

STN[®]

Advanced Structure Searching:
Why did(n't) I get that?

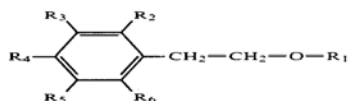
Agenda

- Demonstrate ways to control **precision and recall** in chemical substance searching in CAS REGISTRYSM

STN

2

Example 1: Sample Claim

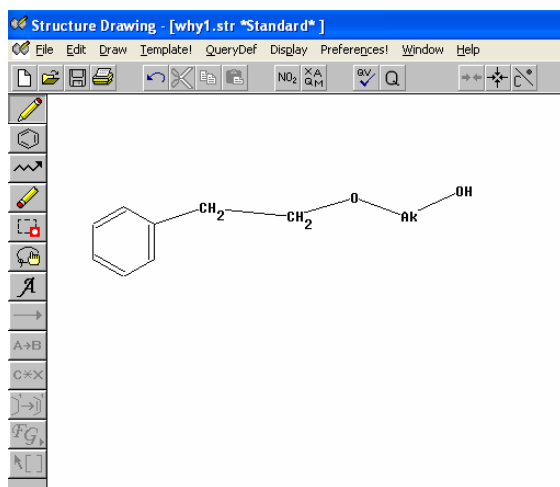


Where R₁ is an alcohol
R₂ – R₆ can be anything

STN

3

Query structure



STN

4

SAM SEARCH

=> FILE REG

Uploading C:\CASNC\STN Express\Queries\529.str

L1 STRUCTURE UPLOADED

=> S L1

SAMPLE SEARCH INITIATED 11:38:41 FILE 'REGISTRY'

SAMPLE SCREEN SEARCH COMPLETED - 38027 TO ITERATE

3.0% PROCESSED 1000 ITERATIONS 4 ANSWERS

INCOMPLETE SEARCH (SYSTEM LIMIT EXCEEDED)

SEARCH TIME: 00.00.01

FULL FILE PROJECTIONS: ONLINE **COMPLETE**

BATCH **COMPLETE**

PROJECTED ITERATIONS: 748891 TO 772189

PROJECTED ANSWERS: 998 TO 2044

L2 4 SEA SSS SAM L1

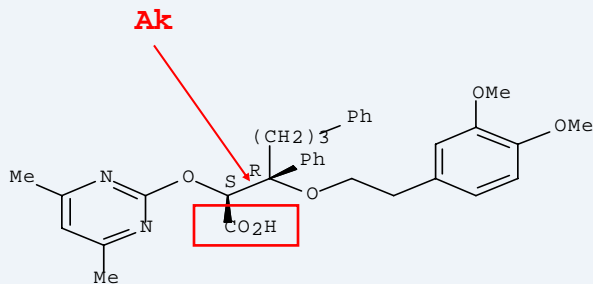
Search projects to run to completion, but...

STN

5

Why did I get that?

=> D SCAN



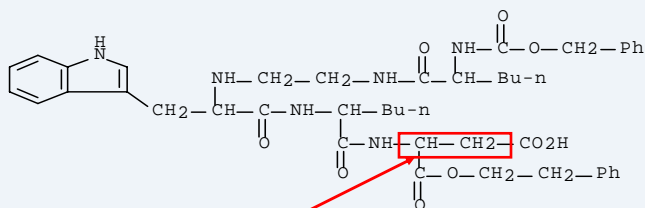
This is not an alcohol but an acid. It is retrieved because there is indeed an OH on the Alkyl group.

STN

6

Why did I get that?

=> D SCAN



Ak is further substituted

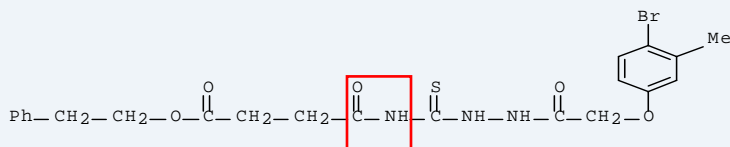
Nothing was done to control the number of additional substitutions on the Alkyl chain, with more than one substituent on the Ak.

STN

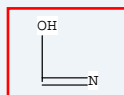
7

Why did I get that?

=> D SCAN



Tautomeric with



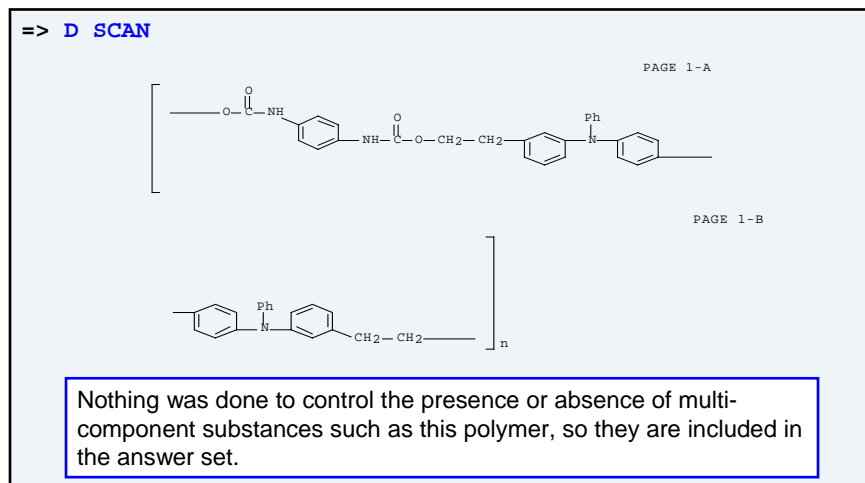
STN Express chose a 'normalized bond' value for the Ak-OH bond. This results in finding potentially tautomeric substances.

STN

8

Why did I get that?

=> D SCAN



STN

9

Revising the query structure

- Use Non-Hydrogen attachments to limit the substituents on the Ak
- Adjust the bond value to EXACT so that only OH groups are found
- Use a filter to eliminate polymers

STN

10

Revising the query structure

Structure Drawing - [why1.str *Standard*]

File Edit Draw Template QueryDef Display Preferences Window Help

Right-click on the bond; then choose EXACT.

Bond Characteristics

Bond Type:	Bond Value:
<input checked="" type="radio"/> Chain	<input type="radio"/> Exact/Norm
<input type="radio"/> Ring/Chain	<input checked="" type="radio"/> Exact
<input type="radio"/> Ring	<input type="radio"/> Normalized
<input type="radio"/> Mixture	<input type="radio"/> Unspecified
	<input type="radio"/> Mixture

Bond Types...
Cancel OK

STN

11

Revising the query structure

Structure Drawing - [why2.str *Standard*]

File Edit Draw Template QueryDef Display Preferences Window Help

Right-click on the Ak; set the non-H attachments to exactly 2.

Non Hydrogen Attachments (Connectivity)

<input checked="" type="radio"/> Specific	<input type="radio"/> Any
<input checked="" type="radio"/> Exact	<input type="radio"/> Chain
<input type="radio"/> Minimum	<input type="radio"/> Ring
<input type="radio"/> Maximum	<input checked="" type="radio"/> Ring/Chain
<input type="radio"/> Mixture	
Count (0 to 16): <input type="text" value="2"/>	

Cancel OK

STN

12

Revising the query structure

Save the structure with Filters- Choose the polymer filter and NOT it out.

Note: STN Express suggested the C-CH2-C filter.

STN

13

Upload and search revised query

```

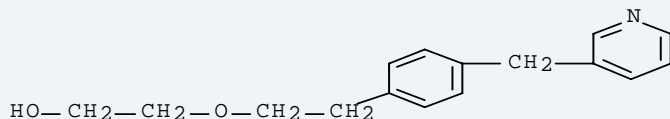
=> SCREEN ...
=> QUE L5 AND L3 NOT L4
L6  QUERY CREATED
=> S L6
SAMPLE SEARCH INITIATED 11:44:11 FILE 'REGISTRY'
SAMPLE SCREEN SEARCH COMPLETED - 21715 TO ITERATE
      5.3% PROCESSED      1000 ITERATIONS              0 ANSWERS
INCOMPLETE SEARCH (SYSTEM LIMIT EXCEEDED)
SEARCH TIME: 00.00.01
FULL FILE PROJECTIONS:  ONLINE  **COMPLETE**
                        BATCH  **COMPLETE**
PROJECTED ITERATIONS:   425481 TO 443119
PROJECTED ANSWERS:     0 TO 0
L7      0 SEA SSS SAM L5 AND L3 NOT L4
=> S L6 FULL
FULL SEARCH INITIATED 11:44:15 FILE 'REGISTRY'
FULL SCREEN SEARCH COMPLETED - 433510 TO ITERATE
      100.0% PROCESSED   380072 ITERATIONS              120 ANSWERS
SEARCH TIME: 00.00.06
L8      120 SEA SSS FUL L5 AND L3 NOT L4
    
```

STN

14

Why did I get that?

=> D SCAN



Some substances have additional rings. These can be eliminated with a 'dictionary' search if desired.

STN

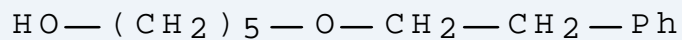
15

Limiting to substances with only 1 ring

=> S L8 AND 1/NR

L9 41 L8 AND 1/NR

=> D SCAN



/NR = NUMBER OF RINGS

STN

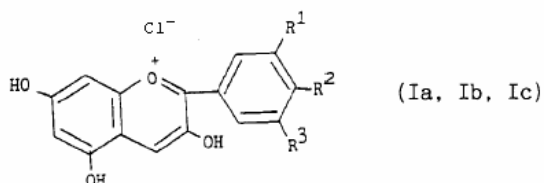
16

Example 2: Sample Claim

A process for producing a compound of formula Ia, Ib or Ic,

5

10



wherein

15

- 1a R¹=R²=OH, R³=H (cyanidin)
- 1b R¹=R³=H, R²=OH (pelargonidin)
- 1c R¹=R²=R³=OH, (delphinidin)

STN

17

Structure Query is drawn like the claim

Structure Drawing

File Edit Draw Template QueryDef Display Preferences Window Help

del3.str *Standard*

Query Verification

G1:

OH
H

OK Cancel

STN

18

Why did I get that? (zero answers!)

```
=> FILE REG
=> S L1
SAMPLE SEARCH INITIATED 13:26:18 FILE
SAMPLE SCREEN SEARCH COMPLETED - 12
100.0% PROCESSED 127 ITERATIONS
SEARCH TIME: 00.00.01
FULL FILE PROJECTIONS: ONLINE **COMPLETE**
                        BATCH **COMPLETE**
PROJECTED ITERATIONS: 1864 TO 3216
PROJECTED ANSWERS:    0 TO 0
L2 0 SEA SSS SAM L1

=> S L1 FULL
FULL SEARCH INITIATED 13:26:25 FILE 'REGISTRY'
FULL SCREEN SEARCH COMPLETED - 2504 TO ITERATE
100.0% PROCESSED 2504 ITERATIONS 0 ANSWERS
SEARCH TIME: 00.00.01
L3 0 SEA SSS FUL L1
```

Sometimes sample searches do not retrieve any answers, but the structure does iterate, so we plunge ahead...

...but the FULL search still finds zero.

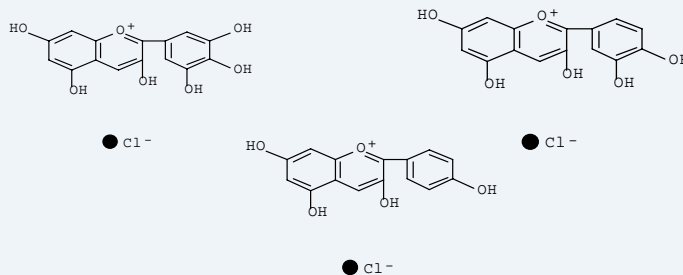
STN

19

Why didn't I get the compounds mentioned in the claim?

```
=> S (CYANIDIN OR PELARGONIDIN OR DELPHINIDIN)/CN
L4 3 (CYANIDIN OR PELARGONIDIN OR DELPHINIDIN)/CN
```

```
=> D SCAN
```

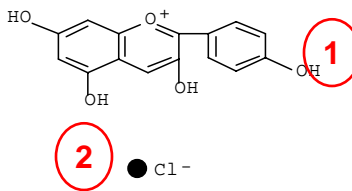


STN

20

Explanation

- These structures **do match** the query as drawn but the problem is how STN processed the search:
 - Structure fragments in a structure query (our query has two fragments) get searched together **as pieces of the same structure**. That is, the search system is looking for the two structural pieces as part of a structure in a single component
 - In the REGISTRY File, salts are found as multi-component substances



STN

21

The Rule

To find salts and other multi-component substances:

A. Draw and upload each fragment in its own L-number and AND them together (S L1 AND L2)

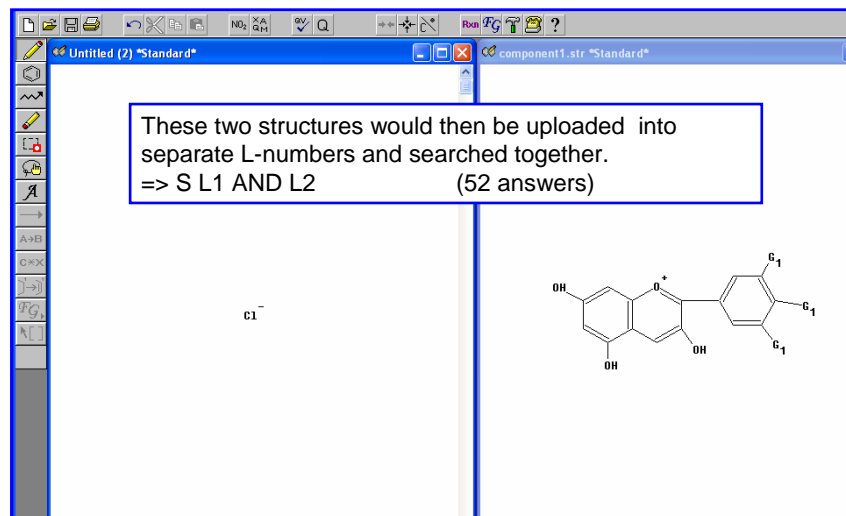
OR

B. Search one fragment and use text terms, or some other means to locate the other component(s)

STN

22

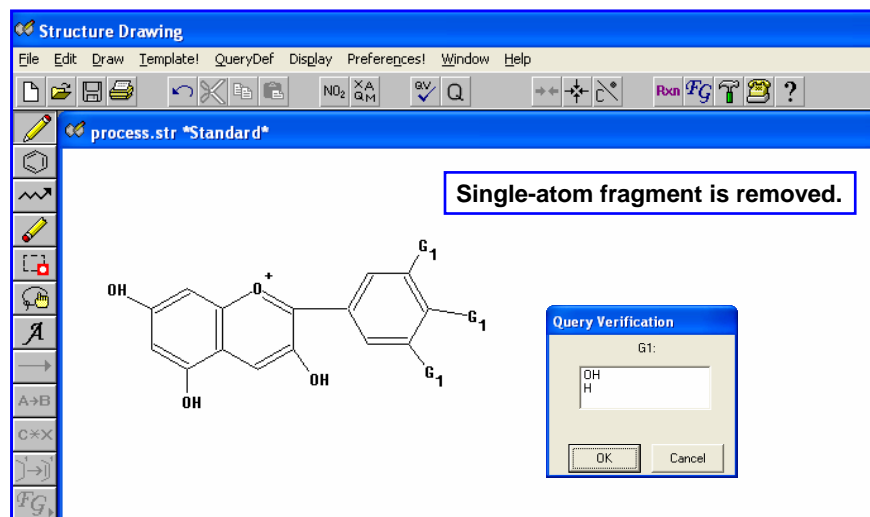
Approach A



STN

23

Approach B



STN

24

Search one fragment

```
L5      STRUCTURE UPLOADED
=> S L5 FULL
FULL SEARCH INITIATED 13:31:57 FILE 'REGISTRY'
FULL SCREEN SEARCH COMPLETED - 2504 TO ITERATE
100.0% PROCESSED 2504 ITERATIONS 91 ANSWERS
SEARCH TIME: 00.00.01
L6      91 SEA SSS FUL L5

=> D HIS
L1      STRUCTURE UPLOADED
L2      0 S L1
L3      0 S L1 FULL
L4      3 S (CYANIDIN OR PELARGONIDIN OR DELPHINIDIN)/CN
L5      STRUCTURE UPLOADED
L6      91 S L5 FULL

=> S L6 AND L4
L7      3 L6 AND L4
```

91 answers are found.

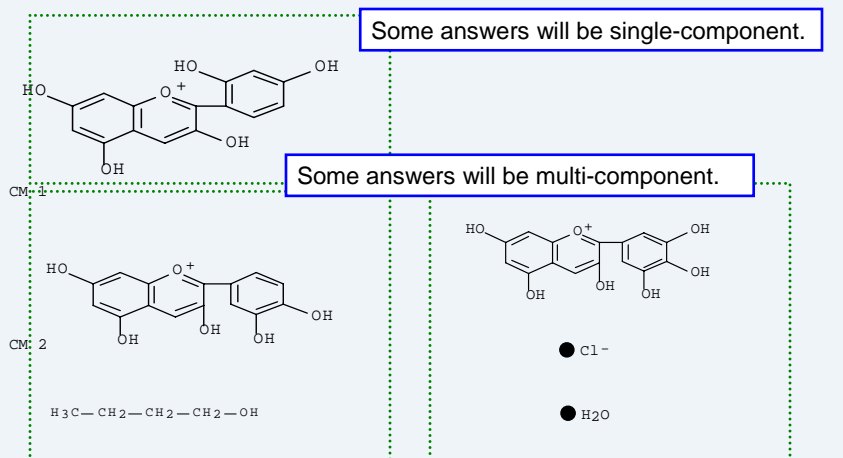
Sanity check- making sure that the three compounds mentioned in the claim were found this time!

STN

25

Scan preliminary results

```
=> D SCAN L6
```



STN

26

Refine the search with text terms

```
=> S L6 AND CL>=1
L8      52 L6 AND CL>=1
```

```
=> S L8 AND 2/NC
L9      32 L8 AND 2/NC
```

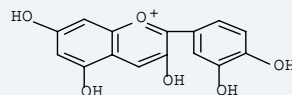
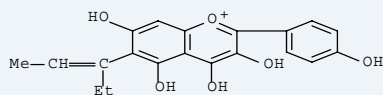
Limit to chlorine-containing compounds.

```
=> D SCAN
```

Limit to 2-component compounds if desired.

/NC = number of components

Note: Hydrates will not be included!



STN

27

Why did I get that?

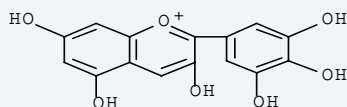
```
HOW MANY MORE ANSWERS DO YOU WISH TO SCAN? (1):1
```

```
L9      32 ANSWERS   REGISTRY   COPYRIGHT 2006 ACS on STN
IN      1-Benzopyrylium, 3,5,7-trihydroxy-2-(3,4,5-trihydroxyphenyl)-,
        chloride, monoarabioside (9CI)
MF      C20 H19 O11 . Cl
CI      IDS
```

```
CM      1
```

```
( C 5 H 9 O 4 ) — OH
CM      2
```

Where's the Chlorine?



STN

28

Why did I get that?

HOW MANY MORE ANSWERS DO YOU WISH TO SCAN? (1):1

L9 29 ANSWERS REGISTRY COPYRIGHT 2005 ACS on STN
 IN 1-Benzopyrylium, 3,5,7-trihydroxy-2-(3,4,5-trihydroxyphenyl)-,
 chloride, monoarabinoside (9CI)

MF C20 H19 O11 . Cl

CI IDS

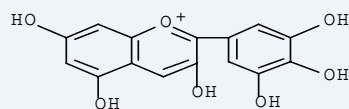
CM 1

Here it is!

(C₅H₉O₄) — OH

CM 2

This substance is an Incompletely Defined Substance (IDS). As a result, not all of the substance shows in the structure diagram- note that Component 1 is really not structured either.



STN

29

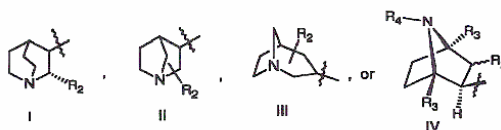
Example 3: Sample Claim

1. A compound of Formula I:

Azabicyclo-N(R₁)-C(=O)-W

Formula I

wherein Azabicyclo is



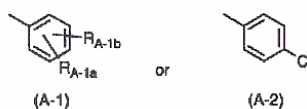
R₁ is H;

R₂ is H or alkyl;

Each R₃ is independently H, alkyl, or substituted alkyl;

R₄ is H, alkyl, an amino protecting group, or an alkyl group having 1-3 substituents selected from F, Cl, Br, I, -OH, -CN, -NH₂, -NH(alkyl), or -N(alkyl);

wherein W is (A):



STN

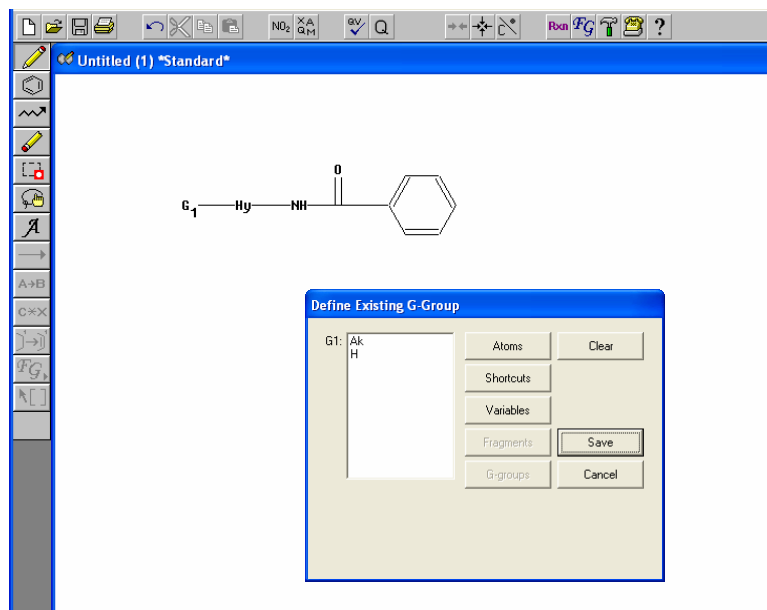
30

The approach

1. Draw a simplified representation of the structure
2. Describe the Azabicyclos in a generic fashion
3. Run a SAM search and check the statistics and answers (if any)
4. Revise and refine as necessary

STN

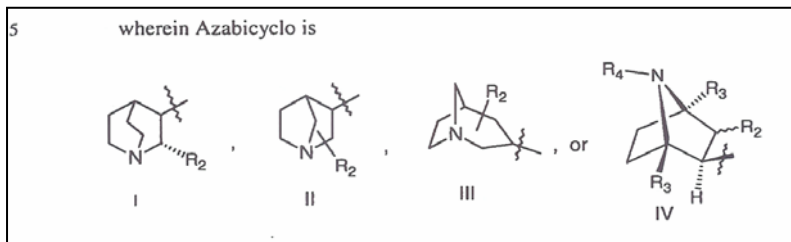
31



STN

32

Using Hy for these rings



- Bicyclic
- Saturated
- One nitrogen
- 6-7 carbons
- No oxygen, no sulfur

STN

33

Assign attributes to the Hy

1. Right-click on the Hy node
2. Select Element Count

STN

34

Assign Element counts

Element Count

<input type="radio"/> C	<input type="radio"/> Ac	<input type="radio"/> Ca	<input type="radio"/> Er	<input type="radio"/> In	<input type="radio"/> Na	<input type="radio"/> Po	<input type="radio"/> Se	<input type="radio"/> Tm	
<input type="radio"/> H	<input type="radio"/> Ag	<input type="radio"/> Cd	<input type="radio"/> Eu	<input type="radio"/> Ir	<input type="radio"/> Nb	<input type="radio"/> Pr	<input type="radio"/> Sm	<input type="radio"/> U	
<input type="radio"/> O	<input type="radio"/> Am	<input type="radio"/> Ce	<input type="radio"/> Fe	<input type="radio"/> K	<input type="radio"/> Nd	<input type="radio"/> Pt	<input type="radio"/> Sn	<input type="radio"/> V	
<input type="radio"/> S	<input type="radio"/> Al	<input type="radio"/> Cf	<input type="radio"/> Fm	<input type="radio"/> Kr	<input type="radio"/> Ne	<input type="radio"/> Pu	<input type="radio"/> Sr	<input type="radio"/> W	
<input checked="" type="radio"/> N	<input type="radio"/> Ar	<input type="radio"/> Cm	<input type="radio"/> Fr	<input type="radio"/> La	<input type="radio"/> Ni	<input type="radio"/> Ra	<input type="radio"/> T	<input type="radio"/> Xe	
<input type="radio"/> P	<input type="radio"/> As	<input type="radio"/> Co	<input type="radio"/> Ga	<input type="radio"/> Li	<input type="radio"/> No	<input type="radio"/> Rb	<input type="radio"/> Ta	<input type="radio"/> Yb	
<input type="radio"/> Si	<input type="radio"/> At	<input type="radio"/> Cr	<input type="radio"/> Ge	<input type="radio"/> Lu	<input type="radio"/> Np	<input type="radio"/> Re	<input type="radio"/> Tb	<input type="radio"/> Y	
<input type="radio"/> Cl	<input type="radio"/> Au	<input type="radio"/> Cs	<input type="radio"/> Gd	<input type="radio"/> Lr	<input type="radio"/> Os	<input type="radio"/> Rh	<input type="radio"/> Tc	<input type="radio"/> Zn	
<input type="radio"/> Br	<input type="radio"/> Ba	<input type="radio"/> Cu	<input type="radio"/> He	<input type="radio"/> Md	<input type="radio"/> Pa	<input type="radio"/> Rn	<input type="radio"/> Te	<input type="radio"/> Zr	
<input type="radio"/> F	<input type="radio"/> Be	<input type="radio"/> D	<input type="radio"/> Hf	<input type="radio"/> Mg	<input type="radio"/> Pb	<input type="radio"/> Ru	<input type="radio"/> Th		
<input type="radio"/> I	<input type="radio"/> Bi	<input type="radio"/> Dy	<input type="radio"/> Hg	<input type="radio"/> Mn	<input type="radio"/> Pd	<input type="radio"/> Sb	<input type="radio"/> Ti		
<input type="radio"/> B	<input type="radio"/> Bk	<input type="radio"/> Es	<input type="radio"/> Ho	<input type="radio"/> Mo	<input type="radio"/> Pm	<input type="radio"/> Sc	<input type="radio"/> Tl		

Exact Range Limited

Minimum Maximum

STN

35

Element Counts will appear in box as they area added

Element Count

<input type="radio"/> C	<input type="radio"/> Ac	<input type="radio"/> Ca	<input type="radio"/> Er	<input type="radio"/> In	<input type="radio"/> Na	<input type="radio"/> Po	<input type="radio"/> Se	<input type="radio"/> Tm	
<input type="radio"/> H	<input type="radio"/> Ag	<input type="radio"/> Cd	<input type="radio"/> Eu	<input type="radio"/> Ir	<input type="radio"/> Nb	<input type="radio"/> Pr	<input type="radio"/> Sm	<input type="radio"/> U	
<input type="radio"/> O	<input type="radio"/> Am	<input type="radio"/> Ce	<input type="radio"/> Fe	<input type="radio"/> K	<input type="radio"/> Nd	<input type="radio"/> Pt	<input type="radio"/> Sn	<input type="radio"/> V	
<input type="radio"/> S	<input type="radio"/> Al	<input type="radio"/> Cf	<input type="radio"/> Fm	<input type="radio"/> Kr	<input type="radio"/> Ne	<input type="radio"/> Pu	<input type="radio"/> Sr	<input type="radio"/> W	
<input type="radio"/> N	<input type="radio"/> Ar	<input type="radio"/> Cm	<input type="radio"/> Fr	<input type="radio"/> La	<input type="radio"/> Ni	<input type="radio"/> Ra	<input type="radio"/> T	<input type="radio"/> Xe	
<input type="radio"/> P	<input type="radio"/> As	<input type="radio"/> Co	<input type="radio"/> Ga	<input type="radio"/> Li	<input type="radio"/> No	<input type="radio"/> Rb	<input type="radio"/> Ta	<input type="radio"/> Yb	
<input type="radio"/> Si	<input type="radio"/> At	<input type="radio"/> Cr	<input type="radio"/> Ge	<input type="radio"/> Lu	<input type="radio"/> Np	<input type="radio"/> Re	<input type="radio"/> Tb	<input type="radio"/> Y	
<input type="radio"/> Cl	<input type="radio"/> Au	<input type="radio"/> Cs	<input type="radio"/> Gd	<input type="radio"/> Lr	<input type="radio"/> Os	<input type="radio"/> Rh	<input type="radio"/> Tc	<input type="radio"/> Zn	
<input type="radio"/> Br	<input type="radio"/> Ba	<input type="radio"/> Cu	<input type="radio"/> He	<input type="radio"/> Md	<input type="radio"/> Pa	<input type="radio"/> Rn	<input type="radio"/> Te	<input type="radio"/> Zr	
<input type="radio"/> F	<input type="radio"/> Be	<input type="radio"/> D	<input type="radio"/> Hf	<input type="radio"/> Mg	<input type="radio"/> Pb	<input type="radio"/> Ru	<input type="radio"/> Th		
<input type="radio"/> I	<input type="radio"/> Bi	<input type="radio"/> Dy	<input type="radio"/> Hg	<input type="radio"/> Mn	<input type="radio"/> Pd	<input type="radio"/> Sb	<input type="radio"/> Ti		
<input type="radio"/> B	<input type="radio"/> Bk	<input type="radio"/> Es	<input type="radio"/> Ho	<input type="radio"/> Mo	<input type="radio"/> Pm	<input type="radio"/> Sc	<input type="radio"/> Tl		

Exact Range Limited

Minimum Maximum

N,Exact,1
O,Exact,0
S,Exact,0
C,Range,6-7

Assign attributes to the Hy

The screenshot shows the STN Structure Drawing software interface. The title bar reads "Structure Drawing" and the menu bar includes "File", "Edit", "Draw", "Template!", "QueryDef", "Display", "Preferences!", "Window", and "Help". The toolbar contains various drawing tools. The main window displays "Untitled (1) *Standard*" and a chemical structure fragment. A context menu is open over a hydrogen atom node (labeled G₁), with the following options: "Markush Attributes...", "Element Count...", "Generic Definition..." (highlighted in blue), and "Non-H Attachments...". A blue box with white text contains the instructions: "1. Right-click on the Hy node" and "2. Select Generic Definitions". The STN logo is in the bottom left corner, and the number "37" is in the bottom right corner.

Set Generic Definitions

The screenshot shows the "Generic Definition" dialog box in the STN software. The dialog has a light beige background and contains several sections with radio button options. The "Saturation" section has "Any", "Unsaturated", and "Saturated" (selected) options. The "Type of Chain" section has "Any", "Branched", "Mixture", and "Linear" options. The "Number of Hetero Atoms" section has "Any", "2 or more", and "Exactly 1" (selected) options. The "Type of Ring System" section has "Any", "Monocyclic", "Mixture", and "Polycyclic" (selected) options. The "Number of Carbon Atoms" section has "Any", "7 or more", "Mixture", and "less than 7" options. "OK" and "Cancel" buttons are located at the bottom right. The STN logo is in the bottom left corner, and the number "38" is in the bottom right corner.

SAM SEARCH – Why did I get that?

```
=> FILE REG
Uploading C:\CASNC\STN Express\Queries\529.str
L1      STRUCTURE UPLOADED
=> S L1
SAMPLE SEARCH INITIATED 22:27:17 FILE 'REGISTRY'
SAMPLE SCREEN SEARCH COMPLETED - 150151 TO ITERATE
0.9% PROCESSED      1000 ITERATIONS                        4 ANSWERS
INCOMPLETE SEARCH (SYSTEM LIMIT EXCEEDED)
SEARCH TIME: 00 00 01
FULL FILE PROJECTIONS:  ONLINE  **INCOMPLETE**
                        BATCH   **INCOMPLETE**
PROJECTED ITERATIONS:   2980174 TO 3025866
PROJECTED ANSWERS:     4967 TO    7045
L2              4 SEA SSS SAM L1
```

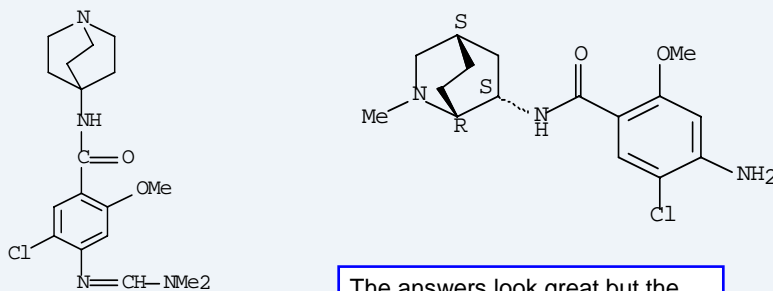
The number of iterations will exceed system limits of 1,000,000.

STN

39

SAM SEARCH

```
=> D SCAN ...
```



The answers look great but the search will not finish...

STN

40

Hitting iteration limits

- Means the number of atom-by-atom bond-by-bond examinations the structure search system has to do is too high
- Means the structure query probably needs to be restricted or simplified in some way. Some ideas:
 - Use specific atoms and/or bonds
 - Use fewer G-groups or variable groups
 - Use fewer atoms/nodes
 - Use fewer fragments
 - Limit the length of repeating groups

STN

41

Revise the query by building a more specific structure

The screenshot displays the STN software interface. The main window shows a chemical structure query with a benzamide-like structure (G₂-NH-C(=O)-C₆H₅) and four bicyclic structures labeled G₁ with atom labels a¹, a², a³, and a⁴. A 'Define Existing G-Group' dialog box is open, showing a list of G₂ groups: [1], [2], [3], and [4]. The dialog box has buttons for 'Atoms', 'Clear', 'Shotcuts', 'Variables', 'Fragments', 'G-groups', 'Save', and 'Cancel'.

STN

42

New SAM SEARCH

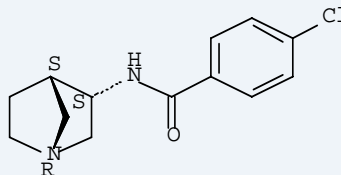
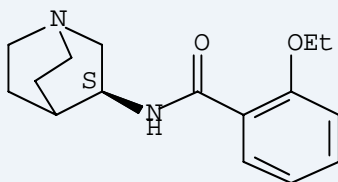
```
Uploading C:\CASNC\STN Express\Queries\529a.str
L3      STRUCTURE UPLOADED
=> S L3
SAMPLE SEARCH INITIATED 23:05:28 FILE 'REGISTRY'
SAMPLE SCREEN SEARCH COMPLETED - 938 TO ITERATE
100.0% PROCESSED 938 ITERATIONS 50 ANSWERS
INCOMPLETE SEARCH (SYSTEM LIMIT EXCEEDED)
SEARCH TIME: 00.00.01
FULL FILE PROJECTIONS: ONLINE **COMPLETE**
                        BATCH **COMPLETE**
PROJECTED ITERATIONS: 16923 TO 20597
PROJECTED ANSWERS: 1795 TO 3125
L4      50 SEA SSS SAM L3
```

STN

43

New SAM SEARCH

=> D SCAN



STN

44

Lessons learned

- There is always some way to control (wholly or partially) what you get from REGISTRY
- Having a full complement of structure searching skills will assure that you can control **precision and recall** in REGISTRY
- Recognizing why an answer was retrieved will add additional insights into how to adjust your query
- Knowing some basic rules about structure building will go a long way towards getting the right answer set from REGISTRY
- There is almost always a way to work within system limits

STN

45



STN[®]

Advanced Structure Searching:
Why did(n't) I get that?