

Scoring Synthetic Accessibility: A very different problem

A Peter Johnson, K. Boda, G.J. Myatt and J.C. Baber
University of Leeds

Problem Specification

De novo design programs such as SPROUT can suggest large sets of entirely novel potential leads

Powerful heuristics are necessary to evaluate (and reduce) large answer sets

Binding Score

Eliminate candidates with poor estimated binding affinity

Synthetic Feasibility Analysis

complex molecular structures

For de novo design prediction of synthetic accessibility is equally important

Hypothetical ligands, including those predicted to bind very strongly, have no practical value unless they can be readily synthesised.

Our Attempts to Provide Solutions:

- ❑ **CAESA** (estimates synthetic accessibility)
- ❑ **Complexity Analysis** (estimates structural complexity and drug-likeness)
- ❑ **SynSPROUT** (avoids the problem by building constraints into the structure generation process)

CAESA

Computer Assisted Estimation of
Synthetic Accessibility

- Glenn Myatt
- Jon Baber

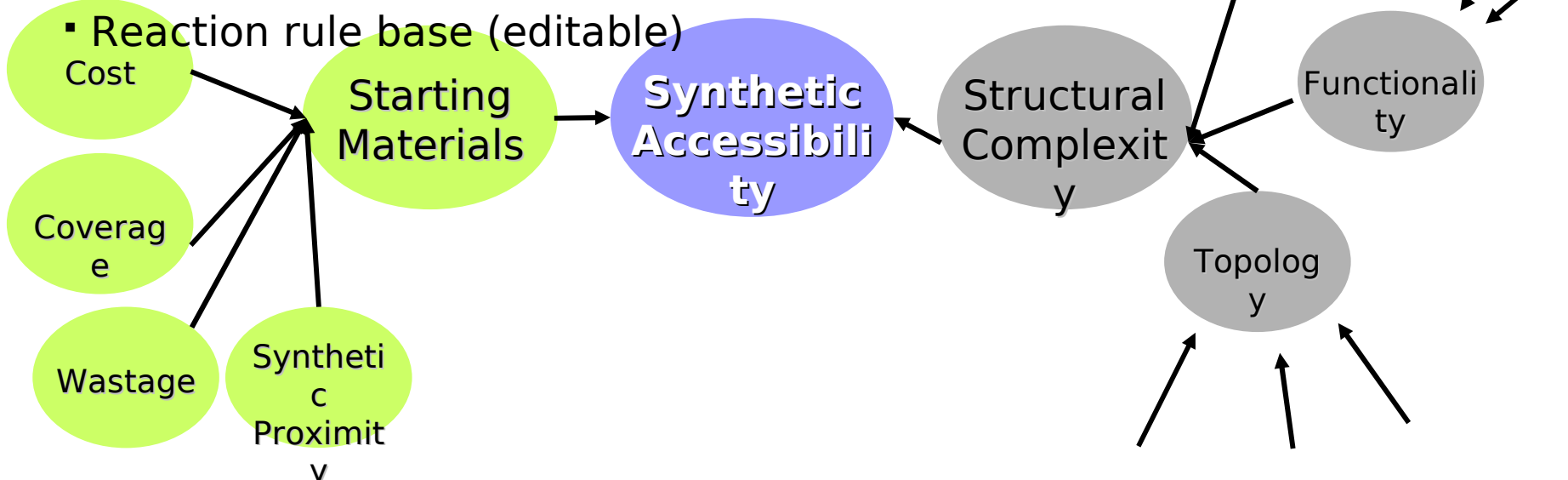
Goals of CAESA Project

- Clear need for automated method of ranking hypothetical compounds according to perceived ease of synthesis
- Good synthetic chemists can do this job themselves on small number of compounds but are unwilling to do it for hundreds or thousands of compounds
- CAESA attempts to do the same job but never gets bored!

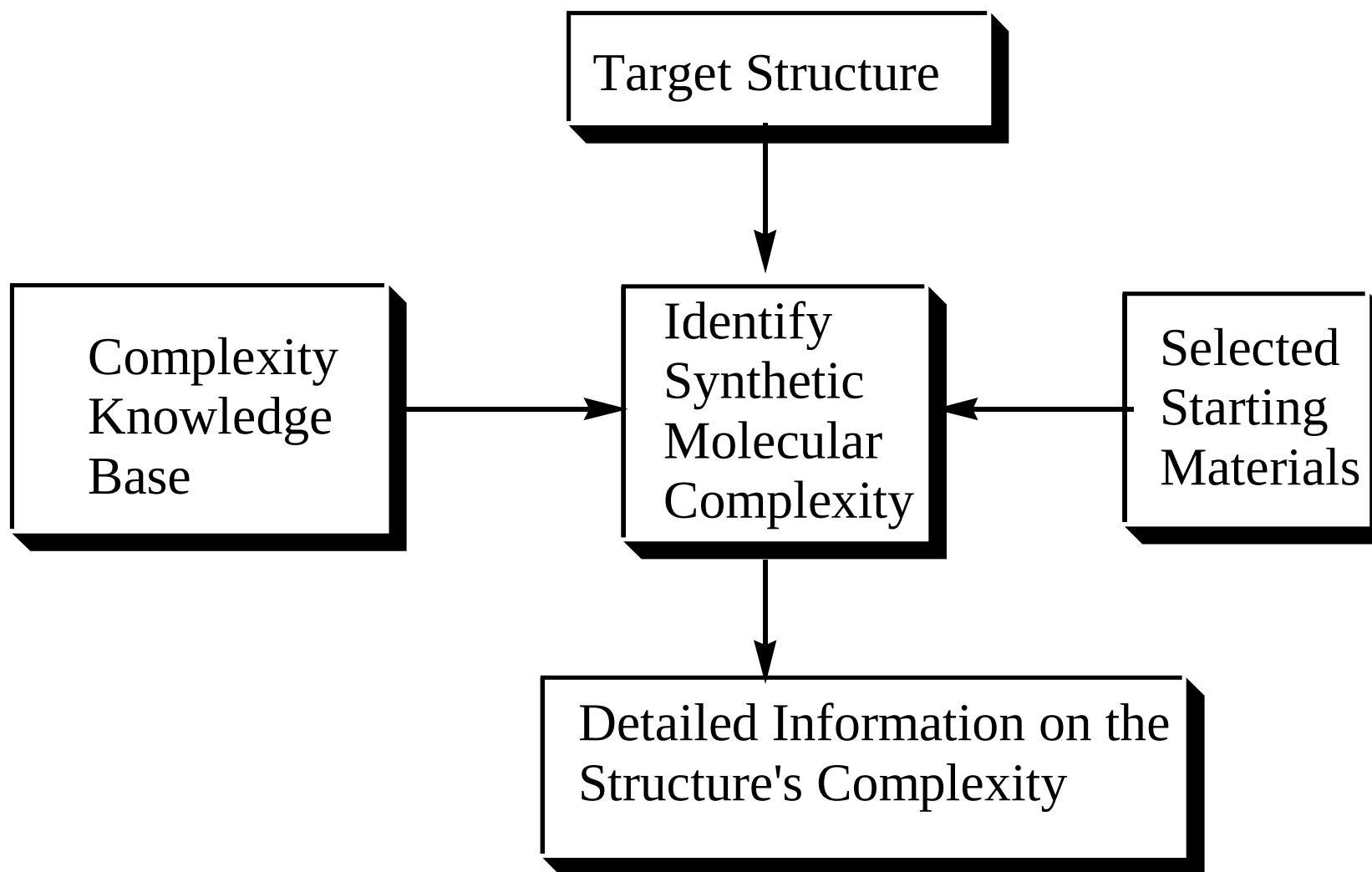
CAESA Overview

Overview of CAESA

- Estimate of structural complexity
- Stereochemical complexity
 - Complex topological features (fusions etc.)
 - Functional group starting materials
- Availability of good materials
- Rapid retrosynthetic analysis
 - Database of commercially available materials



CAESA Components

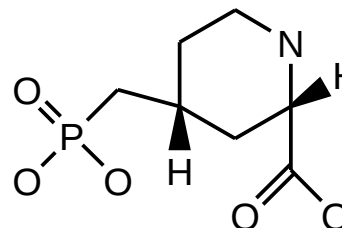


Perception of Structural Features which Reduce Synthetic Accessibility

- ❑ Synthetic chemists can identify such features just by inspection of the structure
- ❑ A computer program can be trained to emulate this process
- ❑ The rules governing the recognition should be stored in one or more (Text) knowledge bases which are separate from the program and are easily editable by chemists
- ❑ The same type of knowledge base drives recognition of all important structural features (rings, aromaticity, functional groups etc)
- ❑ All these rules are coded in a special form of chemical English called PATRAN

Example of recognition of complex structural features

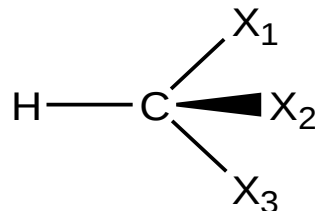
Target Structure



Not just the presence of complex features but also the environment of each one contributes to overall complexity score

<i>Factors of Importance</i>	<i>Explanation of Complex Fragments</i>	<i>Heuristic Values</i>
Stereocentre	Tertiary Stereocentre	10
	Secondary Stereocentre	<u>8</u>
		<u>18</u>
Isolated	Isolated Stereocentre	7
	Isolated Stereocentre	<u>7</u>
		<u>14</u>
Substitution Patterns	Cyclic appendage with no heteroatoms close on the ring	10
	Cyclic appendage with heteroatom alpha	<u>5</u>
		<u>15</u>

Isolated stereocentre identification



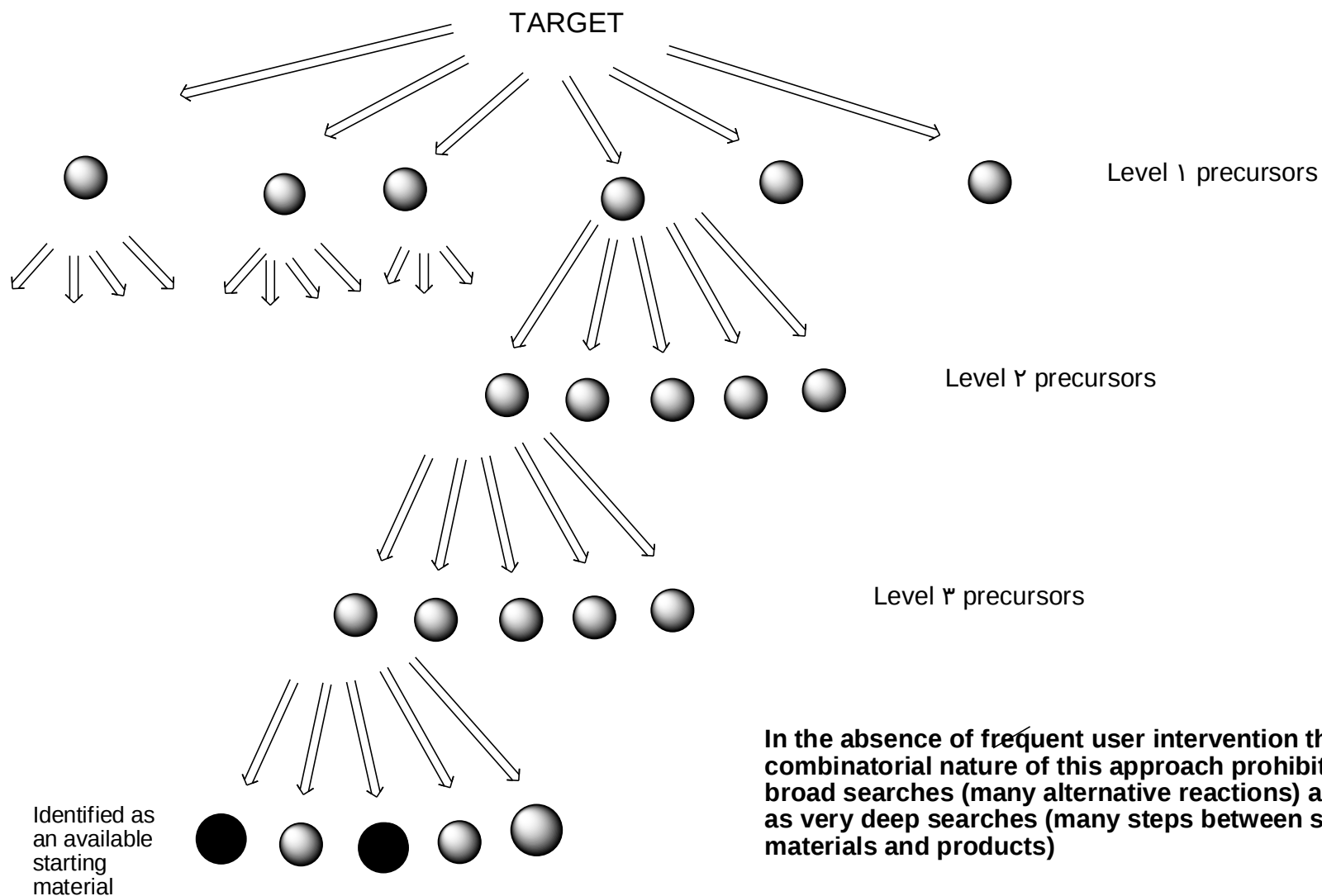
```
CHEMICAL-LABEL <Isolated Tertiary Stereocentre>  
...STARTP  
...C[STEREO=YES];[HS=1](-X[STEREO=NO])(-X[STEREO=NO])(-X[STEREO=NO])  
...{1,2,3,4}{1,2,4,3}{1,3,2,4}{1,3,4,2}{1,4,2,3}{1,4,3,2}  
...ENDP
```

Automatic Selection of Starting Materials

Starting Materials and Synthetic Accessibility

- § Availability of suitable starting materials very important factor - good starting materials can dramatically reduce the difficulty of synthesising a compound.
- § Good starting materials for part of the target molecule means the analysis of structural synthetic difficulty or complexity can be directed to just those portions of the target molecule that cannot be made from available starting materials
- § Finding good starting materials through **retrosynthetic analysis** also provides possible synthetic routes as a byproduct

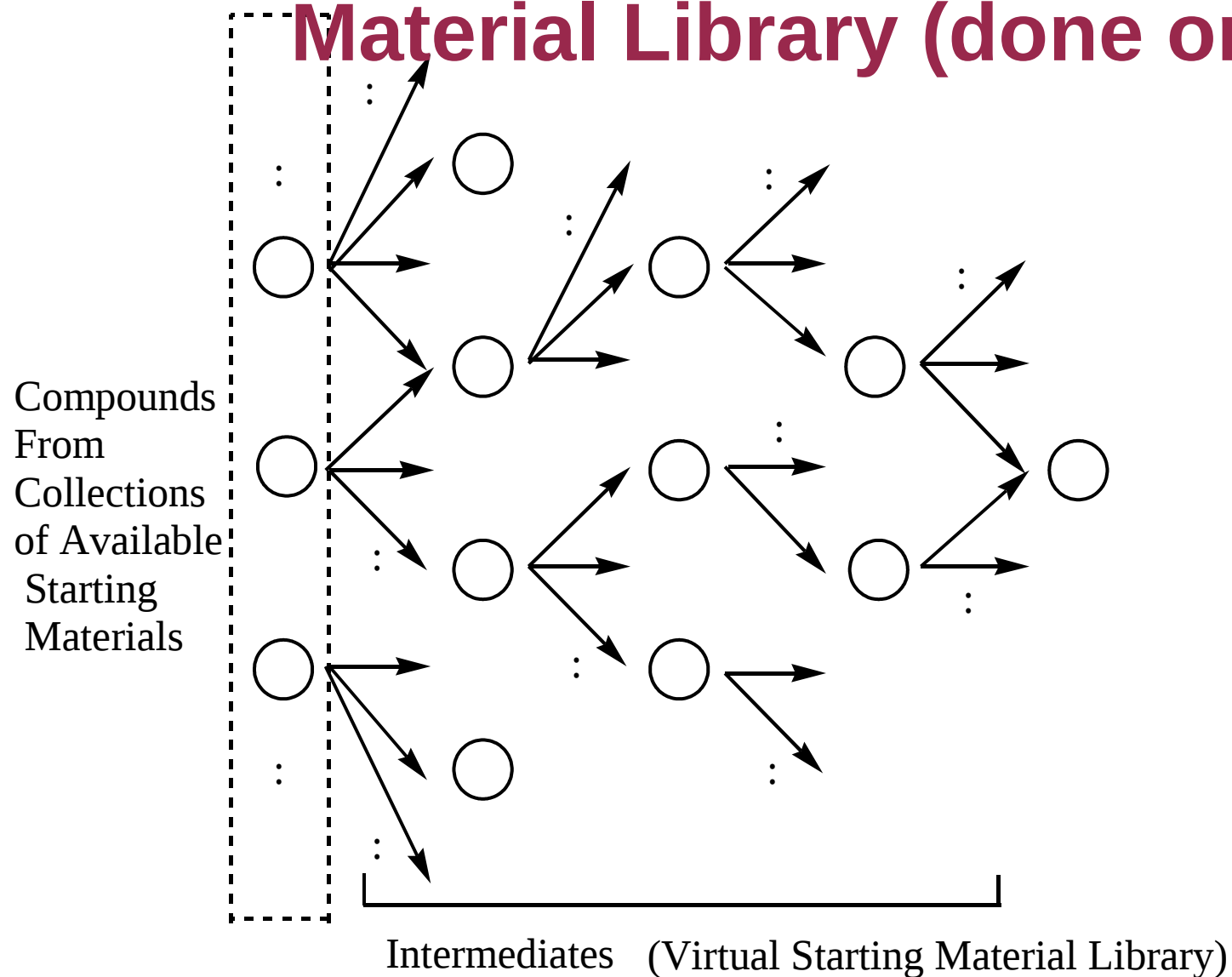
Traditional Retrosynthetic Analysis



Bidirectional search for synthetic routes

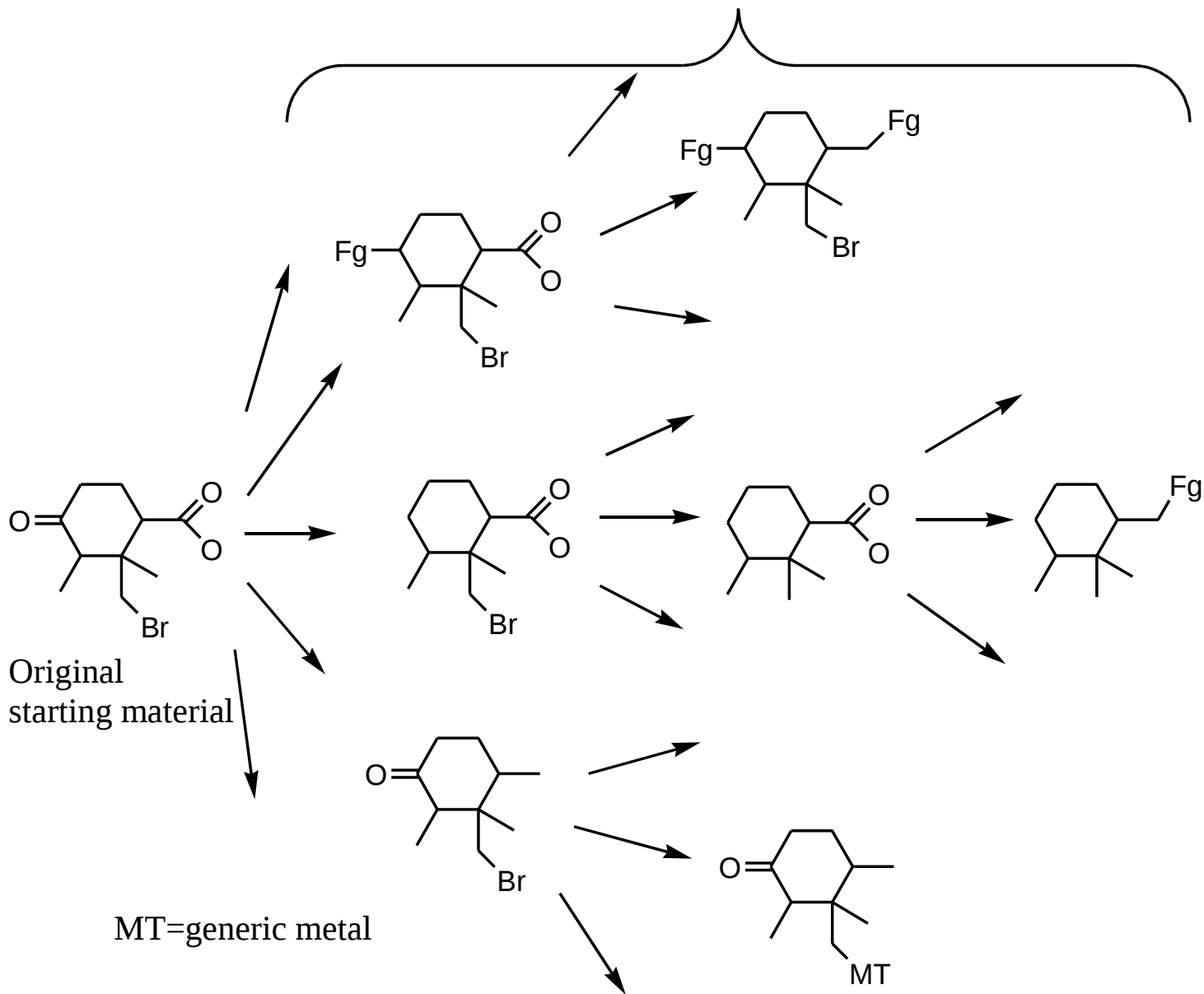
- ❑ The CAESA approach includes an opportunistic **synthetic analysis** of all the compounds in the starting materials databases, which is only performed once. The structures generated from this synthetic analysis are stored in a relational database of virtual starting materials.
- ❑ For each target molecule, a focused **retrosynthetic analysis** is performed on-line.
- ❑ Once a precursor is generated retrosynthetically, a quick direct access search of the database of intermediates is performed to verify if the precursor has been generated synthetically.
- ❑ If it has, then there exists a synthetic route and the compound from which the intermediate was generated is selected as a potential starting material.

Generation of Virtual Starting Material Library (done once)

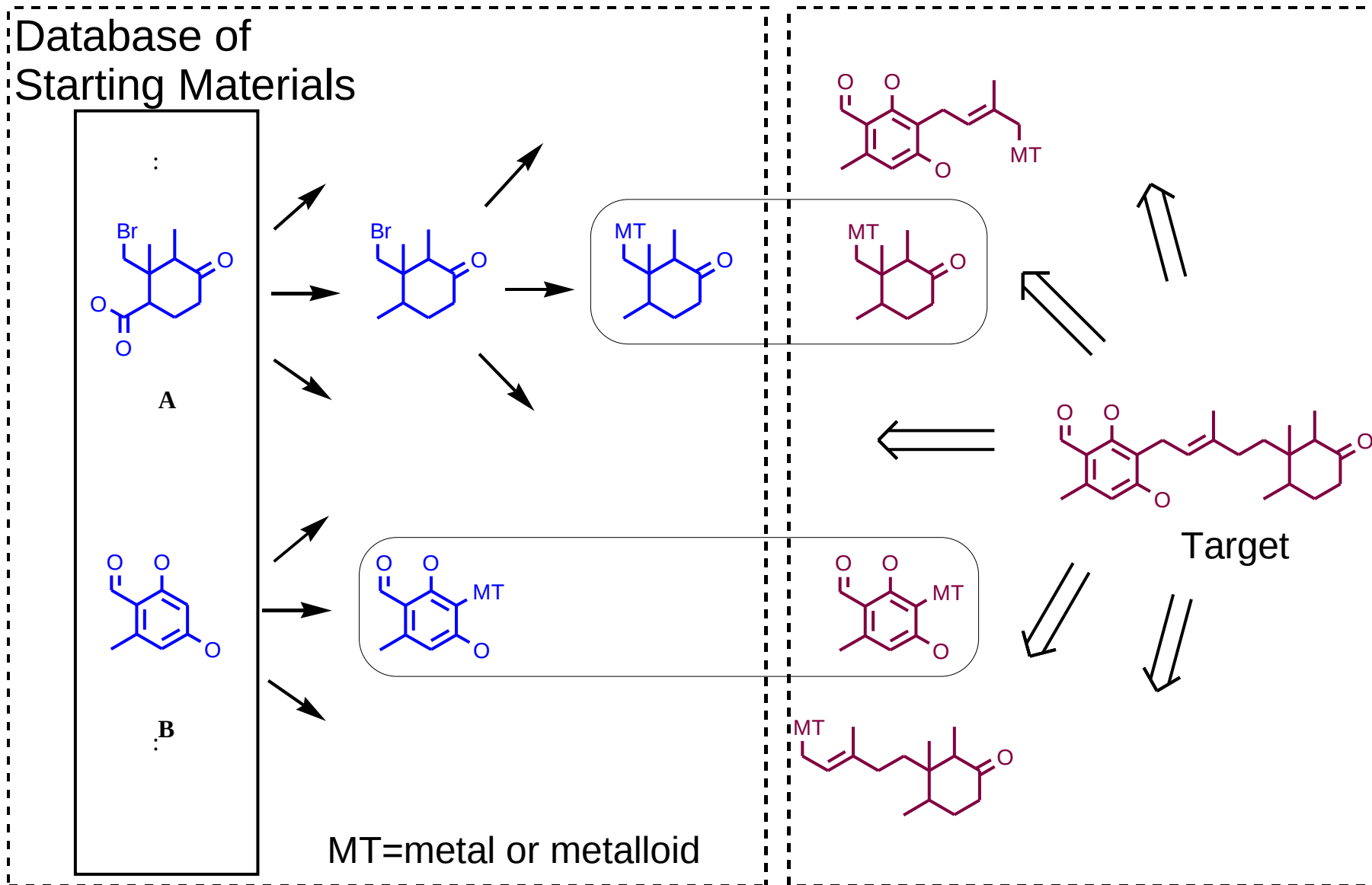


Arrows represent *synthetic* transformations

Virtual Starting Materials Generated and Stored



Bidirectional Search for Synthetic Routes

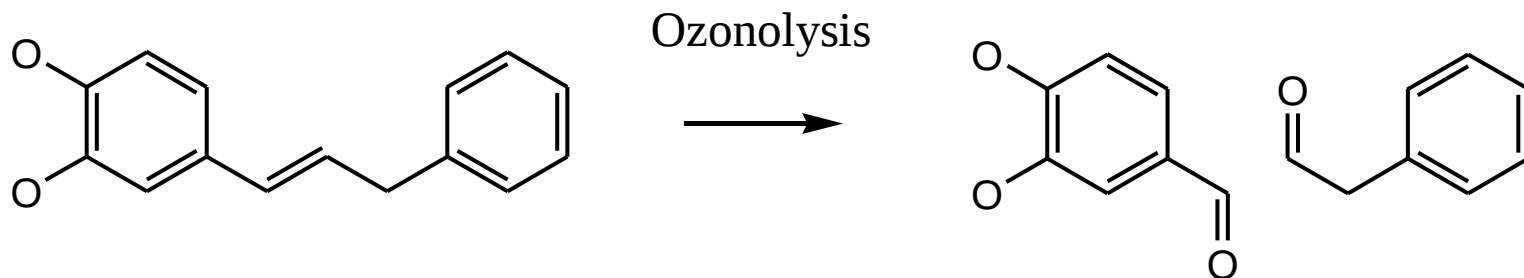


Synthetic off-line
generation of virtual SMs

Retrosynthetic on-line
Generation of Precursors

Why Bother with the Synthetic Analysis? Some Reactions are Best Treated in Synthetic Direction

SYNTHETIC ANALYSIS

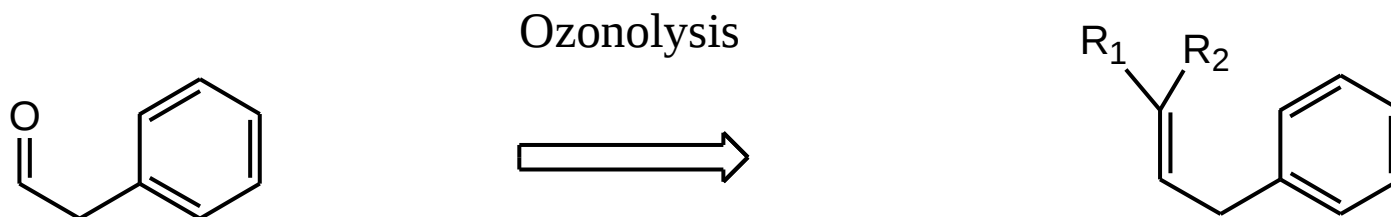


Available starting material

Target

Ozonolysis is an example of a FRAGMENTATION reaction in the synthetic direction

RETROSYNTHETIC ANALYSIS



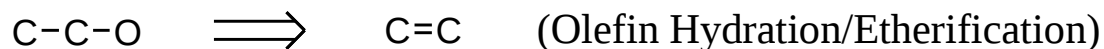
Target

Requires substructure
search to find starting
material

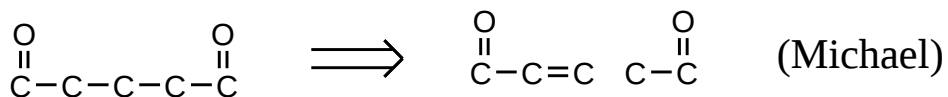
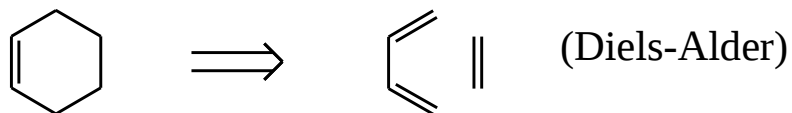
In the retrosynthetic direction ozonolysis is a
RECONNECTIVE transform

Majority of conversions treated in retrosynthetic direction

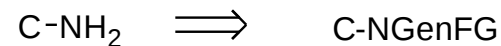
General Transforms



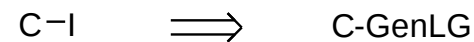
Highly Simplifying Transforms



Generic Carbon Centred Transforms

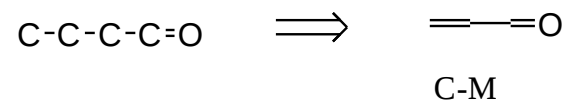


Leaving Group Interconversion



(generic leaving group)

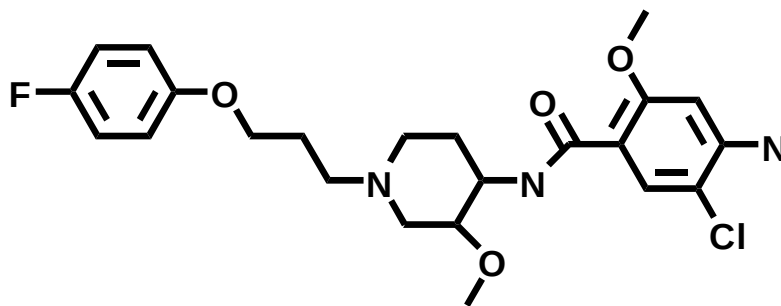
Organometallic Preparation



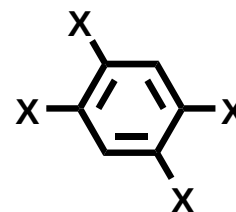
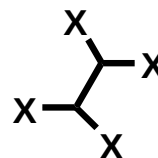
(* M generic metal)

Example of Starting Material Selection

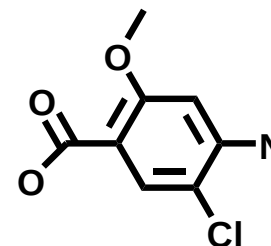
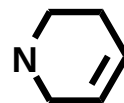
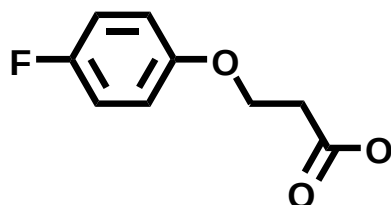
Target Structure



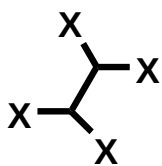
Complexity Analysis



Starting Materials Selected



Residual Complexity

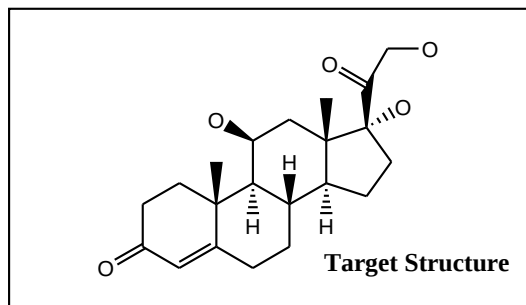


adjacent stereocentres on ring -
relatively easy to control stereochemistry

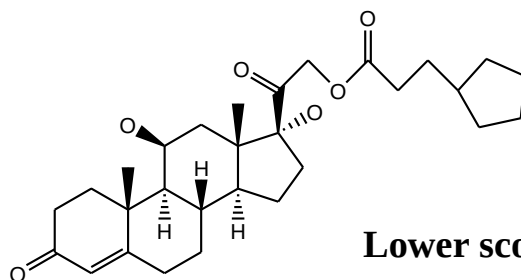
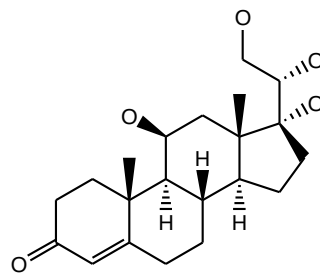
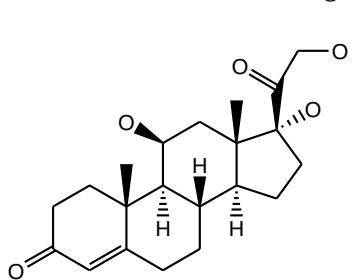
Scoring alternative starting materials

- ❑ There are numerous factors that need to be considered in the identification of good starting materials.
- ❑ There should be a feasible synthetic route between the starting material and the target structure.
- ❑ The number and difficulty of the reaction steps, from the starting material to the target structure, should be kept to a minimum.
- ❑ The physical coverage of the target by the selected starting materials, i.e. the number of atom to atom mappings, is also an important criterion.
- ❑ This number is one aspect of the coverage, but probably even more important is the amount of synthetic difficulty the starting materials remove from the target compound.
- ❑ Good starting materials also result from their complementarity to others. For example, one starting material may have a greater coverage than any single alternative starting material, however two starting materials with individually lower coverage, but greater overall coverage may be more applicable.
- ❑ Other factors such as the price and the availability of the starting materials, that are not directly dependent on the molecular structure, may also play a role.
- ❑ The automatic selection of starting materials should attempt to optimise all of these criteria.

Effect of Starting Material Wastage on Score

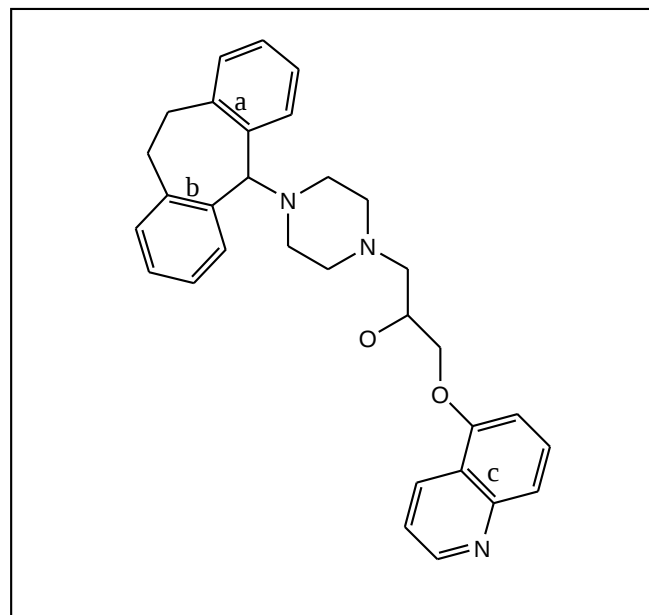


Starting Materials Selected

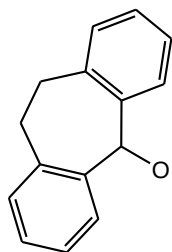


Lower score - more wastage

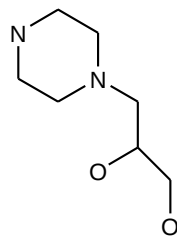
Target Structure



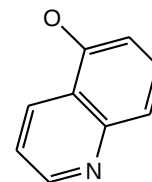
Starting Materials Selected



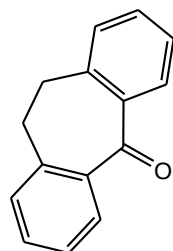
A



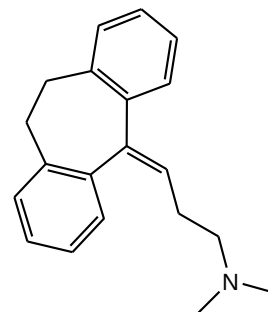
B



C



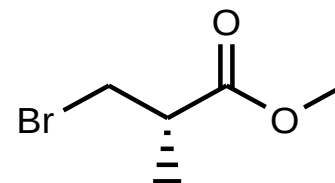
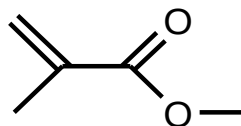
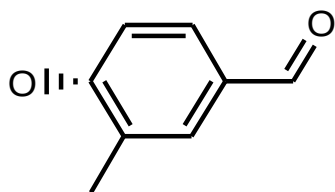
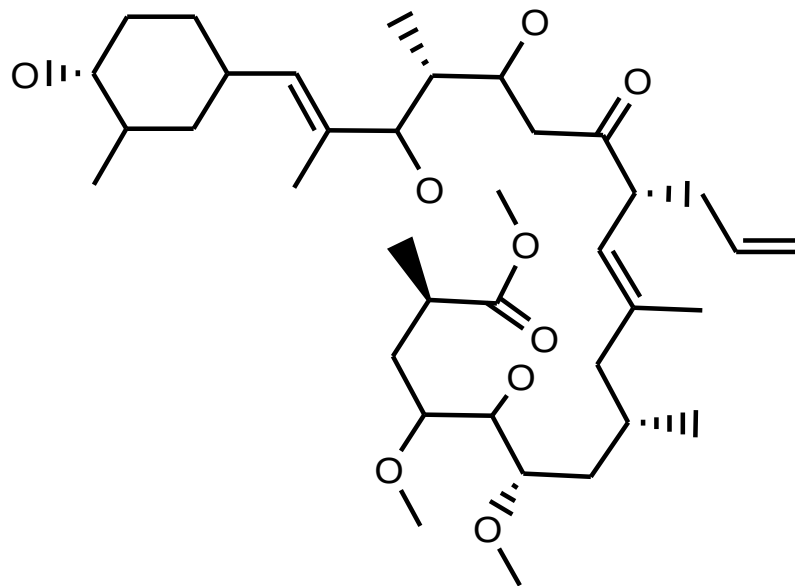
D



E

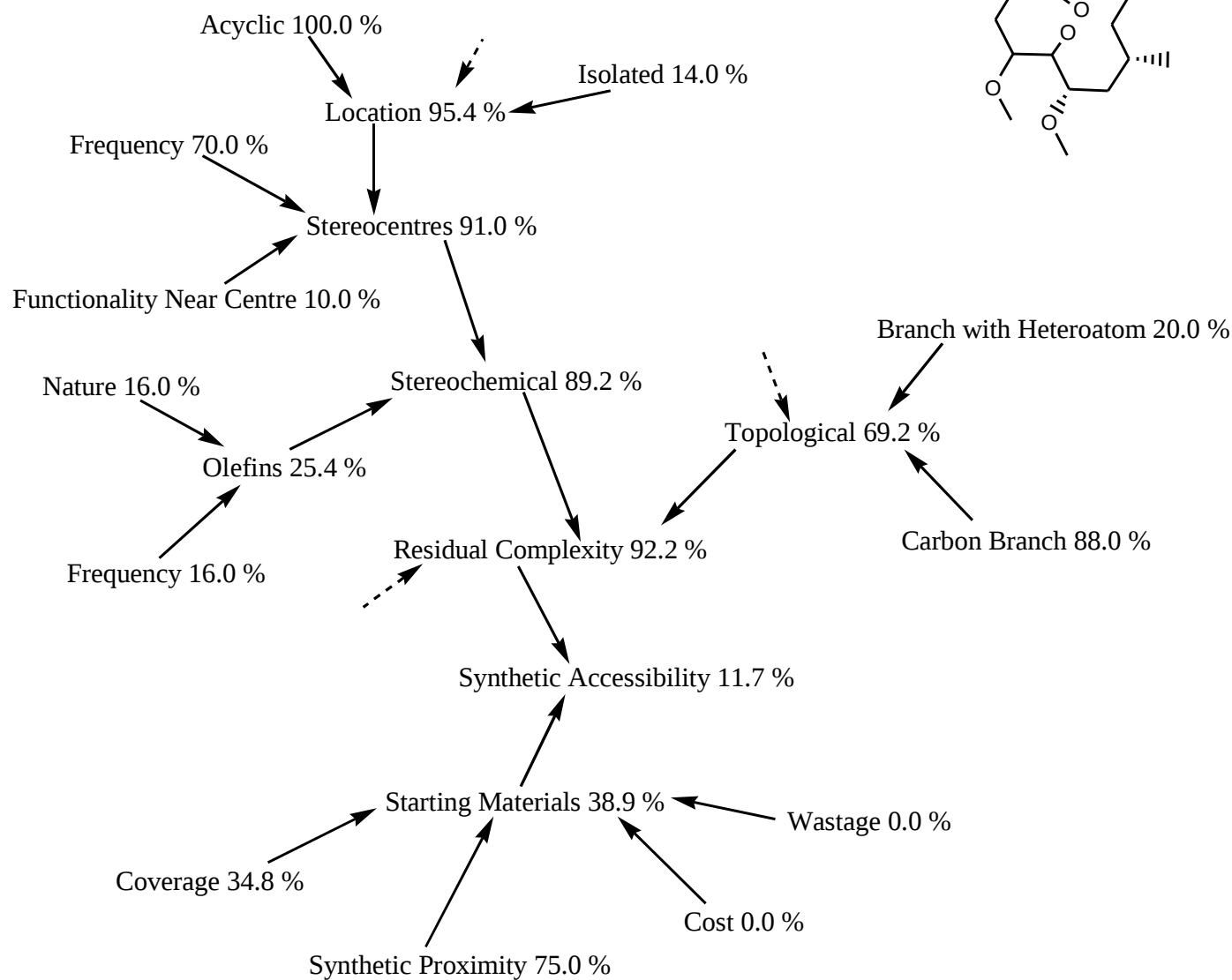
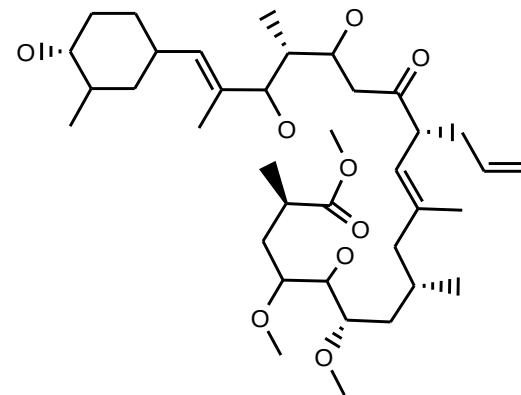
Starting Material Search With Poor Returns

Target Structure I

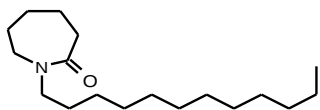


Starting Materials Selected

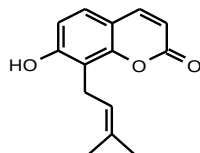
Causal Network for Complexity



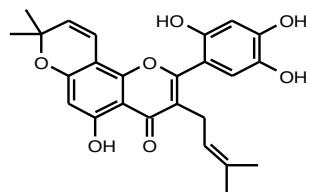
Effect of Starting Material Selection on Synthetic Accessibility Rank



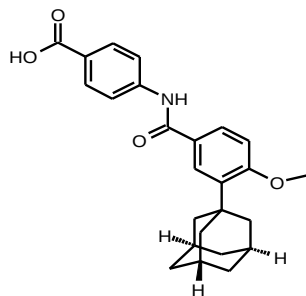
Structure B (Score: 82)



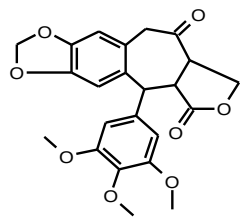
Structure C (Score: 63)



Structure F (Score: 49)

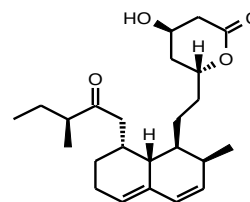


Structure E (Score: 48)

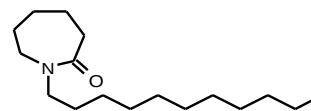


Structure G (Score: 36)

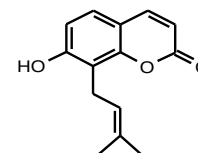
Ordering based on complexity alone



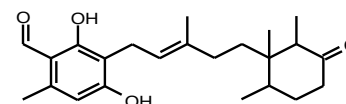
Structure A (Score: 100)



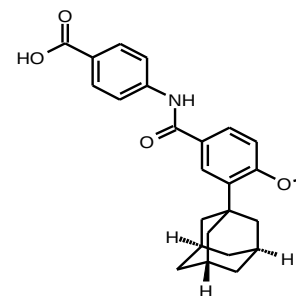
Structure B (Score: 96)



Structure C (Score: 95)



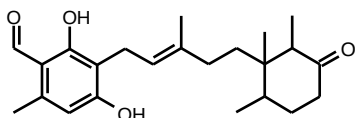
Structure D (Score: 85)



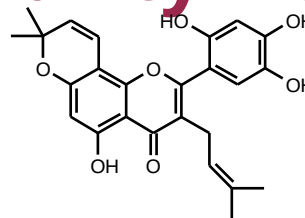
Structure E (Score: 84)

Ordering taking account of available starting materials

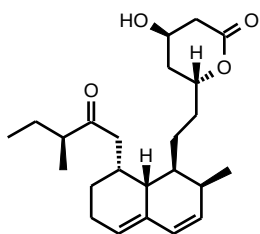
Effect of Starting Material Selection on Synthetic Accessibility Rank



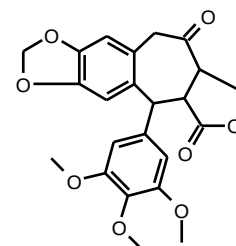
Structure D (Score: 36)



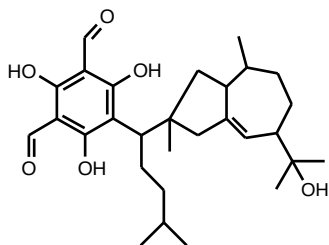
Structure F (Score: 65)



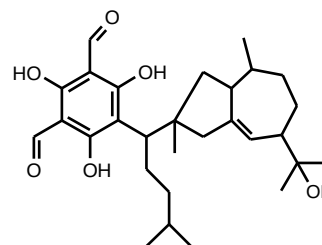
Structure A (Score: 32)



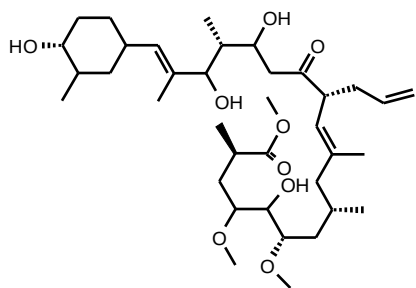
Structure G (Score: 44)



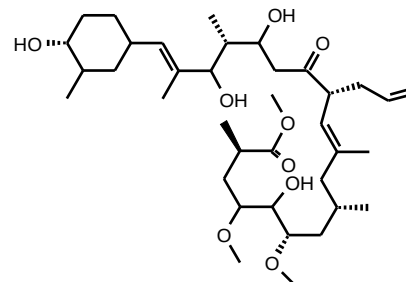
Structure H (Score: 19)



Structure H (Score: 21)



Structure I (Score: 9)

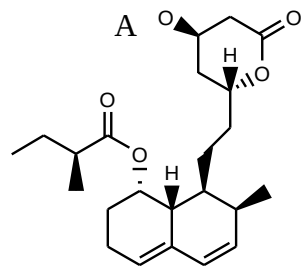


Structure I (Score: 12)

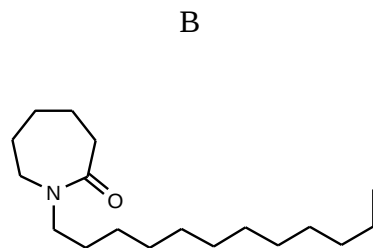
Ordering based on complexity alone

Ordering taking account of available starting materials

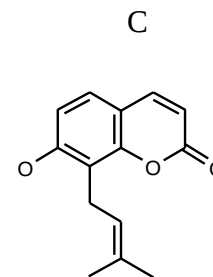
Final CAESA Rank



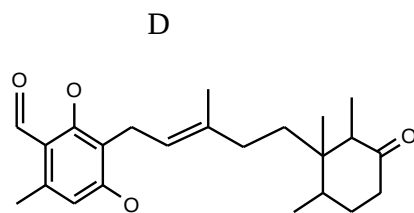
100.0 %



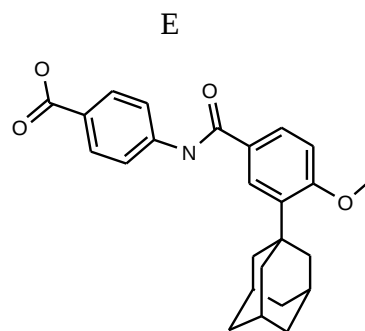
95.9 %



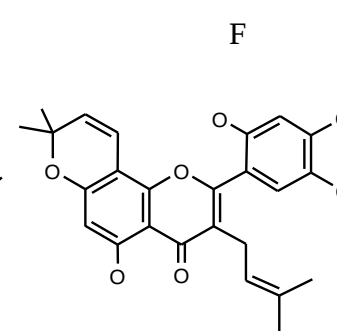
94.5 %



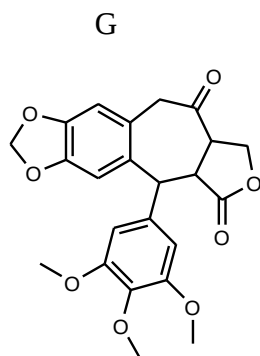
84.6 %



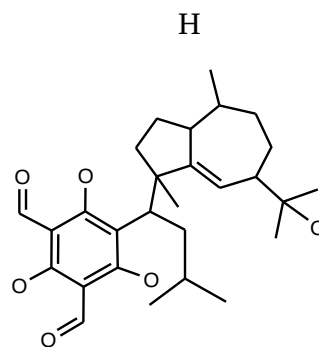
83.5 %



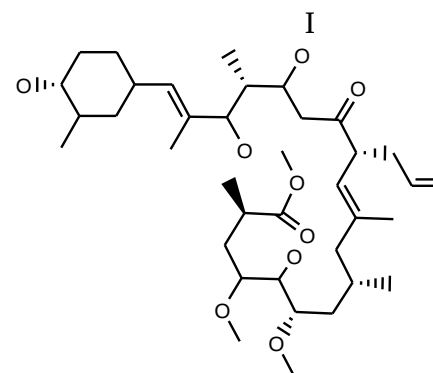
64.8 %



43.6 %



20.6 %



11.7 %

Summary of CAESA Features

- CAESA carries out a retrosynthetic analysis which terminates when a starting material from a database (such as ACD) is found
- Found starting materials are scored according to length and difficulty of reaction sequence and coverage of target compound
- All chemistry rules and transformations are described in editable text knowledge bases easily modified by chemists
- Quality of the analysis depends on the chemistry included in the knowledge bases and the comprehensiveness of the starting material libraries
- But CAESA is relatively slow and speedier methods needed for pruning of large data sets

Complexity Analysis by similarity comparison

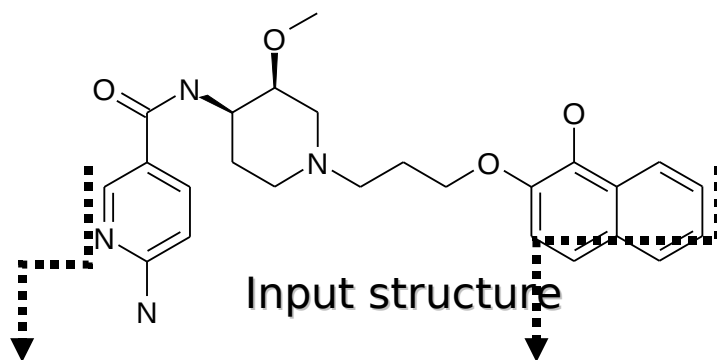
If a molecular structure contains ring and chain substitution patterns which are common in existing drugs than the structure is more likely to be “drug-like” as well as readily synthesisable

starting materials, than the structure is more likely to be readily synthesisable

Complexity analysis based on statistical distribution of various substitution patterns

Molecular Complexity Analysis of de Novo Designed Ligands
Krisztina Boda and A. Peter Johnson
J. Med. Chem. **49**, 5869-5879, 2006

Building Complexity Database



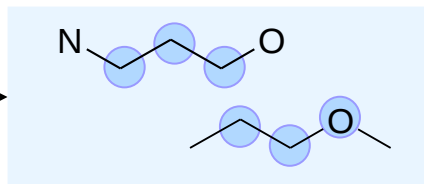
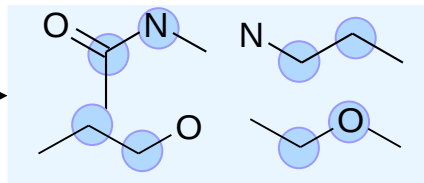
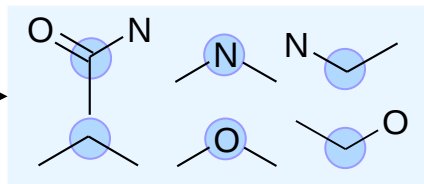
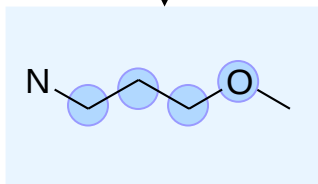
Enumerate chain patterns

• 1-centred

• 2-centred

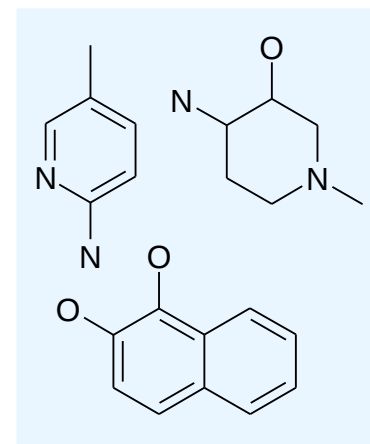
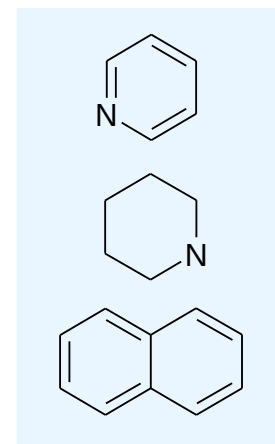
• 3-centred

• 4-centred



Database of chains

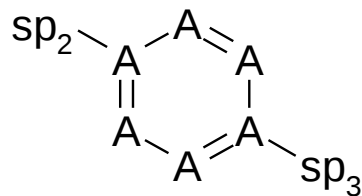
Enumerate ring/ring substitution patterns



Database of rings/ring substitutions

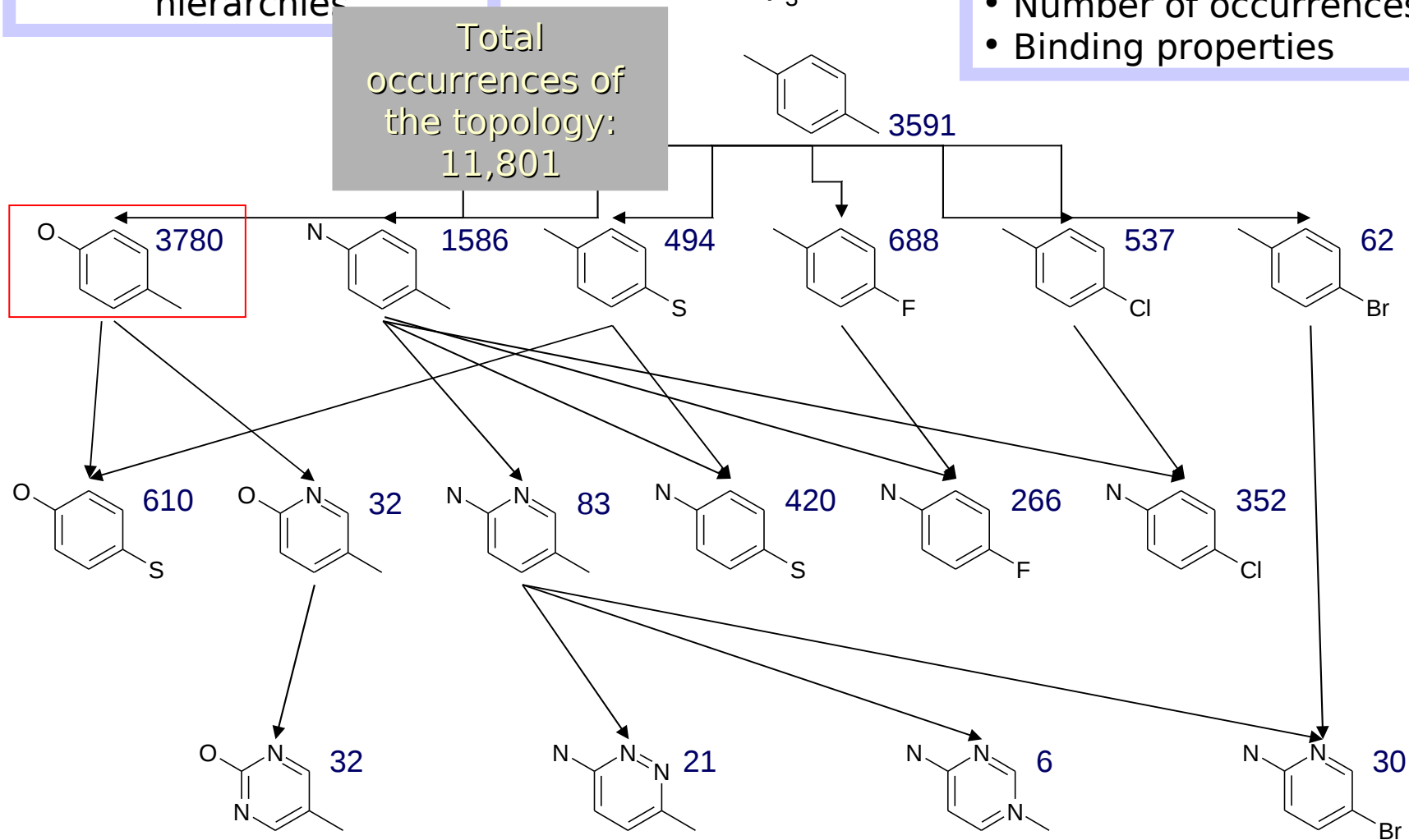
Atom Substitution Hierarchy

Ring (and chain) substitutions are organised in hierarchies



The hierarchy stores:

- Atom type sequence
- Number of occurrences
- Binding properties



Building Complexity Database

MDDR
~110.000

(Aldrich+Maybridge+
Lancaster)
~170.000

Filters

- Molecular weight ≤ 700
- Allowed atom types:
H, B, C, N, O, F,
P, S, Cl, Br, I
+ (therapeutic class filter
for MDDR)

Perceiving Atom & Ring
Properties

Enumerating Chain & Ring
Patterns

Perception Knowledge Base

- Aromatic
- Hybridisation
- H-bonding
properties

MDDR

SM

	Unique Topology	Unique Atom Subs.	Unique Topologies	Unique Atom Subs.
1-centred	144	937	185	1,359
2-centred	656	3,524	801	4,453
3-centred	2,392	9,108	2,609	9,059
3-centred Ring	5,918	16,931	5,454	12,646
	2,689	5,085	1,853	3,340

Total Elapsed
Time:
~ 6 hours
on Linux PC
(3GHz)

Calculation of Complexity Score

Penalise atom patterns which are infrequent or not present in the complexity database.

$$\text{SCORE}_{\text{TOTAL}} = \frac{\sum \text{SCORE}_{\text{TOPOLOGY}} + \sum \text{SCORE}_{\text{ATOM SUBS.}}}{\text{Num of Patterns}} + P_{\text{stereo}} + P_{\text{rotbond}}$$

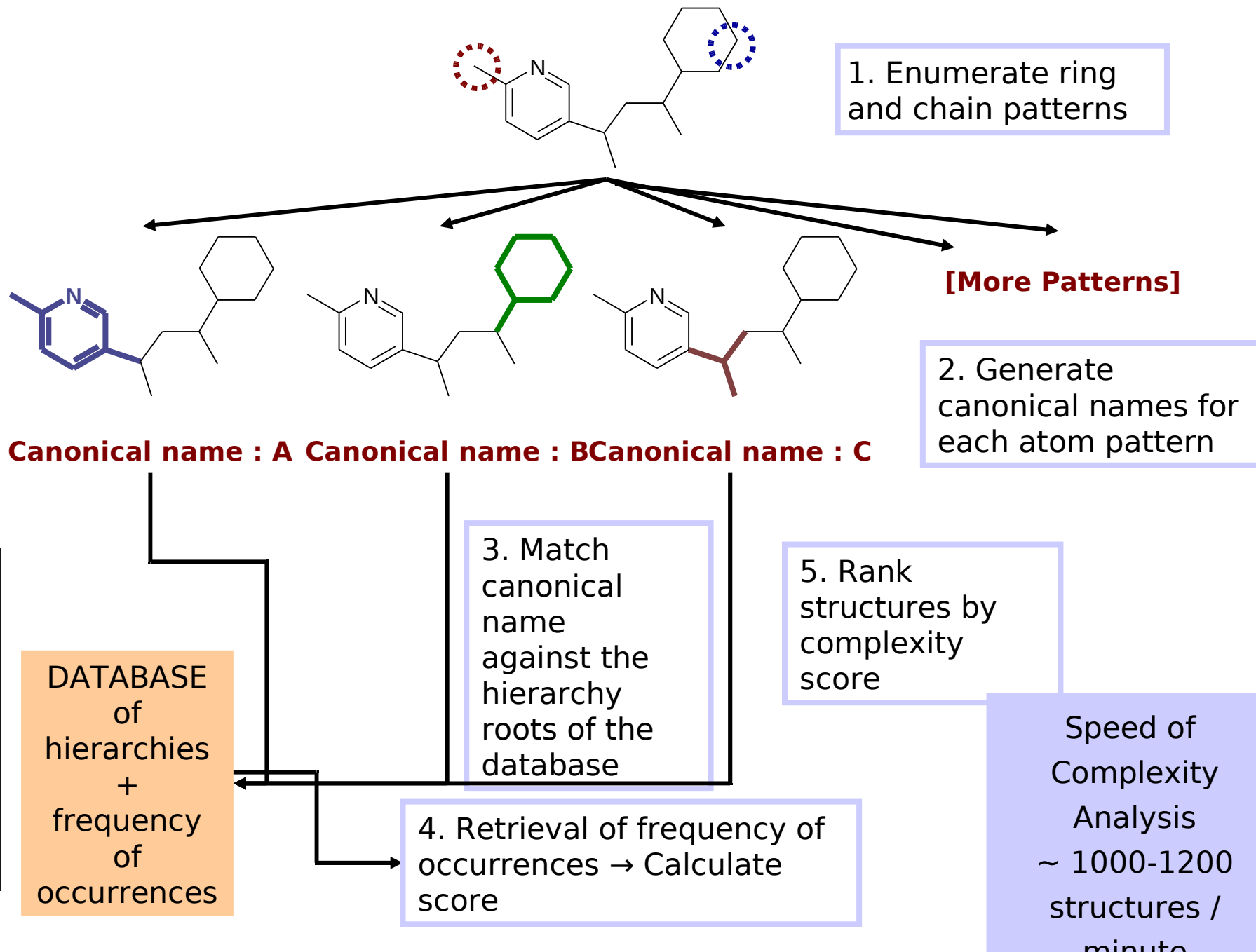
$$\text{SCORE}_{\text{TOPOLOGY}} = \begin{cases} \left(1 - \frac{\ln(\text{No. occur. matched topology})}{\ln(\text{No. occur. most common topology})}\right) * \text{Penalty} & \text{, if topology exists} \\ \gamma * \text{Penalty} & \text{, if topology missing from database} \end{cases}$$

$$\text{SCORE}_{\text{ATOMS SUBS}} = \begin{cases} \left(1 - \frac{\ln(\text{No. occur. best matched atom subs.})}{\ln(\text{No. occur. most common atom subs.})}\right) * \text{Penalty} & \text{, if matching atom subs exists} \\ 2 * \text{Penalty} & \text{, if topology or atom subs. missing} \end{cases}$$

Penalty values can be altered to tailor the system for different applications.
In SPROUT the complexity analysis is followed by ranking the putative ligands according to their evaluated complexity score.

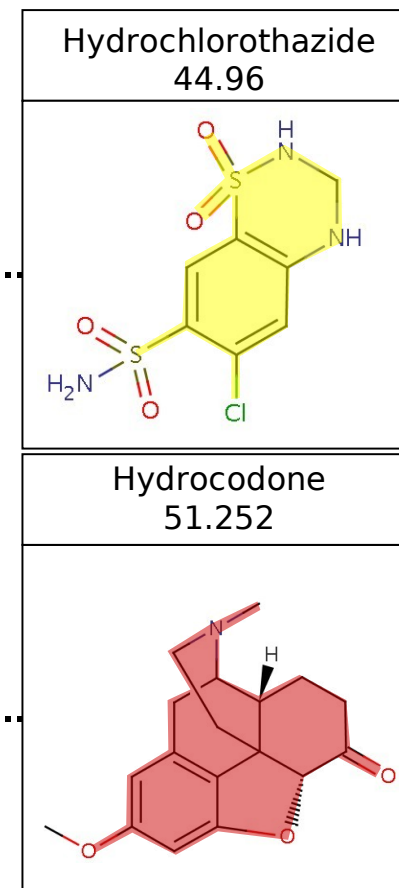
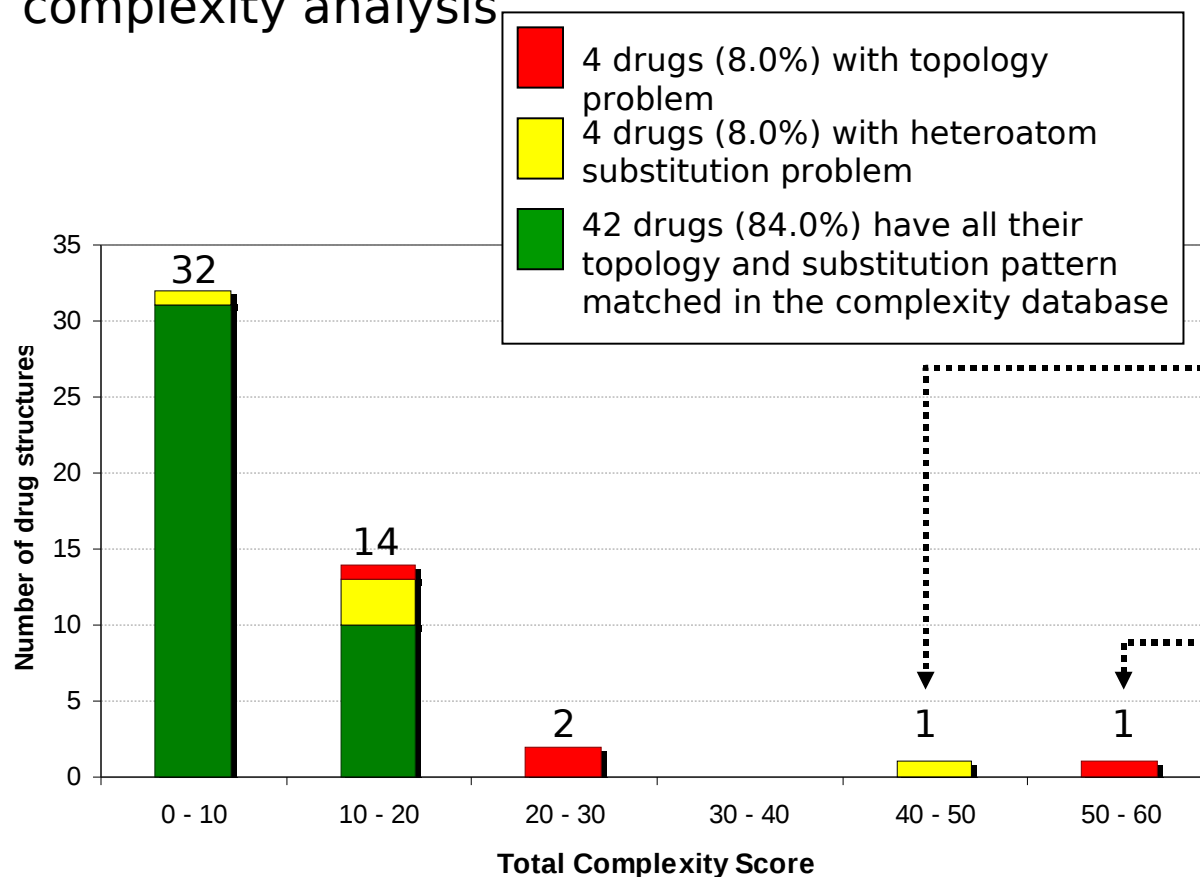
The penalty values used in the examples presented here are 25, 20, 15, 10 for 1-,2-,3- and 4-centred chain patterns, 40 and 30 for rings and ring substitutions.

Complexity Analysis



Validation I.

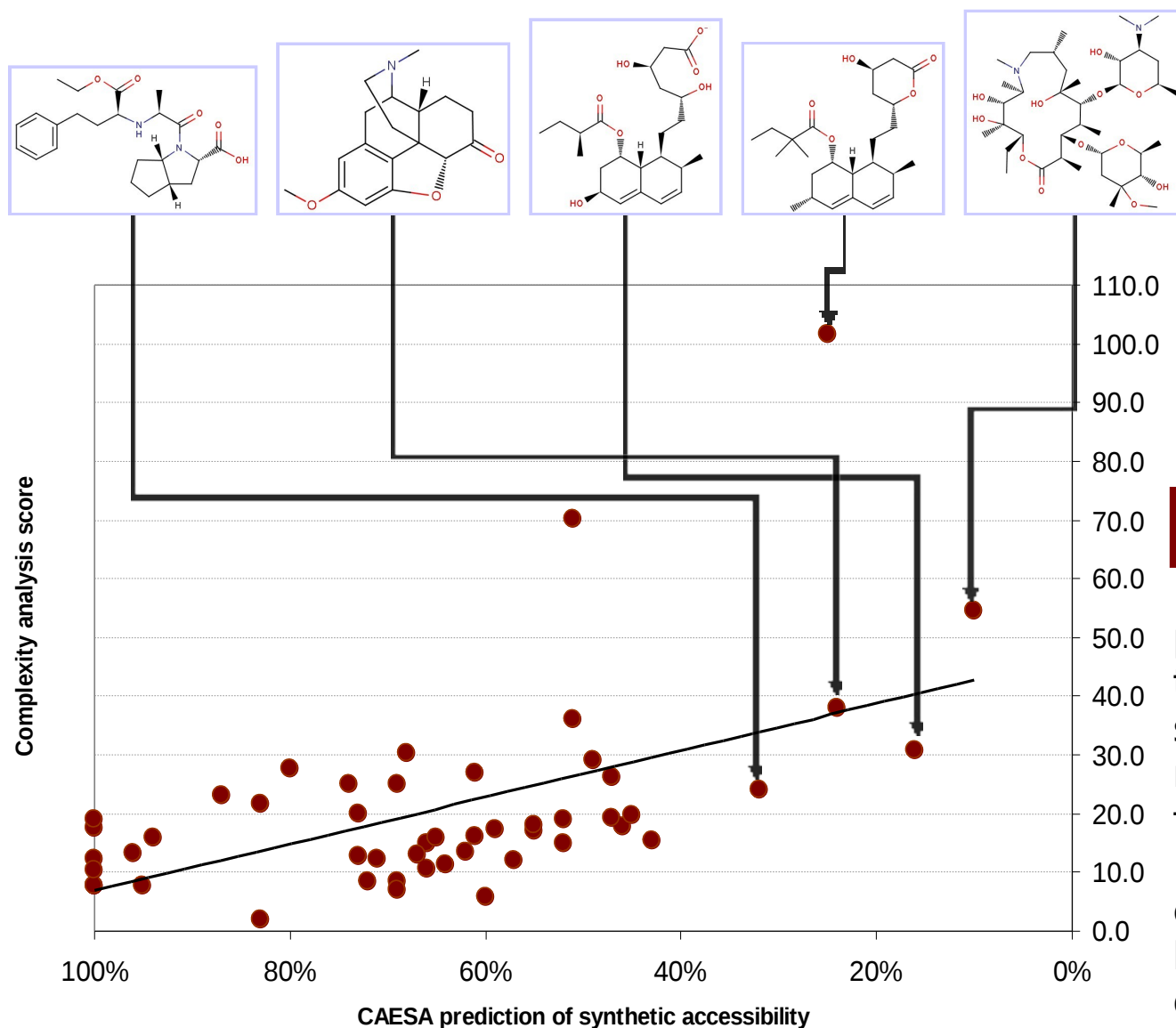
50 of the top selling (non-steroid) drugs^[1] were subjected to this complexity analysis



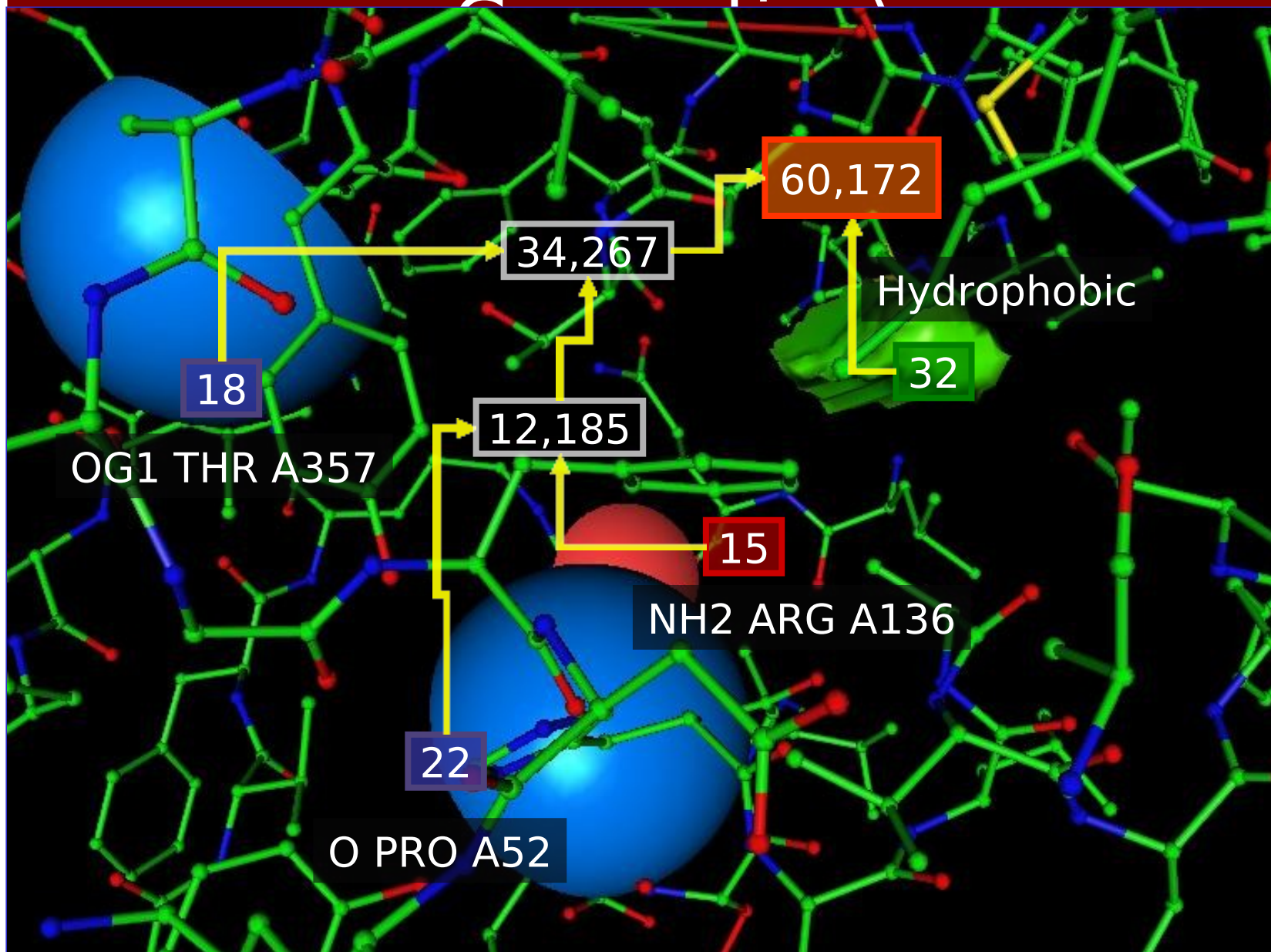
Distribution of the estimated complexity scores of 50 top selling drugs (using the complexity database derived from MDDR)

[1] RxList LLC "The Top 200 Prescription for 2003 by number of US Prescriptions Dispensed <http://www.rxlist.com/top200.html>"

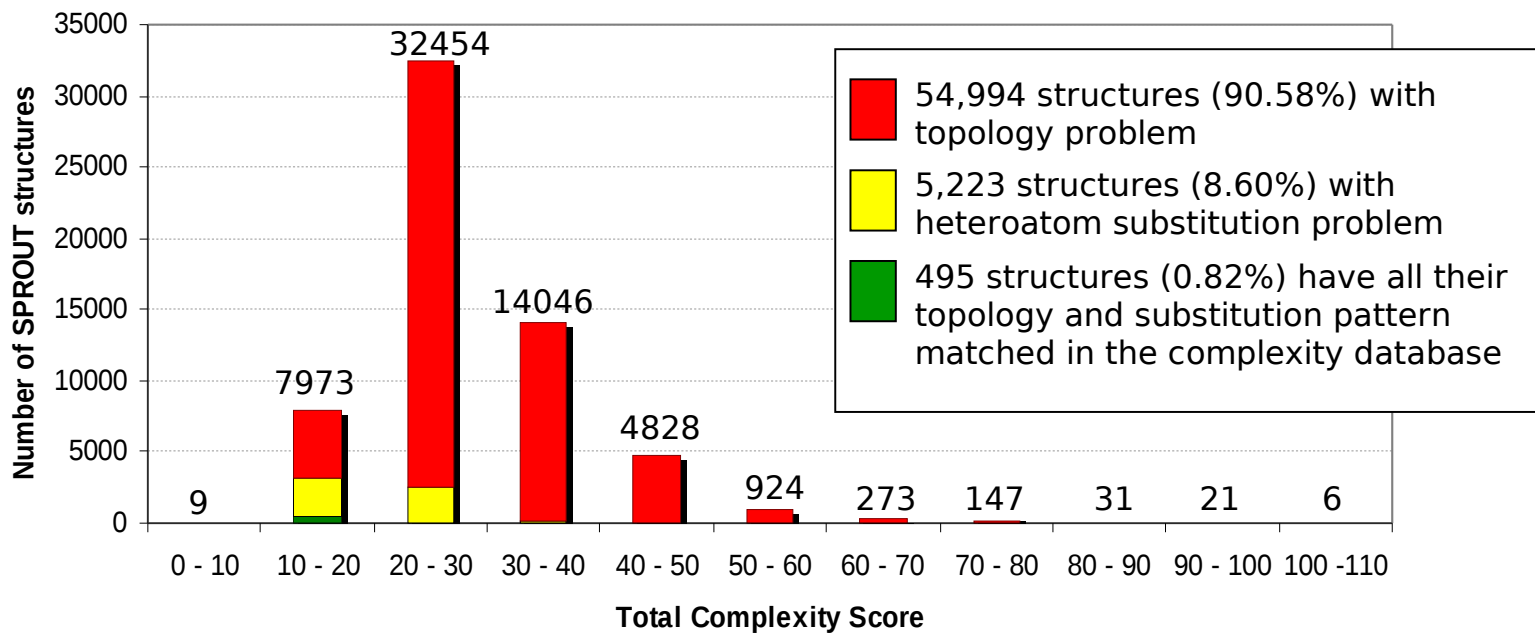
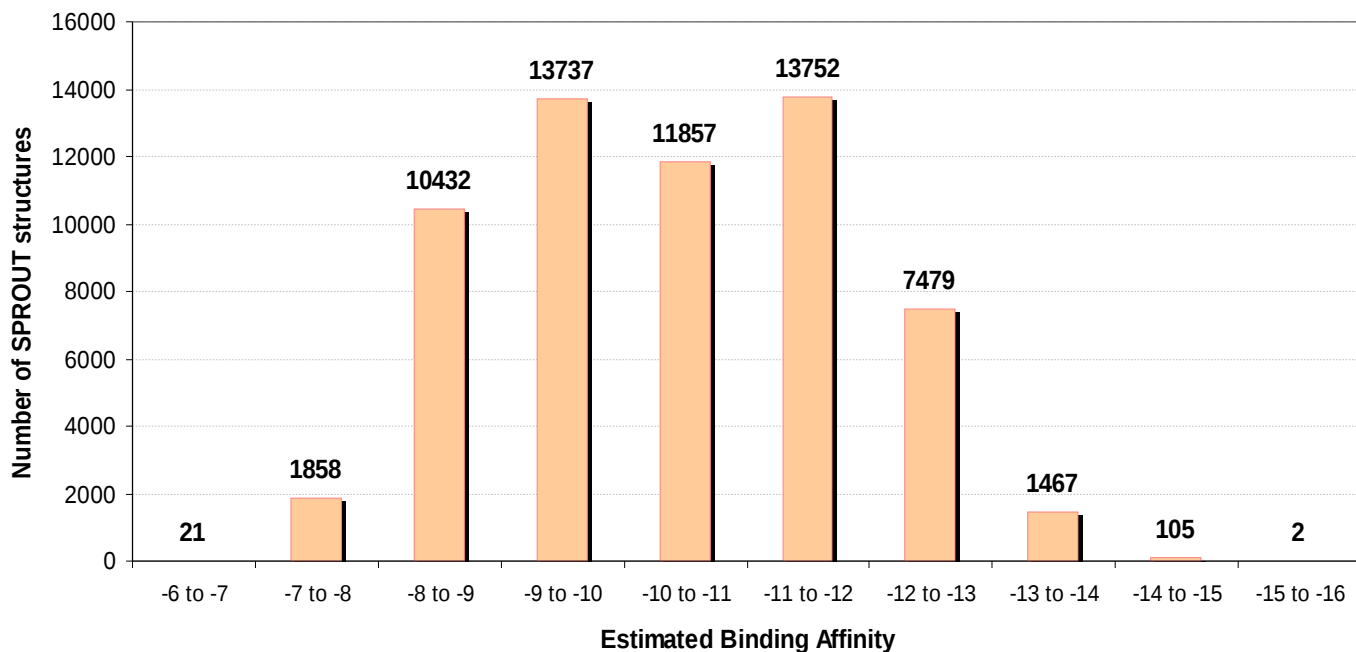
CAESA vs. Complexity Analysis





Case Study (Structure

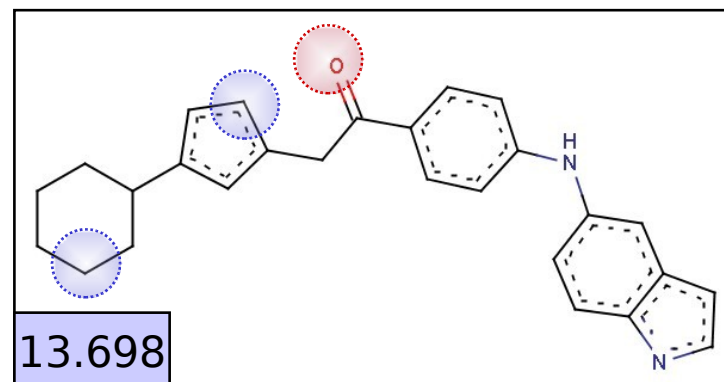
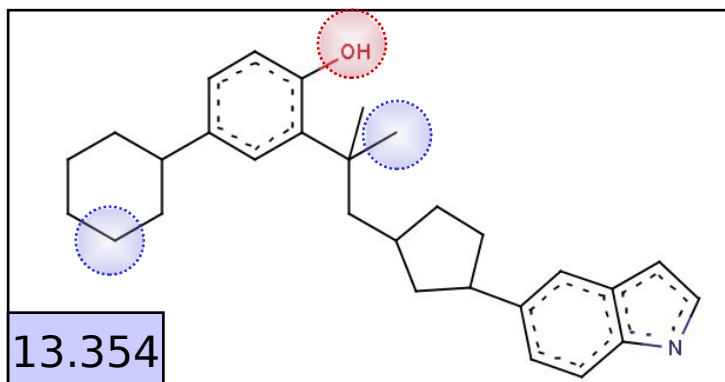
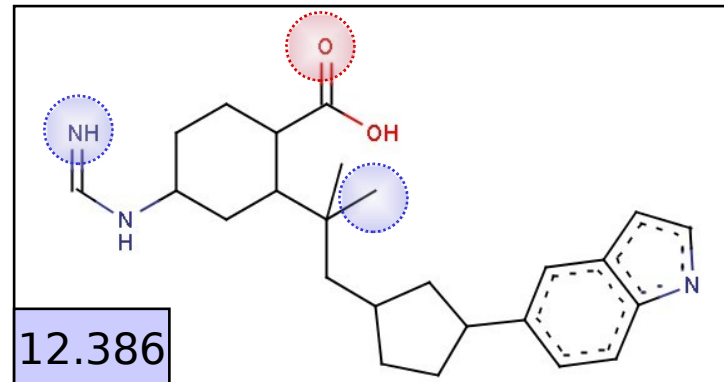
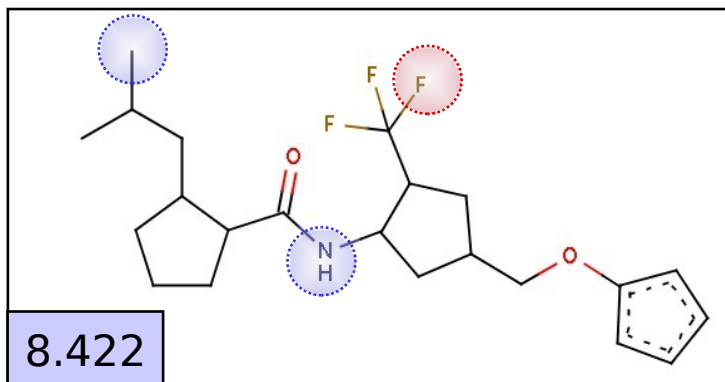
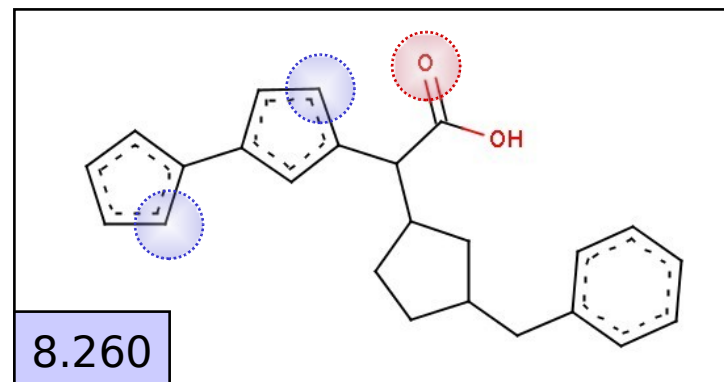
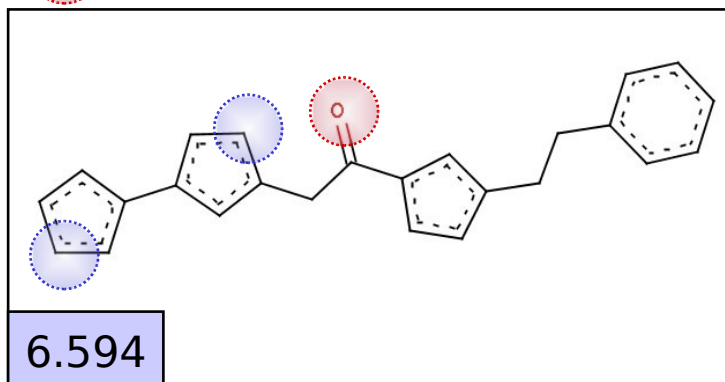


Case Study Results



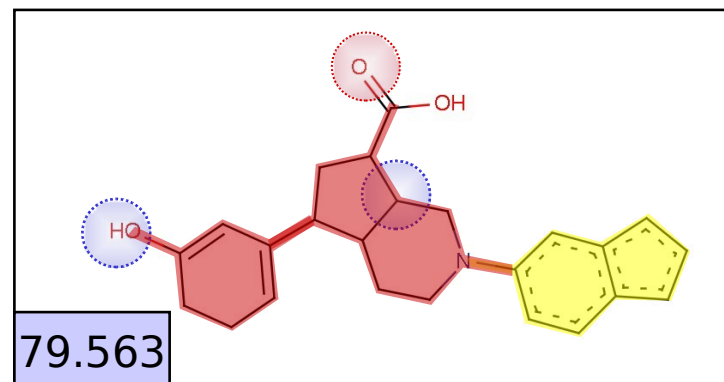
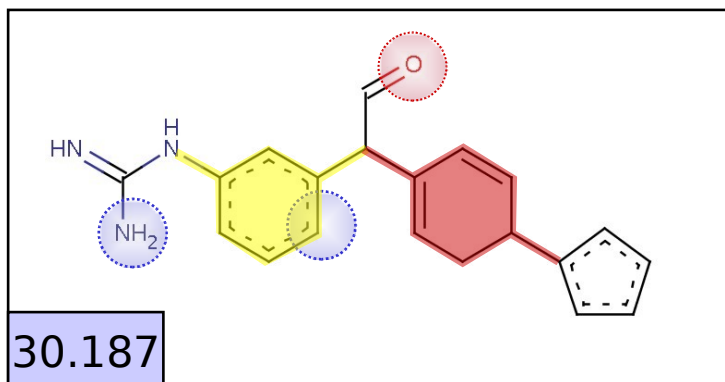
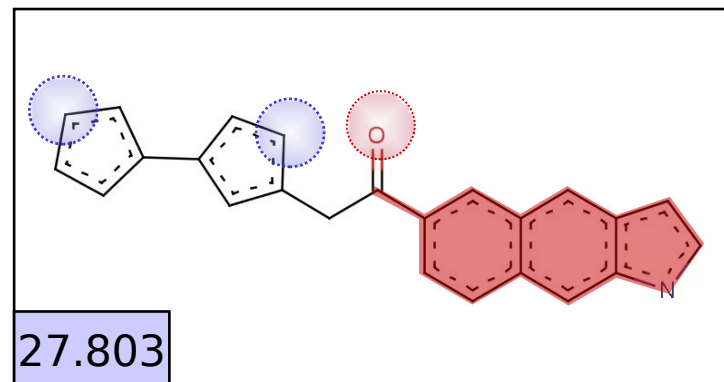
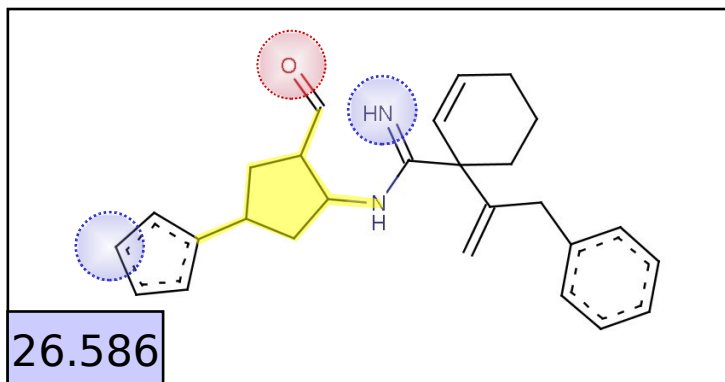
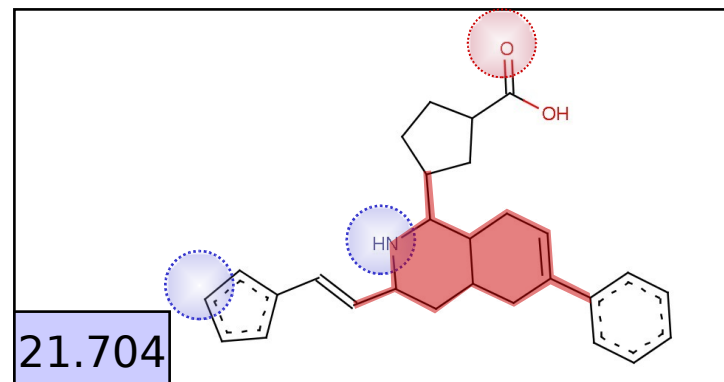
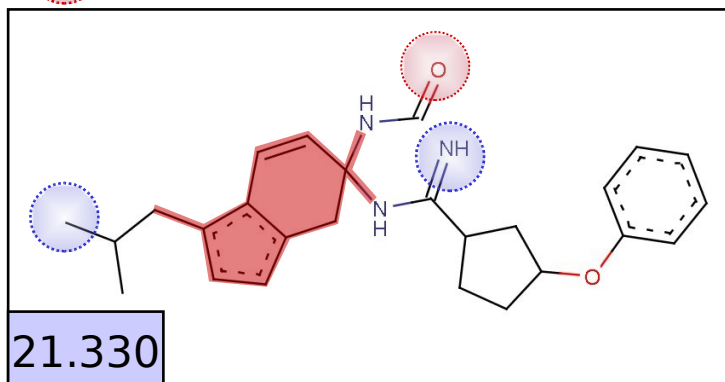
Case study ("Simple" Structures)

-  = Docked to donor target site
-  = Docked to acceptor target site



Case study (Complex Structures)

-  = Docked to donor target site
-  = Docked to acceptor target site
-  = Topology problem
-  = Heteroatom substitution problem



Conclusion

Complexity analysis based on structural motifs of existing drugs and compounds provides a fast and effective method to rank structures and eliminate complex structures prior to the computationally more expensive estimation of binding affinity.

Warning

This approach is based on characteristics of existing drugs and compounds



Structures with simple but novel structural features may be incorrectly identified as complex

Acknowledgements

 for support of the CAESA project

CAESA and Complexity Analysis are now developed and supported by Keymodule Ltd, Leeds, UK

In North America they are available from Simbiosys Inc

