



# Control Substitution and Ring Fusion in STNext Structure Searches

**Jan Baur and Ernst Aichinger**

**STNext**<sup>®</sup>

 **FIZ Karlsruhe**  
Leibniz Institute for Information Infrastructure

 **CAS**<sup>®</sup>  
A DIVISION OF THE  
AMERICAN CHEMICAL SOCIETY

# Agenda

- Structure databases and structure search options on STNext
- Node and bond attributes
- Opening and closing structure queries: SSS versus CSS
- Summary

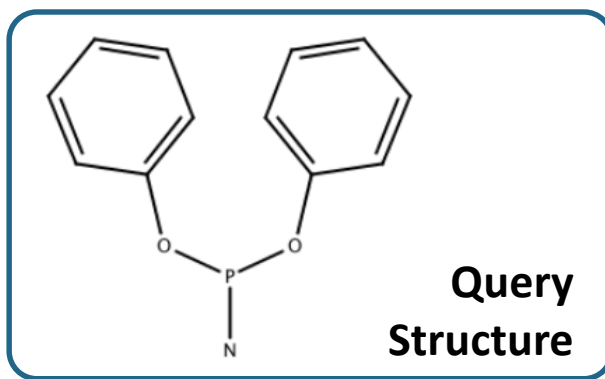


# Agenda

- **Structure databases and structure search options on STNext**
- Node and bond attributes
- Opening and closing structure queries: SSS versus CSS
- Summary



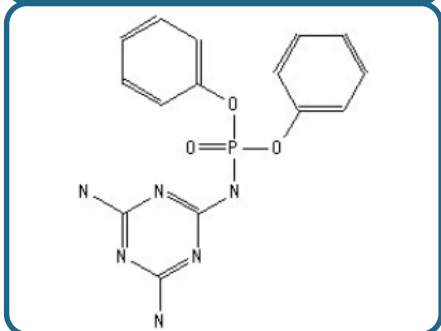
# Structure databases on STNNext



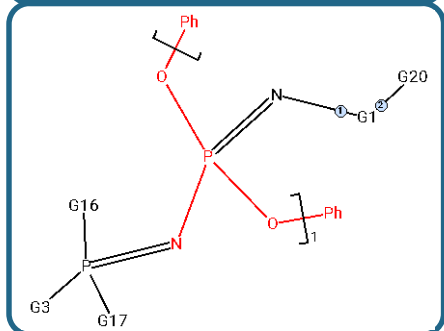
Each file has  
unique content

Adjust query structure for  
comprehensive retrieval

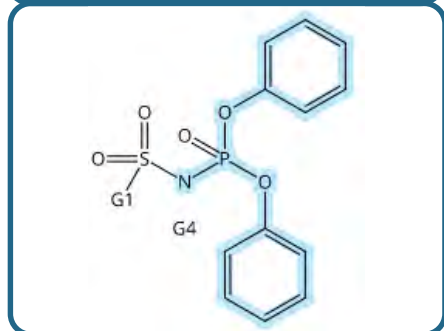
**DCR**  
≈ 3.7 M structures



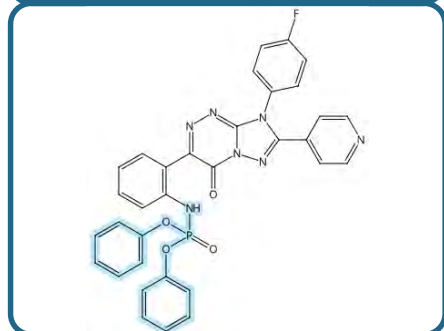
**DWPIM**  
≈ 2.3 M structures



**MARPAT**  
≈ 1.3 M structures



**CAS Registry**  
≈ 156 M structures



# Structure search options on STNnext

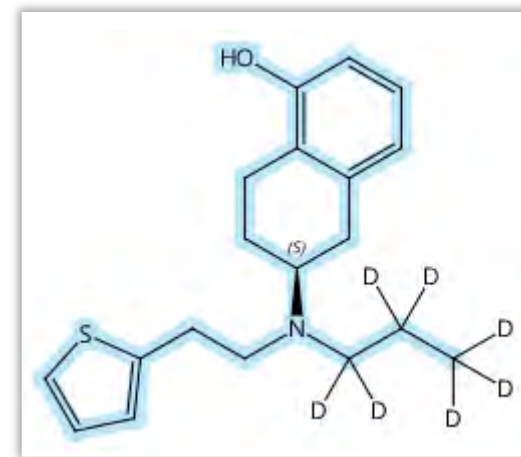
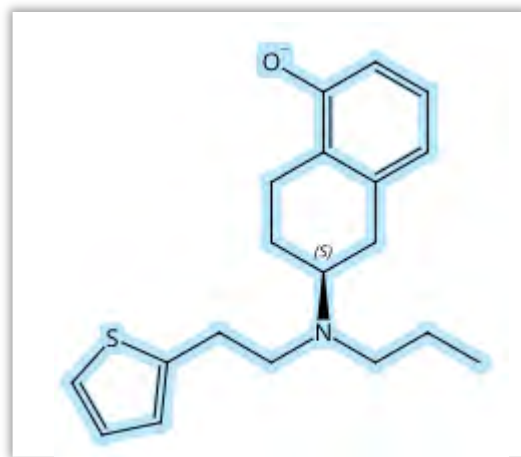
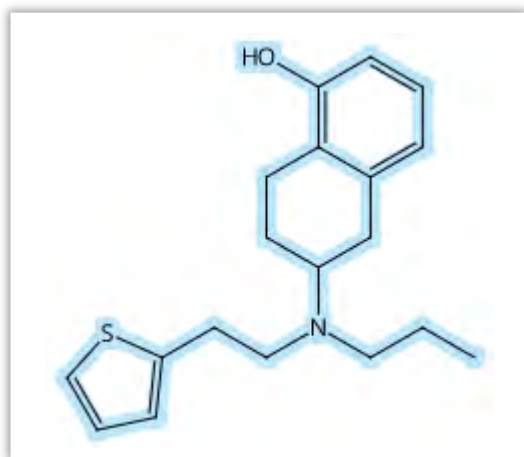
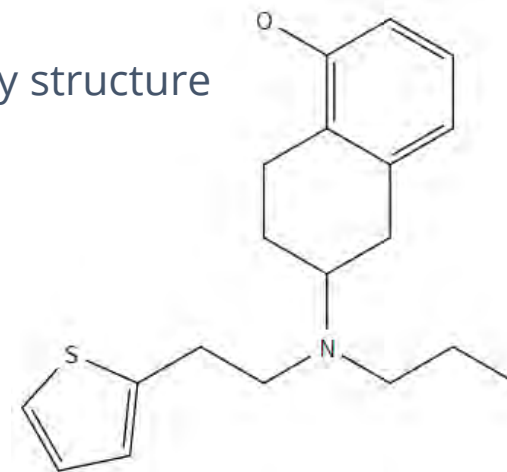
Structure Search Type	Retrieval	Use variables
EXA (exact)	Stereoisomers, charged, and isotopically labeled structures	No
FAM (family)	EXA + salts, mixtures	No
CSS (closed substructure search)	EXA + FAM + substances with no substitutions at open nodes unless the query is specified for substitution	Yes
SSS (substructure search)	EXA + FAM + CSS + analogs and derivatives of a core structure	Yes



# Structure search options on STNnext: EXACT

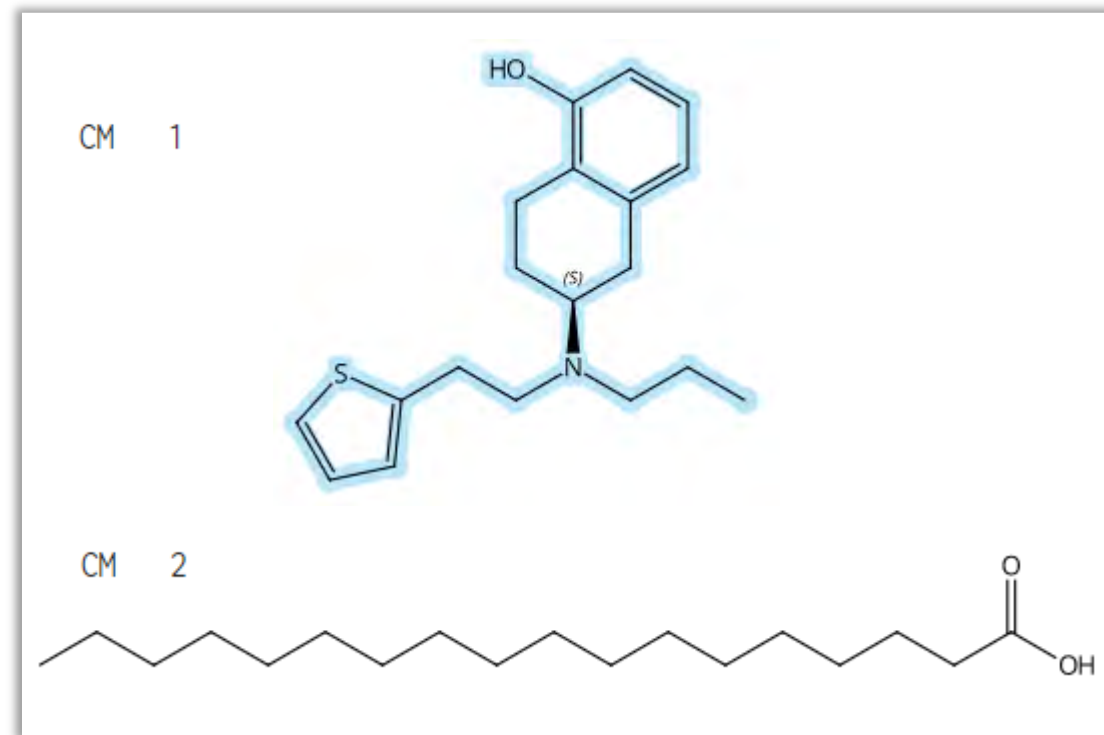
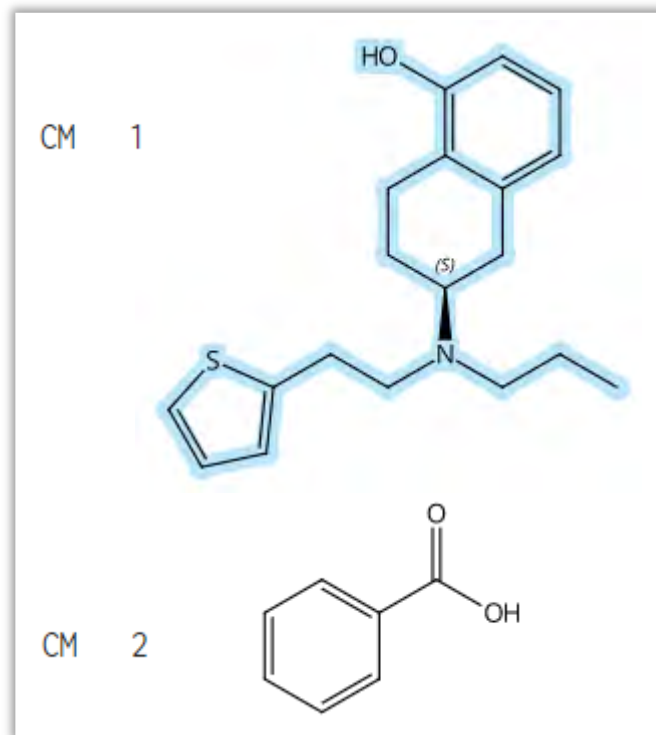
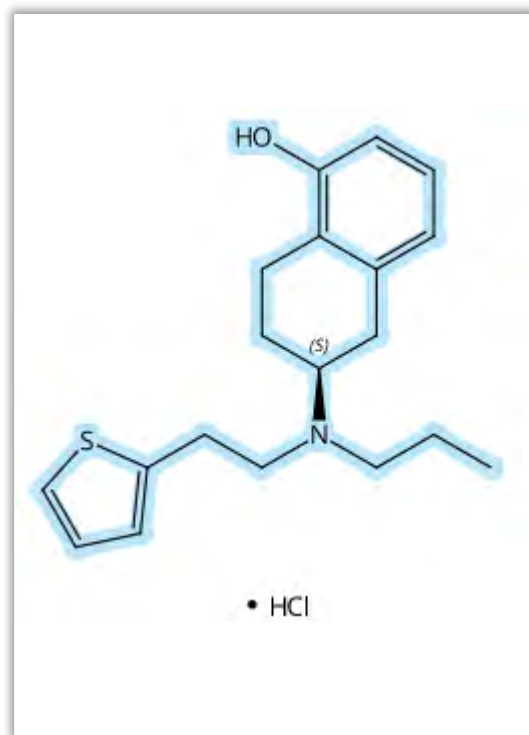
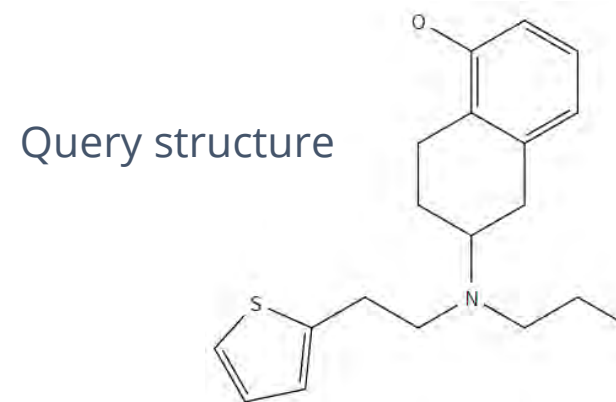
Structure Search Type	Retrieval
EXA (exact)	Stereoisomers, charged, and isotopically labeled structures

Query structure



# Structure search options on STNNext: FAMILY

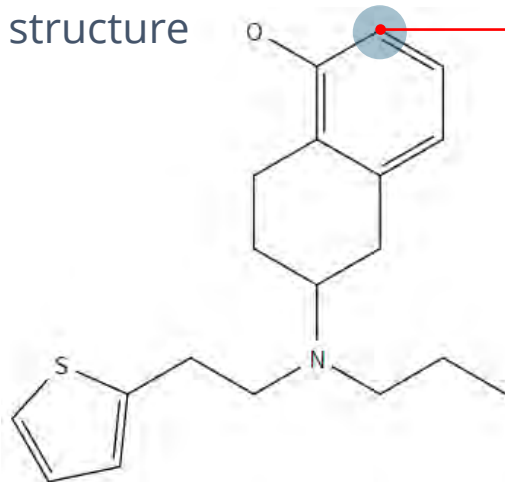
Structure Search Type	Retrieval
FAM (family)	EXA + salts, mixtures



# What is the non-hydrogen count?

- The non-hydrogen count is the number non-hydrogen atoms attached to a specific atom (node)
- Existing connections are considered, e.g. every carbon atom in a cyclohexane ring has a non-hydrogen count of 2, because it has 2 neighbouring atoms.
- You can allow for or force substitution in **CSS** or **SSS** searches by altering the non-hydrogen count of one or more atom nodes

Query structure



This carbon is connected to two other carbon atoms, so its non-H count is 2

If we set this count to **min. 2**, we will also be able to retrieve substances with a substituent at this position

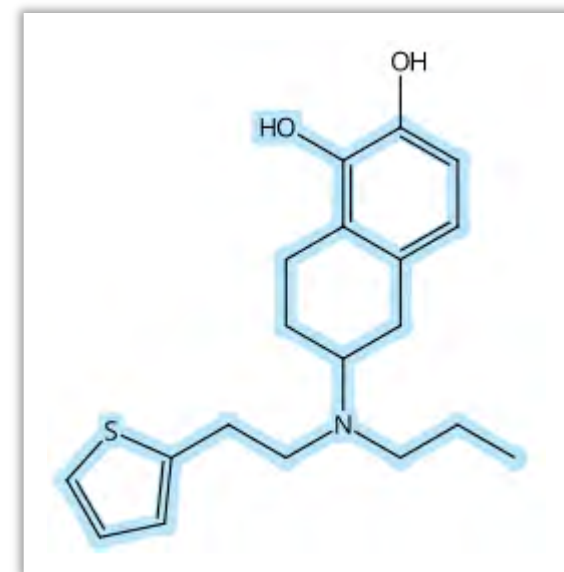
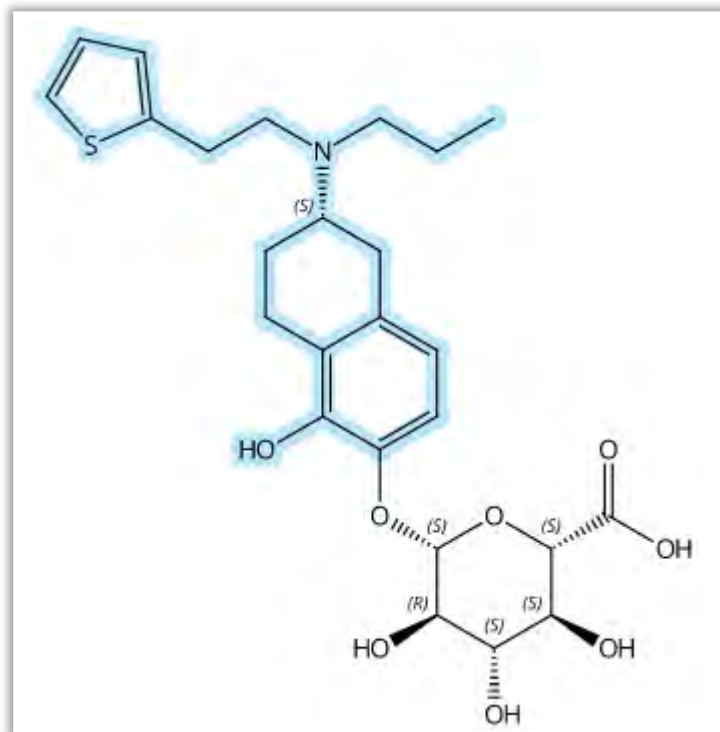
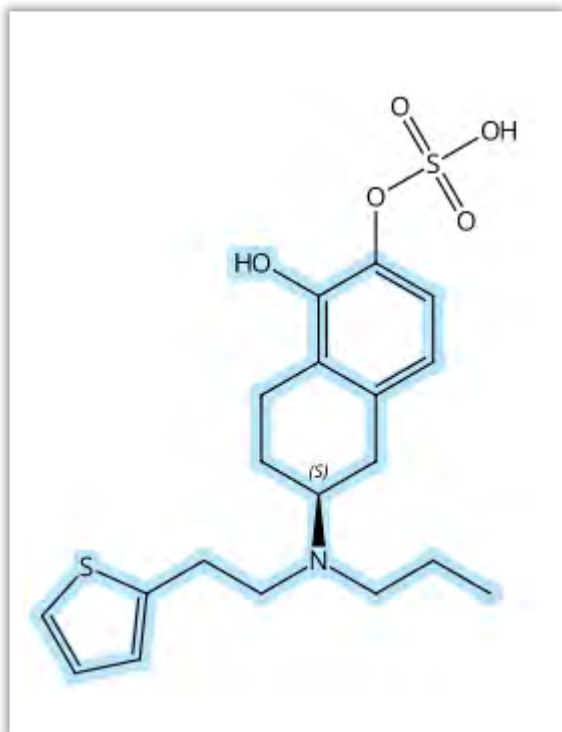
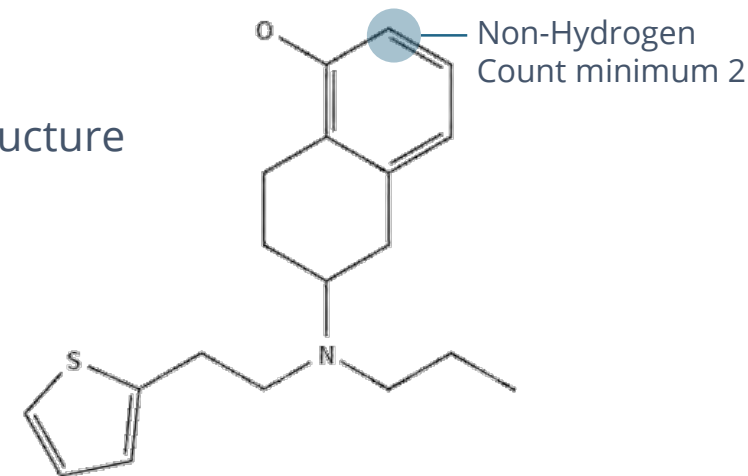
Node Attributes	
Hydrogen Count	<input type="radio"/> Any
Markush Attributes	<input checked="" type="radio"/> Specific
Mass	1. Type <input type="radio"/> Chain <input type="radio"/> Ring <input checked="" type="radio"/> Ring/Chain
Node Type	2. Count <input type="radio"/> Exact <input checked="" type="radio"/> Minimum <input type="radio"/> Maximum
<b>Non-Hydrogen Count</b>	} <input type="text" value="2"/> (0 to 16)
Valency	



# Structure search option: CLOSED SUBSTRUCTURE

Structure Search Type	Retrieval
CSS (closed sub-structure search)	EXA + FAM + substances with no substitutions at open nodes unless the query is specified for substitution

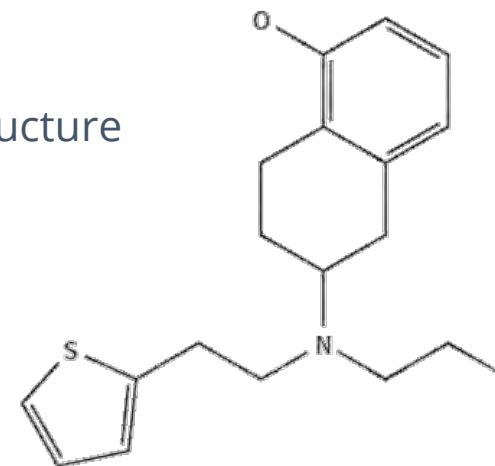
Query structure



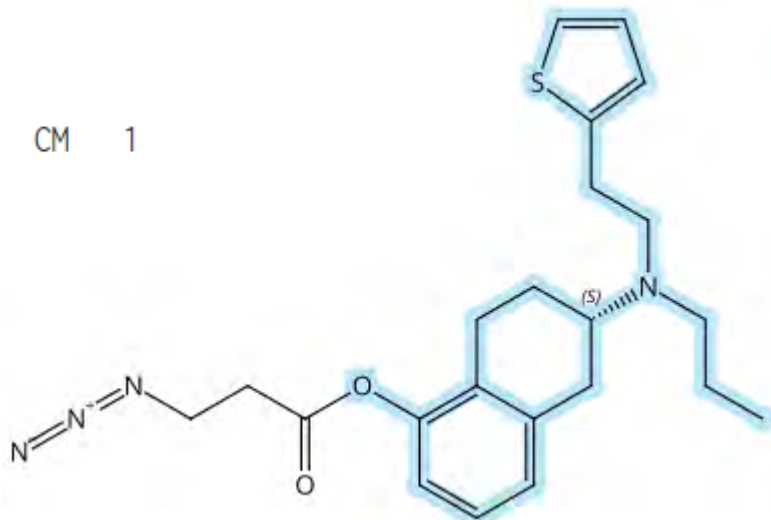
# Structure search option: SUBSTRUCTURE SEARCH

Structure Search Type	Retrieval
SSS (substructure search)	EXA + FAM + CSS + analogs and derivatives of a core structure

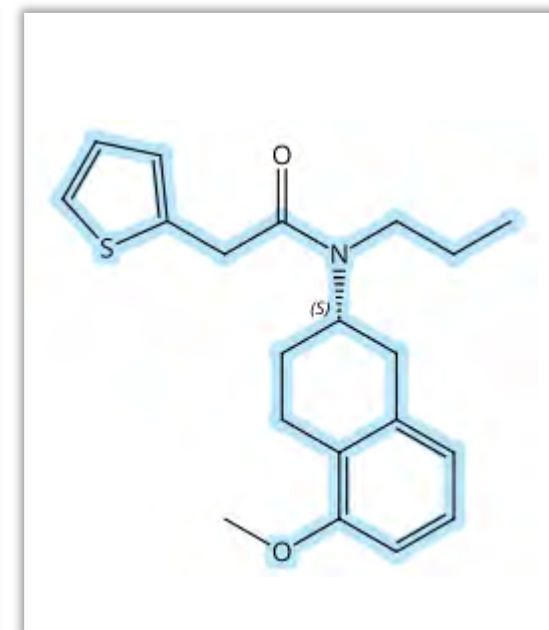
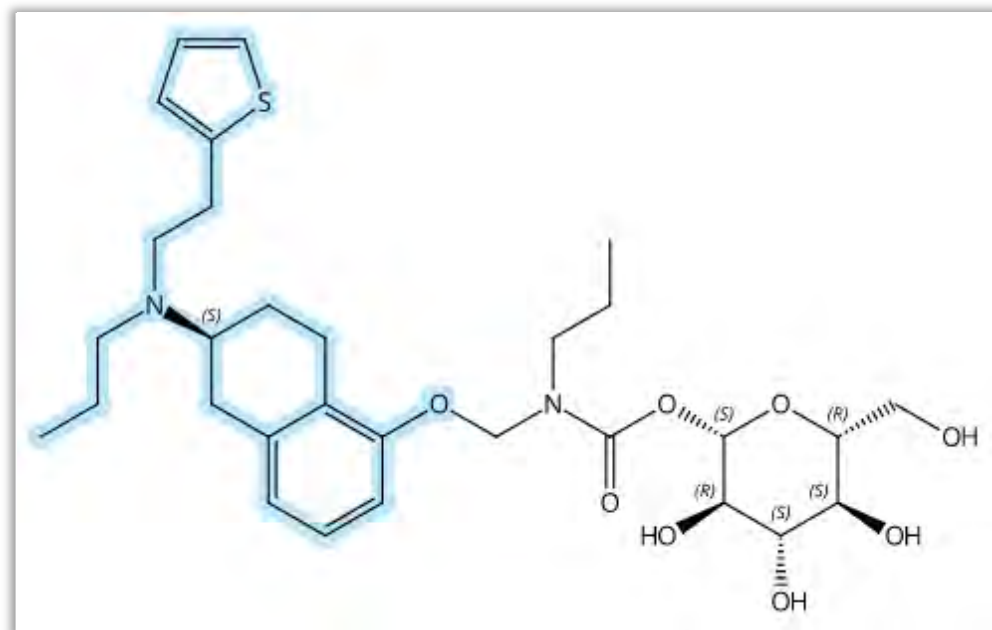
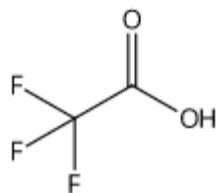
Query structure



CM 1



CM 2



# Agenda

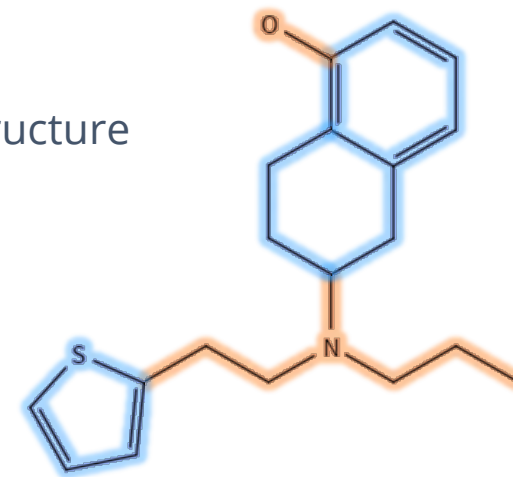
- Structure databases and structure search options on STNext
- **Node and bond attributes**
- Opening and closing structure queries: SSS versus CSS
- Summary



# Substructure search option: node attributes

Structure Search Type	Retrieval
SSS (substructure search)	EXA + FAM + CSS + analogs and derivatives of a core structure

Query structure

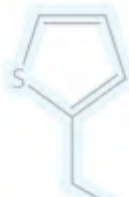


The default assumptions for atom retrieval depends on the location of an atom in the substructure query:

**Atom of a chain** will always be part of a chain.

**Atom of a ring** will always be part of a ring.

CM 1

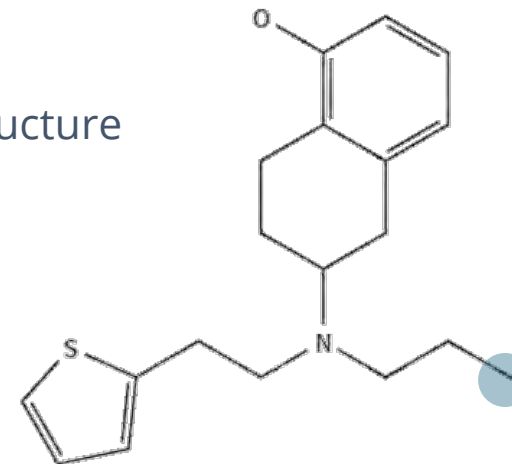


CM 2

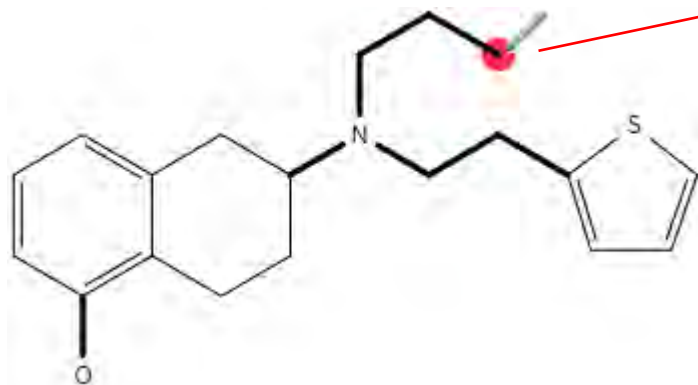


# Assign non-default attributes: Node attributes

Query structure



Structure Editor



Node Attributes

Hydrogen Count

Markush Attributes

Mass

Node Type

Non-Hydrogen Count

Valency

Chain

Ring

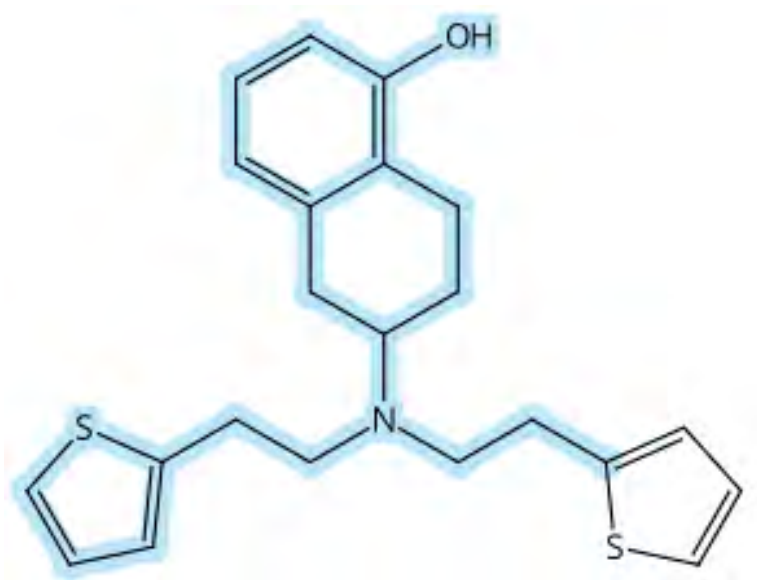
Ring/Chain

Apply

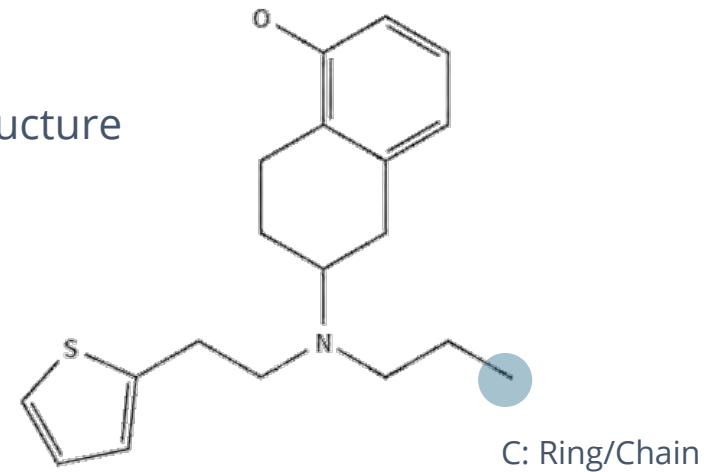
OK

Cancel

# Assign non-default attributes: Node attributes



Query structure



Additional substance records if chain atoms are set to ring/chain!

# Bonding attributes

- Each bond in a substructure query has three parameters assigned to it:
  - **Bond type**
  - **Single/double/triple/unspecified bond**
  - **Bond value**
- Default settings can be modified!

# Bonding attributes

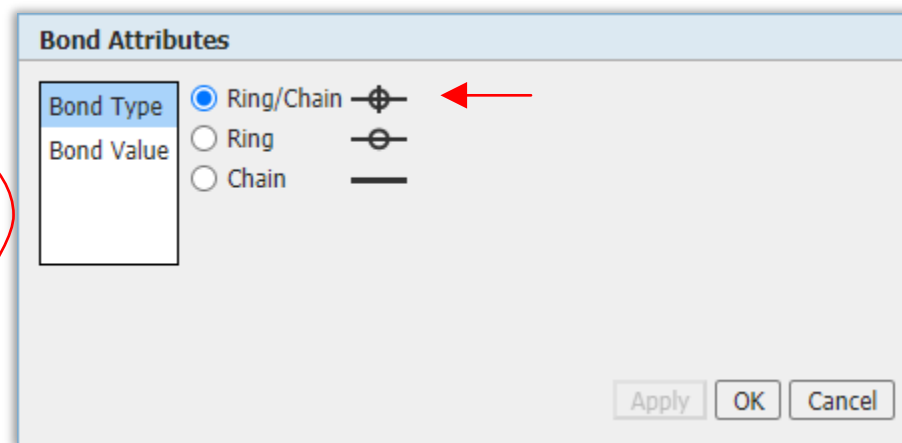
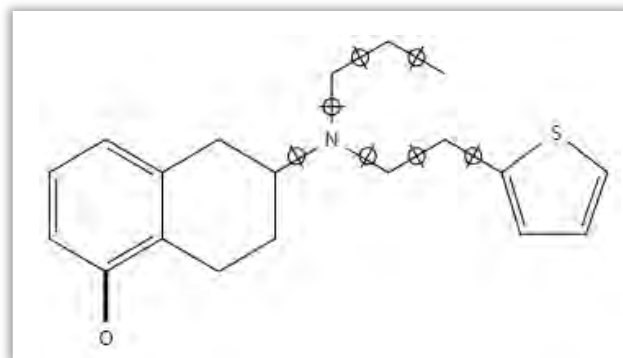
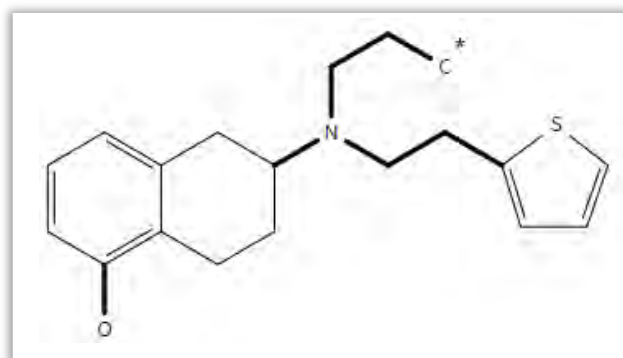
- Each bond in a substructure query has three parameters assigned to it:
  - **Bond type**
  - Single/double/triple/unspecified bond
  - Bond value

Bond type	RETRIEVED STRUCTURES
Chain	The bond will always be part of a chain
Ring	The bond will always be part of a ring
Ring/Chain	The bond will be part of a ring or chain



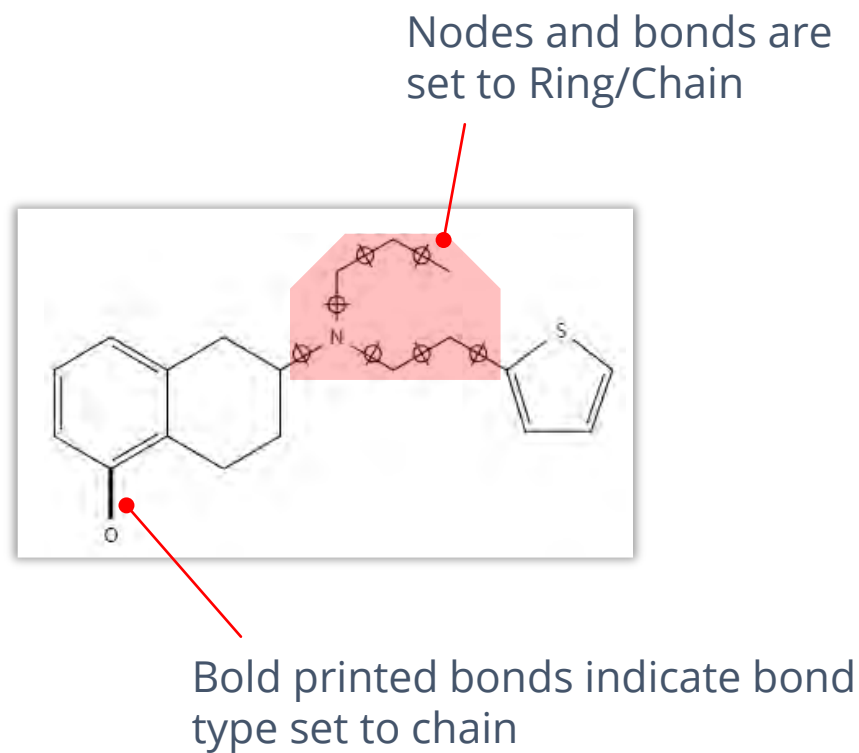
# Bonding attributes

- To allow ring formation, the bond attributes can be changed with a right mouse click on the bond to be changed (you can use the lasso or marquee tools to select a multiple bonds at once).

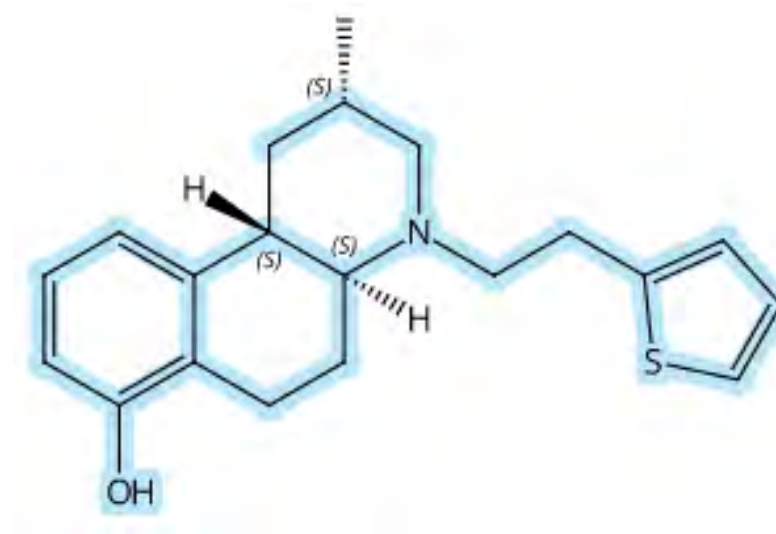


Changing the bond type to Ring/Chain also automatically changes the atom nodes at either end of the bond to Ring/Chain.

# Substructure search option: bond attributes



Additional records

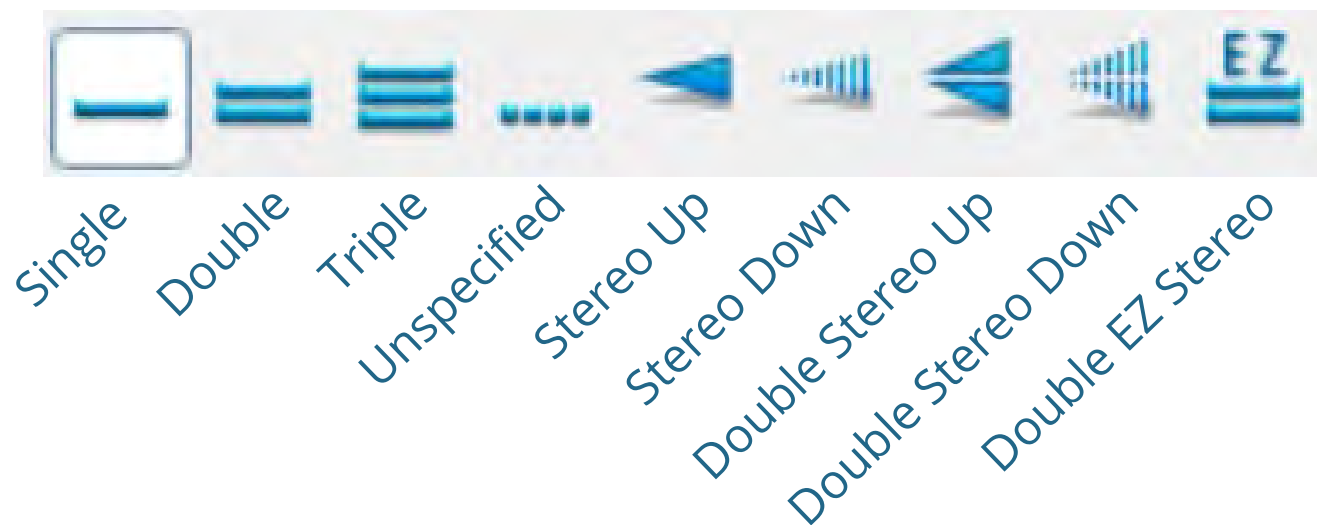


(!) Marpat does not allow for ring/chain bonds. Draw two structures, one with chain the other one with ring bonds

Check the settings and consider what might be lost or would be gained when changing bond characteristics

# Bond attributes

- Each bond in a substructure query has three parameters assigned to it:
  - Bond type
  - **Single/double/triple/unspecified bond**
  - Bond value



Use unspecified bonds whenever unsure about the bonding pattern

# Bond attributes

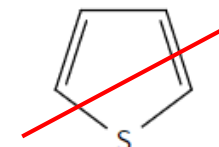
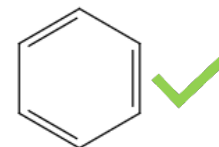
- Each bond in a substructure query has three parameters assigned to it:
  - Bond type
  - Single/double/triple/unspecified bond
  - **Bond value**  
The bond values are automatically adjusted by the structure editor to maximize retrieval of a structure query. Check the settings and consider what might have been lost or would be gained when changing bond characteristics

Bond value	RETRIEVED STRUCTURES
Exact	Exact bonds will be retrieved (either single, double or triple)
Normalized	Normalized bonds will be retrieved
Exact/Normalized	Exact or Normalized bonds will be retrieved (either single, double, triple or normalized)

# What are normalized bonds?

Certain structural moieties in the structure databases get normalized bonds assigned.  
The below definitions are true for **CAS Registry and MARPAT**:

- Normalized bonds are assigned to rings with an even number of atoms that contain alternating single and double bonds



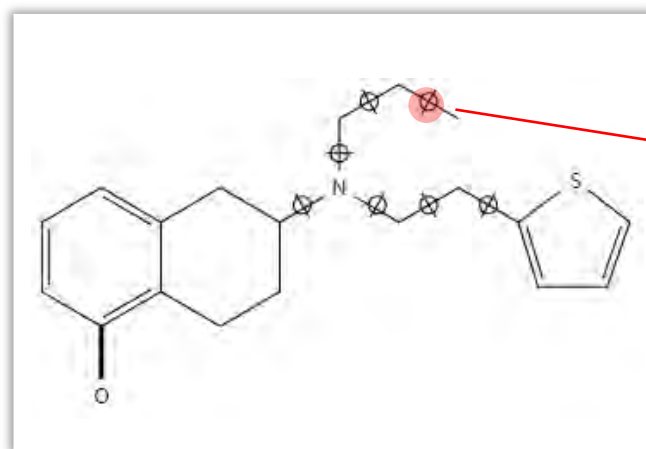
Normalized is not synonymous with aromatic!

- Tautomers:  $H(1)-(2)=(3) \leftrightarrow (1)=(2)-(3)H$  | Central atom (2) is connected to two hetero atoms (1) and (3)  
2 may be: C, N, P, As, Sb, S, Se, Te, Cl, Br, I  
1 and 3 may be: N, O, S, Se, Te
  - A double bond can be drawn between one of the hetero atoms (1 and 3) and the central atom (2)
  - A single bond can be drawn between the central atom (2) and the other hetero atom (3 or 1)
  - One of the hetero atoms (1 or 3) has at least one hydrogen, hydrogen isotope or charge
- (!) Keto-Enol tautomers are not indexed with normalized bonds, both variants should be searched

**DCR** applies similar rules as above (some exceptions, e.g. nitro group with normalized bonds in DCR).  
**DWPIM** definitions in the DWPIM manual: <https://stn.products.fiz-karlsruhe.de/sites/default/files/X/4.pdf>  
For questions please contact the STN helpdesk.

# Bonding attributes

- Each bond in a substructure query has three characteristics assigned to it:
  - Bond type
  - Single/double/triple/unspecified bond
  - **Bond value**  
The bond values are automatically adjusted to maximize the answers by a query.  
Check the settings and consider what might have been lost or would be gained when changing bond characteristics



**Bond Attributes**

Bond Type	<input checked="" type="radio"/> Current value based on structure (Exact/Normalized)
Bond Value	<input type="radio"/> Exact/Normalized
	<input type="radio"/> Exact
	<input type="radio"/> Normalized

Apply OK Cancel

Changing the bond type to Ring/Chain automatically adjusts the bond value to Exact/Normalized

# Summary: Default assumptions for nodes and bonds

- Default assumptions are made about the rings and chains in a structure query
- These assumptions determine the types of answers retrieved by a substructure search
- Default assumptions can be modified
- Changes of bonds from Chain to Ring/Chain automatically changes atoms from chain to Ring/Chain

FOR THE STRUCTURAL FEATURE IN THE SEARCH QUERY	THE DEFAULT ASSUMPTION IS TO RETRIEVE STRUCTURE MATCHES WITH THE
Ring systems	Ring system as drawn Ring system as part of a larger ring system
Atom in a chain	Atom in a chain only
Bond in a chain	Bond in a chain only

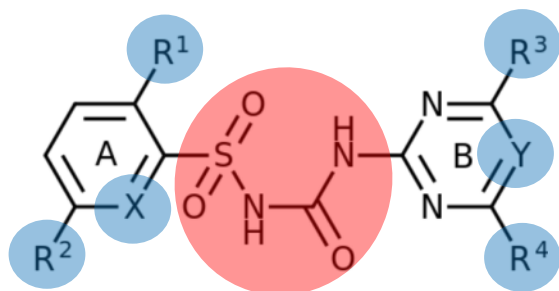
# Agenda

- Structure databases and structure search options on STNext
- Node and bond attributes
- **Opening and closing structure queries: SSS versus CSS**
- Summary





# Close a structure query: From SSS to CSS



Formula of active sulfonylurea herbicides, showing the sulfonylurea backbone itself in red and the side chains that distinguish each compound in blue (X, Y = CH or N).

- Sulfonylureas are a class of organic compounds used in medicine and agriculture
- As herbicides they mostly act by interfering with plant biosynthesis of certain amino acids (e.g. valine, isoleucine, and leucine)

Search Question:

Explore sulfonylureas derived from the above structure  
Substitutions on ring A are only allowed at position R1



Solution 1:

Perform SSS, 4 nodes on ring A are locked

Solution 2:

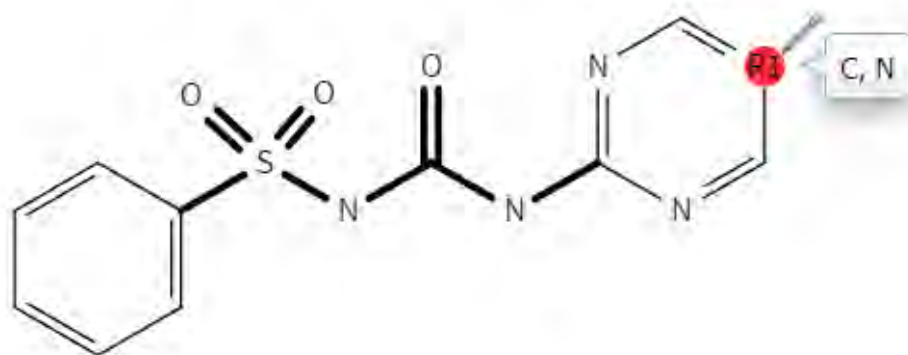
Perform CSS, open specific sites to allow for substitution

# Options to prevent substitution

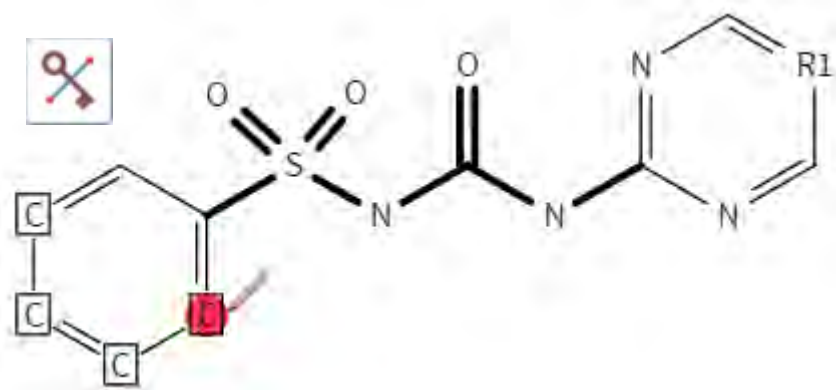
- Use **Lock Atoms**  to block substitution in a **SSS search**
- Use appropriate Shortcuts  , they often have implicit hydrogens and block substitution in an SSS search
- Modify the Hydrogen count attribute
- Modify the **Non-Hydrogen Count** attribute
- Perform a **CSS search** and open specific atom nodes only

# Use lock atoms to block substitution | SSS

1



2



Use lock atoms to block any substituents except hydrogen

## Node Attributes

Hydrogen Count  
Markush Attributes  
Mass  
Node Type  
Non-Hydrogen Count  
Valency

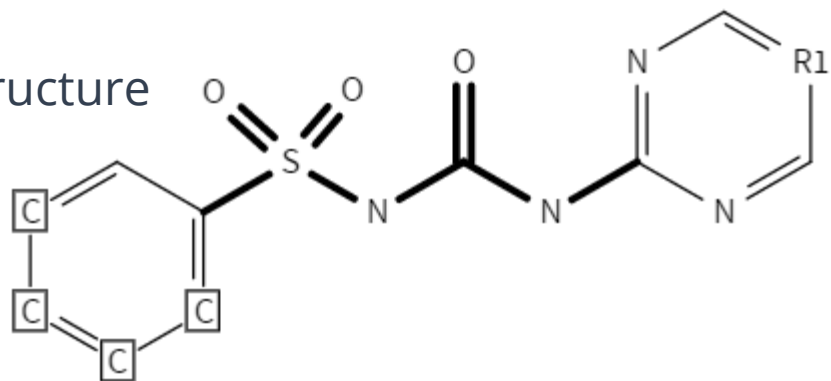
Any  
Specific  
1. Type  
2. Count  
 Chain  Exact  
 Ring  Minimum  
 Ring/Chain  Maximum  
0 (0 to 16)

Non-Hydrogen count attribute is not available if one or more nodes are locked.

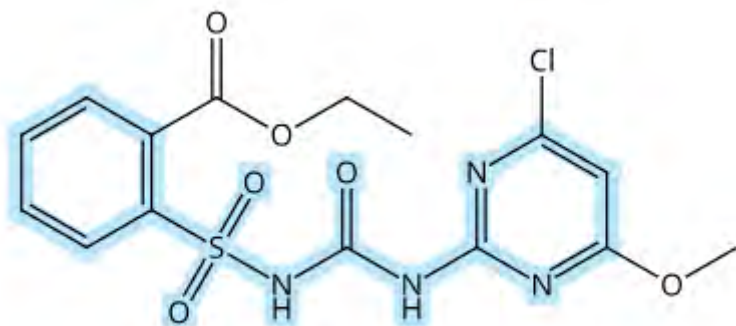
**Lock atoms** automatically sets a non-hydrogen count (exact=2 in this case), so this parameter is not available anymore

# Adjust attributes for enhanced retrieval | SSS

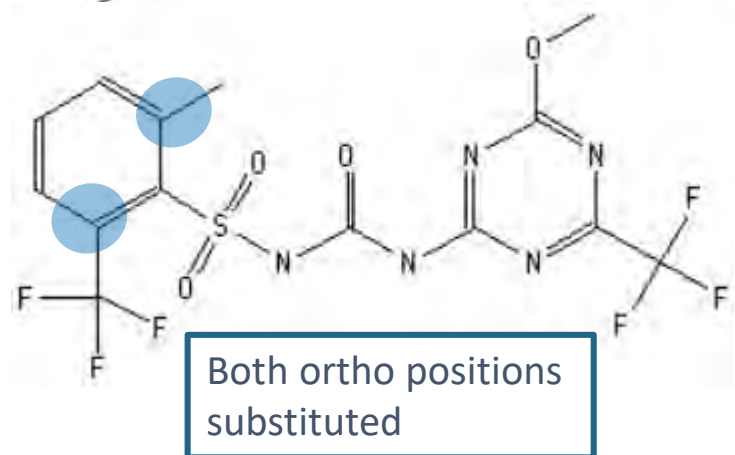
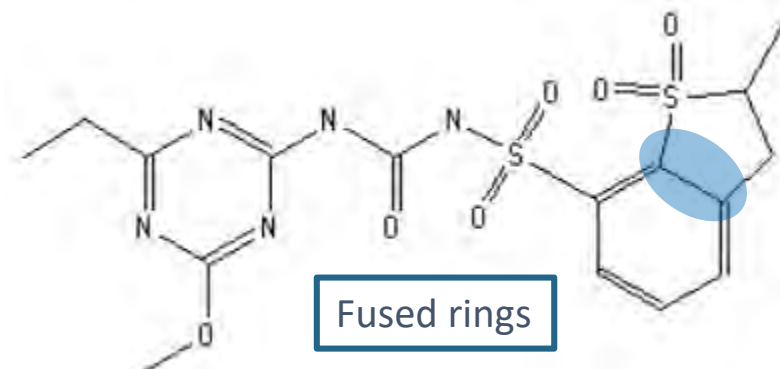
Query structure



Hits

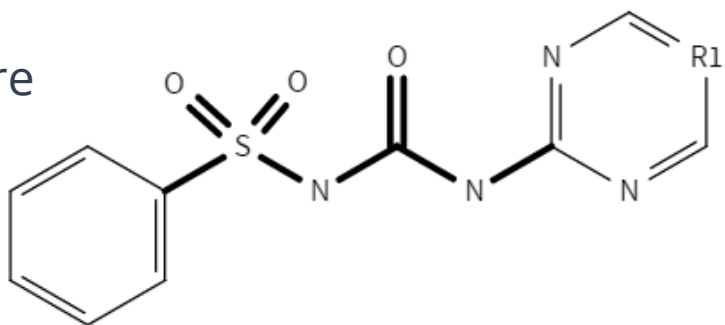


Additional records if the 4 nodes are not blocked:

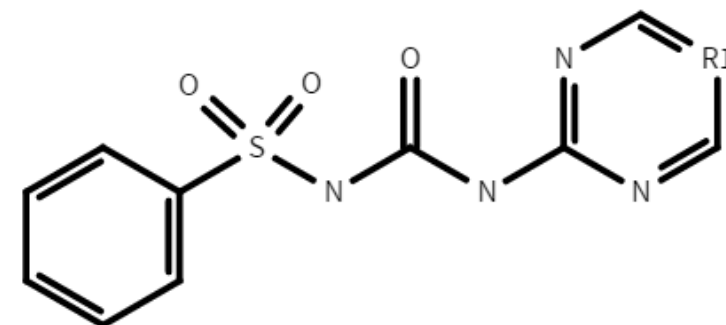


# Ring Lock prevents ring fusion or ring formation | SSS

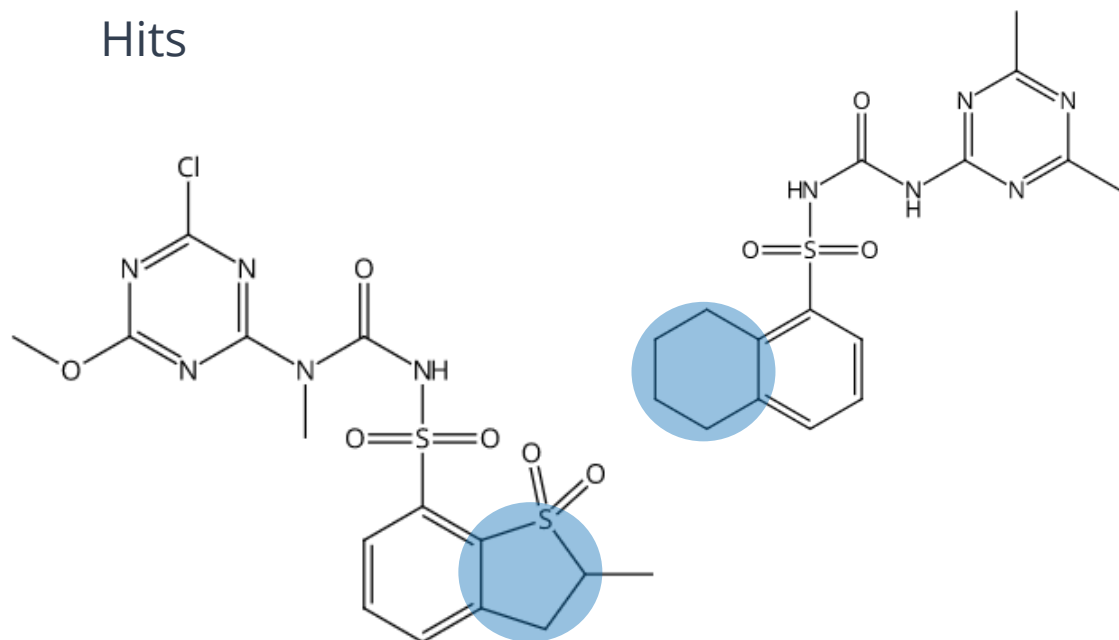
Query structure



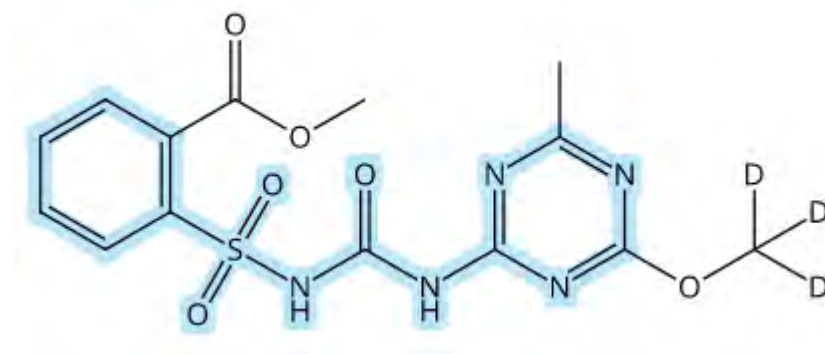
Query structure with ring lock:



Hits

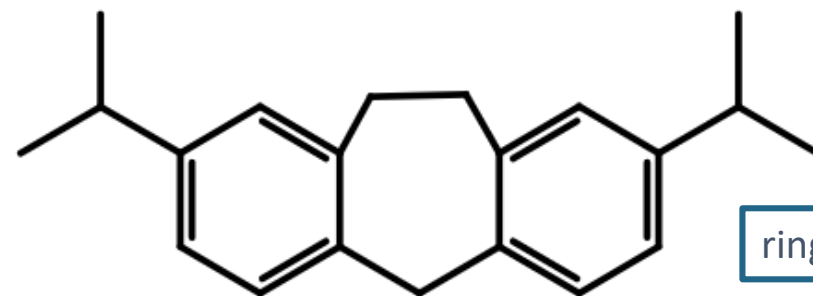
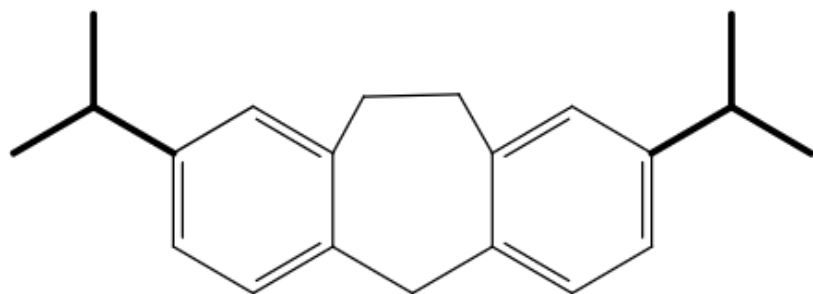


Hits:

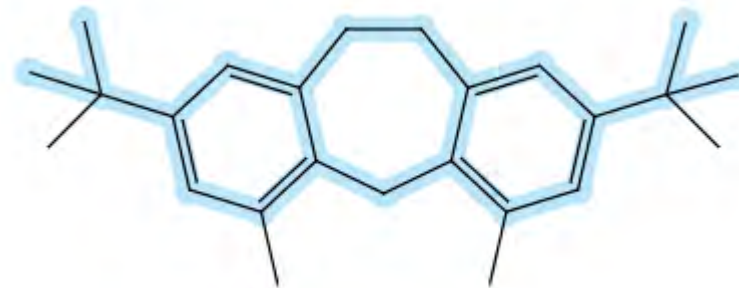
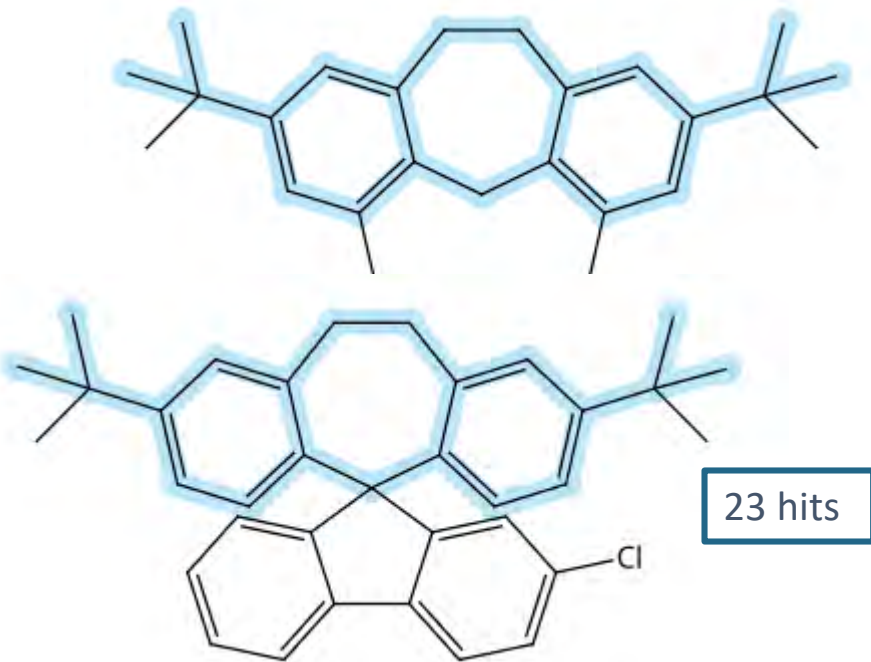


# Ring lock also blocks spiro systems | SSS

Query structure

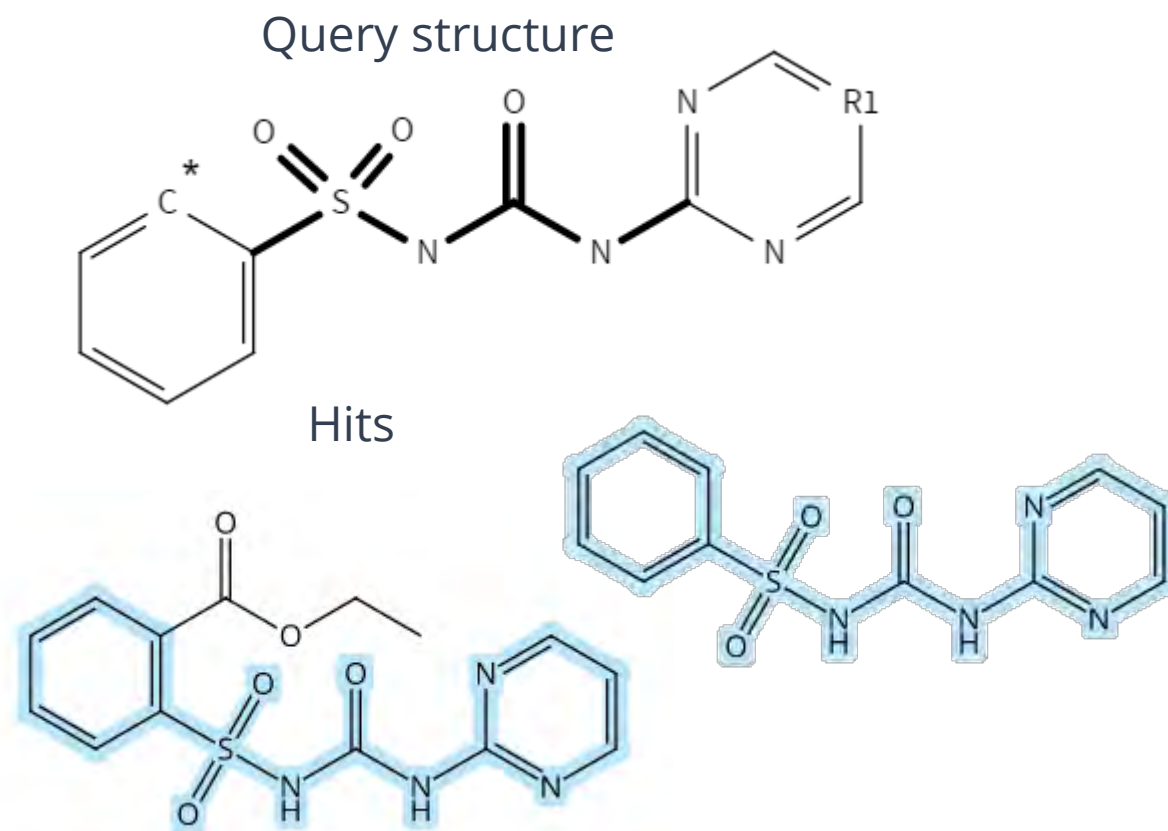
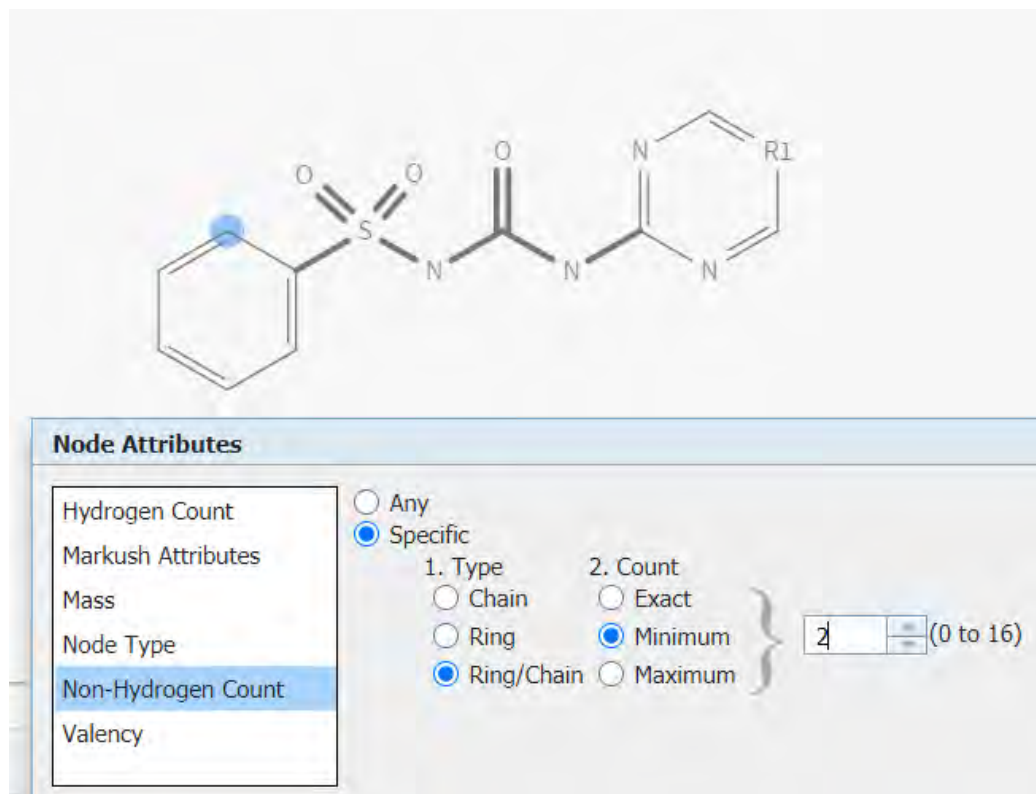


Hits



# Specify open positions with CSS

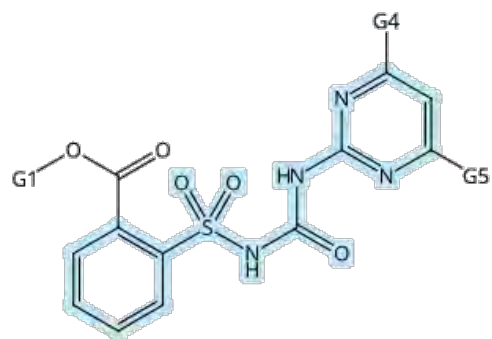
- A CSS by default blocks all non-hydrogen substituents but allows to use variables such as user-defined R-groups or STN variables. This query uses user-defined  $R1=\{C,N\}$
- If substitution is wanted, you must set it in the editor



# CSS is the first best option for Markush searches

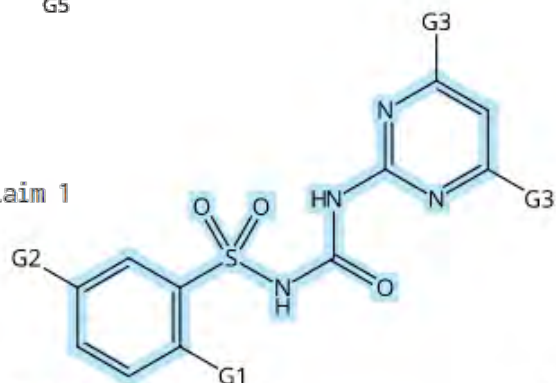
- Consider using CSS in Markush searches
- Results will be very close to the drawn structure
- You can still use variables
- Define open positions if desired

MSTR 1 Assembled



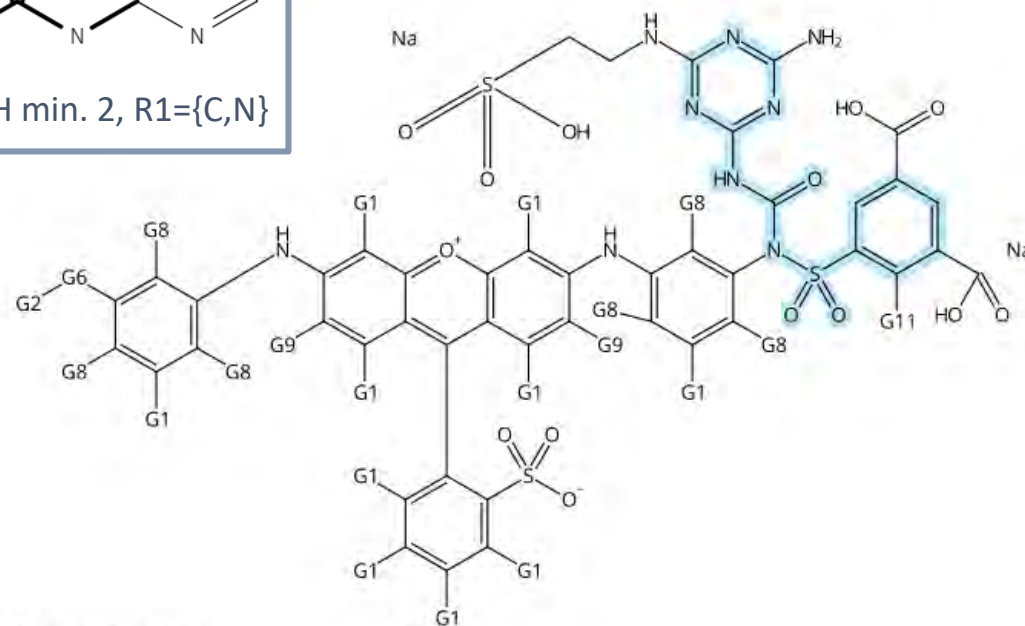
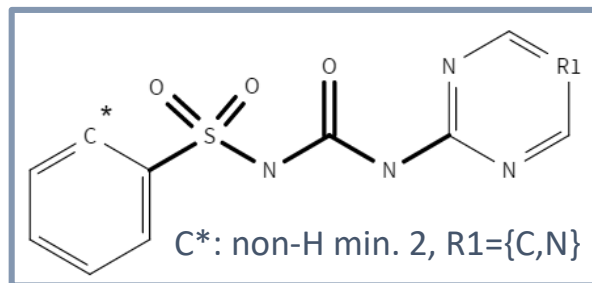
Patent location:

claim 1



Patent location:

claim 1



Patent location:

claim 1

CSS search in Marpat: 107 patents

SSS search in Marpat: 523 patents



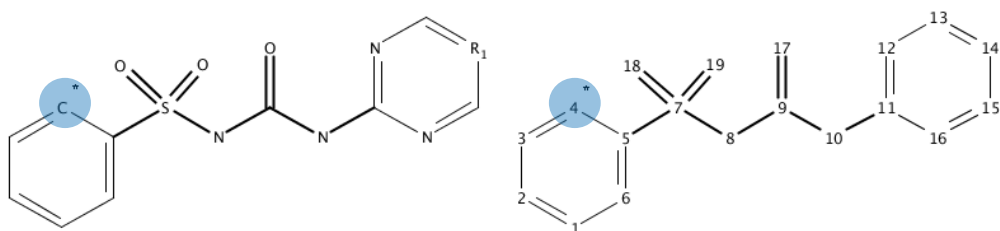
# Where to check structural parameters?

All attributes are displayed after uploading the structure

Use the Attribute Values Panel to check structural attributes when editing the structure

=>

Uploading structure file: 2020\_0144\_Structure



R-Group Definitions

R1: C,N

Node Attributes

Ring Nodes : 1 2 3 4 5 6 11 12 13 15 16

Chain Nodes : 7 8 9 10 17 18 19

Non H Count

Ring/Chain, Minimum (2) : 4

Bond Attributes

Ring Bonds : 1-2 2-3 3-4 4-5 5-6 6-1 11-12 12-13 13-14 14-15 15-16 16-11

Chain Bonds : 5-7 7-8 7-18 7-19 8-9 9-10 9-17 10-11

Normalized Bonds : 1-2 2-3 3-4 4-5 5-6 6-1

Exact/Normalized Bonds : 5-7 7-8 7-18 7-19 8-9 9-10 9-17 10-11 11-12 12-13 13-14 14-15 15-16 16-11

Markush Attributes

Match Level (ATOM) : 1 2 3 4 5 6 11 12 13 15 16

Match Level (CLASS) : 7 8 9 10 17 18 19

Element Count Level (LIMITED) : 1 2 3 4 5 6 7 8 9 10 11 12 13 15 16 17 18 19

# Summary

- STNext offers precision tools to adjust substitution in structure queries
- Ring and bond parameters will influence your structure search results
- Use lock atoms and lock rings to block substitution and prevent ring formation/fusion in substructure searches
- Keep in mind that structure shortcuts (e.g. Me) might block substitution
- Use predefined or self-defined variables in SSS and CSS searches
- Closed substructure searches will block substitution unless specified otherwise
- Make use of the non-hydrogen count to precisely define substitution patterns and selectively open positions in CSS searches

# Contact Us



**CAS** [help@cas.org](mailto:help@cas.org)  
[www.cas.org](http://www.cas.org)

**FIZ Karlsruhe**

[helpdesk@fiz-karlsruhe.de](mailto:helpdesk@fiz-karlsruhe.de)  
[www.stn-international.de](http://www.stn-international.de)

**STNnext**<sup>®</sup>

 **FIZ Karlsruhe**  
Leibniz Institute for Information Infrastructure

 **CAS**<sup>®</sup>  
A DIVISION OF THE  
AMERICAN CHEMICAL SOCIETY