Computational Sciences CoE

Analyzing Exemplified and Markush Structures in Patents

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

Pfizer Worldwide Research and Development, Worldwide Medicinal Chemistry
*Pfizer Worldwide Research and Development, Business Technology

Pfizer

Competitive Intelligence and Drug Design

Computational Sciences CoE

A thorough understanding of the competitive landscape is an important aspect of drug design.

- 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
- Protect Pfizer's chemical equity

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

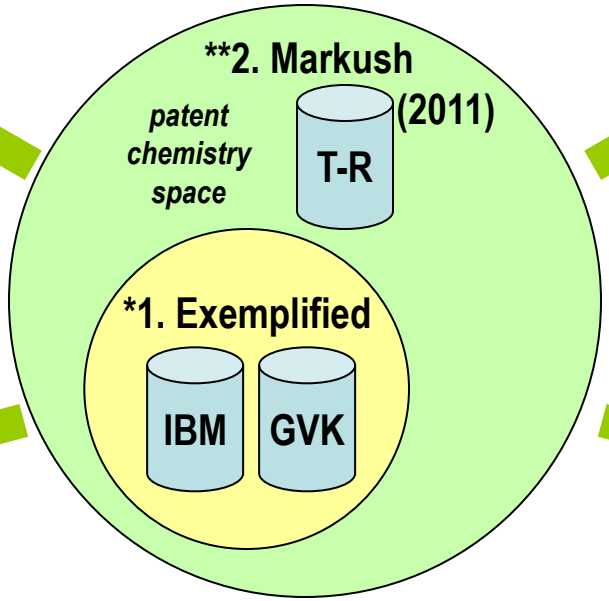
- Access to electronic structures from patents and appropriate tools for analysis is critical to success

Pfizer

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

Computational Sciences CoE

Chemistry Landscape
- Substructure search
- Similarity search
- Exact-match search
- IP assessment



Patent Ranking
- Rank order a series of patents by feature similarity to a query structure

Project Datasets
Download structures for specific patents related to a medicinal chemistry project

Feature Analysis/Visualization
- Detailed analysis and visualization of substructure features within patent structures compared with those for a set of query structures

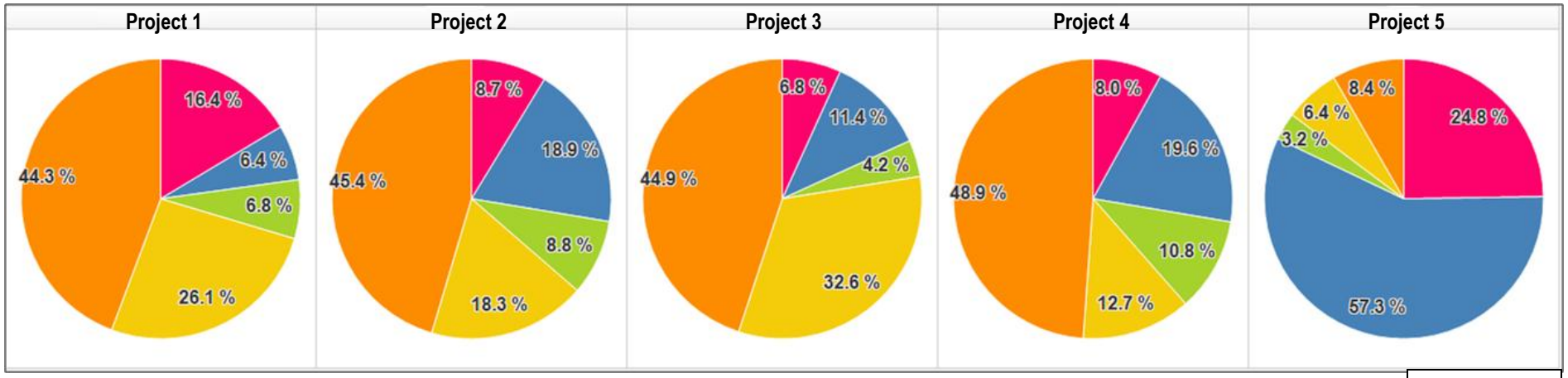
*IBM database of ~8 million exemplified structures from patents (nightly updates)
*GVK database of ~3 million exemplified structures from patents (quarterly updates)
**Thomson-Reuters database of 1.2 million Markush structures from patents (weekly updates)

Pfizer

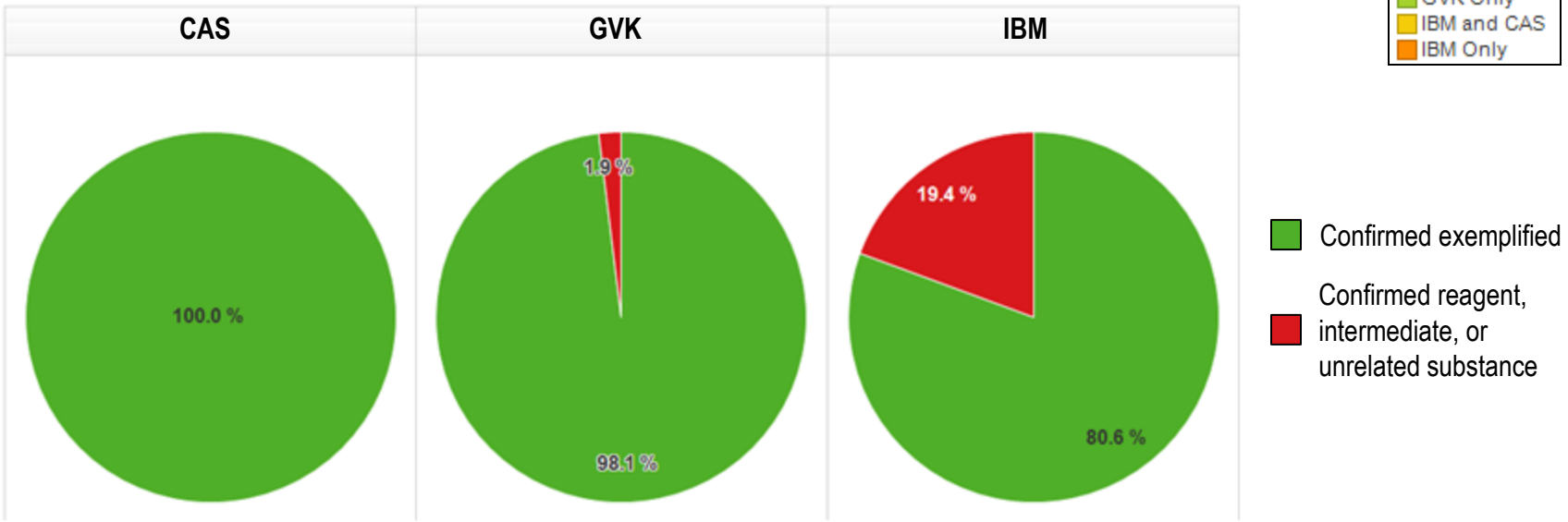
Comparison of patent structures from commercial providers

Computational Sciences CoE

Distribution of exemplified structures (by vendor) across five medicinal chemistry projects



Confirmation of exemplified structures from commercial providers

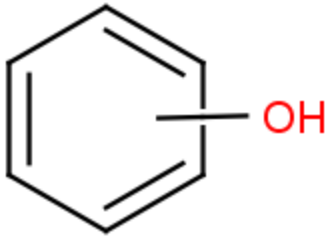


Pfizer

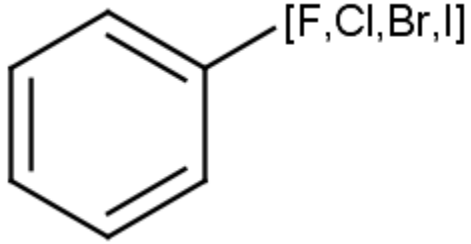
Types of structure variation encoded in exemplified and Markush structures in patents

Computational Sciences CoE

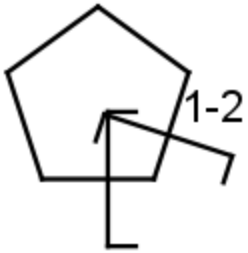
Position (p-variation)



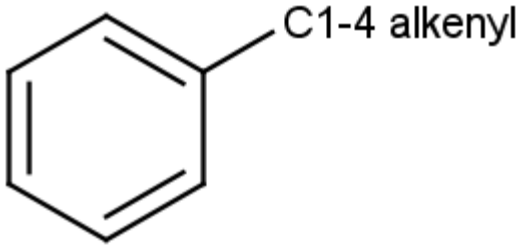
Substitution (s-variation)



Frequency (f-variation)



Homology (h-variation)

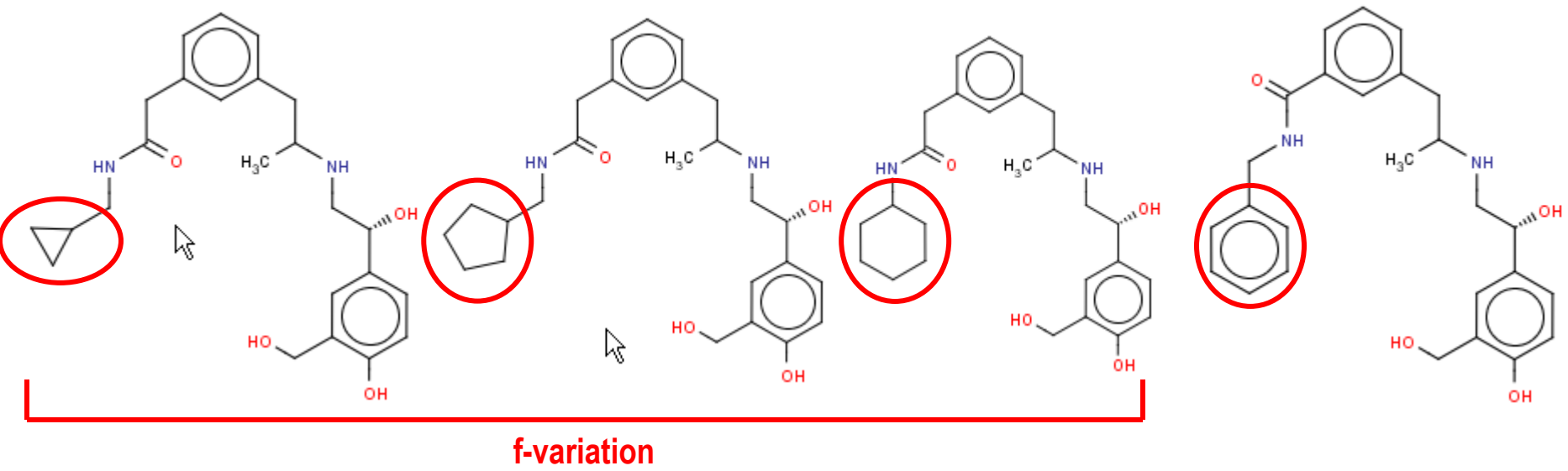


Pfizer

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

Computational Sciences CoE

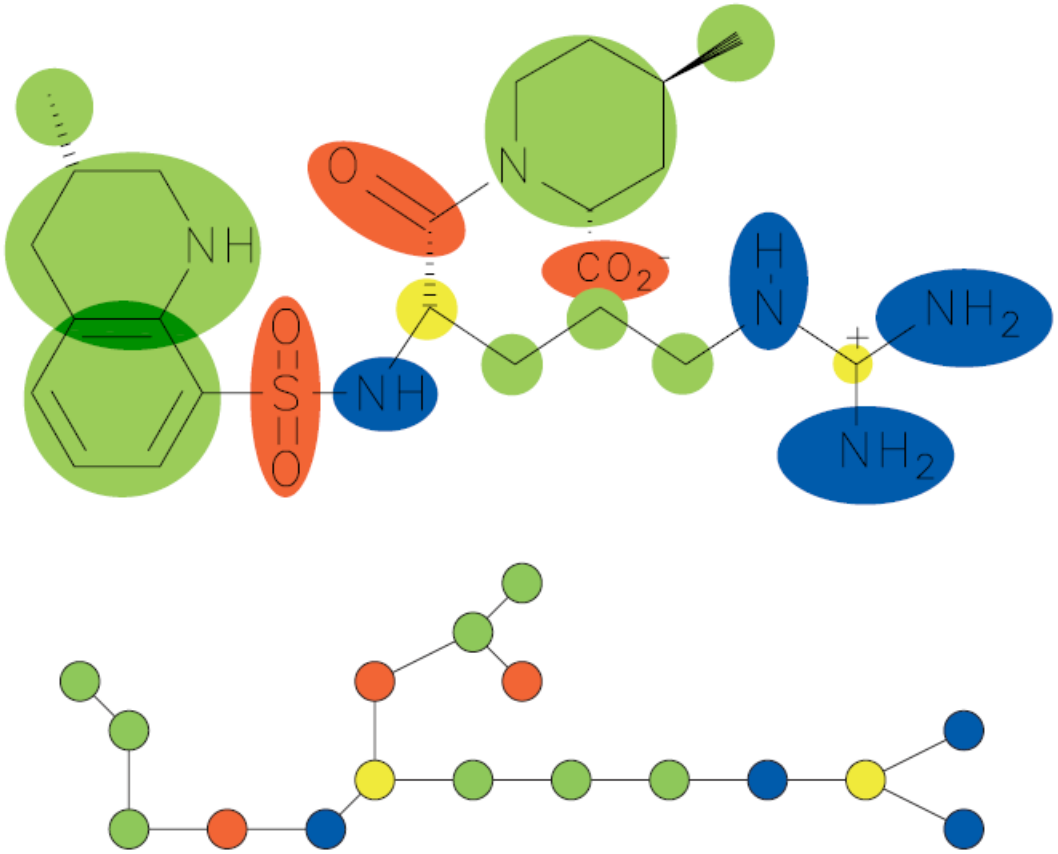
Substructure search cannot account for h-, s-, p-, and f-variation



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

Computational Sciences CoE



Hydrophobic

No direct interaction

H bond acceptor

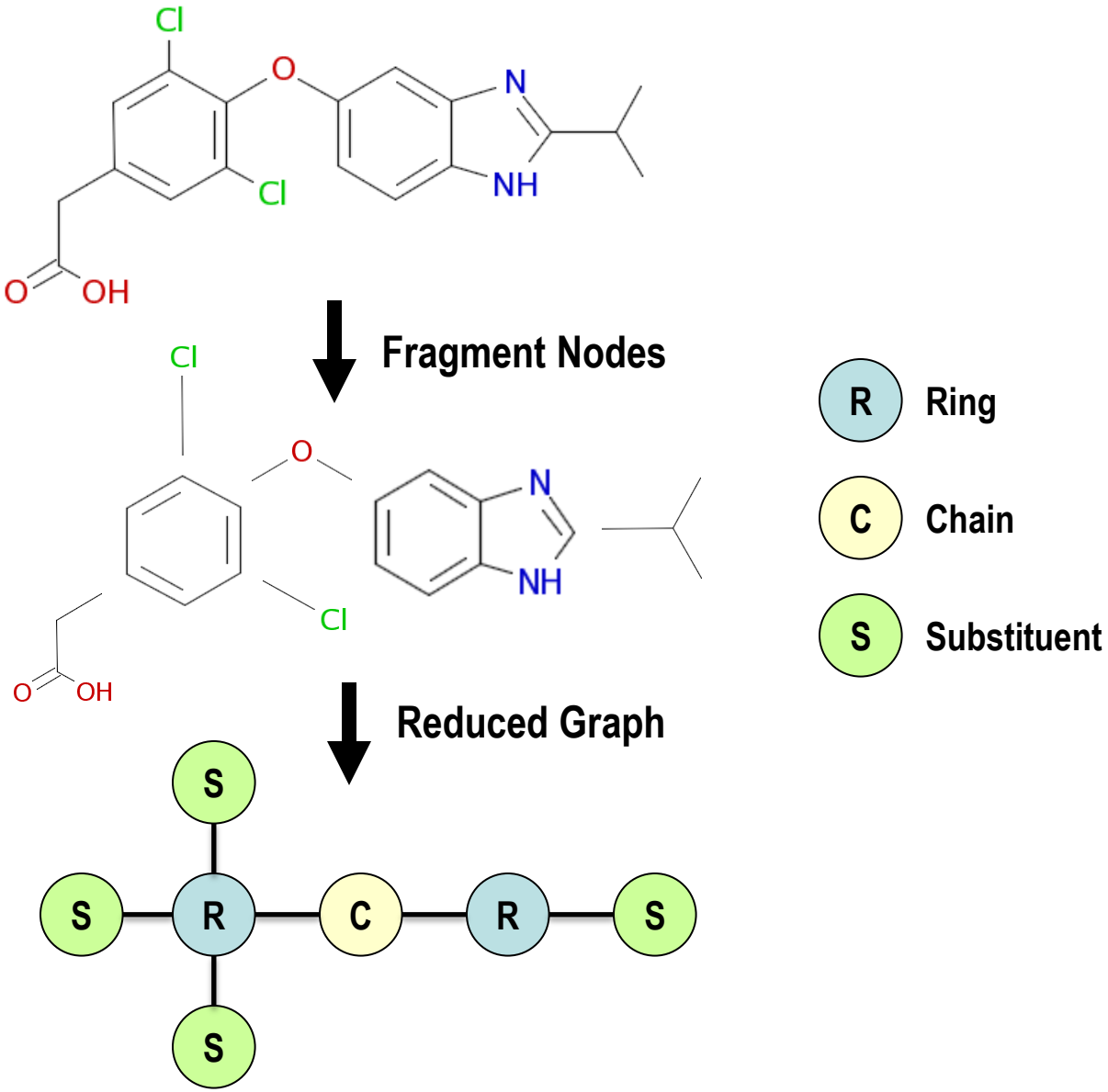
H bond donor

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

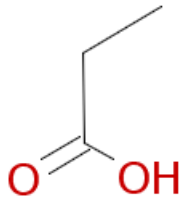
Computational Sciences CoE



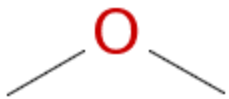
Feature Analysis - Encoding node fingerprints

Computational Sciences CoE

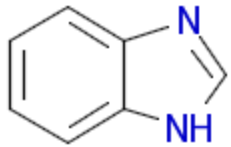
Substituents-S	Chains-C	Rings-R
Contains [F,Cl,Br,I]	Contains [F,Cl,Br,I]	Aromatic
Contains C	Contains C	Fused
Contains N	Contains N	Fused and Fully Aromatic
Contains O	Contains O	Contains C
Contains S	Contains S	Contains N
Has Single Bond	Has Single Bond	Contains O
Has Double Bond	Has Double Bond	Contains 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



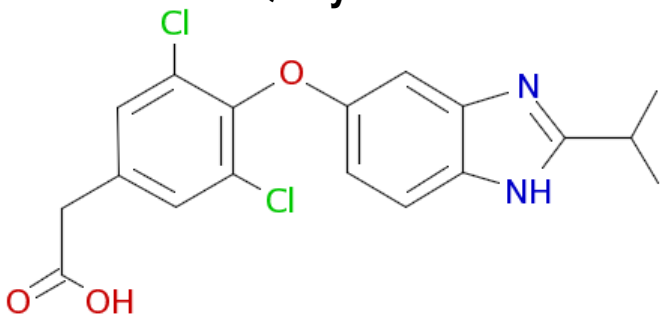
R11111000000

Pfizer

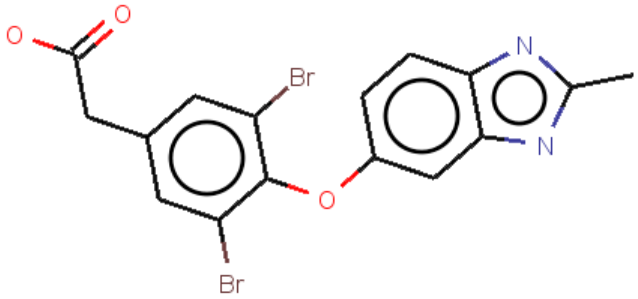
Feature Analysis - Overlaying reduced graphs and scoring

Computational Sciences CoE

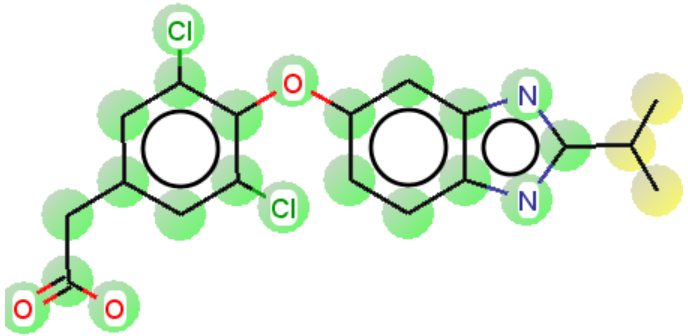
Query



Target



*Score = 0.97



1.00 – 0.75

0.75 – 0.25

0.25 – 0.00

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

Pfizer

Advantages of applying Feature Analysis to structures in patents

Computational Sciences CoE

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)

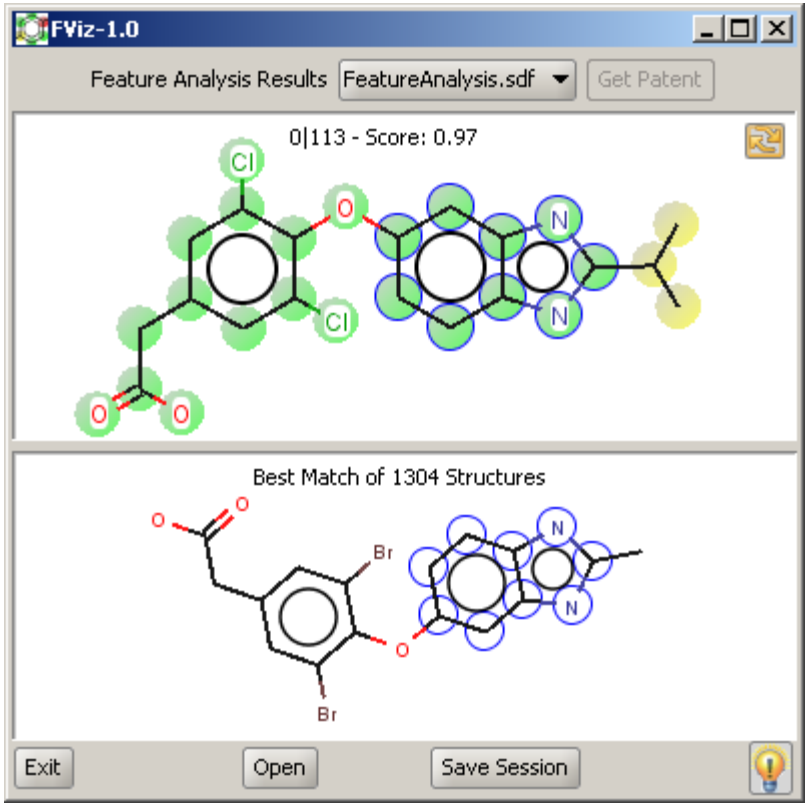
- f-variation: allows for variation in chain, ring and substituent size by atom type
- h-variation: various homology group definitions can be encoded in node fingerprints
- p-variation: algorithm ignores attachment geometry among R, C, S nodes
- s-variation: algorithm accommodates differences in substitution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

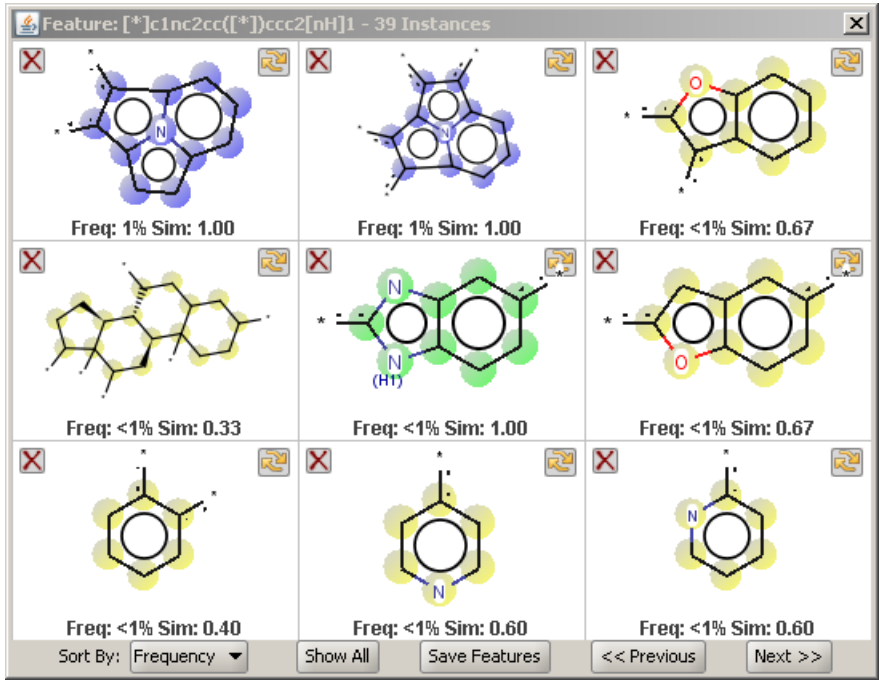
Pfizer

Visualizing the results of Feature Analysis

Computational Sciences CoE



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

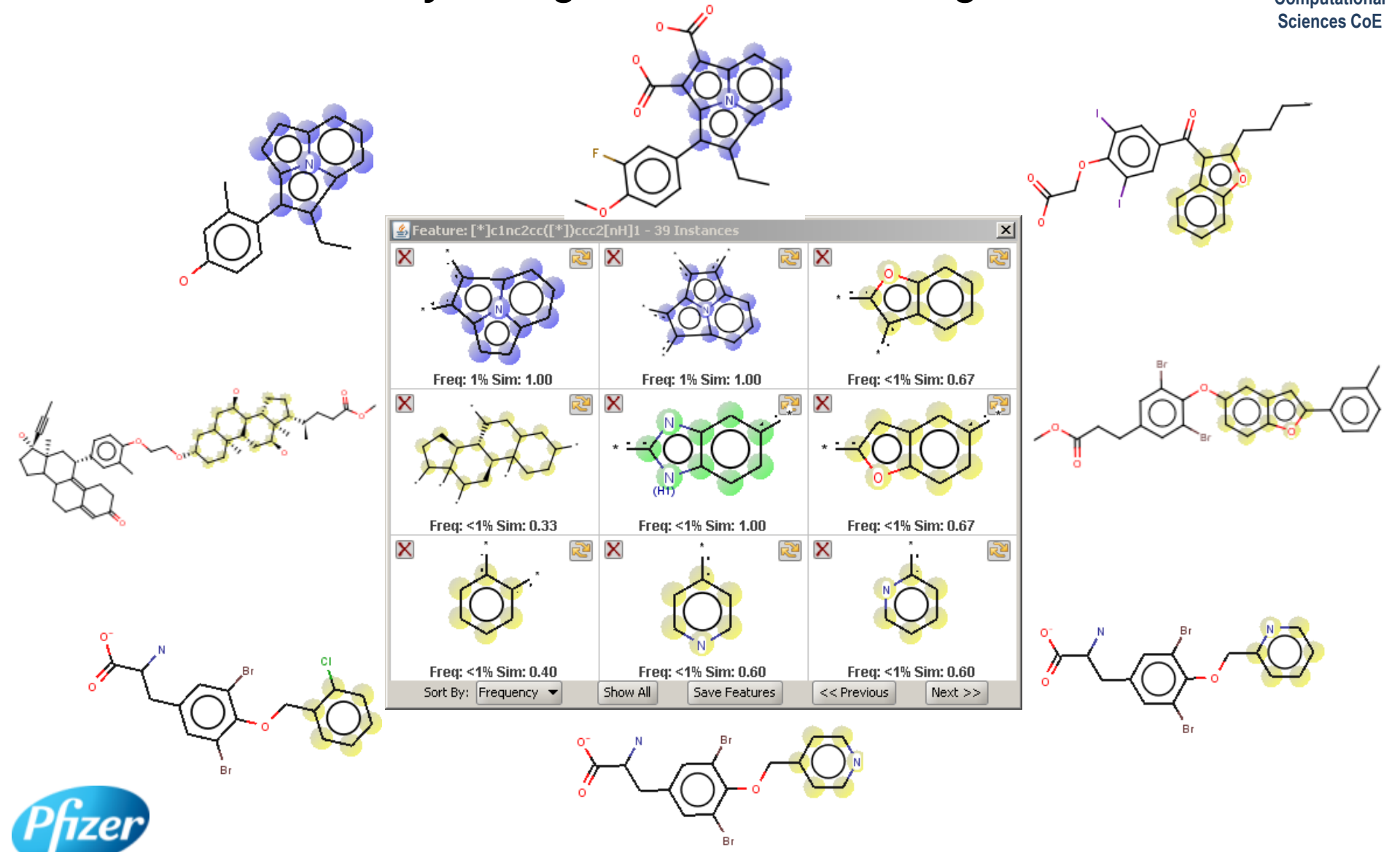


Select a query feature to display corresponding sub-structure features for exemplified compounds

Pfizer

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

Computational Sciences CoE



Pfizer

Feature Analysis applied to ranking patents

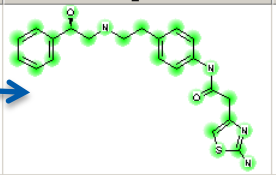
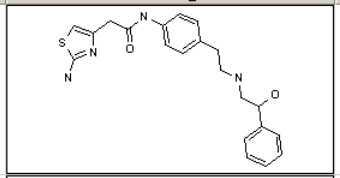
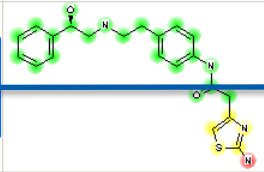
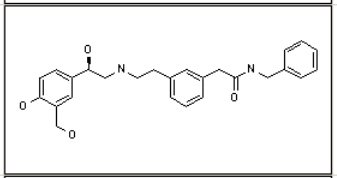
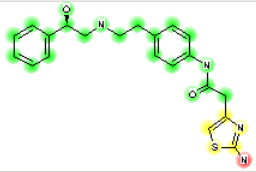
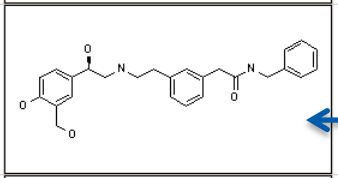
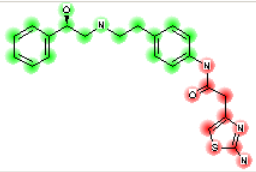
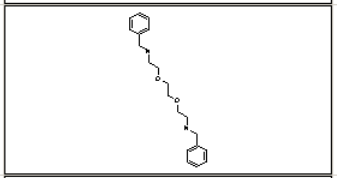
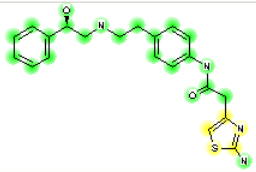
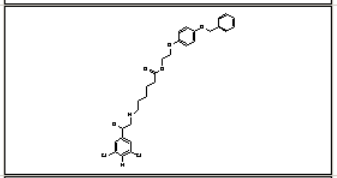
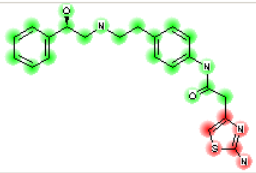
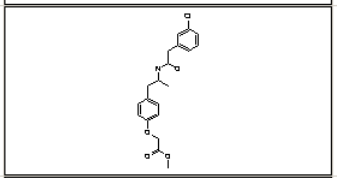
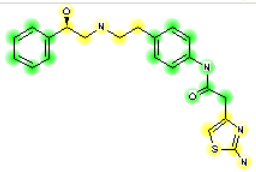
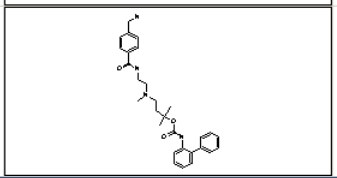
Computational Sciences CoE

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		US6346532	[0.44, 0.21, 0.24, 0.03, 0.07, 0.00, 0.00, 0.00, 0.00, 0.00]
		0.81		WO2005090286	[0.00, 0.02, 0.72, 0.22, 0.04, 0.00, 0.00, 0.00, 0.00, 0.00]
Molecule1 Instance 3		0.81		US2005023409	[0.00, 0.03, 0.67, 0.17, 0.12, 0.01, 0.00, 0.00, 0.00, 0.00]
Molecule1 Instance 4		0.76		US6362371	[0.00, 0.00, 0.04, 0.04, 0.33, 0.44, 0.07, 0.04, 0.04, 0.00]
Molecule1 Instance 5		0.74		US5599966	[0.00, 0.00, 0.04, 0.18, 0.58, 0.09, 0.11, 0.02, 0.00, 0.00]
Molecule1 Instance 6		0.74		US5153210	[0.00, 0.00, 0.31, 0.42, 0.16, 0.05, 0.06, 0.00, 0.00, 0.00]
Molecule1 Instance 7		0.7		US2006011639	[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

Computational Sciences CoE

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?

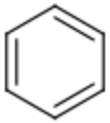
2. Facilitate identification of unexplored areas of chemical space
    - What regions of the Markush are under-represented by the exemplified structures in the patent?
    - Are these under-represented regions covered by granted claims?

Pfizer

Challenges to Markush analysis

Computational Sciences CoE

Substructure and exact match searches against Markush structures can be problematic

- Use of specialized drawing rules to encode Markush structures
  - Normalized bonds (e.g.  as  and  as )
- Limitations imposed by indexing system rules
  - May leave gaps in chemistry space covered by patent
- Inconsistent handling of stereochemistry in patents

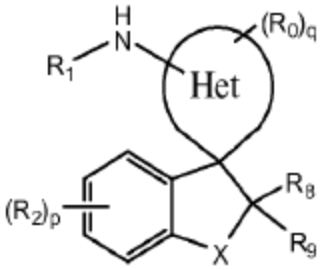
No commercial solution exists for similarity search against Markush structures

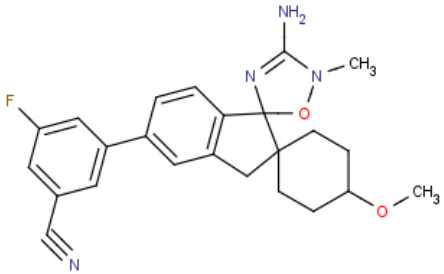
Pfizer

Example: Substructure search fails due to Markush indexing

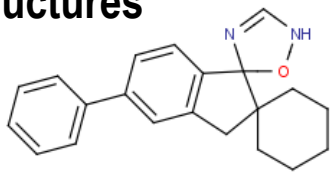
Computational Sciences CoE

WO2010105179 – 24 Markush Structures

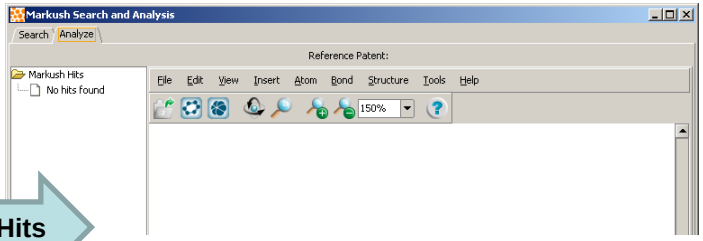


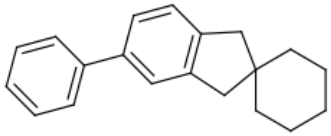


Exemplified Structure

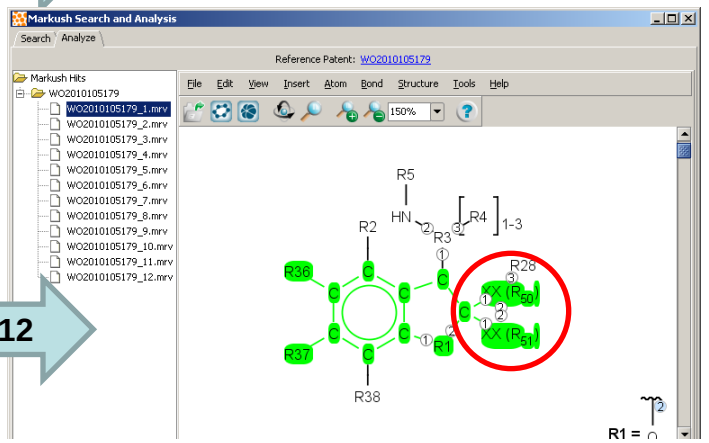


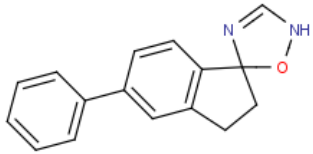
Substructure Search → **No Hits**



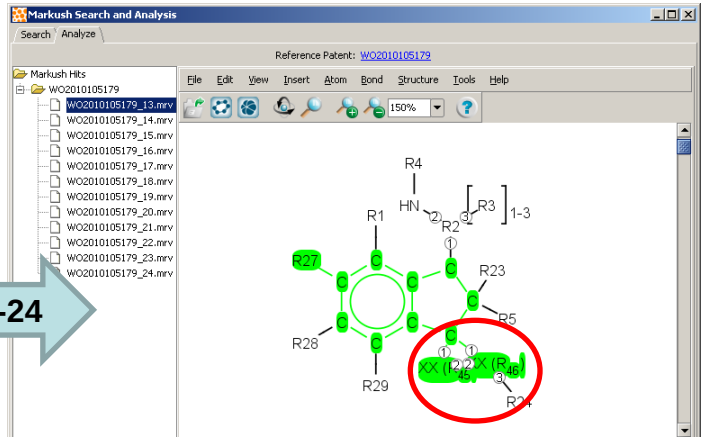


Substructure Search → **1-12**





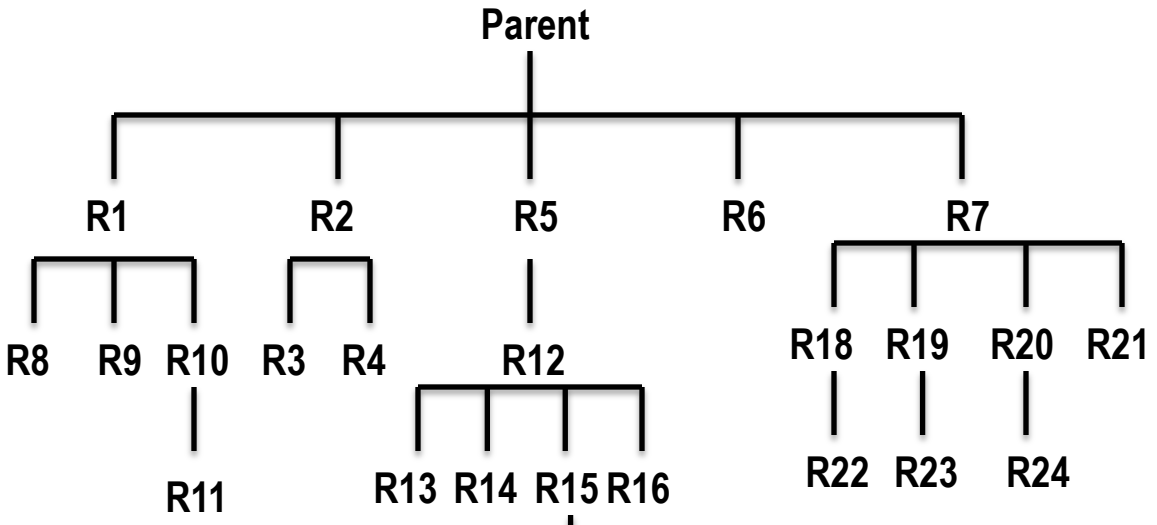
Substructure Search → **13-24**

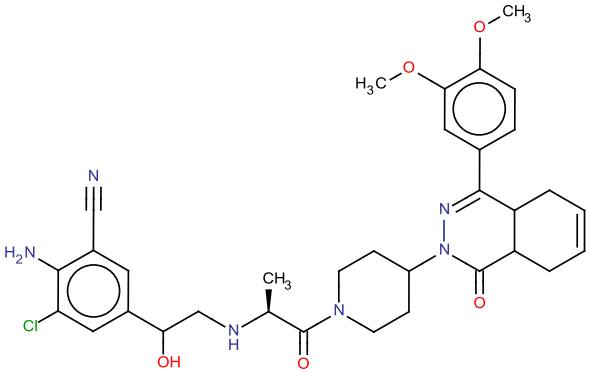


Enumeration of virtual libraries from a Markush is an ineffective strategy for (similarity) analysis

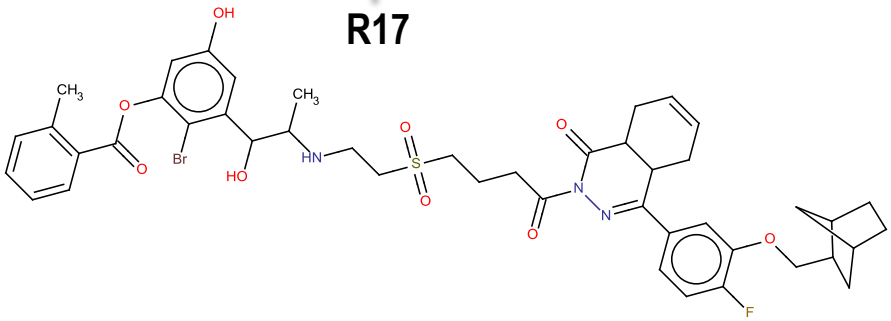
Computational Sciences CoE

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





WO2001094319 – Exemplified



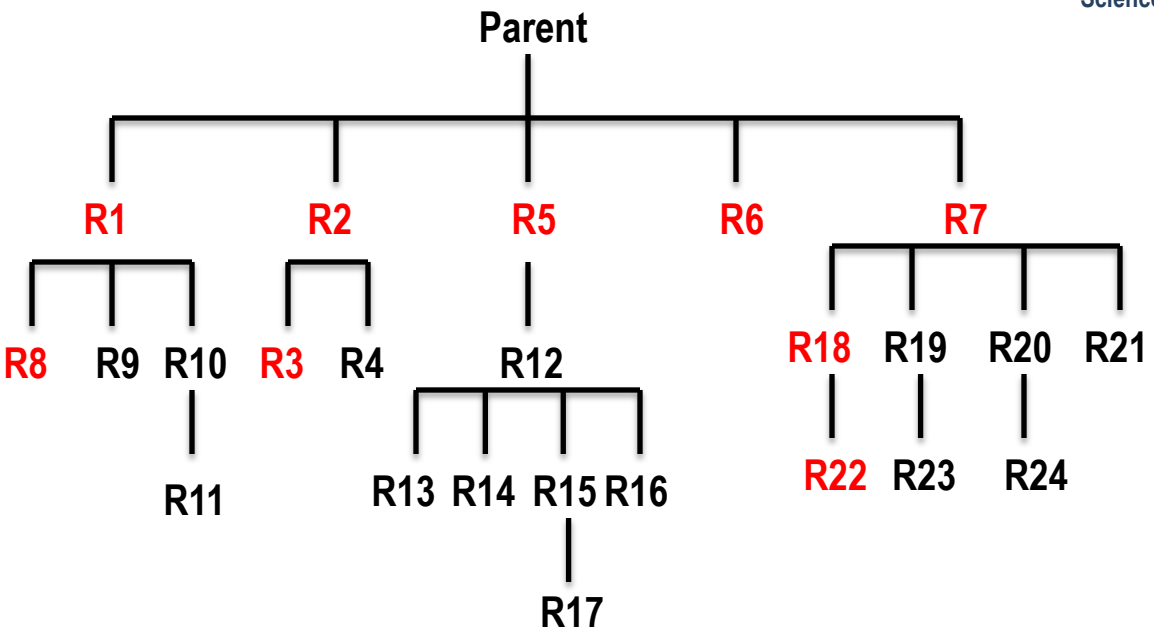
Random enumeration yields structures with variation far-away from the core scaffold within an extremely large virtual space (e.g., ~10^{15})

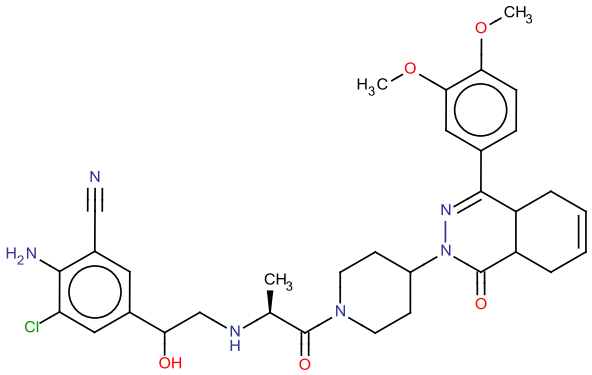
Pfizer

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

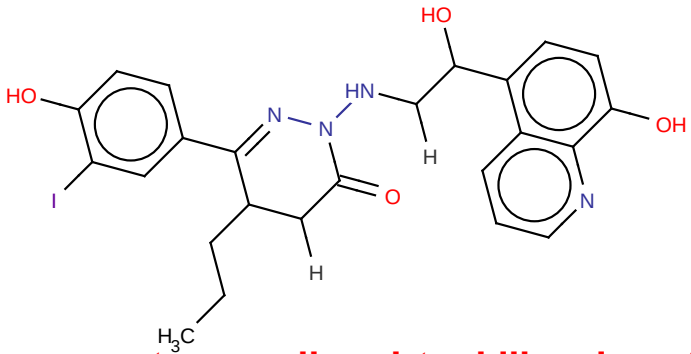
Computational Sciences CoE

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





WO2001094319 – Exemplified

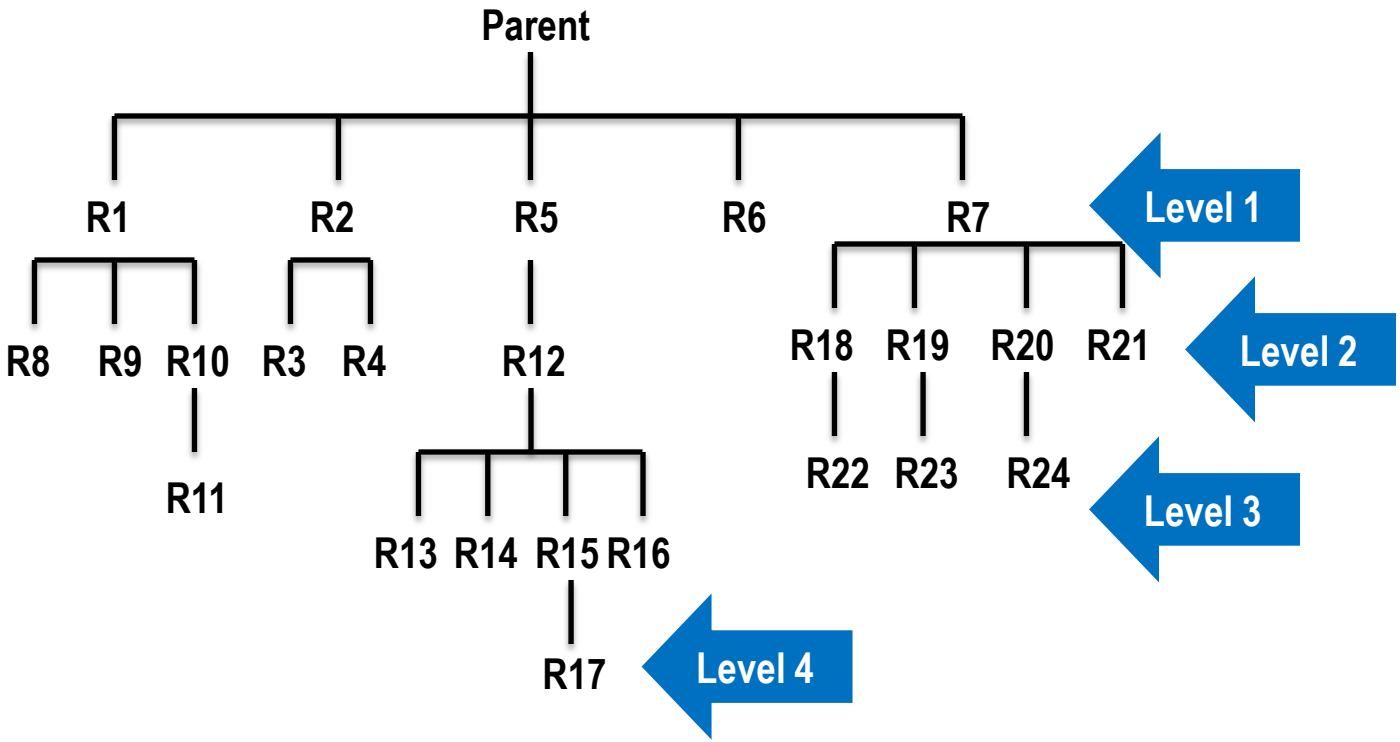


Sequential enumeration generates smaller virtual libraries of close-in structures, but ignores much of the chemical space within the Markush.

Pfizer

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

Computational Sciences CoE



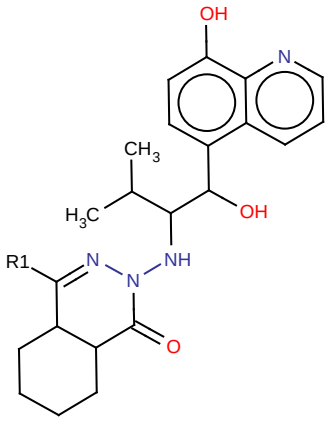
MMS allows up to 50 Rgroups with 4 levels of nesting per Markush structure

Pfizer

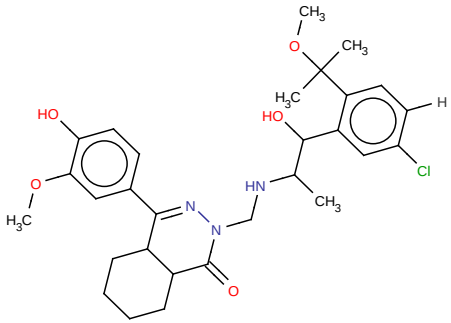
Level enumeration yields smaller libraries of close-in structures while maintaining representative Rgroup coverage

Computational Sciences CoE

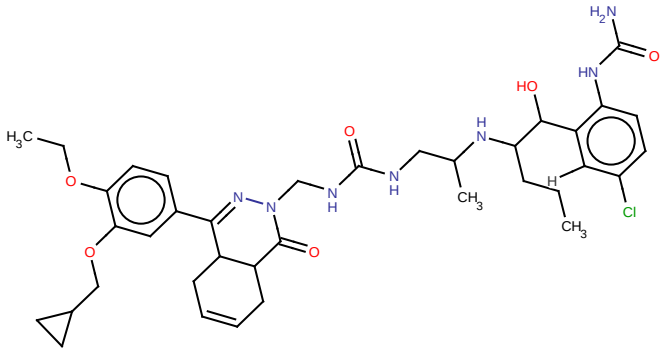
Level 1
Library Size = 24



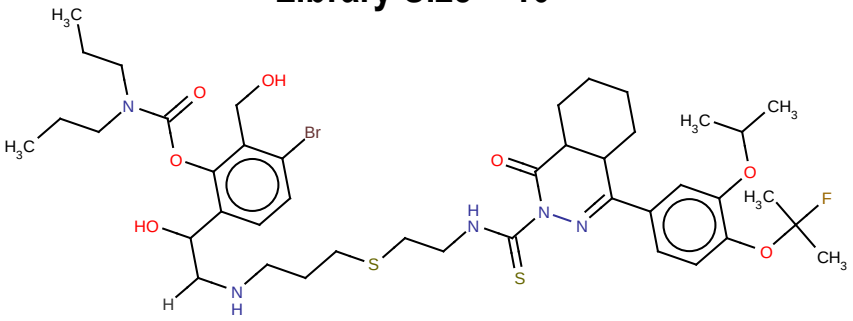
Level 2
Library Size = 10^{11}



Level 3
Library Size = 10^{14}



Level 4
Library Size = 10^{15}



Strategy for similarity search against Markush (Level enumeration + Feature Analysis)

Computational Sciences CoE

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

Pfizer

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