



Structure-based approaches to the indexing and retrieval of patent chemistry

Tim Miller
Head of Research
May 2010

Donald Walter
Product Specialist



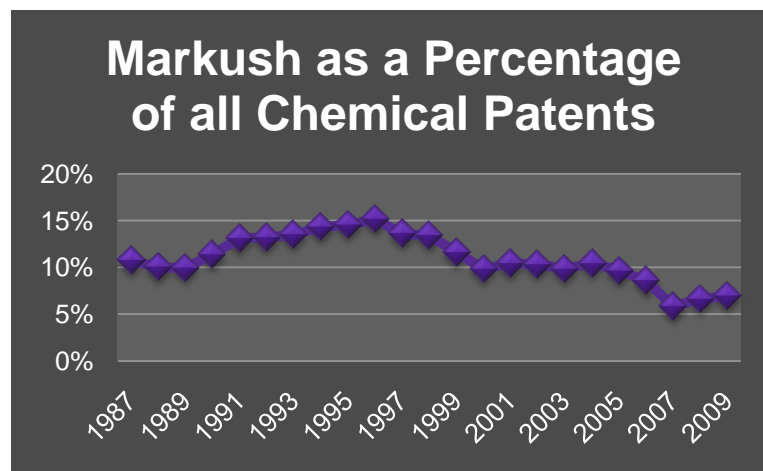
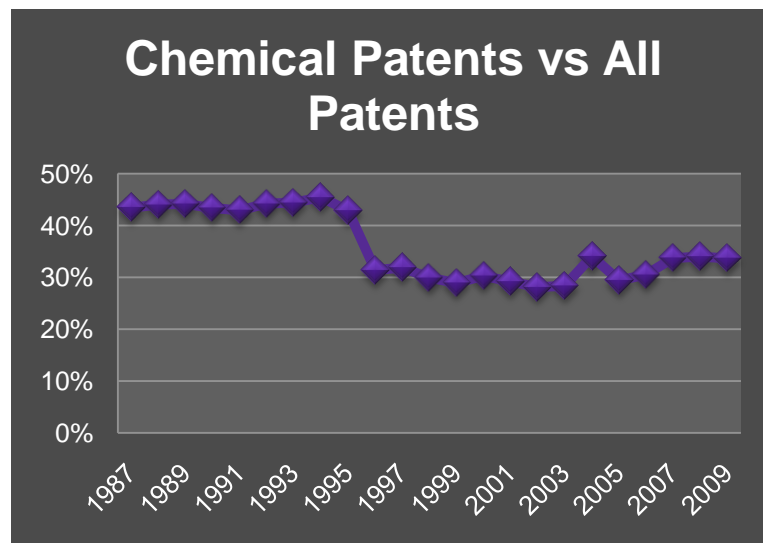
THOMSON REUTERS

TOPICS

- Chemistry in Patents
- Structure Indexing of patents
- New developments
- Challenges yet to be overcome

Chemistry in Patents

- Importance
 - Patents are essential to protect chemical inventions
 - 70% of patent information is never published elsewhere.
- Volumes
 - Chemical patenting is steady
 - Markush patenting as a percentage of all chemical patents is decreasing



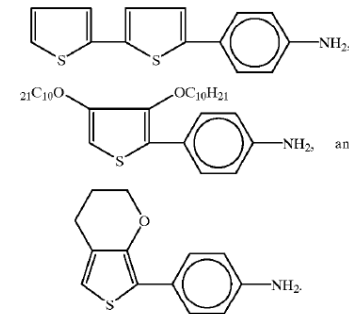
Chemistry in Patents

- Specific Compounds
- Markush
- Reactions & new Intermediates
- Polymers
- Inorganics

(57)

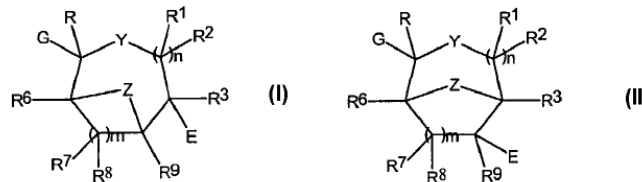
ABSTRACT

Provided are polymers prepared by the polymerization of monomers of the formulae:



1 Claim, No Drawings

(54) Title: PROCESS FOR PREPARATION OF BICYCLIC AND POLYCYCLIC MOLECULES



(57) Abstract: A method of synthesis of a bicyclic or polycyclic compound of formula (I) or formula (II) in which: E represents an electrophile; each of R, R1, R2, R3, R6, R7, R8, R9 and X independently represents the common organic substituent groups defined in claim 1; Y represents C(r12)R13, O, NR14, or S; Z represents O, NR15, S or CR16W; G represents W or X; W

represents an electron withdrawing group; X has the same definition as R, and W= W; and each of n and m represents an integer from 0 to 100. The method comprises the steps of (a) activating a compound of formula (III); 8b) subjecting a compound of formula (IV) to nucleophilic addition with the activated form of compound III; (c) subjecting the product of step (b) to ring closing metathesis; and (d) subjecting the product of step (c) to stereoselective ring closure. The methods of invention are useful in the synthesis of candidate pharmaceutical agents or intermediates in drugs synthesis.

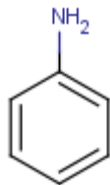
Indexing Patents, the Early Days

- Fragmentation Codes

- GREMAS
- DWPI
- CLAIMS/IFI

- Full Structures

- CAS Registry
- Beilstein



Aniline is the only specific compound given in the patent

The original Markush claim

Claim 1. The process for the manufacture of dyes which comprised coupling with a halogen-substituted pyrazolone, a diazotized unsulphonated material selected from the group consisting of aniline, homologues of aniline and halogen substitution products of aniline.

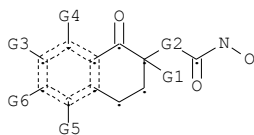
Code	Fragment
H1	Amine essentially present
M521	1 mononuclear heterocyclic ring
M520	No mononuclear heterocyclic ring
M320	No multivalent carbon chains
M210	C1-6 alkyl chain
M270	Alkyl attached to heteroatom
M273	Heteroatom is N
F011	Substitution on 1-position of heterocycle
H6411	Halogen linked to aromatic ring
H6422	Halogens linked to aromatic ring

Starting material indexed using
Derwent
Fragmentation codes

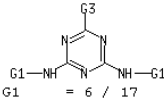
The 1980s – Topological Markush Searching Arrives

- University of Sheffield
 - GENSAAL
- Derwent, INPI and Questel
 - Markush DARC
- CAS
 - MARPAT

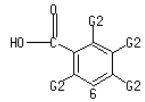
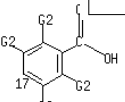
Screenshot of Markush DARC

MARKUSH/DARC	1 / 1	CN :0069-16801	MMS
-FG: 0-	-GM: 3/ 6-		AV NU CR
		H	
		O	
		N	
		CN	
		HAL	
1-O-CHK	1-N-CHK	CHK	ARY
1-O-ARY	1-N-CHK	1-F	HEA
1-O-CHK-ARY	1-N-CO1-CHK	F	HEF
		1-N	
		1-N	
		1-N	
CHK3=C1-4.			

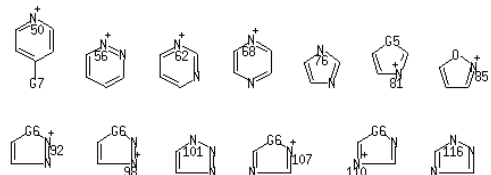
MSTR 1



G1-NH G3
G1 = 6 / 17

G2 = H / R / (Specifically claimed: alkyl (opt. substd. by 1 or more G4))
G3 = heterocycle <containing 1 or more N, attached through 1 or more N> (opt. substd.) / (Specifically claimed: 50 / 56 / 62 / 68 / 76 / 81 / 85 / 92 / 98 / 101 / 107 / 110 / 116 / 120 / 126 / 129 / 136 / 143 / 146 / 151 / 157 / 162 / 181 / 168 / 189 / 192 / 203) / (Examples: 235 / morpholino / pyrrolidino / piperidino / piperazino)



G7 G8 G9 G10 G11 G12 G13 G14 G15 G16

Screenshot of MARPAT

The New Millennium – Technology to Extract Structures from Patents

- Text mining
 - Temis, IBM, ReelTwo, etc.
- Name to Structure
 - ACD/Labs, CambridgeSoft, ChemAxon, etc.
- Chemical OCR
 - CLiDE, Kekule, SCAI
- New Products & Services
 - Elsevier's Patent Chemistry Database
 - SureChem
 - ChemSpider

Today – A New Generation of Markush Tools?

- ChemAxon
 - Search, Enumerate
 - Available in current release of Marvin/JChem
- Digital Chemistry
 - Search, Enumerate
 - In prototype in TORUS
- DecrIPt
 - Enumerate, Rank overlaps
 - Available as service
- Symyx

Look at the cool things we can do!

- Quick, easy searches to establish where best to focus my efforts
- Overlap and difference analysis to find holes in an IP portfolio
- IP Screening of my combinatorial libraries

Are we done then?

Can the combination of new Markush-capable systems and smart text + image processing open the world of patent chemistry to the masses?



Challenge #1

THERE'S A LOT OF IT

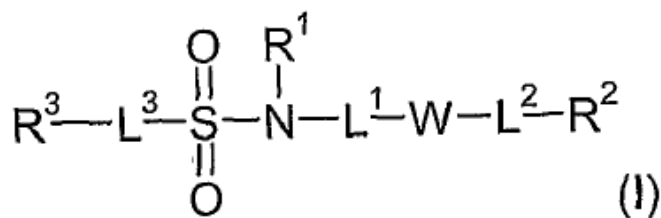
Size Matters

- The original Markush was comparatively small
- Recent Markush are somewhat larger

The screenshot displays the 'Markush Enumeration' software interface. At the top, the window title is 'Markush Enumeration' and the 'Markush library size' is 7776. The main area shows a Markush structure of a benzimidazole derivative with five substituent positions labeled R1 through R5. To the right of the structure, the definitions for R1 through R5 are listed: R1 = H, C, halogen; R2 = H, C, halogen; R3 = H, C, halogen; R4 = H, C, halogen; R5 = H, C, halogen. Below the main window, a smaller window titled 'Enumerate a Markush structure' is open, showing the 'Enumeration options' (Full enumeration selected, Max structures: 12, Output to file checked) and 'Display options' (Rows: 3, Columns: 4, Show R-groups checked, Colouring unchecked). A 'Full enumeration of this structure produces ~ 10^36 structures' message is displayed. An 'Enumerate' button is present, and below it, a grid of 12 enumerated chemical structures is shown.

Sneaky, hidden Markush

(54) Title: NOVEL SULPHONAMIDE DERIVATIVES AS GLUCOCORTICOID RECEPTOR MODULATORS FOR THE TREATMENT OF INFLAMMATORY DISEASES



(57) Abstract: A compound of formula (I) or a pharmaceutically acceptable salt thereof; compositions comprising them, processes for preparing them and their use in medical therapy (for example modulating the glucocorticoid receptor in a warm blooded animal).

Enumerate a Markush structure

Enumeration options:

- Full enumeration
- Random enumeration
- Markush reduction according to the hit

Max structures:

Output to file

Display options

Rows:

Columns:

Show R-groups

Colouring

Full enumeration of this structure produces ~ 10²⁶ structures

12 structures enumerated

Challenge #2

CONTEXTUAL INFORMATION ASSOCIATION



Connecting Contextual Information

ID	Title	Description	Assignees
119			
vmns			
Markush structure			
1	0327-34308	New hydroxybenzoate salts formed as the reaction product of E-metanicotine compound and hydroxybenzoic acid, useful for treating e.g. CNS disorder, HIV-dementia, epilepsy, mania, and depression	A hydroxybenzoate salt formed as the reaction product of an E-metanicotine compound of formula $Cy-C(E)=CE-(C(E)Z)_m-CEE1-N(Z)1-Z2$ (where Z is a hydroxybenzoic acid of formula (II) (where the hydroxy group can be present at a position ortho, meta or para to the carboxylic acid group), is new. The molar ratio of (I) to (II) is 1:2 - 2:1. Cy = 5- or 6-membered heterocyclic ring other than 5-isopropoxy-pyridine (preferably rings of formula (i) or (ii)); E, E1 = H or alkyl (optionally halo substituted); Z1 = H or alkyl; m = 1 - 6; Z = non-hydrogen substituent selected from alkenyl, heterocyclyl, (cyclo)alkyl, aryl, alkylaryl, arylalkyl (all optionally substituted), F, Cl, Br
2	0327-34309		
3	0327-34310		
4	0327-34311		
5	0327-34312		
	Patents	Use	Activity
	WO2006053039-A2 * US20060122238-A1 EP1814853-A2 AU2005304575-A1 IN200701331-P2 CN101068784-A JP2008519766-W US20080249142-A1	In the preparation of a disorder that results in normal neurotransmission (claimed), of the symptoms and conditions, disease, neurological disorder	CNS-Gen.; Neuroprotective; Antiinflammatory; Tranquilizer; Neuroleptic; Nootropic; Anti-HIV; Antiparkinsonian; Cerebroprotective; Anticonvulsant; Muscular-Gen.; Antimanic; Antidepressant; Vasotropic; Antismoking; Antiaddictive; Antialcoholic; Anorectic; Gastrointestinal-Gen.; Antidiarrheic; Antiulcer; Vulneryary.
		Mechanism Of Action	
		CNS nicotinic receptor binder. The CNS nicotinic receptor binding efficacy of hydroxybenzoate salts with relevant receptor sites can be determined according to the methods described in US5597919. The results showed low binding constants thus the salts exhibits good high affinity binding to certain	

Active material, reagent, intermediate?

Provisos, other information that can't be represented in the structure

Biological information

How we do it in DWPI

Indexing
"paragraphs"
based on
IBM card
records

CMC UPB 20060724

M2 *01* C316 D022 D029 D621 D699 F011 F012 F013 F014 F015 F016
 F021 F029 F111 F199 F431 F432 F499 F530 F541 F542 F552 F553
 F599 G001 G002 G003 G010 G011 G012 G013 G014 G015 G016 G017 G018
 G019 G020 G021 G022 G029 G030 G031 G032 G039 G040 G050 G051 G100
 G111 G112 G113 G221 G299 G553 G563 H100 H101 H102 H103 H121 H122
 H141 H142 H143 H161 H181 H182 H183 H321 H341 H342 H343 H401 H402
 H403 H404 H405 H421 H441 H442 H443 H444 H481 H482 H483 H494 H521
 H522 H523 H541 H542 H543 H561 H581 H582 H583 H592 H594 H599 H601
 H602 H603 H604 H608 H609 H621 H641 H642 H643 H681 H682 H683 H685
 H689 H715 H721 H722 H723 J011 J012 J013 J014 J111 J131 J132 J133
 J171 J172 J173 J211 J221 J222 J231 J232 J241 J242 J261 J271 J272
 J273 J311 J321 J322 J331 J332 J341 J342 J361 J371 J372 J373 J411
 J431 J432 J471 J521 J581 J582 J583 J592 K442 K499 K510 K599 K711
 K799 L142 L143 L199 L462 L463 L499 L921 L922 L941 L943 L999 M111
 M112 M113 M115 M116 M119 M121 M122 M123 M124 M125 M126 M129
 M132 M133 M135 M136 M137 M139 M141 M142 M143 M150 M210 M211 M212
 M213 M214 M215 M216 M220 M221 M222 M223 M224 M225 M226 M227 M232
 M233 M240 M273 M280 M281 M311 M314 M315 M316 M320 M321 M322 M323
 M331 M332 M333 M334 M340 M342 M343 M344 M349 M352 M353 M354 M391
 M392 M393 M412 M413 M414 M510 M511 M512 M520 M521 M522 M523 M530
 M531 M532 M533 M540 M541 M542 M543 M630 M650 M720 N221 N225 N242
 N311 N421 N422 N512 N513 P210 P420 P442 P444 P446 P448 P451 P510
 P517 P617 P625 P641 P642 P646 P714 P731 P735 P738 P942 M905
 M904
 RIN: 00210 00211 00212
MCN: 0327-34301-K 0327-34301-P

Structural

Synthesis

Descriptors

M2 *07* M510 M511 M520 M521 M530 M531 M540 M541 M630 M650 M720 N221 N225
 N242 N311 N421 N422 N512 N513 P210 P420 P442 P444 P446 P448 P451
 P510 P517 P617 P625 P641 P642 P646 P714 P731 P735 P738 P942
 M905 M904
MCN: 0327-34307-K 0327-34307-P

M2 *08* M531 M540 M630 M650 M720 N221 N225 N242 N311 N421 N422 N512 N513
 P210 P420 P442 P444 P446 P448 P451 P510 P517 P617 P625 P641 P642
 P646 P714 P731 P735 P738 P942 M905 M904
MCN: 0327-34308-K 0327-34308-P

M2 *09* M531 M540 M630 M650 M720 N221 N225 N242 N311 N421 N422 N512 N513
 P210 P420 P442 P444 P446 P448 P451 P510 P517 P617 P625 P641 P642
 P646 P714 P731 P735 P738 P942 M905 M904
MCN: 0327-34309-K 0327-34309-P

M2 *10* M630 M650 M720 N221 N225 N242 N311 N421 N422 N512 N513 P210 P420
 P442 P444 P446 P448 P451 P510 P517 P617 P625 P641 P642 P646 P714
 P731 P735 P738 P942 M905 M904
MCN: 0327-34310-K 0327-34310-P

M2 *11* M521 M530 M531 M540 M630 M650 M720 M800 N221 N225 N242 N311 N421
 N422 N512 N513 P210 P420 P442 P444 P446 P448 P451 P510 P517 P617
 P625 P641 P642 P646 P714 P731 P735 P738 P942 M905 M904
 DCN: RAMSYO-K RAMSYO-P
DCR: 1310317-K 1310317-P

M2 *12* M391 M413 M510 M520 M521 M530 M531 M540 M630 M650 M720 M800 N221
 N225 N242 N311 N421 N422 N512 N513 P210 P420 P442 P444 P446 P448
 P451 P510 P517 P617 P625 P641 P642 P646 P714 P731 P735 P738 P942
 M905 M904
 DCN: RAMSY-P-K RAMSY-P
DCR: 1310318-K 1310318-P

M2 *13* N512 N513 P210 P420 P442 P444 P446 P448 P451 P510 P517 P617 P625
 P641 P642 P646 P714 P731 P735 P738 P942 M905 M904
 DCN: RAMSYQ-K RAMSYQ-P
DCR: 1310319-K 1310319-P

M2 *14* M520 M521 M522 M530 M531 M532 M540 M541 M720 N209 N221 N225 N231
 N242 N309 N311 N361 N421 N422 N512 N513 P446 M905 M904
 RIN: 00210 00211 00212
MCN: 0327-34311-K 0327-34311-P

M2 *15* M393 M413 M415 M510 M520 M521 M530 M540 M541 M720 N209 N221 N225
 N231 N242 N309 N311 N361 N421 N422 N512 N513 P446 M905 M904
 RIN: 00060 00061 00074 00076 00081 00083 00084 00087 00088 00089
 00094 00096 00102 00105 00110 00115 00131 00133 00135 00137
 00138 00734 009740 11555 41038 41246 42005 45813
MCN: 0327-34312-K 0327-34312-P

M2 *16* F013 F015 F620 H1 H102 H181 H7 H721 M210 M211 M240 M273 M281
 M314 M321 M332 M342 M373 M391 M413 M510 M521 M530 M540 M720 N209
 N221 N225 N231 N242 N309 N311 N361 N421 N422 N512 N513 P446
 M905 M904
 DCN: RAMSYR-K RAMSYR-P
DCR: 1310320-K 1310320-P

M2 *17* F013 F015 F620 H1 H103 H181 H7 H721 M210 M211 M240 M273 M281
 M282 M314 M321 M332 M342 M373 M391 M413 M510 M521 M530 M540 M720
 N209 N221 N225 N231 N242 N309 N311 N361 N421 N422 N512 N513 P446
 M905 M904
 DCN: RAMSYS-K RAMSYS-P
DCR: 1310321-K 1310321-P

Specific
compounds

Interpretation

- The indexing in Paragraph 1 describes the synthesis of a known compound

M2 *01* C316 D022 D029 D621 D699 F011 F012 F013 F014 F015 F016 F019 F020
 F021 F029 F111 F199 F431 F432 F499 F530 F541 F542 F551 F552 F580
 F599 G001 G002 G003 G010 G011 G012 G013 G014 G015 G016 G017 G018
 G019 G020 G021 G022 G029 G030 G031 G032 G039 G040 G050 G051 G100
 G111 G112 G113 G221 G299 G553 G563 H100 H101 H102 H103 H121 H122
 H141 H142 H143 H161 H181 H182 H183 H321 H341 H342 H343 H401 H402
 H403 H404 H405 H421 H441 H442 H443 H444 H481 H482 H483 H494 H521
 H522 H523 H541 H542 H543 H561 H581 H582 H583 H592 H594 H599 H601
 H602 H603 H604 H608 H609 H621 H641 H642 H643 H681 H682 H683 H685
 H689 H715 H721 H722 H723 J011 J012 J013 J014 J111 J131 J132 J133
 J171 J172 J173 J211 J221 J222 J231 J232 J241 J242 J261 J271 J272
 J273 J311 J321 J322 J331 J332 J341 J342 J361 J371 J372 J373 J411
 J431 J432 J471 J521 J581 J582 J583 J592 K442 K499 K510 K599 K742
 K799 L142 L143 L199 L462 L463 L499 L921 L922 L941 L943 L999 M111
 M112 M113 M115 M116 M119 M121 M122 M123 M124 M125 M126 M129 M131
 M132 M133 M135 M136 M137 M139 M141 M142 M143 M150 M210 M211 M212
 M213 M214 M215 M216 M220 M221 M222 M223 M224 M225 M226 M231 M232
 M233 M240 M273 M280 M281 M311 M314 M315 M316 M320 M321 M322 M323
 M331 M332 M333 M334 M340 M342 M343 M344 M349 M352 M353 M373 M391
 M392 M393 M412 M413 M414 M510 M511 M512 M520 M521 M522 M523 M530
 M531 M532 M533 M540 M541 M542 M543 M630 M650 M720 N221 N225 N242
 N311 N421 N422 N512 N513 P210 P420 P442 P444 P446 P448 P451 P510
 P517 P617 P625 P641 P642 P646 P714 P731 P735 P738 P942 M905
 M904
 RIN: 00210 00211 00212
MCN: 0327-34301-K 0327-34301-P

P210 : Antiviral
 P420 : Antiinflammatory
 P442 : Anticonvulsant
 P444 : Antiparkinson
 P446 : Neuroleptic
 P448 : Anxiolytic
 P451 : Antidepressant
 P510 : Autnomic Nervous System
 P517 : Muscle Relaxant
 P617 : Antimetabolite
 P625 : Hormone Activity
 P641 : Alcoholism
 P642 : Smoking
 P646 : Antidote
 P714 : Anabolic
 P731 : Anorectic
 P735 : Antidiarrhoeal
 P738 : Ulcers
 P942 : Wound Treatment

N221 : C=C-H H bond broken
 N225 : C-Hal bond broken
 N242 : C-O bond broken
 N311 : C-C bond formed
 N421 : Acid conditions
 N422 : Basic conditions
 N512 : 10-30°C
 N513 : 30-200°C

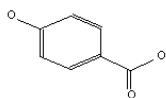
A: substance analysed/detected	Q: product defined by starting materials
C: catalyst	R: removing/purifying agent
D: detecting agent	S: starting material
E: excipient	T: therapeutic
K: known (always output for recent data if not N)	U: use of single compound
M: component of a mixture	V: reagent
N: new compound	X: substance removed
P: known compound produced	Z: miscellaneous

Specific Compounds

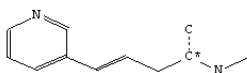
CN.S METHYL- ((E)- (S)-1-METHYL-4-PYRIDIN-3-YL-BUT-3-ENYL)-AMINE
4-HYDROXY-BENZOATE

SDCN RAMSYO

CM 1



CM 2

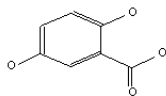


AN.S DCR-1310318

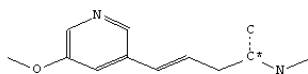
CN.S [(E)- (S)-4- (5-METHOXY-PYRIDIN-3-YL)-1-METHYL-BUT-3-ENYL]-METHYL-AMINE
2,5-DIHYDROXY-BENZOATE

SDCN RAMSYP

CM 1

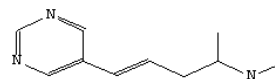


CM 2

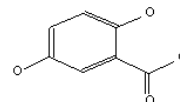


AN.S DCR-1310319

CN.S METHYL- ((E)-1-METHYL-4-PYRIMIDIN-5-YL-BUT-3-ENYL)-AMINE
2,5-DIHYDROXY-BENZOATE

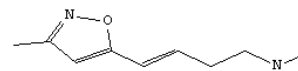


CM 2



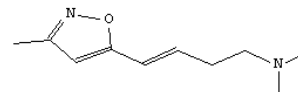
AN.S DCR-1310320

CN.S Methyl-[(E)-4- (3-methyl-isoxazol-5-yl)-but-3-enyl]-amine
SDCN RAMSYR



AN.S DCR-1310321

CN.S Dimethyl-[(E)-4- (3-methyl-isoxazol-5-yl)-but-3-enyl]-amine
SDCN RAMSYS



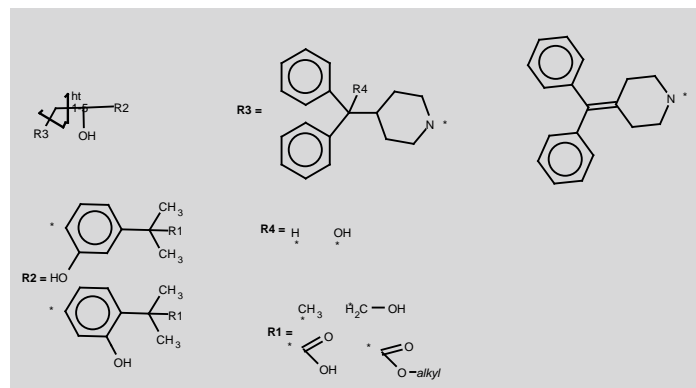
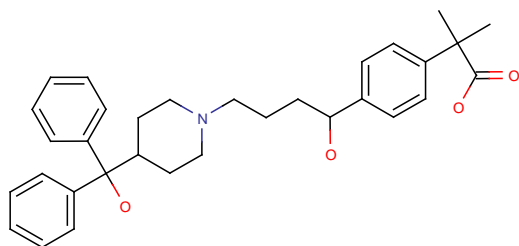


Challenge #3

WHAT DOES IT MEAN?

Where is my hit – and does it matter?

- Ranking of hits based on “nearness to the core of the invention”



Enumeration options:

- Full enumeration
- Random enumeration
- Markush reduction according to the hit

Max. structures:

Output to file

Display options

Rows:

Columns:

Show R-groups

Colouring

Enumerate

4 structures enumerated

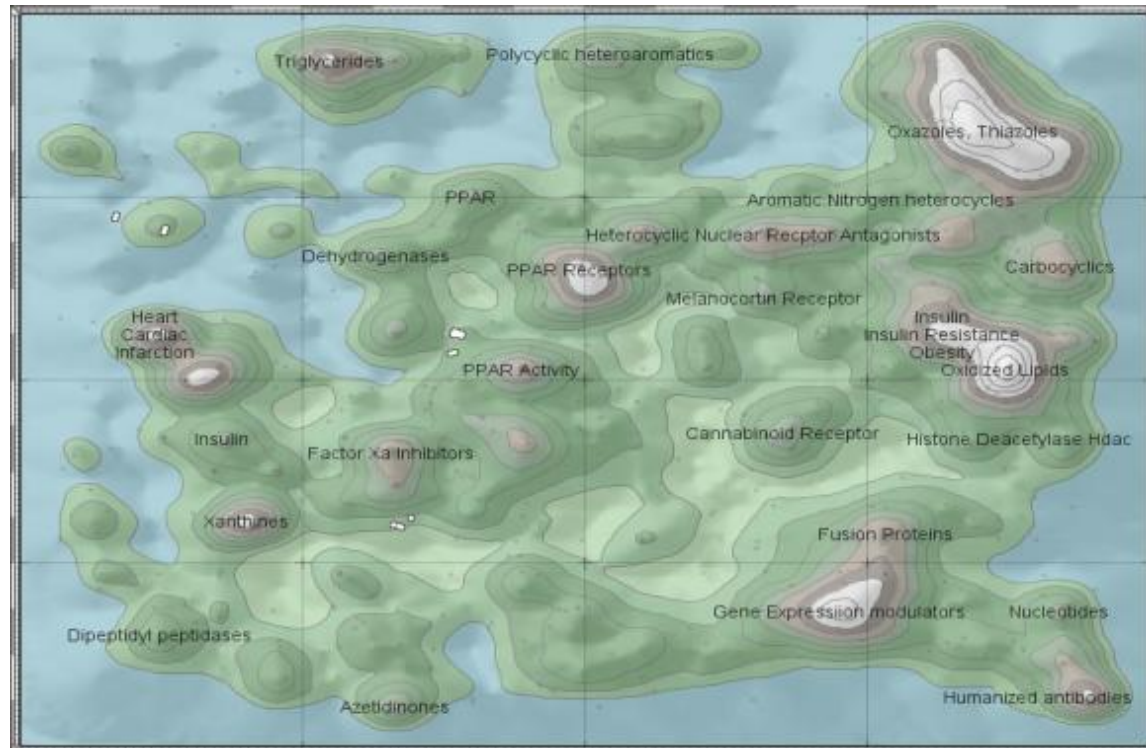
Full enumeration of this structure produces 240 structures

It's a hit – but is it a “good” hit?

JChem's useful selective enumeration

Visualising Results

- Can we rank results in terms of “nearness”?
- Can we visualise the patent landscape in some way?



Smarter Querying

- Can we formulate sophisticated, specialist queries?
- Can we do so without requiring extensive training?

