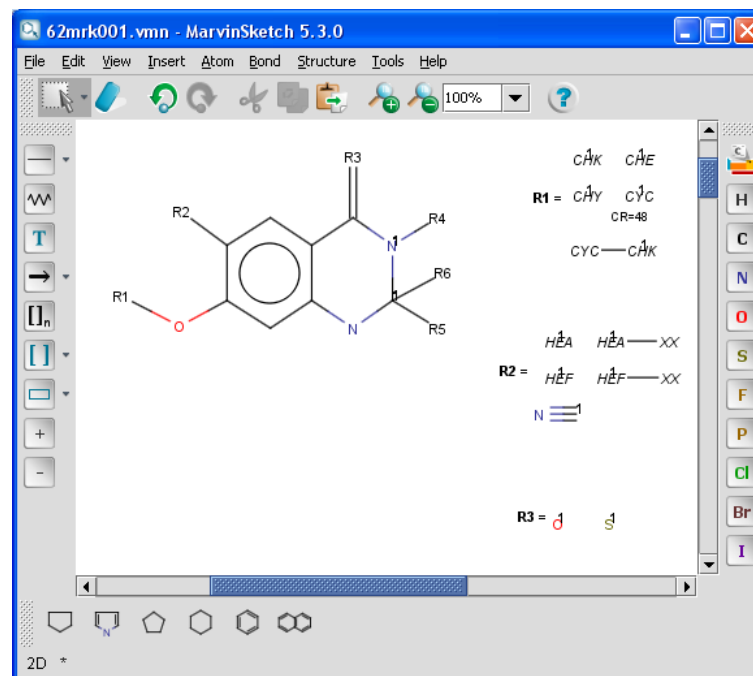
- Markush DARC
- Tools
 - Enumeration
 - Storage, search
- Challenges in chemical representation

Markush (generic) input

Markush DARC file format

Why?

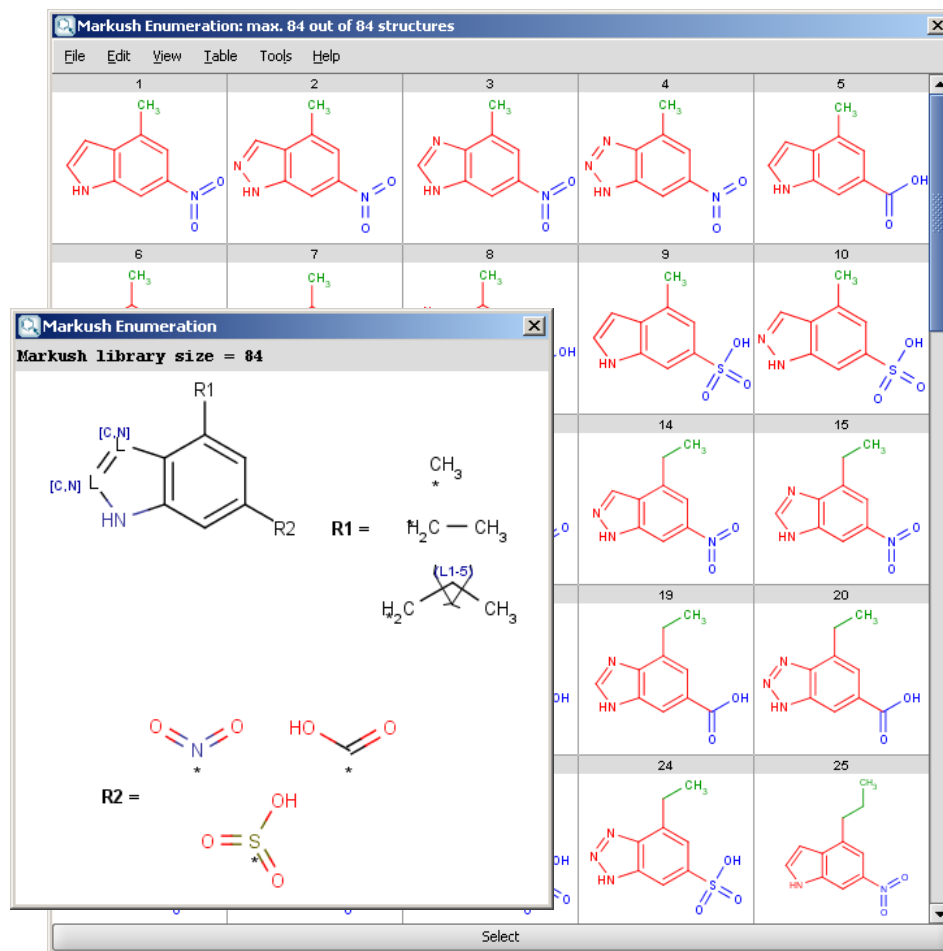
- MMS (Merged Markush Service)
- Patent data is available (soon) from Thomson Reuters
 - Markush structures
 - Associated patent data



What to do with them?

Markush Enumeration

- Markush enumeration plugin
 - Full enumeration
 - Selected parts only
 - Random enumeration
 - Calculate library size
 - Scaffold alignment and coloring
 - Markush code
 - Optional example homology group enumeration



Markush storage & search

- JChem Base and Instant JChem
- No enumeration involved
- Can handle complex Markush structures (10^{40} or more)

The screenshot shows the Instant JChem 2.5.2 interface. The main window displays a table with the following data:

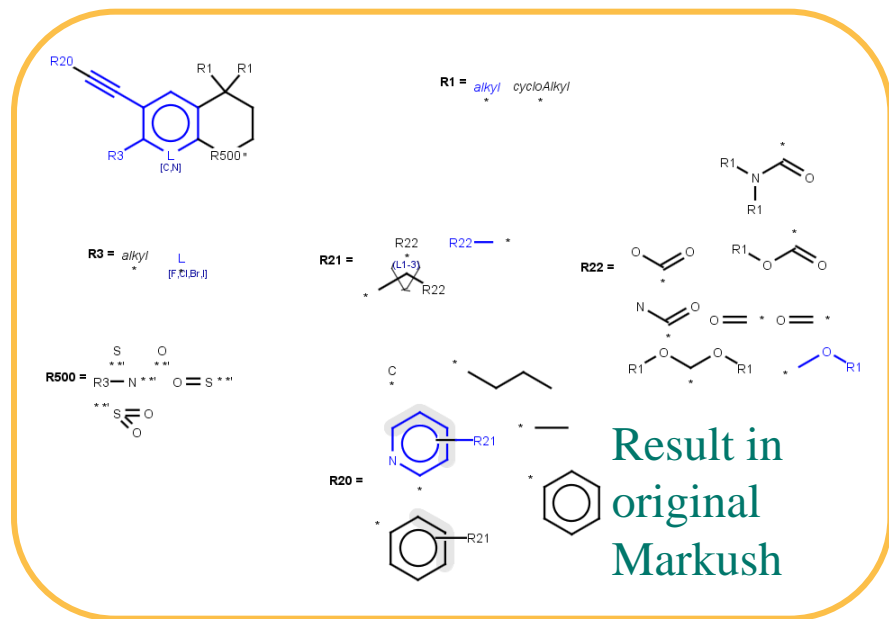
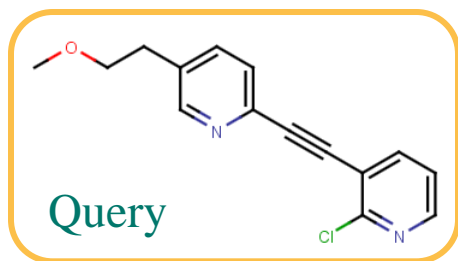
Cdid	Markush structure	Library Size	Library size (calculated)
1		2,735,568.00	
2		2,445,552.00	2445552
3		5,934,096.00	5934096
4		7.63E09	7834640000
5			

The 'Query - Grid view for Markush_table' window shows a substructure search interface with a chemical structure of a pyridine ring and a methyl group.

- Substructure and Full structure search
- Broad translation of homology groups is supported. (Homology in DB, specific in query.)

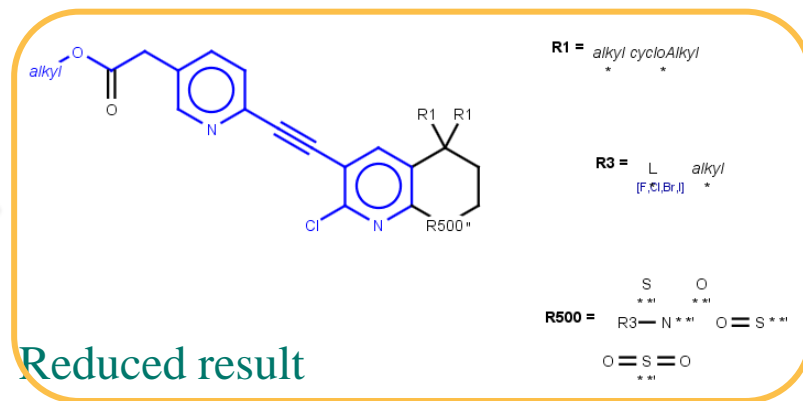
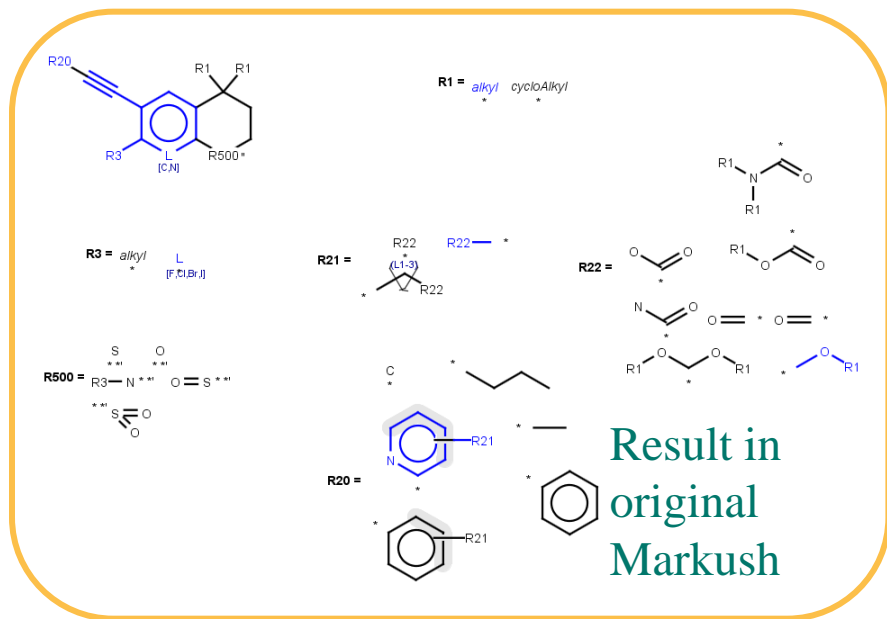
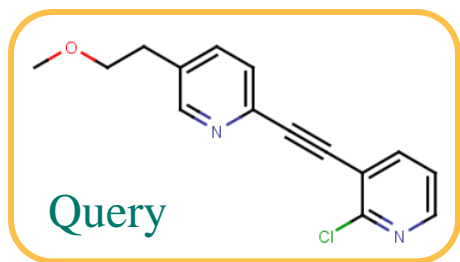
Markush storage & search

Substructure hit visualization



Markush storage & search

Substructure hit visualization:
„Markush structure reduction”



Main use cases

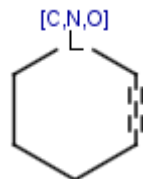
- Patent search hits refining / visualization,
- White space analysis,
- Patent busting,
- Markush structure curation,
- In-house storage of small Markush DB,
- etc...

Challenges in chemical representation (solved)

Representation - What we already had

Generic notation in queries:

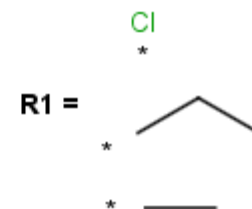
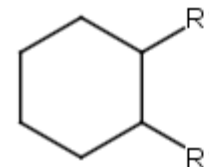
- Atom lists, bond lists



Single or double

- R-group queries

(Problem: RGFile R-logic and patent R-logic are different! - Solution: Just ignore R-logic.)



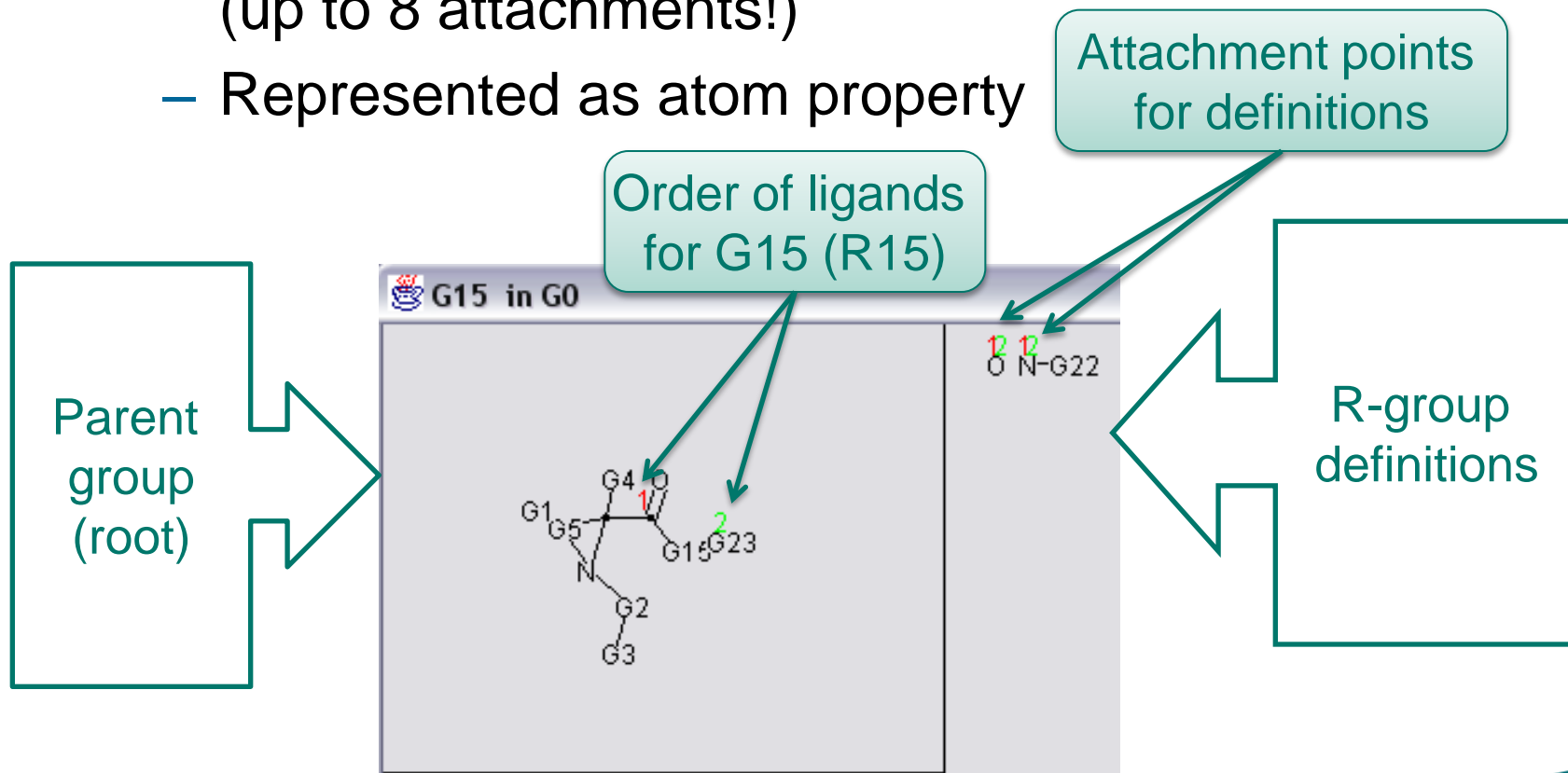
- Link nodes $\text{O} - \left\{ \begin{matrix} \text{L1-3} \\ \text{C} \end{matrix} \right\} - \text{N}$

- Some generic atoms (X) – represented as pseudo atoms.

Challenge 1: Attachment point

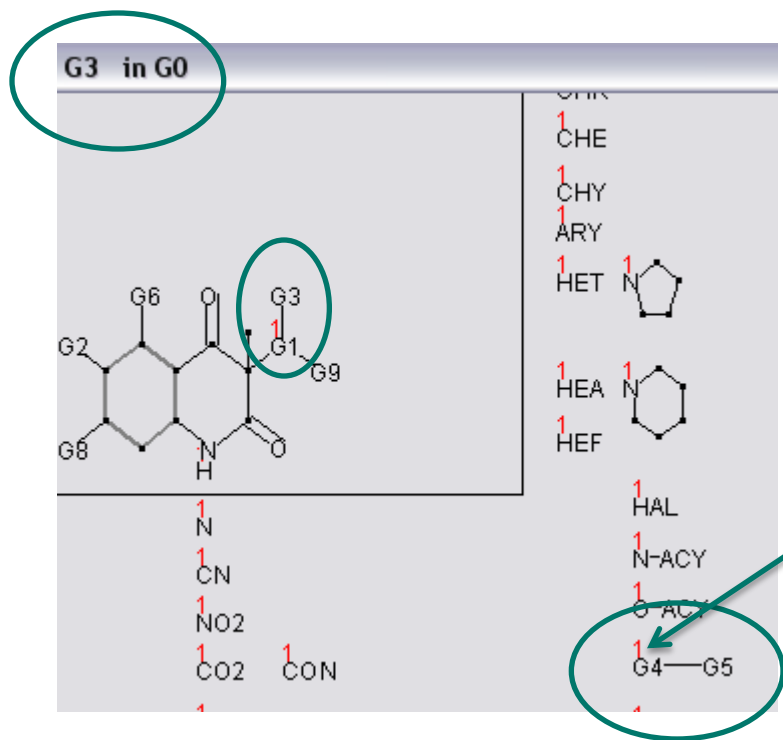
Multiple

- Heavily used in Markush DARC (up to 8 attachments!)
- Represented as atom property



Challenge 1: Attachment point

Embedded R-groups: Grandparent relations may be needed between attachment points:



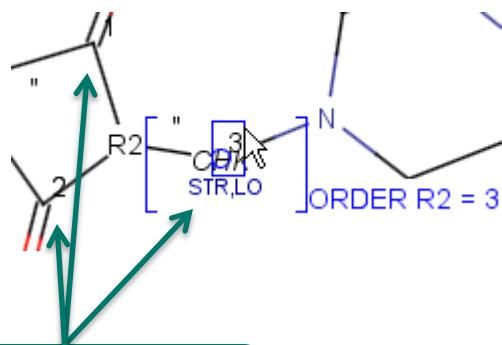
G3's attachment point „1” is mapped to G4's attachment point „1”



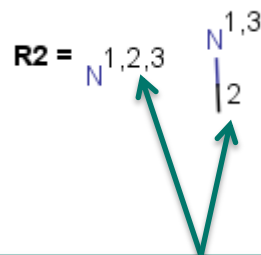
Challenge 1: Attachment point

Temporary representation: attached data

- ligand order
- attachment point in R-group definition
- still an atom property
- ligand order sometimes in parent group (grandparent relation)



Order of ligands
for R2

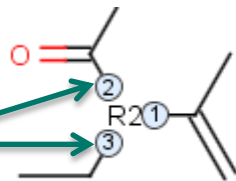


Attachment points
for definitions

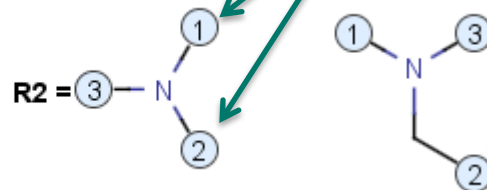
Challenge 1: Attachment point

Real attachment object with bond
(under development)

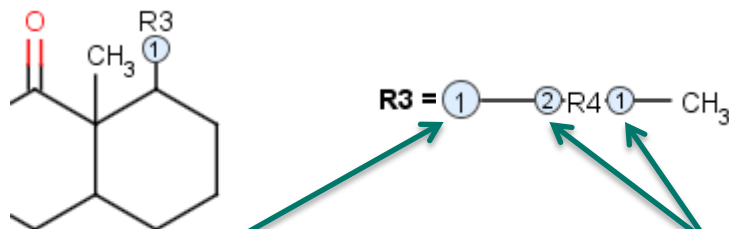
Order of ligands
for R2



Attachment points
for definitions



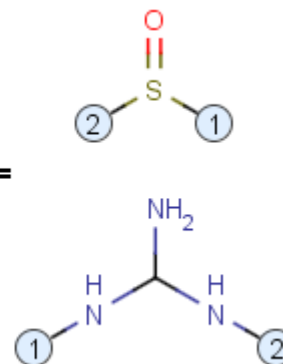
Eliminates need for grandparent relations table:



Attachment point
for R3

Order of ligands
for R4

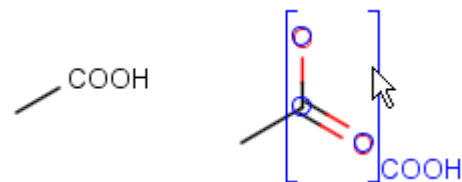
R4 =



Challenge 2: Abbreviations

- Were originally in Marvin

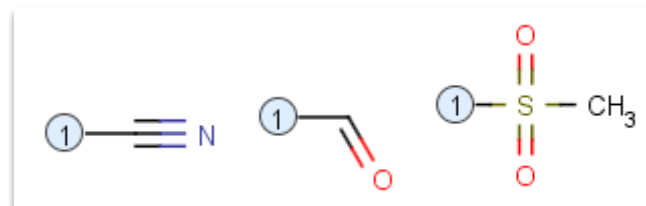
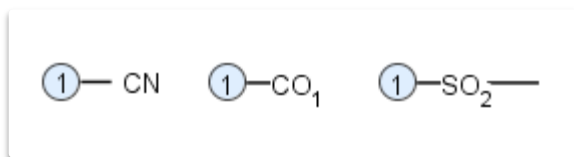
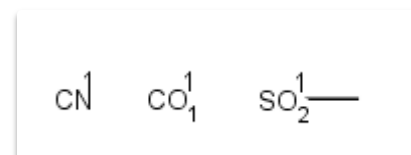
- ~700 built-in shortcuts
- Expand / Contract
- Search understands them



- M. DARC had 21 shortcuts + 31 peptides.

- Attachment point

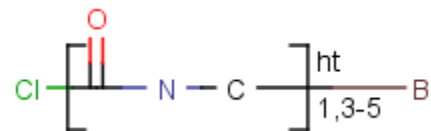
- Visible „outside” and connected „inside”.
- New attachment point solves this also:



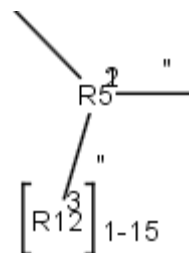
Challenge 4: Frequency variation

- Link nodes 

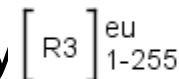
- Repeating units: modified SRU



- Multipliers:
 - special SRU, 1 outer bonds.
(Currently visualization only.)



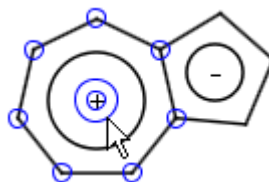
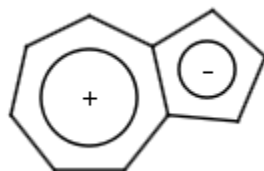
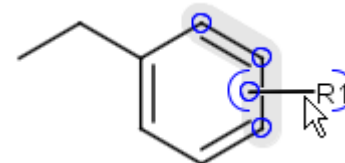
- Moieties:
 - special SRU, 0 outer bonds
 - to describe (variable) stoichiometry
(Currently visualization only.)



R1 = TRM LAN
ACT

Challenge 5: Position variation bond

- New special S-group type
- Relocatable multicenter atom
- Also useful to represent multicenter charge and coordination compounds:



What (else) keep us busy

Under development

Further improvements in Markush DARC support:

Ring segment groups (XX form a ring)

More robust representation for attachment points

Homology codes (low and high order, C1, B, N, 5, C)

Ranking of results

New ways to navigate zoom Markush structures

Maximum common substructure search

Biased enumeration and covering Markush – based on examples

Improve search speed to handle larger Markush sets.

Other Markush formats – Markush InChI standard committee

Overlap analysis of Markush structures

Conditions for Markush variables

Please come to the Markush forum tomorrow Morning

Summary

- Markush structure storage, search and enumeration at ChemAxon now patent coverage
- Compatible patent data is available from Thomson Reuters
- Well thought out chemical representation
- Continuous development, improvements in the pipeline