Structure Search in MARPAT®
Agenda

• MARPAT® database content
• Review MARPAT structure representations
• Use CASLINK to conduct a comprehensive substructure search
• Discuss Match Level for search precision
Markush structures are prevalent in chemical claims

• Searching Markush structures is useful
  – Assessing patentability of a substance or substance class
  – Evaluating double patenting
  – Exploring freedom-to-operate
  – Assessing the value of a patent portfolio
What is a Markush structure?

A Markush structure is a single structure used to define a set of implied structures.

G1     = H / R "substituent"
G2     = heterocycle <containing 5-6 atoms, 5- to 6-membered monocyclic ring> (opt. substd.) /
        (Specifically claimed: 145 / 153 / 162 / 171 / 180 / pyridyl (opt. substd.))

...
Eugene A. Markush is the inventor on the original Markush patent granted in 1924.

The claims in this patent include "Markush" language, listing alternative chemical substances, but no Markush structures.
MARPAT is a structure-searchable document database

- Covers Markush structures from patents from 1961 to present
- Covers patents from 63 patent authorities
- Over 404,000 records, containing over 980,000 Markush structures
- Updated daily by CAS with 150-200 Markush structures
MARPAT records provide a concise representation of structure information from patents

- A searchable structure that includes
  - Variable groups
  - Optional substitution within the variable groups
  - Occurrence counts for groups and substituents

- Patent location and derivative information

- Detailed description of substituents
MARPAT records include bibliographic information and abstracts from CAplus

<table>
<thead>
<tr>
<th>AN</th>
<th>153:589331 MARPAT</th>
<th>Full-text</th>
</tr>
</thead>
<tbody>
<tr>
<td>TI</td>
<td>Glycine-substituted thieno [2,3-d] pyrimidines with combined LH and FSH agonistic activity</td>
<td></td>
</tr>
<tr>
<td>IN</td>
<td>Hanssen, Robert Gerard Jules Marie; Timmers, Cornelis Marius; Kelder, Jan</td>
<td></td>
</tr>
<tr>
<td>PA</td>
<td>N. V. Organon, Neth.</td>
<td></td>
</tr>
<tr>
<td>SO</td>
<td>Taiwan., 4pp.</td>
<td></td>
</tr>
<tr>
<td>CODEN</td>
<td>TWXXA5</td>
<td></td>
</tr>
<tr>
<td>DT</td>
<td>Patent</td>
<td></td>
</tr>
<tr>
<td>LA</td>
<td>Chinese</td>
<td></td>
</tr>
<tr>
<td>FAN.CNT</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>PATENT NO.</th>
<th>KIND</th>
<th>DATE</th>
<th>APPLICATION NO.</th>
<th>DATE</th>
</tr>
</thead>
<tbody>
<tr>
<td>TW 314861</td>
<td>B</td>
<td>20090921</td>
<td>TW 2002-115229</td>
<td>20020709</td>
</tr>
<tr>
<td>TW 2002-115229</td>
<td>20020709</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| AB    | The invention provides novel glycine-substituted thieno [2,3-d] pyrimidines with combined LH and FSH agonistic activity. |
G1 = C(O) / CH2
G2 = alkyl <containing 1-4 C> /
    alkenyl <containing 2-4 C> / Ph (opt. substd. by 1 or more G3) / furyl (opt. substd. by 1 or more G3) /
    thienyl (opt. substd. by 1 or more G3) / pyridyl (opt. substd. by 1 or more G3) / 28
G3 = OH / F / Cl / Br / I / NO2 / CF3 / CN / NH2 /
    alkylamino <containing 1-4 C> /
    dialkylamino <each alkyl containing 1-4 C>

Patent location: claim 1
Note: or pharmaceutically acceptable salts
Patent documents are evaluated for specific and Markush structures

- Specific Structures
- Patent Document*
- Markush Structures

**CAS REGISTRY**
- A substance database
- Structure-searchable

**MARPAT**
- A document database
- Structure-searchable

**CAplus**

* Must meet CAS selection criteria
Specific substances from patents are added to REGISTRY and CAplus

• Specific substances are registered in REGISTRY
  – Specific substances can be described by name or structure in the source patent, but must be unambiguous
  – Can be single-component or multi-component

• Specific substances described in the patent claims and examples are indexed in CAplus
  – Products, starting materials, catalysts, etc.
Patents indexed for CAplus are examined for Markush structures

• MARPAT is designed as an extension of REGISTRY in coverage of chemical structural information in patents
  – REGISTRY contains specific structures from journals, patents, and other documents
  – MARPAT contains generic structures from patents
• A single structure query can be searched in both REGISTRY and MARPAT
CASLINK is a multifile search environment

Structure query

Real substances from REGISTRY

Generic substances from MARPAT

Records from CAILplus

Records from MARPAT

Combined records

De-duplication

Searches run automatically in CASLINK.
MARPAT is designed for generic structures

• Markush structures include
  – Organic structures
  – Organometallics
  – Oligomeric substances having up to ten Structural Repeating Units (SRUs)

• Criteria include
  – Symbolically depicted formulas or structures
  – Defined variable alternatives
Markush structures can also be selected from the disclosure

- If the disclosed Markush structure is broader than the claimed Markush structure
  - Claims indicate $R = O$ or $S$
    Disclosure describes $R = O$, $S$, or NH
- If there is no Markush structure in the claims
- Others based on analyst judgment
  - Additional preferred embodiment
How does MARPAT handle polymers?

- Oligomers having up to ten SRUs are included

Claimed polymer starting materials e.g., “Claim 1. A monomer of the formula…”

Polymerization catalysts and additives

- Not selected for MARPAT

- Longer polymers having more than ten SRUs

\[ G_1 \cdot G_2 \cdot G_4 \cdot G_8 \cdot G_6 \cdot G_2 \cdot G_8 \cdot G_6 \cdot G_{10} \cdot G_{12} \cdot G_{24} \cdot G_{14} \]

\[ G_1 = \text{alkyl } <\text{containing 1-20 C, DC } (0^-) \text{ M3} > / \text{Et} / \text{Bu-t} / 28 / \text{dodecyl} / \text{tetradecyl} / \text{H} \]
Metal-containing substances require a separate search in MARPAT

- Metal ions are often shown as separate fragments in the structure file in MARPAT
- Metal-containing substances should not be searched in the CASLINK environment

G19 = R <"signal"> / (Specifically claimed: 205)
MARPAT can include generic sequences

- MARPAT covers amino acid sequences of up to ~150 residues and nucleic acid sequences of up to ~30 residues
- Markush structures in patents containing longer biological sequences are uncommon
The following substance types are not covered in the MARPAT database

- Alloys, inorganic salts, intermetallics, metal oxides and polymers
- Substances with no searchable variability
  - Example: “Ph-R, where R is any substituent”
- Table structures comprised entirely of substances covered in REGISTRY
- Text describing generic substance classes
  - Example: “halogenated alkenediols”
How does CAS take a Markush structure from a patent and convert it into a MARPAT file structure?

Claim 1.

In which:
A represents an alkylene group that can be substituted by up to three halogen atoms or is alkenyene substituted by 1-2 hydroxy groups
D represents optionally substituted aryl or heteroaryl group such as phenyl
X represents halo, alkyl, or alkoxy, and \( n = 0-5 \); or salts thereof
Three G-groups will be created for this structure

G1 will represent the A variable

G2 will represent the D variable

G3 will represent the X variable
Comparison of a patent Markush structure and a MARPAT display

Patent

\[(X)n - D - A\]

In which:
- **A** represents an alkylene group that may be substituted by up to three halogen atoms or is alkenylene substituted by 1-2 hydroxy groups
- **D** represents optionally substituted aryl or heteroaryl group such as phenyl
- **X** represents halo, alkyl, or alkoxy, and **n** = 0-5; or salts thereof

MARPAT

\[\text{G1} = \text{alkylene (opt. subst. by up to 3 halo) / alkenylene (subst. by 1-2 OH)}\]
\[\text{G2} = \text{aryl (opt. subst. by up to 5 G3) / heteroaryl (opt. subst. by up to 5 G3) / Ph (opt. subst. by up to 5 G3)}\]
\[\text{G3} = \text{halo/alkyl/alkoxy}\]

Patent Location: claim 1
Note: or salts
Specific claim and disclosure information are added to the variable group descriptions

- Specific groups from dependent claim(s)
  - Independent Claim: alkylene
    Dependent Claim: CH2

- Preferred embodiments from the disclosure
  - Claim: aryl, heteroaryl
    Disclosure: preferably phenyl or pyridyl

- Author definitions of terms from the disclosure
  - “Alkylene and alkenylene are meant to include straight, branched, or cyclic groups of up to 10 carbon atoms.”
Markush structures in patents are often intricate and extensive

- Patent claim limitations can have varying levels of specificity
  - “methyl – alkyl – carbon chain”
- MARPAT is indexed with varying levels of specificity
  - “Me – alkyl – carbon chain”
- Search queries can be constructed with varying levels of specificity

Different portions of a structure query can be searched with different levels of specificity!
MARPAT is primarily accessible via structure search

- Bibliographic and abstract information are
  - Displayable in MARPAT
  - Searchable in CAplus and in CASLINK
- Limited text information can be searched in the MARPAT Basic Index
  - Patent location, derivatives, physical properties, stereochemistry, and other notes
MARPAT allows comprehensive and precise structure searches

• Why does precision matter?
  – All MARPAT answers are valid answers, regardless of precision settings
  – When building a MARPAT query, anticipate acceptable candidate answers

• MARPAT precision search tools
  – Match Level: ATOM, CLASS, ANY
  – Generic Definitions
  – Element Count: Limited versus Unlimited
Match Levels in MARPAT mirror the specificity typical of patent Markush structures

<table>
<thead>
<tr>
<th>Most specific</th>
<th>ATOM</th>
<th>Methyl, Phenyl</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CLASS</td>
<td>Ak, Cb</td>
</tr>
<tr>
<td>Least specific</td>
<td>ANY</td>
<td>R</td>
</tr>
</tbody>
</table>

- Ak = alkyl
- Cb = carbocyclic
- R represents groups that cannot be drawn and are described in the patent using text, such as “electron-withdrawing group”, “leaving group”, “linker”, “group to form a ring”, etc.
Match Level functions the same for real atoms or generic variables in queries

- **ATOM** matches at the specific atom level
- **CLASS** matches at both the generic CLASS level and specific atoms
- **ANY** matches at ATOM and CLASS, including indefinitely defined substitutions

<table>
<thead>
<tr>
<th>MATCH LEVEL</th>
<th>RETRIEVAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATOM</td>
<td>Specific</td>
</tr>
<tr>
<td>CLASS</td>
<td>Specific or Generic</td>
</tr>
<tr>
<td>ANY</td>
<td>Specific or Generic or Indefinite</td>
</tr>
</tbody>
</table>
Structure queries automatically include a default Match Level assignment

- This assignment is only taken into account when searching MARPAT
- The default settings for Match Levels are

<table>
<thead>
<tr>
<th>QUERY NODES</th>
<th>DEFAULT MATCH LEVEL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ring atoms</td>
<td>ATOM</td>
</tr>
<tr>
<td>Cy, Cb, Hy</td>
<td>ATOM</td>
</tr>
<tr>
<td>Chain atoms</td>
<td>CLASS</td>
</tr>
<tr>
<td>Ak</td>
<td>CLASS</td>
</tr>
</tbody>
</table>
Search Example

Run a comprehensive search on the following structure and its use as a bronchodilator:

R1 = Any non-hydrogen substituent (in a ring or chain)
R2 = Any ring
No additional fusion on the nitrogen-containing ring.
Additional substitution allowed at all open sites.
Build and save the structure query

• Ring is isolated and terminal node A is designated as ring/chain.
• By default, ring atoms and Cy are set to Match Level ATOM and all chain nodes are set to Match Level CLASS.
• These settings have no effect in REGISTRY.
Enter CASLINK and upload the structure

=> FILE CASLINK
FILE 'CAPLUS' ENTERED AT ...
FILE 'MARPAT' ENTERED AT ...
FILE 'REGISTRY' ENTERED AT
CLUSTER 'CASLINK' ENTERED

Predefined command sequences will be executed in REGISTRY, MARPAT, and CAPLUS.

Uploading C:● ● ●
L1 STRUCTURE UPLOADED

=> D L1
L1 HAS NO ANSWERS
L1

![Chemical structure diagram](attachment:image.png)
Run a SAMPLE search

\[
=> \text{S SSS SAM L1}
\]

\[
\text{S L1 SSS SAM FILE=REGISTRY}
\]

\[
\text{FILE 'REGISTRY} \bullet \bullet \bullet
\]

\[
\text{SAMPLE SCREEN SEARCH COMPLETED} - \text{2196 TO ITERATE}
\]

\[
100.0\% \text{ PROCESSED} \quad 2196 \text{ ITERATIONS}
\]

\[
\text{FULL FILE PROJECTIONS: \ ONLINE  \ **COMPLETE**}
\]

\[
\text{BATCH \ **COMPLETE**}
\]

\[
\text{PROJECTED ITERATIONS: \ 41109 \ TO \ 46731}
\]

\[
\text{PROJECTED ANSWERS: \ 2 \ TO \ 124}
\]

\[
\text{L2} \quad 2 \ \text{SEA SSS SAM L1}
\]

\[
\text{S L2 SSS SAM FILE=MARPAT}
\]

\[
\text{FILE 'MARPAT} \bullet \bullet \bullet
\]

\[
\text{SAMPLE SCREEN SEARCH COMPLETED} - \text{1810 TO ITERATE}
\]

\[
100.0\% \text{ PROCESSED} \quad 1810 \text{ ITERATIONS}
\]

\[
\text{SEARCH TIME: 00.00.02}
\]

\[
\text{FULL FILE PROJECTIONS: \ ONLINE  \ **COMPLETE**}
\]

\[
\text{BATCH \ **COMPLETE**}
\]

\[
\text{PROJECTED ITERATIONS: \ 33761 \ TO \ 38639}
\]

\[
\text{PROJECTED ANSWERS: \ 3 \ TO \ 164}
\]

\[
\text{L3} \quad 3 \ \text{SEA SSS SAM L1}
\]
Use D SCAN to evaluate REGISTRY search

=> D SCAN L2

L2 2 ANSWERS REGISTRY

IN 3-Pyrroolidinecarboxylic acid, 1-(1,1-dimethylethyl)-4,4-dimethyl-2,5-dioxo-, 4-nitrophenyl ester

MF C17 H20 N2 O6

**PROPERTY DATA AVAILABLE IN THE 'PROP' FORMAT**

HOW MANY MORE ANSWERS DO YOU WISH TO SCAN? (1): 0
Use D SCAN to evaluate MARPAT search

=> D SCAN L3
L3  3 ANSWERS  MARPAT
IC  ICM  C07D453-02
    ICS  A61K031-435
CC  27-16 (Heterocyclic Compounds (One Hetero Atom))
    Section cross-reference(s): 1
TI  Preparation of R-3-quinuclidinol esters with
    heterocyclylcarboxylic acids as muscarinic M1 antagonists
ST  heterocyclylcarbonyloxyazabicyclooctane prepn muscarinic M1
    antagonist;…
IT  Ulcer inhibitors
    (quinuclidinol esters)
IT  Emphysema
    Heart, disease or disorder
    (treatment of, with quinuclidinol esters)
IT  Bronchodilators
    (antiasthmatics, quinuclidinol esters)
•••••
Use D SCAN to evaluate MARPAT search (cont.)

IT 79232-58-9
RL: RCT (Reactant); RACT (Reactant or reagent)
(transesterification reaction of, with quinuclidinol)

MSTR 1 Assembled

Patent location: claim 1
Stereochemistry: diastereoisomers and mixtures

HOW MANY MORE ANSWERS DO YOU WISH TO SCAN? (1): END
Run a FULL file search

=> S SSS FULL L1

S L1 SSS FUL FILE=REGISTRY
FILE 'REGISTRY'
FULL SEARCH INITIATED 09:36:14
FULL SCREEN SEARCH COMPLETED - 46441 TO ITERATE
100.0% PROCESSED 46441 ITERATIONS 49 ANSWERS
L4 49 SEA SSS FUL L1

S L4 SSS FUL FILE=MARPAT
FILE 'MARPAT'
FULL SEARCH INITIATED 09:36:20
FULL SCREEN SEARCH COMPLETED - 39211 TO ITERATE
100.0% PROCESSED 39211 ITERATIONS 38 ANSWERS
L5 38 SEA SSS FUL L1

S L4 FILE=CAPLUS
L6 10 FILE CAPLUS

DUP REM L5 L6
L7 47 DUP REM L5 L6 (1 DUPLICATE REMOVED)
   ANSWERS '1-38' FROM FILE MARPAT
   ANSWERS '39-47' FROM FILE CAPLUS

Duplicates are removed between MARPAT and CAplus.
Refine search with text terms in CASLINK

\[ S \text{ L7 AND (BRONCHODILAT? OR ASTHMA? OR ANTIASTHMA? OR (PULMONAR? OR LUNG?) (L) DILAT?)} \]

\[ S \text{ L6 AND (BRONCHODILAT? OR ASTHMA? OR ANTIASTHMA? OR (PULMONAR? OR LUNG?) (L) DILAT?) FILE=CAPLUS} \]

L8 0 FILE CAPLUS

\[ S \text{ L5 AND (BRONCHODILAT? OR ASTHMA? OR ANTIASTHMA? OR (PULMONAR? OR LUNG?) (L) DILAT?) FILE=CAPLUS} \]

L9 3 FILE CAPLUS

\[ S \text{ L9 AND L5 FILE=MARPAT} \]

L10 3 FILE MARPAT

DUP REM L10 L8

L8 HAS NO ANSWERS

PROCESSING COMPLETED FOR L10

PROCESSING COMPLETED FOR L8

L11 3 DUP REM L10 L8 (0 DUPLICATES REMOVED)

ANSWERS '1-3' FROM FILE MARPAT

Text terms are searched in CAplus and then recombined with the MARPAT L-number to restore structure highlighting.

- L5 = MARPAT answer set
- L6 = CAplus answer set
- L7 = Multifile answer set
MARPAT display options

• Abstract graphic image (GI) is chosen by the CAS analyst and can be helpful in determining the structure of interest in a patent
• FQHIT and QHIT assembled displays focus on the G-Groups that were part of your structure query
• FQHITEXG and QHITEXG define G-Groups that display in an assembled structure, but were not in the search query
• Use Patent Location to find the structure in the original patent document
CLAIMS:
What we claim is:

1. A compound of the formula I or a pharmaceutically acceptable salt or an in vivo hydrolysable ester thereof

wherein
ABS displays can include a graphic image
FQHIT shows G-Groups that matched the structure query

**MSTR 1 Assembled**

```
N    O     O
   \   /   G4
  \ /   \   
   O     N
     CH   =  O
```

- **Patent location:** claim 1
- **Stereochemistry:** diastereoisomers and mixtures

- **FQHIT** is a condensed format that shows only the “hit” portions of the answer that match the query.
- **MARPAT display algorithms** create assembled Markush structures.
FQHITEXG displays additional definitions of G-groups

Additional displayed G-groups:
G4 = cycloalkyl / aryl / heterocycle <5- to 6-membered monocyclic ring> / (Specifically claimed: Ph / thienyl)

Patent location: claim 1
Stereochemistry: diastereoisomers and mixtures
Revisiting Match Level

• Match Level ATOM only retrieves atom matches
  – For generic rings such as Hy, Cb, and Cy, this means only complete rings defined by real atoms

• Match Level CLASS finds atom matches and broader Markush structures
  – For Hy, this means definitions such as “heterocyclic ring systems”
Consider changing the Match Level for generic rings Hy, Cb, and Cy

Search Example

Compare retrieval when the match level on the Cy (R2) is changed from the default of ATOM to CLASS.
Changing Match Level Assignments

• Step 1: Right-click the atom of interest. Select **Markush Attributes** from the pop-up menu.

• Step 2: Select **Match Level: CLASS**, and then click **OK**.
Compare retrieval when Match Level is changed

=> S L12 SSS FULL

L12     STRUCTURE UPLOADED
L13     49 SEA FILE=REGISTRY SSS FUL L12
L14     123 SEA FILE=MARPAT SSS FUL L12
L15     10 SEA FILE=CAPLUS SPE=ON ABB=ON PLU=ON L13
L16     132 DUP REM L14 L15 (1 DUPLICATE REMOVED)

ANSWERS '1-123' FROM FILE MARPAT
ANSWERS '124-132' FROM FILE CAPLUS

=> S L16 NOT L7

Changing Match Level on the Cy to CLASS retrieved 85 additional answers.

L17     0 SEA FILE=CAPLUS SPE=ON ABB=ON PLU=ON L15 NOT L6
L18     85 SEA FILE=MARPAT SPE=ON ABB=ON PLU=ON L14 NOT L5
L19     85 DUP REM L18 L17 (0 DUPLICATES REMOVED)

ANSWERS '1-85' FROM FILE MARPAT
Refine with text to see additional answers

```plaintext
=> S L19 AND (BRONCHODILAT? OR ASTHMA? OR ANTIASTHMA? OR (PULMONAR? OR LUNG?) (L) DILAT?)

S L17 AND (BRONCHODILAT? OR ASTHMA? OR ANTIASTHMA? OR (PULMONAR? OR LUNG?) (L) DILAT?) FILE=CAPLUS
L20          0 FILE CAPLUS

S L18 AND (BRONCHODILAT? OR ASTHMA? OR ANTIASTHMA? OR (PULMONAR? OR LUNG?) (L) DILAT?) FILE=CAPLUS
L21          6 FILE CAPLUS

S L21 AND L18 FILE=MARPAT
L22          6 FILE MARPAT

DUP REM L22 L20
L22 HAS NO ANSWERS
PROCESSING COMPLETED FOR L22
PROCESSING COMPLETED FOR L20
L23          6 DUP REM L22 L20 (0 DUPLICATES REMOVED)
            ANSWERS '1-6' FROM FILE MARPAT
```

Six additional patents in MARPAT were retrieved that might be relevant to the question.
Review results

=> D SCAN

TI 2-Oxo-4-phenyl-3-pyrrolidinecarboxylic acid hydrazide derivatives as PAR 2-modulating compounds and their preparation, pharmaceutical compositions and use in the treatment of various diseases

MSTR 1 Assembled

\[
\begin{align*}
\text{G1} & = (1-3) \text{ CH2} \\
\text{G5} & = 0 \\
\text{G6} & = \text{carbocycle} \ <\text{containing 3-8 C, non-aromatic, 0 or more double bonds, no triple bonds}> \ (\text{opt. substd.}) \\
\text{G15} & = \frac{49}{101} 4^9 G5-G6 10^1 G31-G5-G6
\end{align*}
\]

Patent location: claim 1

Note: or pharmaceutically acceptable salts or prodrugs
Summary

• MARPAT contains generic chemical structure information
• Use CASLINK to conduct a comprehensive substructure search across the CAS REGISTRY, CAplus, and MARPAT databases
• Generic Definitions, Element Count, and Match Level tools add precision and comprehensiveness to MARPAT searches
Resources

• Online
  => Help Directory
  => Help CASLINK

• MARPAT User Guide (2013)
  – www.cas.org

• CAS Customer Center
  – 1-800-753-4227 option 3
  – help@cas.org