

# Tools for Analyzing Exemplified and Markush Structures in Chemical Patents

Christopher E. Kibbey, Jacquelyn L. Klug-McLeod, Bruce A. Lefker,  
Mark A. Mitchell, and Robert Owen

# Applying chemical patents to drug design

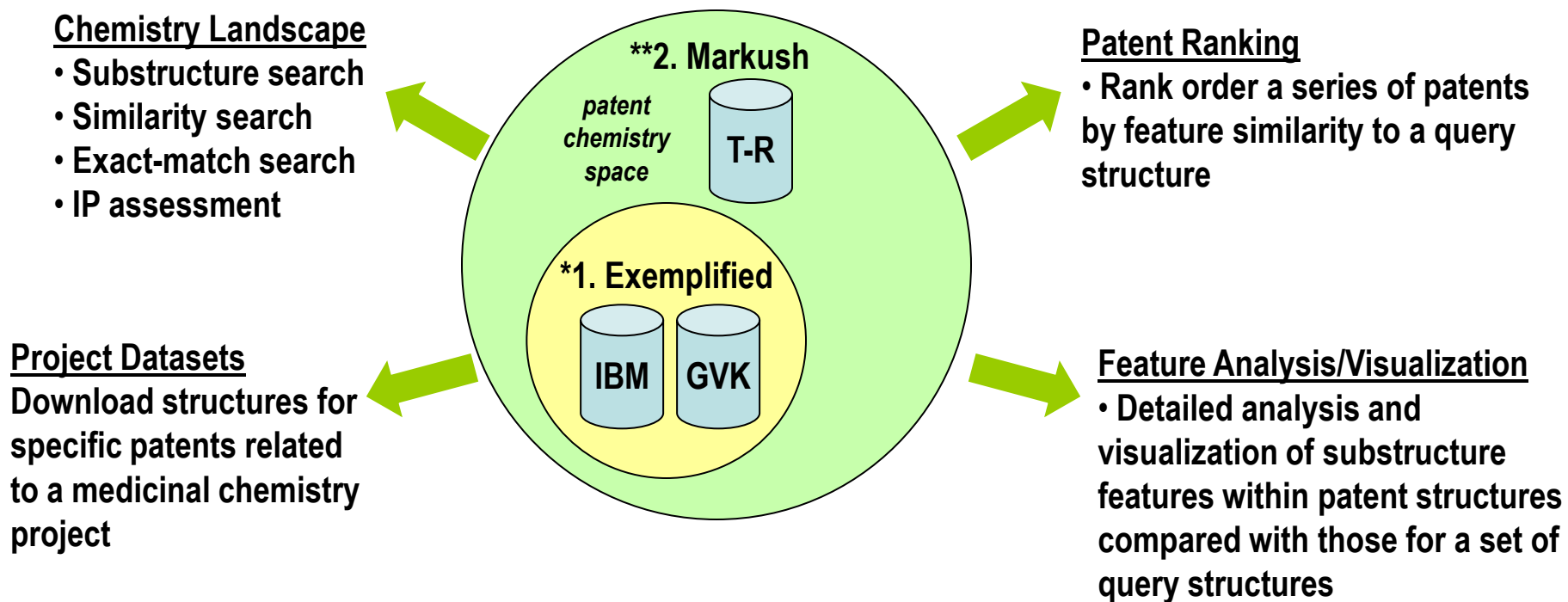
**Understanding how a medicinal chemistry project's chemical matter relates to the external landscape is an essential component of an overall design strategy:**

- Influence the assessment of multiple series for early projects
- Assess the strengths and weaknesses of our chemical matter relative to our competitors
- Facilitate identification of unexplored areas of chemical space in a competitor's IP
- Drive patent strategies to strike a balance between cost and maintaining our competitive advantage

**However, efficiently obtaining this information can be tricky and tedious, especially in crowded or rapidly changing environments.**

- Access to electronic structures from patents and appropriate tools for analysis is critical to success

# Patent structure databases in use at Pfizer and their application to drug design



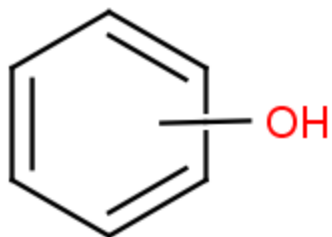
\*IBM database of ~8 million exemplified structures from patents (nightly updates)

\*GVK database of ~3 million exemplified structures from patents (quarterly updates)

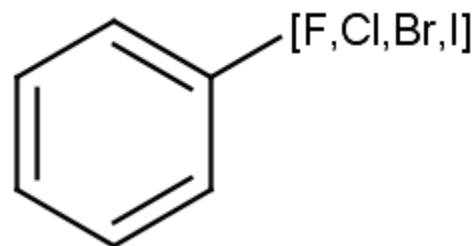
\*\*Thomson-Reuters database of 1.2 million Markush structures from patents (weekly updates)

# Types of structure variation encoded in exemplified and Markush structures in patents

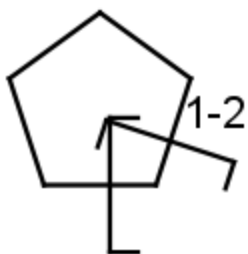
Position (p-variation)



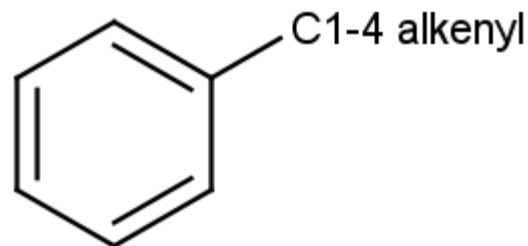
Substitution (s-variation)



Frequency (f-variation)

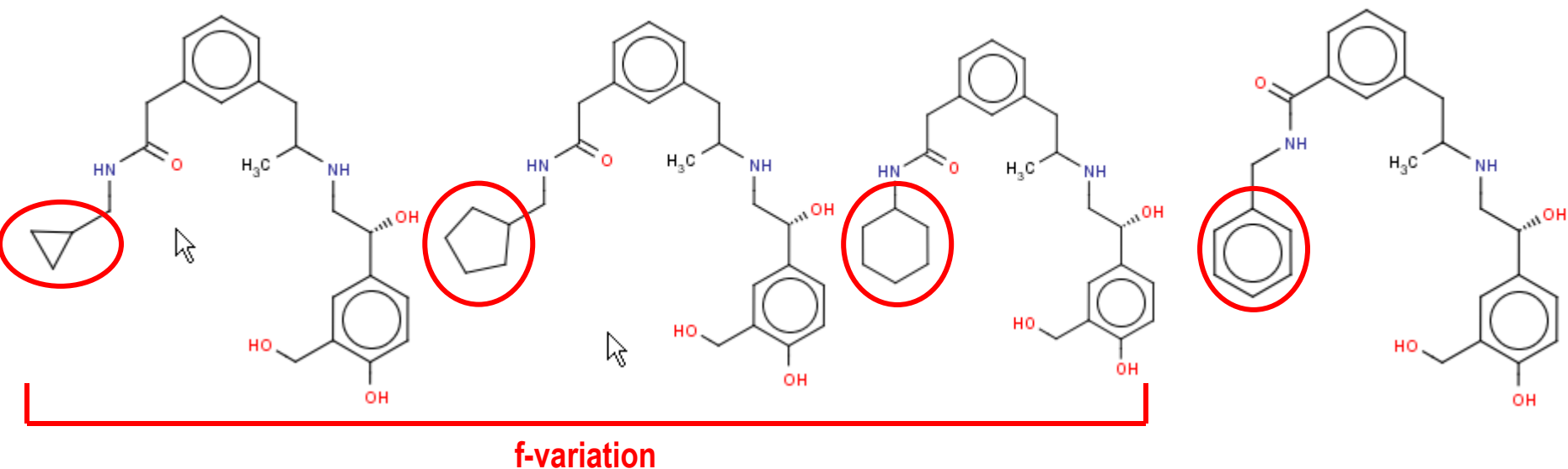


Homology (h-variation)



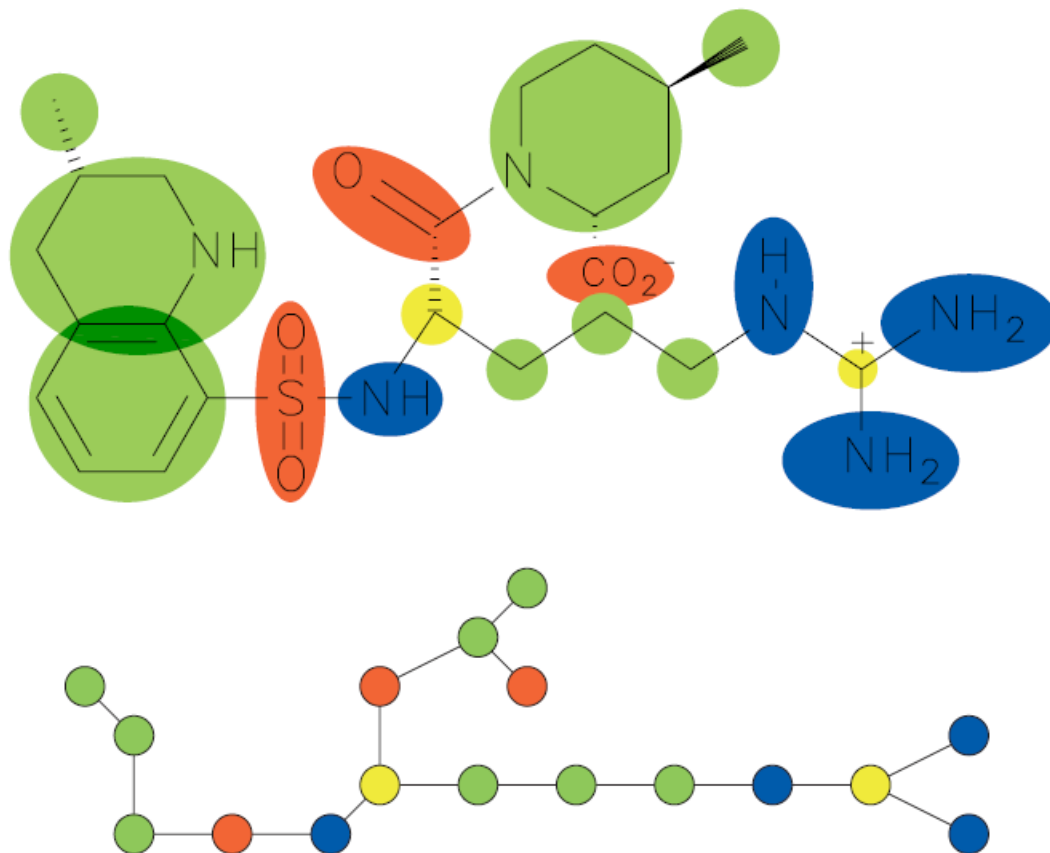
# Substructure and fingerprint-based similarity search results can be misleading when applied to structures in patents

Substructure search cannot account for h-, s-, p-, and f-variation



Furthermore, fingerprint-based similarity search algorithms may not take connectivity of fingerprint features into account

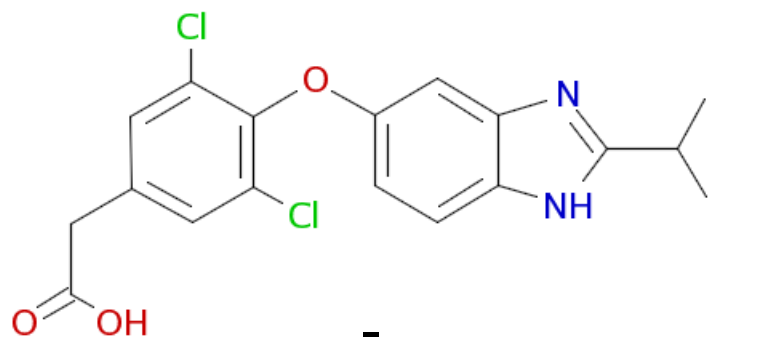
# Reduced graphs offer an alternative to fingerprint-based similarity search algorithms



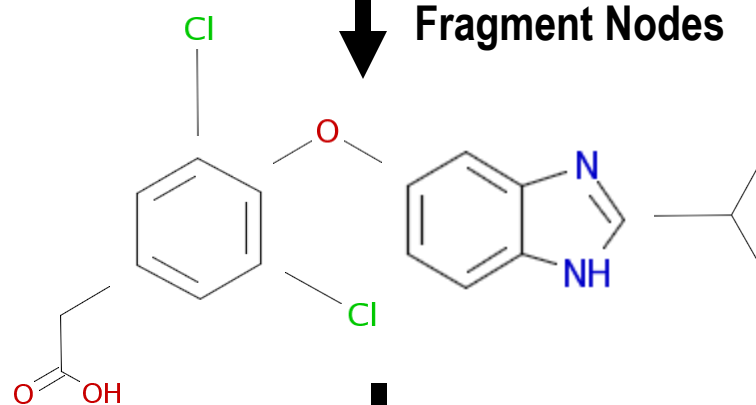
Feature trees: A new molecular similarity measure based on tree matching

M. Rarey, J. Dixon, J. Computer-Aided Molecular Design, 1998, 12, 471-490

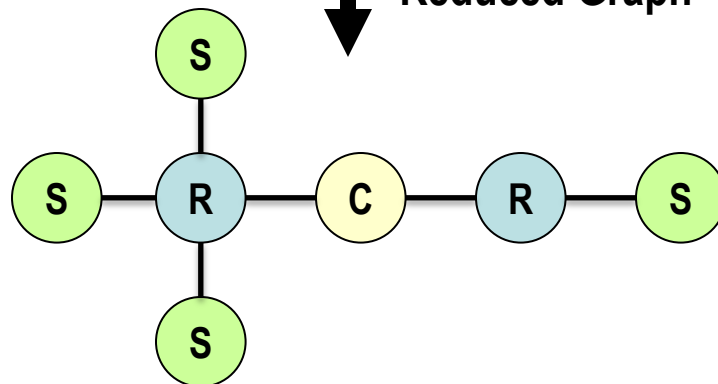
# Feature Analysis – Reduced graph representation of a structure



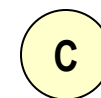
Fragment Nodes



Reduced Graph



Ring



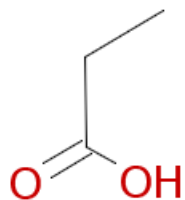
Chain



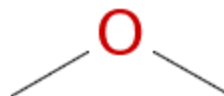
Substituent

# Feature Analysis - Encoding node fingerprints

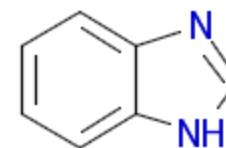
Substituents-S	Chains-C	Rings-R
Contains [F,Cl,Br,I]	Contains [F,Cl,Br,I]	Aromatic
Contains C	Contains C	Fused
Contains N	Contains N	Fused and Fully Aromatic
Contains O	Contains O	Contains C
Contains S	Contains S	Contains N
Has Single Bond	Has Single Bond	Contains O
Has Double Bond	Has Double Bond	Contains S
Has Carbonyl/Sulfonyl Bond	Has Carbonyl/Sulfonyl Bond	Has Single Bond
Has Triple Bond	Has Triple Bond	Has Double Bond
Is Branched	Is Branched	Has Carbonyl/Sulfonyl Bond
Has Charged Atoms		Has Triple Bond



S01010101000

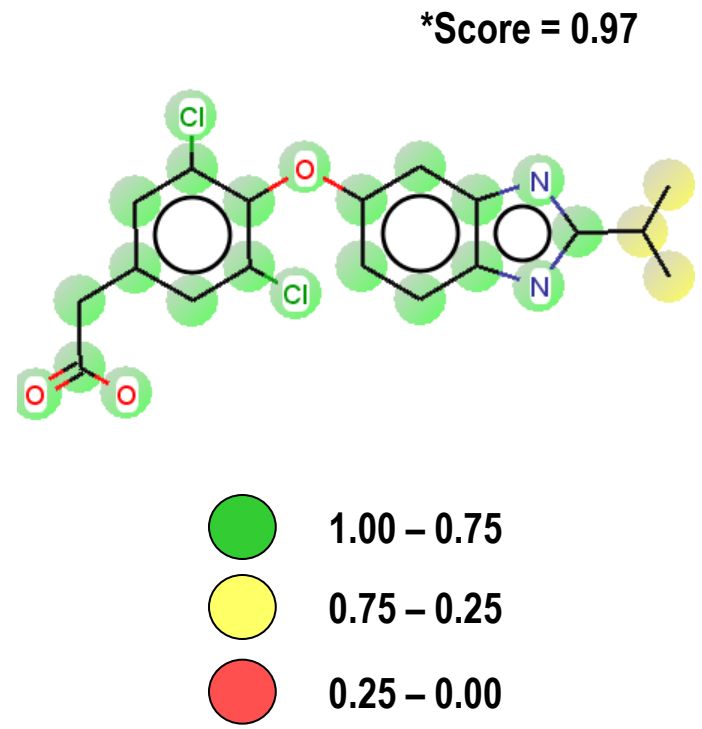
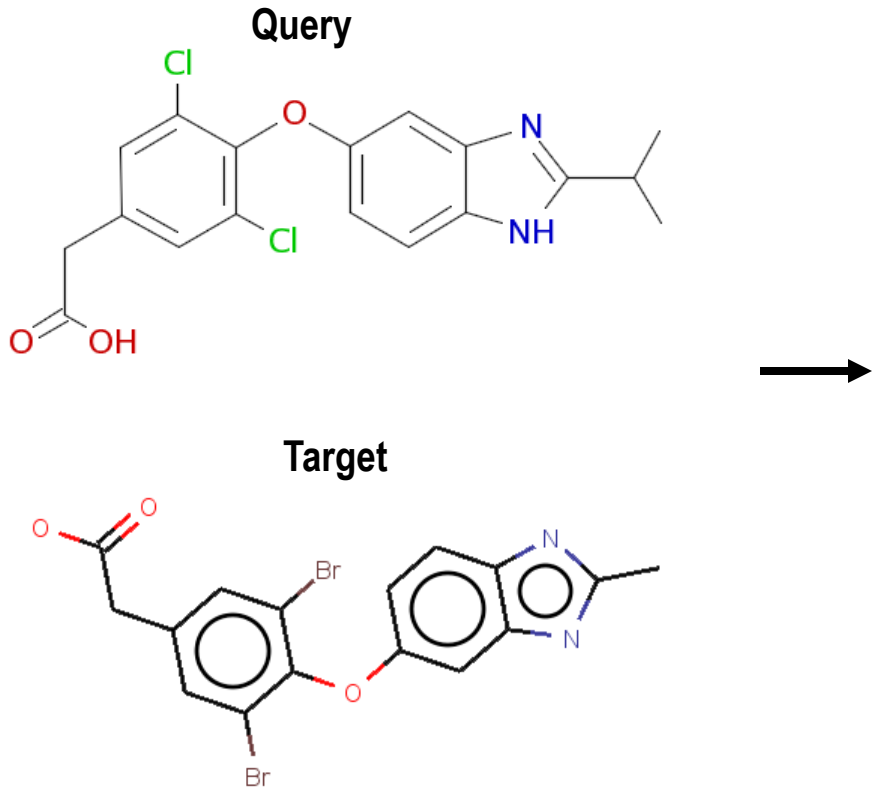


C0101010000



R11111000000

# Feature Analysis - Overlaying reduced graphs and scoring



\*Score = weighted average similarity for matched nodes – penalty for unmatched nodes

# Advantages of applying Feature Analysis to structures in patents

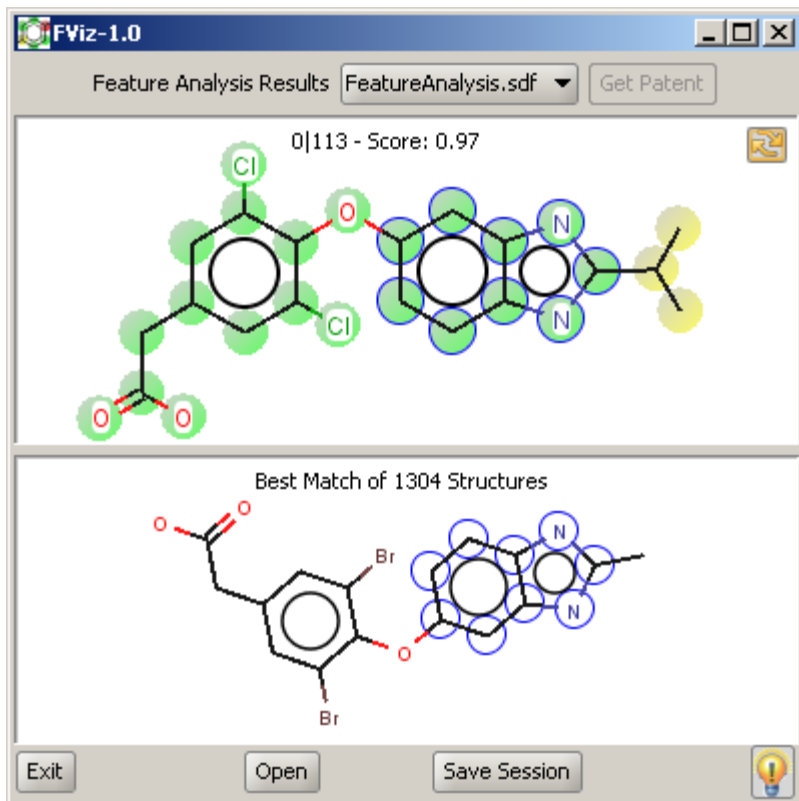
Exemplified structures can be represented by reduced graphs consisting of inter-connected substituent, chain, and ring nodes

FA algorithm is compatible with f-, h-, p- and s-variation represented in patent structures (both exemplified and Markush)

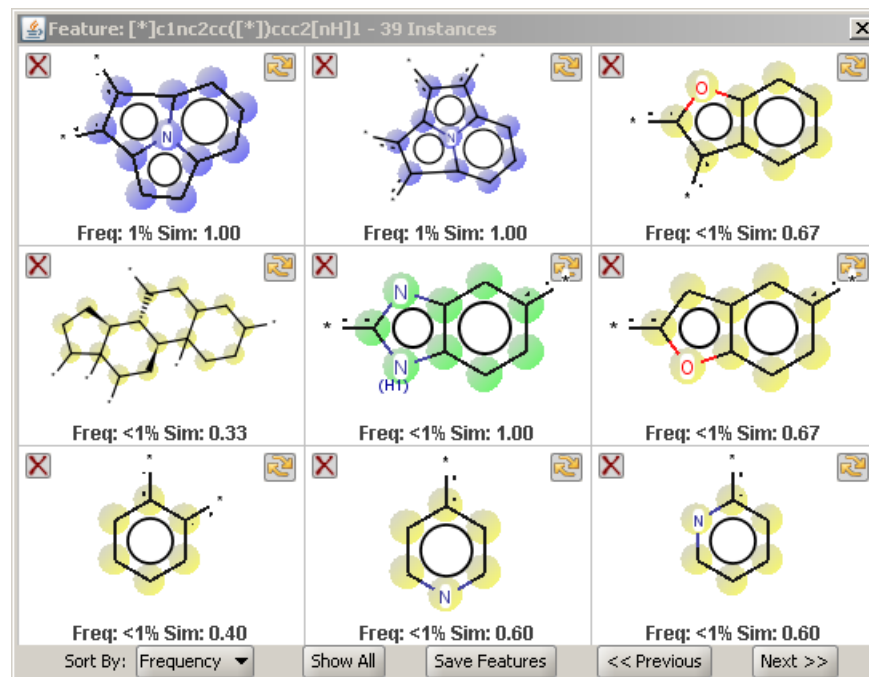
- f-variation: allows for variation in chain, ring and substituent size by atom type
- h-variation: various homology group definitions can be encoded in node fingerprints
- p-variation: algorithm ignores attachment geometry among R, C, S nodes
- s-variation: algorithm accommodates differences in substitution pattern between pairs of reduced graphs being compared

The output of Feature Analysis provides a “similarity-like” score and a “substructure-like” match of ring, chain, and substituent features

# Visualizing the results of Feature Analysis

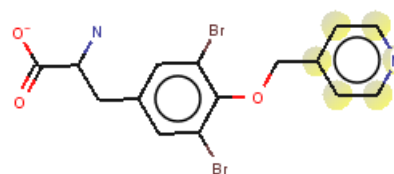
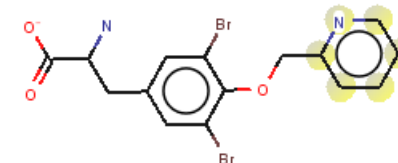
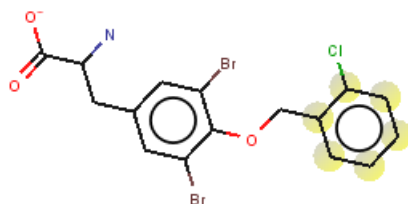
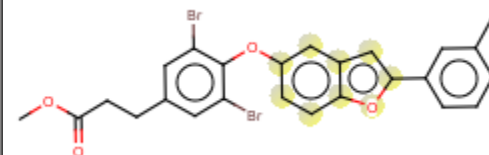
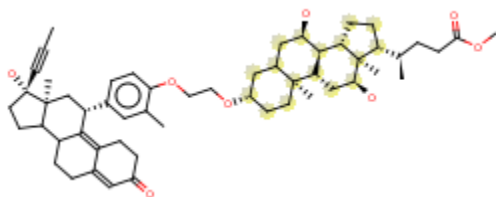
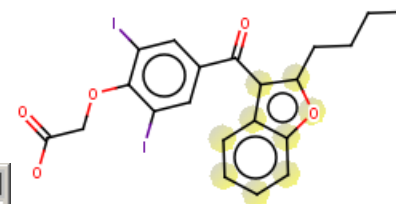
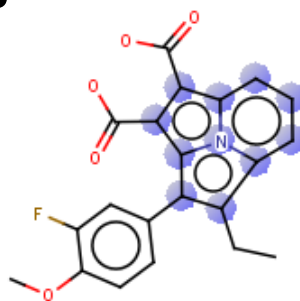
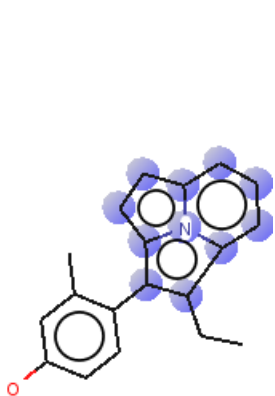


Per-atom differences highlighted  
in blue for 100% similar fragments



Select a query feature to display  
corresponding sub-structure  
features for exemplified compounds

# Benzimidazole ring in the query structure maps to a variety of ring nodes within the target set



Feature: [\*]c1nc2cc([\*])ccc2[nH]1 - 39 Instances

Freq: 1% Sim: 1.00	Freq: 1% Sim: 1.00	Freq: <1% Sim: 0.67
Freq: <1% Sim: 0.33	Freq: <1% Sim: 1.00	Freq: <1% Sim: 0.67
Freq: <1% Sim: 0.40	Freq: <1% Sim: 0.60	Freq: <1% Sim: 0.60

Sort By: Frequency Show All Save Features << Previous Next >>

# Feature Analysis applied to ranking patents

Color-coding of query features to best match in patent

Feature Score for best match in patent

Distribution of scores for all structures in patent

Structure of best match in patent

Id	FEATURES_ATOMPROP	BestScore	BestMatch_stx	PatentNumbe...	BinnedScores
		1		<a href="#">US6346532</a>	[0.44, 0.21, 0.24, 0.03, 0.07, 0.00, 0.00, 0.00, 0.00, 0.00]
		0.81		<a href="#">WO200509026</a>	[0.00, 0.02, 0.72, 0.22, 0.04, 0.00, 0.00, 0.00, 0.00, 0.00]
<a href="#">Molecule1 Instance 3</a>		0.81		<a href="#">US2005023406</a>	[0.00, 0.03, 0.67, 0.17, 0.12, 0.01, 0.00, 0.00, 0.00, 0.00]
<a href="#">Molecule1 Instance 4</a>		0.76		<a href="#">US6362371</a>	[0.00, 0.00, 0.04, 0.04, 0.33, 0.44, 0.07, 0.04, 0.04, 0.00]
<a href="#">Molecule1 Instance 5</a>		0.74		<a href="#">US5599966</a>	[0.00, 0.00, 0.04, 0.18, 0.58, 0.09, 0.11, 0.02, 0.00, 0.00]
<a href="#">Molecule1 Instance 6</a>		0.74		<a href="#">US5153210</a>	[0.00, 0.00, 0.31, 0.42, 0.16, 0.05, 0.06, 0.00, 0.00, 0.00]
<a href="#">Molecule1 Instance 7</a>		0.7		<a href="#">US2006011636</a>	[0.00, 0.00, 0.14, 0.86, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00]



# Markush analysis in drug design

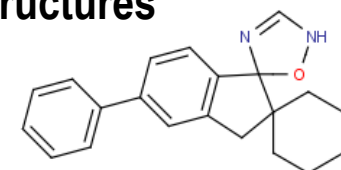
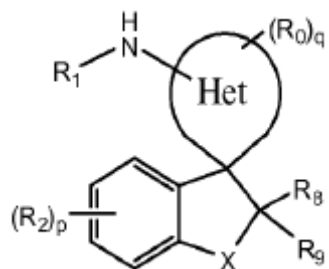
1. Influence the assessment of multiple series for early projects
  - What is the overlap of the project series with the patent Markush?
  - What regions of the project series exhibit potential novelty?
2. Facilitate identification of unexplored areas of chemical space
  - What regions of the Markush are under-represented by the exemplified structures in the patent?
  - Are these under-represented regions covered by granted claims?

Substructure and exact match searches may be ineffective when applied to patents represented by multiple Markush structures

Similarity search against Markush structures is not supported by the JChem toolkit

# Multiple Markush structures exist in the T-R database for a given patent, which can make substructure searching problematic

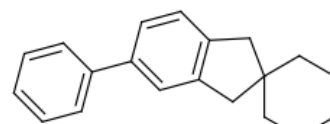
## WO2010105179 – 24 Markush Structures



Substructure Search

No Hits

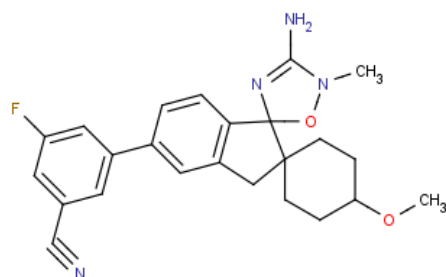
Markush Search and Analysis software interface showing 'No hits found' for the first substructure search.



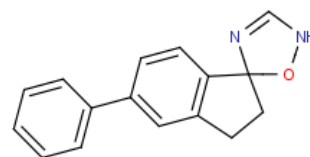
Substructure Search

1-12

Markush Search and Analysis software interface showing 12 hits for the second substructure search. A red circle highlights a specific hit.



Exemplified Structure



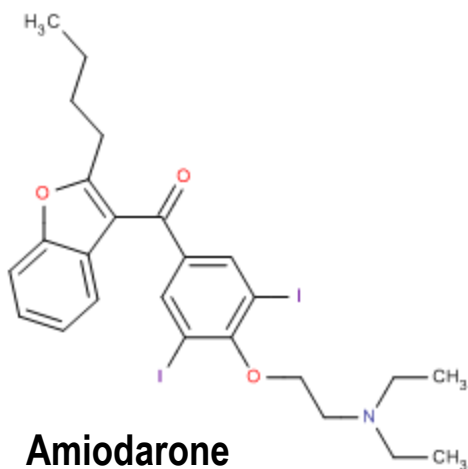
Substructure Search

13-24

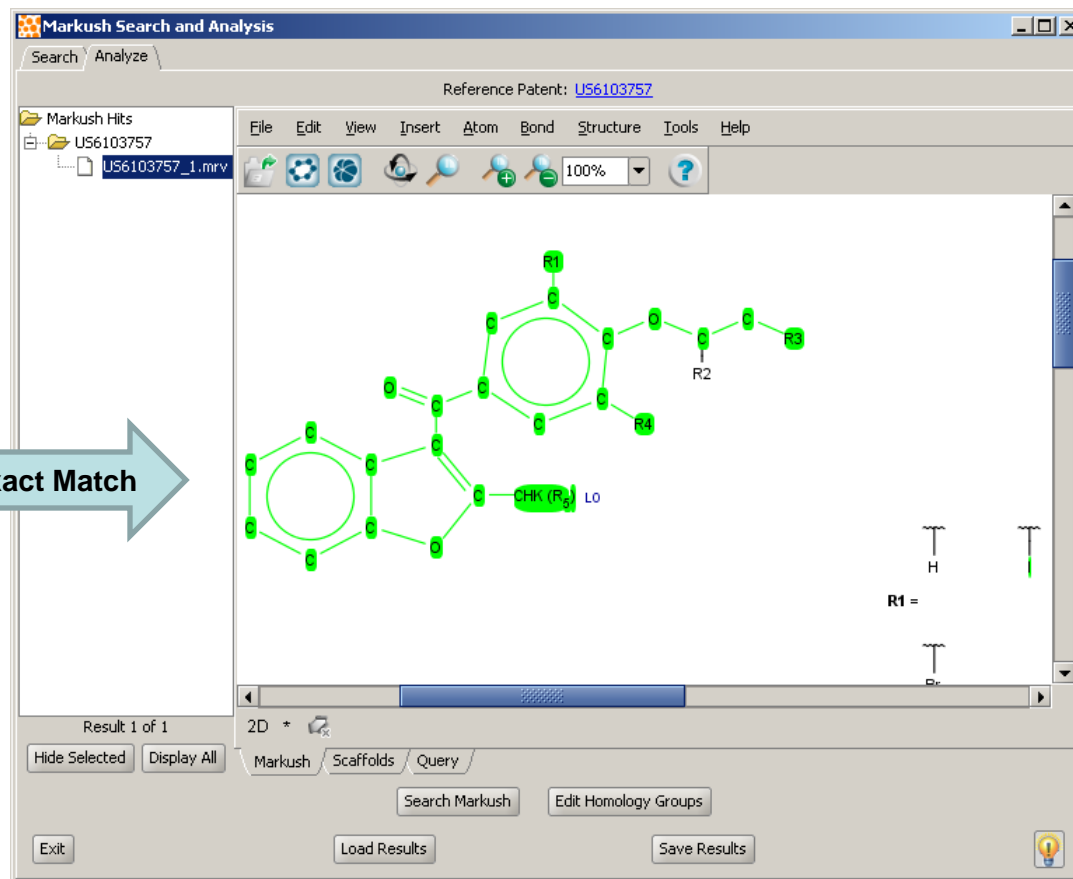
Markush Search and Analysis software interface showing 12 hits for the third substructure search. A red circle highlights a specific hit.

# Patents for which a single Markush structure is defined are amenable to substructure and exact match searching

US6103757 – Methods for treating arrhythmia using acetate buffer solutions of amiodarone



Exact Match

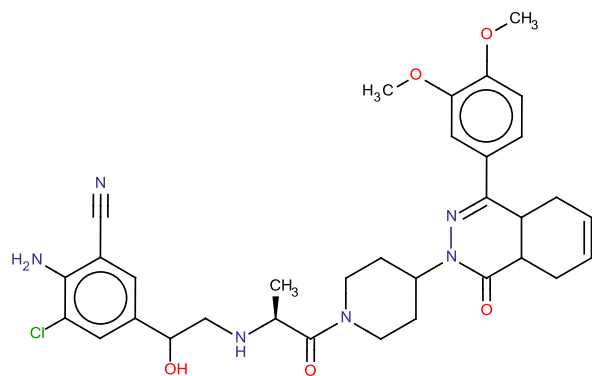
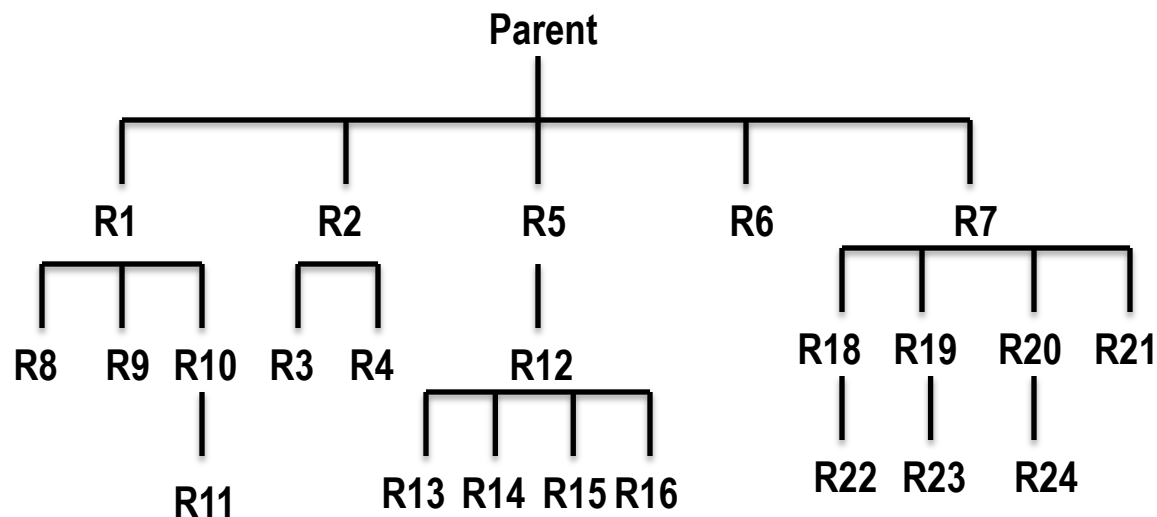


The screenshot shows the 'Markush Search and Analysis' software interface. The window title is 'Markush Search and Analysis'. The 'Reference Patent' is listed as 'US6103757'. The 'Markush Hits' list on the left shows 'US6103757' and 'US6103757\_1.mrv'. The main display area shows a 2D chemical structure of a Markush structure, which is a benzofuran derivative with various substituents labeled R1, R2, R3, and R4. The structure is highlighted in green. The interface includes a menu bar (File, Edit, View, Insert, Atom, Bond, Structure, Tools, Help), a toolbar with icons for search, zoom, and other functions, and a status bar at the bottom with buttons for 'Hide Selected', 'Display All', 'Markush', 'Scaffolds', 'Query', 'Search Markush', 'Edit Homology Groups', 'Exit', 'Load Results', and 'Save Results'.

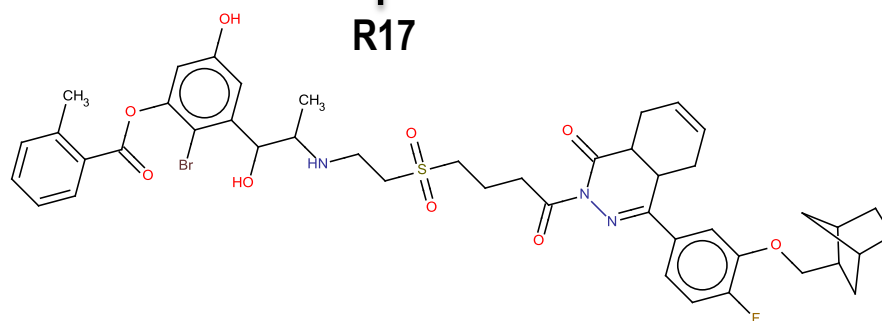
51% of the patents in the T-R database are encoded as a single Markush structure

# Enumeration of virtual libraries from a Markush is an ineffective strategy for (similarity) analysis

Random enumeration generates complete structures using randomly chosen Rgroup instances from the Markush



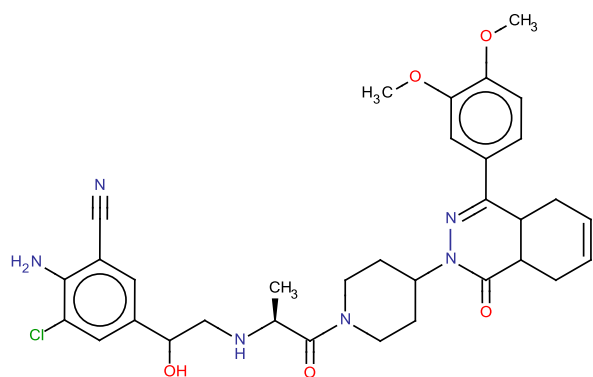
WO2001094319 – Exemplified



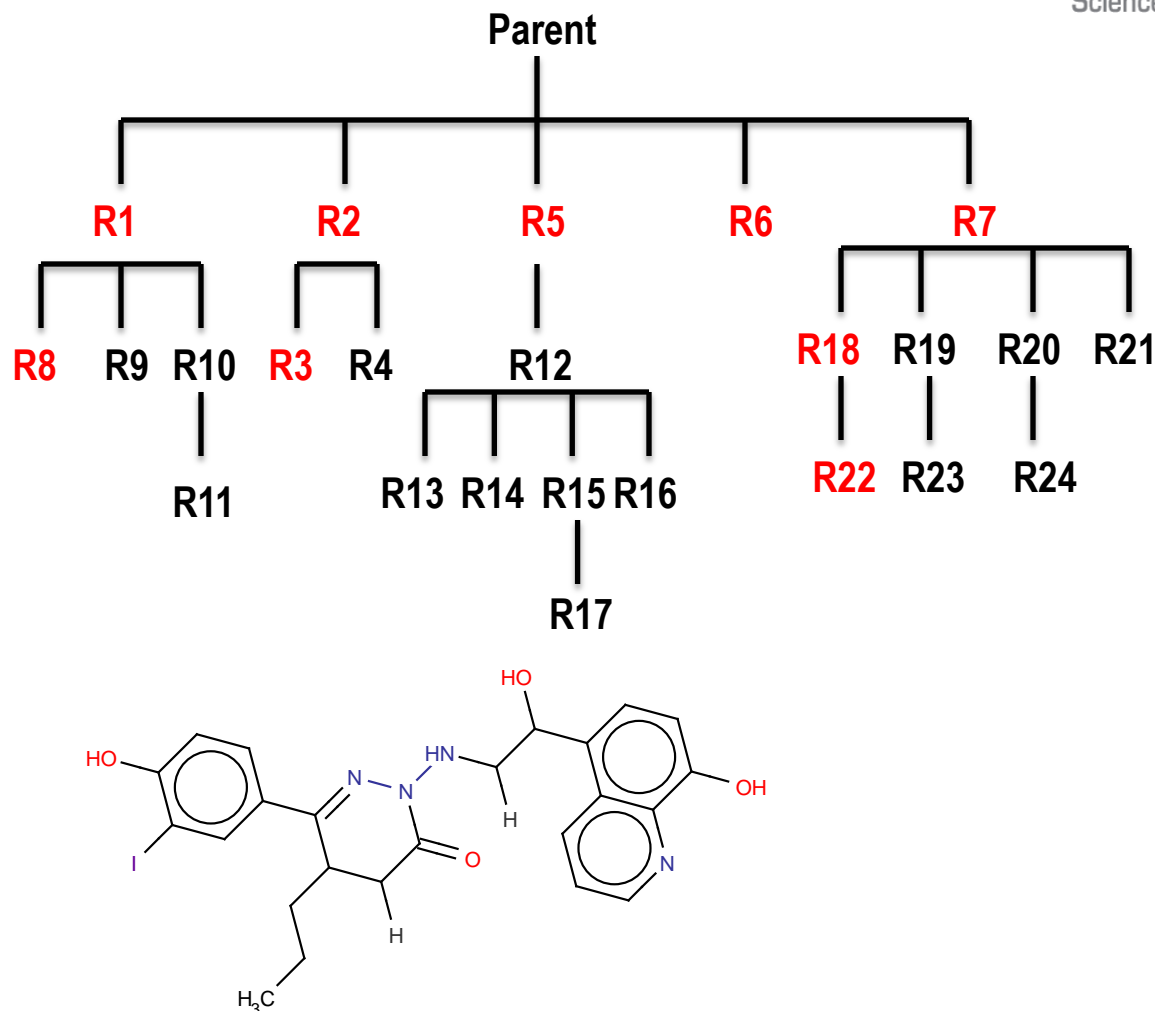
Random enumeration yields structures with variation far-away from the core scaffold within an extremely large virtual space (e.g.,  $\sim 10^{15}$ )

# Enumeration of virtual libraries from a Markush is an ineffective strategy for (similarity) analysis

Sequential enumeration generates structures using only the first instance defined at each Rgroup in the Markush

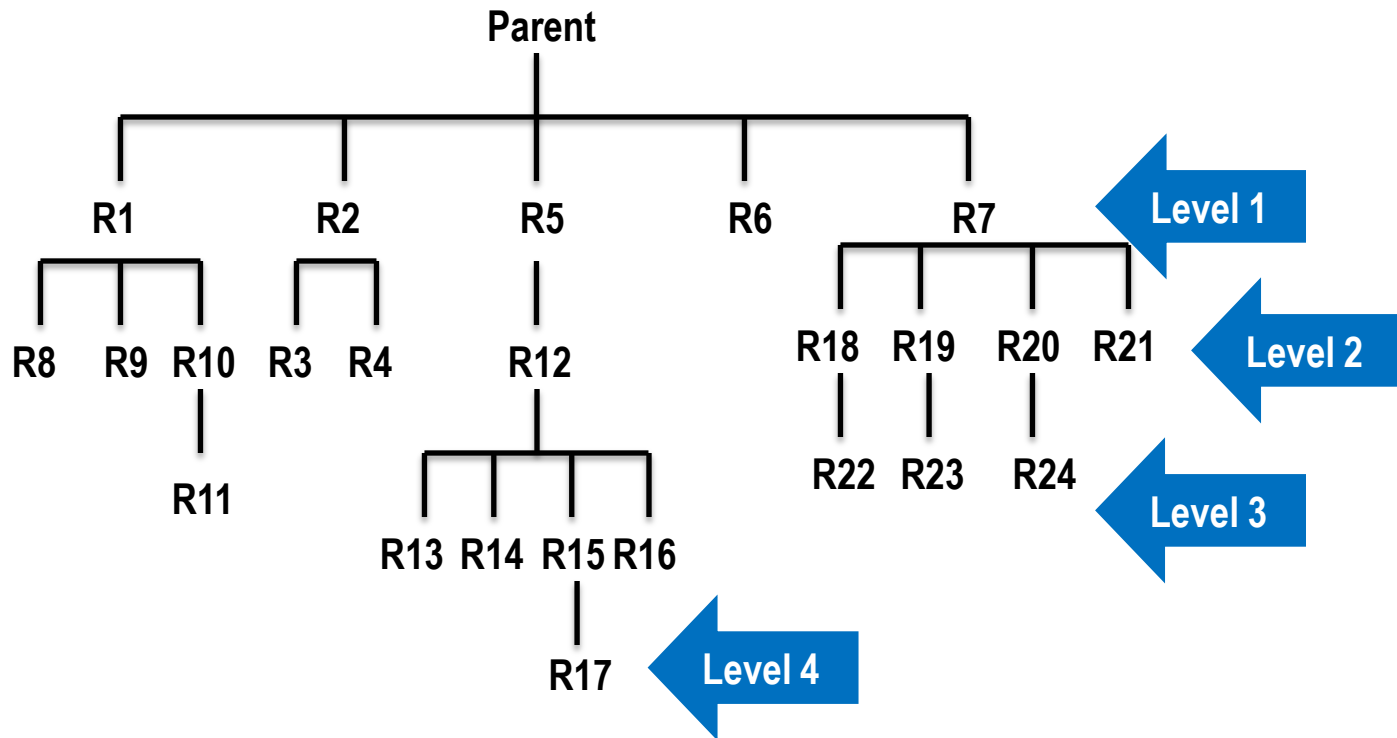


WO2001094319 – Exemplified



Sequential enumeration generates smaller virtual libraries of close-in structures, but ignores much of the chemical space within the Markush.

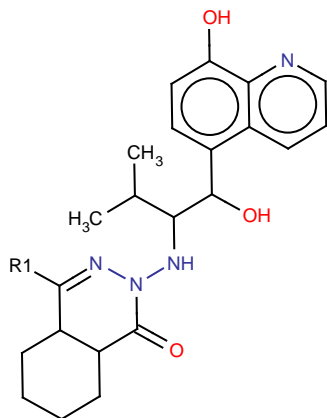
# Level enumeration generates structures through random enumeration of Rgroups up to a maximum specified nesting depth



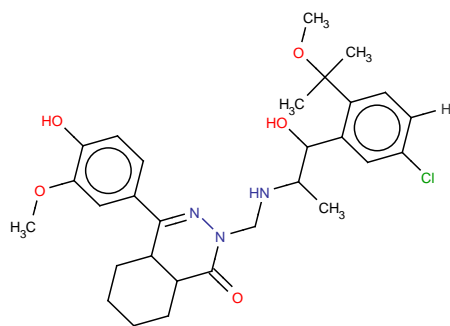
MMS allows up to 50 Rgroups with 4 levels of nesting per Markush structure

# Level enumeration yields smaller libraries of close-in structures while maintaining representative Rgroup coverage

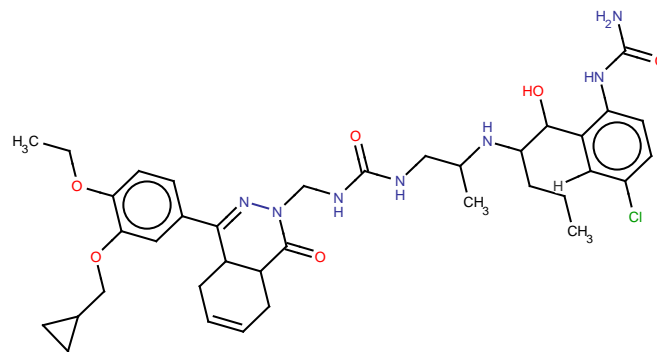
Level 1  
Library Size = 24



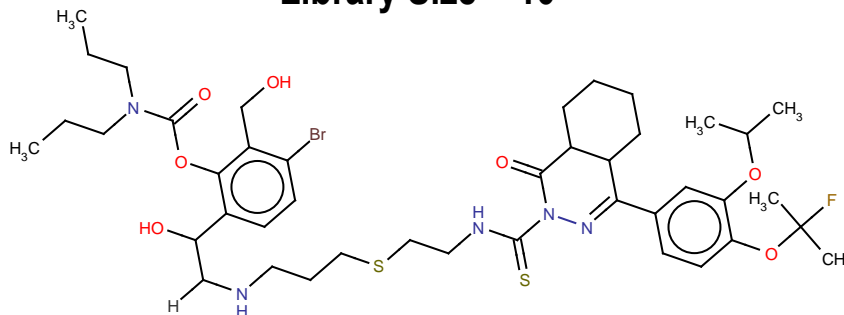
Level 2  
Library Size = 10<sup>11</sup>



Level 3  
Library Size = 10<sup>14</sup>



Level 4  
Library Size = 10<sup>15</sup>



# Improved strategy for Markush analysis (Level enumeration + Feature Analysis)

**Apply level enumeration within a Markush to generate a representative set of close-in structures**

**Encode the enumerated structures as a set of reduced-graphs**

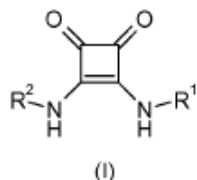
- **Examine Rgroup definitions to encode ring, chain, substituent nodes**
- **Eliminate duplicate Rgroups that map to the same node-type**

**Compare the Markush reduced-graphs with that for a query structure**

- **Score the best match (IP Assessment)**
- **Encode target node mapping for the best match within the query structure**
- **Return the corresponding structure from level enumeration as the best match for the patent Markush structure**

# Additional use case for Markush analysis

## Generate a Markush structure directly from patent text

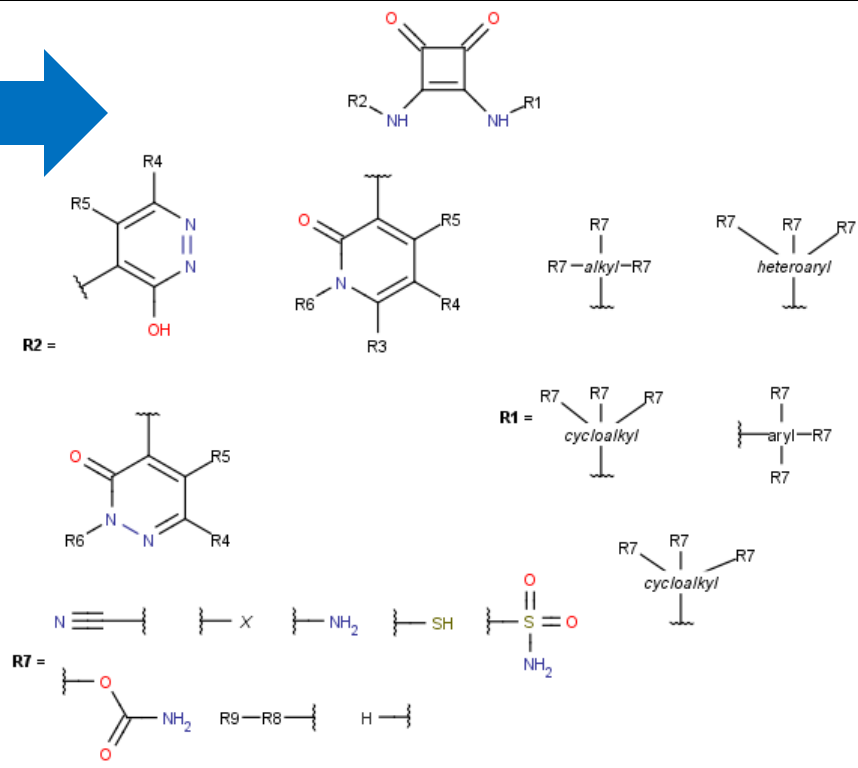
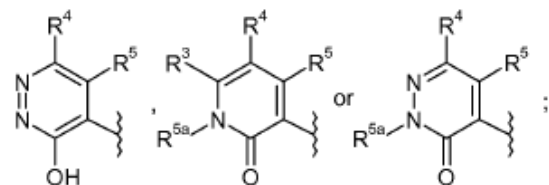


5

or a pharmaceutically acceptable salt thereof, or a pharmaceutically acceptable solvate of said compound or salt, wherein:

R<sup>1</sup> is C<sub>1</sub>-C<sub>8</sub> alkyl, C<sub>3</sub>-C<sub>8</sub> cycloalkyl, C<sub>6</sub>-C<sub>12</sub> bicycloalkyl, Aryl<sup>1</sup>, Aryl<sup>2</sup>, Het<sup>1</sup>, Het<sup>2</sup>, Het<sup>3</sup> or Het<sup>4</sup>, said C<sub>1</sub>-C<sub>8</sub> alkyl, C<sub>3</sub>-C<sub>8</sub> cycloalkyl and C<sub>6</sub>-C<sub>12</sub> bicycloalkyl being optionally substituted by 1 to 3 substituents independently selected from -CN, halo, -NH<sub>2</sub>, -SH, -SO<sub>2</sub>NH<sub>2</sub>, -OCONH<sub>2</sub> and -X-R<sup>2</sup>, with the proviso that the R<sup>1</sup> moiety may not be attached through a methylene (-CH<sub>2</sub>-) group;

R<sup>2</sup> is



Perform substructure and exact match search within the Markush to determine:

- Does our Markush cover all relevant project compounds?
- Are new compounds we plan to make covered?

# Acknowledgements

Bonnie Bacon

Greg Bakken

Marudai Balasubramanian (Balu)

Markus Boehm

Rajiah Denny

Klaus Dress

Anton Fliri

Kevin Foje

Katelin Grover

Robert Goulet

Kazu Hattori

Steven Heck

Andrew Hopkins

Xinjun Hou

Greg Kauffman

Christopher Kibbey

Jacquelyn Klug-McLeod

Bruce Lefker

Jens Loesel

Scot Mente

Mike Miller

James Mills

Mark Mitchell

Israel Nissenbaum

Matthias Nolte

David O'Neill

Robert Owen

Martin Pettersson

Gena Poda

Gaia Paolini

Usa Datta Reilly

Vineet Sardar

Keith Schreiber

Meihua Tu

David Walsh

Simon Xi

Christoph Zapf