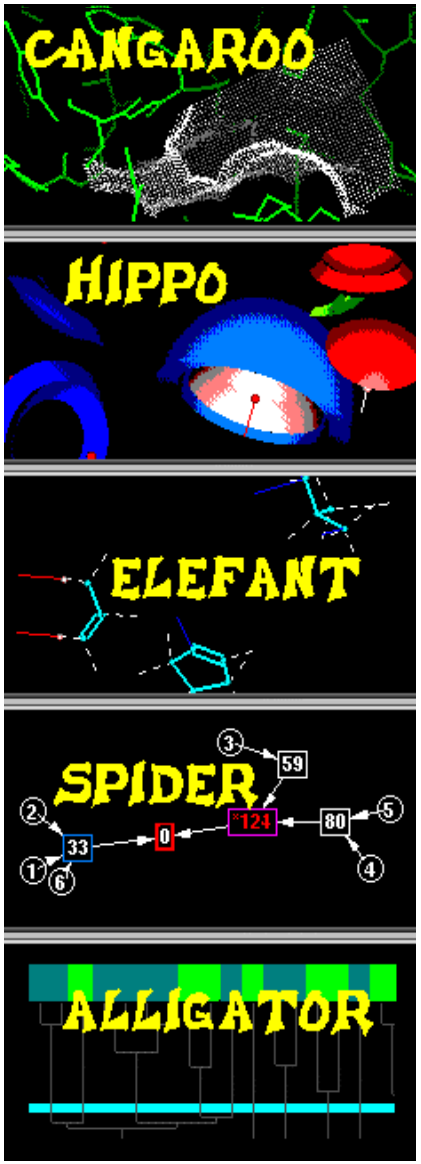
Complexity Analysis of *De Novo* Designed Ligands

Krisztina Boda
Prof. A Peter Johnson

Topics of discussion

- De novo structure generation in SPROUT
- Main concept of structural complexity analysis
  - Generating complexity database
  - Predicting structural complexity
- Case study
- Future work

SPROUT Components



Detecting potential binding pockets of the protein structures

Identify favourable hydrogen bonding interaction sites (H-bonding, hydrophobic, covalent, metal, user defined)

Docking structures to target interaction sites

Generating 3D molecular structures of novel ligands by linking the docked starting fragments together in an incremental construction scheme

Scoring, sorting and clustering the solutions for an efficient means of evaluating the results

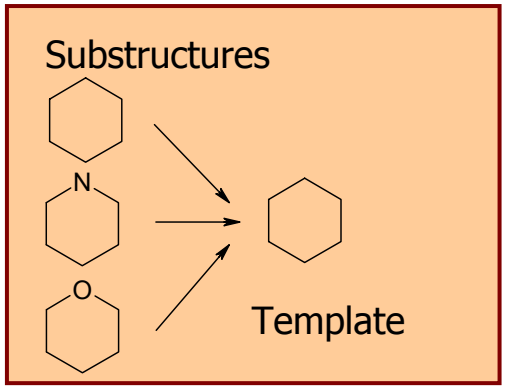
Structure Generation in SPROUT

1st Phase

Primary molecular structure generation

Library:

- Specific functional groups
- Generalised fragments
  - generic atom
  - focus on hybridisation

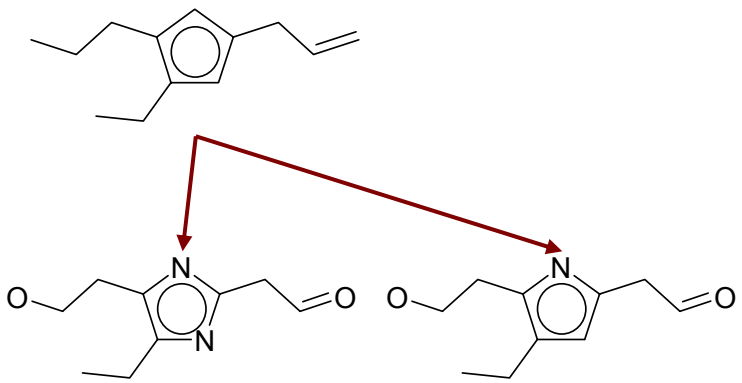


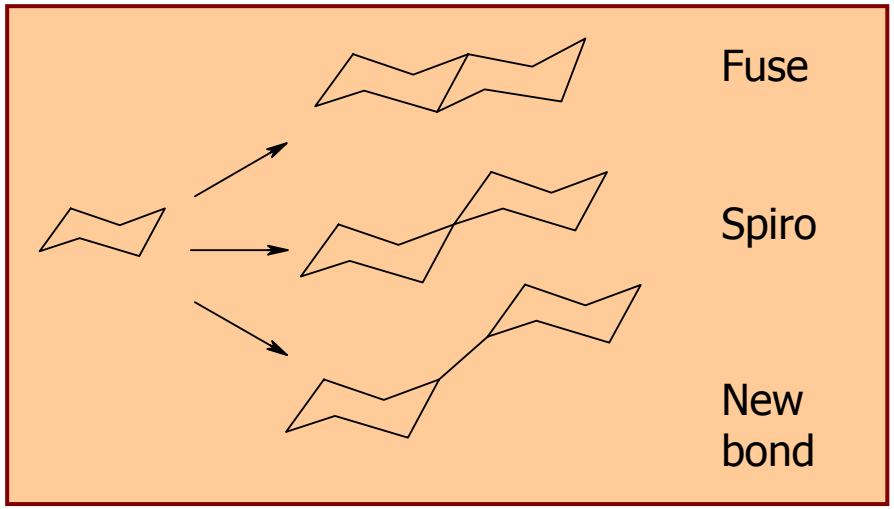
2nd Phase

Conversion of structure graph into molecule

- Sequential method to build structure graph
- Heuristics to avoid combinatorial explosion

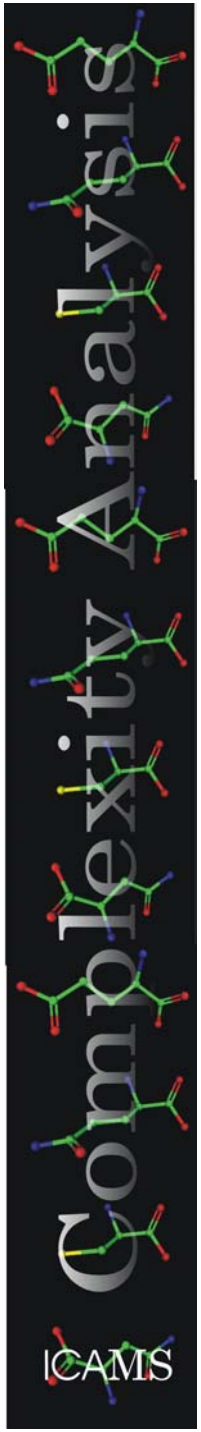
- Atom substitution





ClassicSPROUT

ICAMS



Problem Specification

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

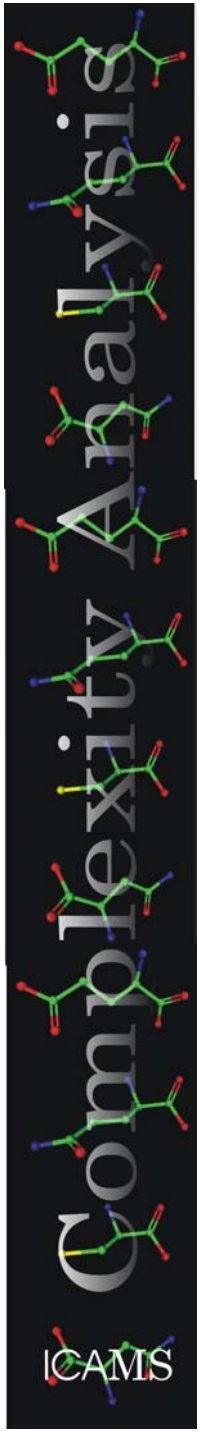
Powerful heuristics are necessary to evaluate (and reduce) large answer sets

Binding Score

Eliminating structures with poor estimated binding affinity

Complexity Analysis

Eliminating structures with complex molecular structure



Assumption

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"

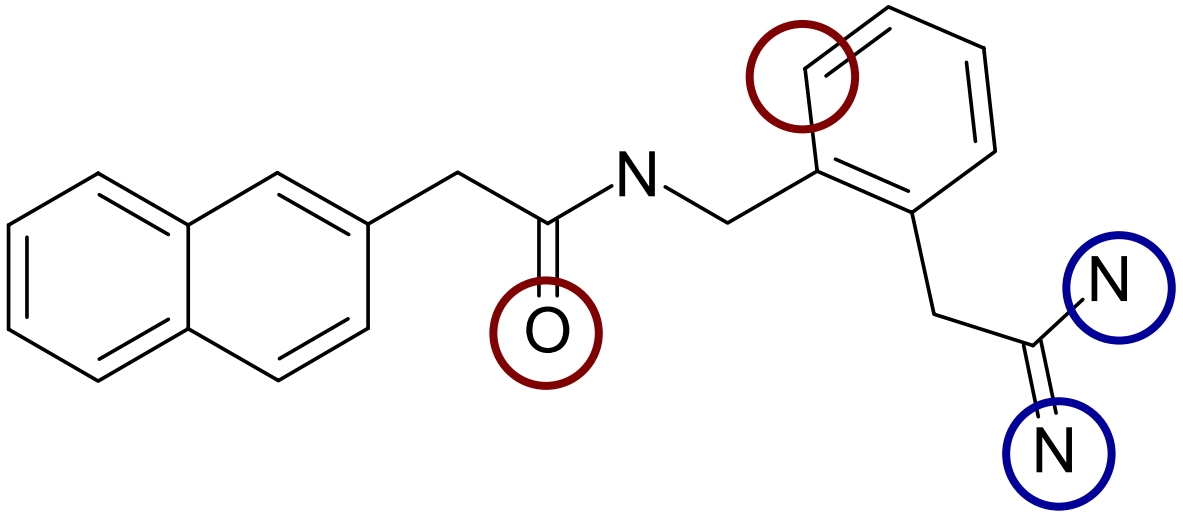
starting materials,
than the structure is more
likely
to be readily synthesisable

Complexity analysis based on statistical distribution of various substitution patterns

Multi-level Complexity Analysis

= docked to acceptor target site

= docked to donor target site



- Topological matching (1st level)
  - considering hybridization
- Atom substitution matching (2nd level)
  - matching specific atom type if present
  - matching binding property if present

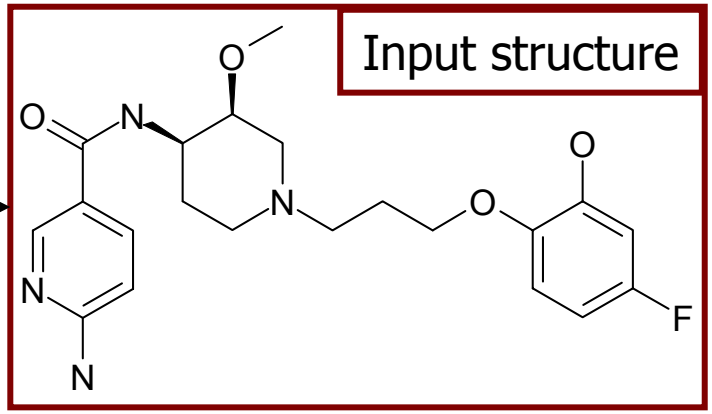
Complexity Analysis

ICAMS

Building Complexity Database

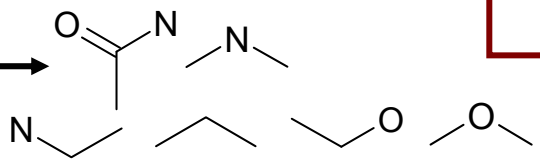
Original Database

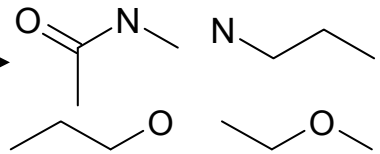
Input structure

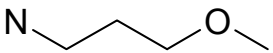


Enumerating chain patterns

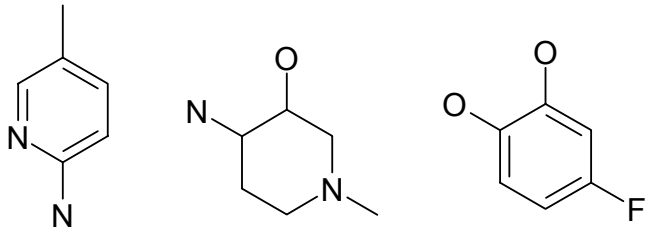
- 1-centred
- 2-centred
- 3-centred
- 4-centred







Enumerating ring patterns



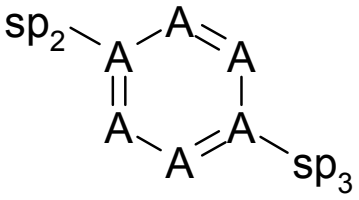
Database of chain substitutions

Database of ring substitutions

ICAMS

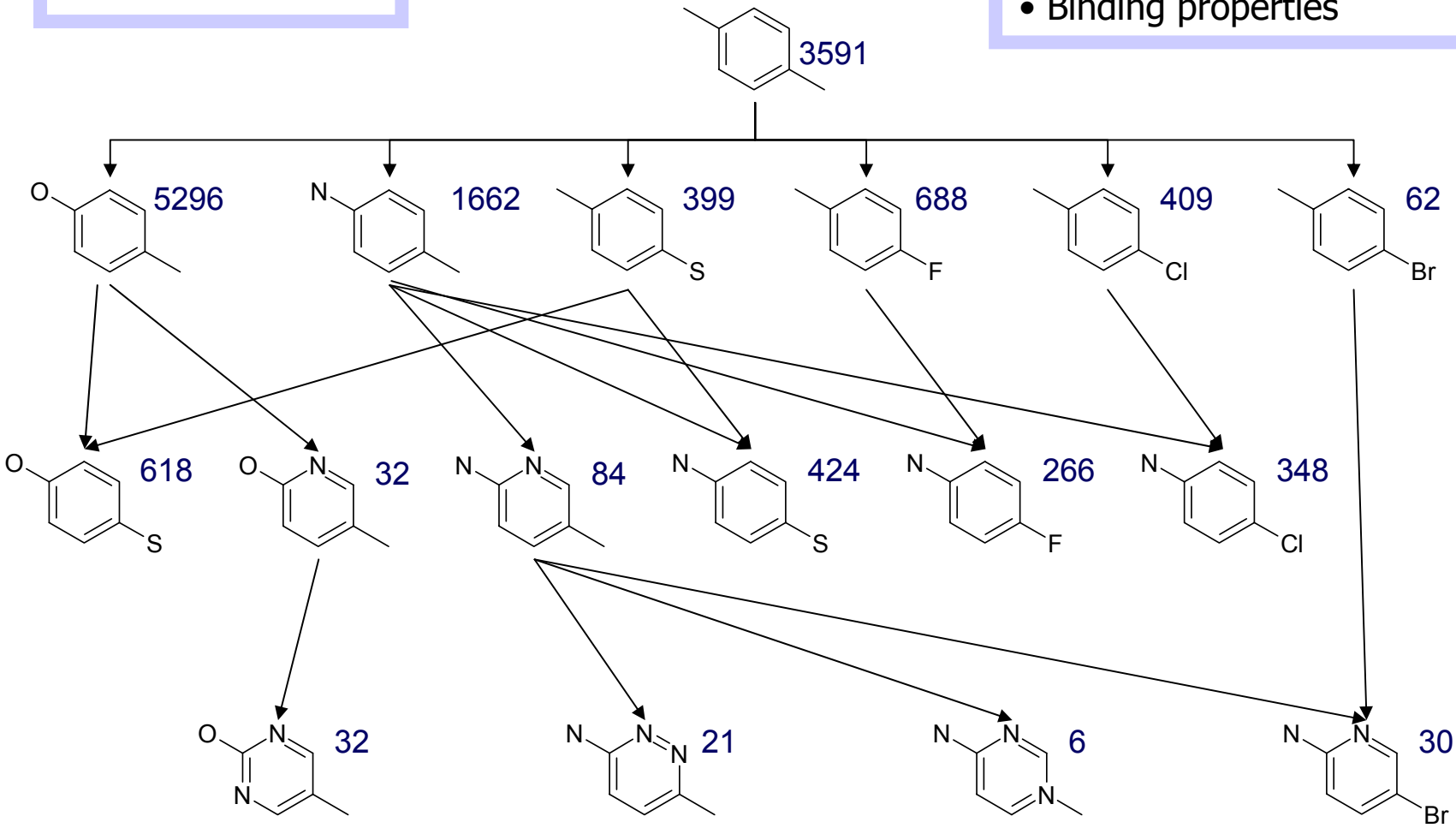
Atom Substitution Hierarchy

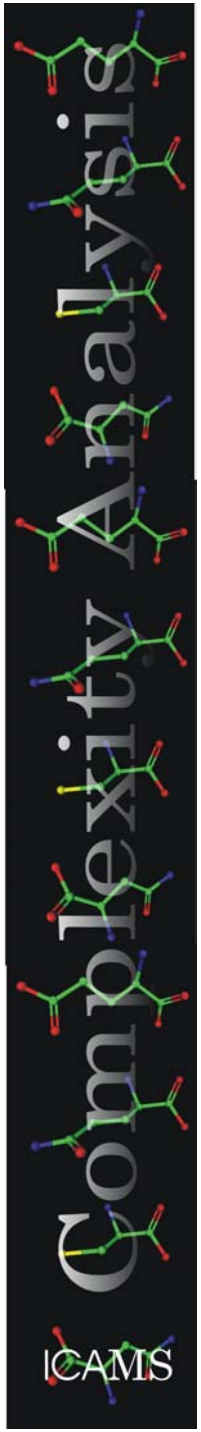
Ring (and chain) substitutions are organised in hierarchies



The hierarchy stores:

- Atom type sequence
- Number of occurrences
- Binding properties





Building Complexity Database

MDDR +
Aldrich + Maybridge
~ 250.000 2D structures

Filters

- Molecular weight ≤ 700
- Allowed atom types:
  H, B, C, N, O, F,
  P, S, Cl, Br, I

4% of structures were filtered out

Perceiving Atom & Ring Properties

Enumerating Chain & Ring Patterns

Perception Knowledge Base

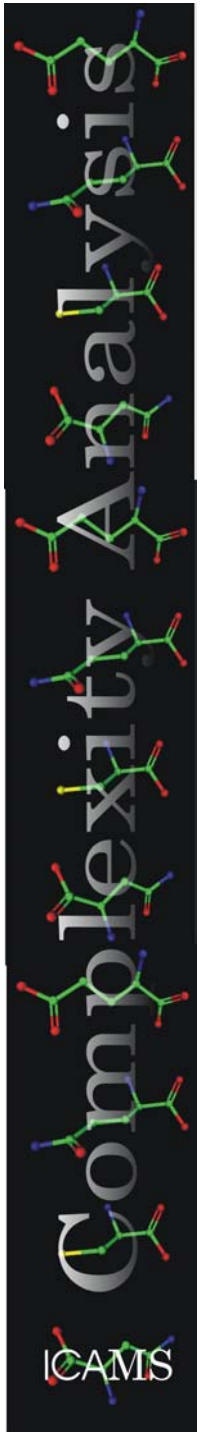
- Aromatic
- Hybridisation
- H-bonding properties

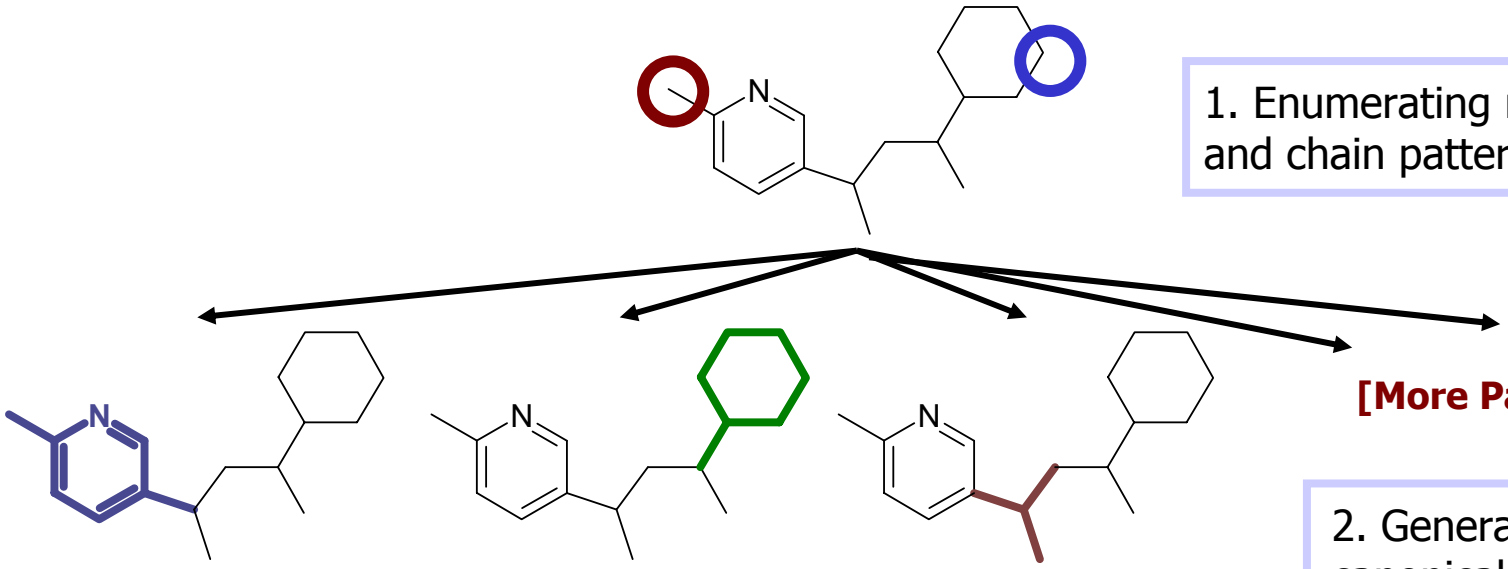
Complexity Database

	Unique Topology	Unique Atom Substitution
1-centred	102	1,000
2-centred	530	4,354
3-centred	2,235	12,035
3-centred	6,513	22,820
Ring substitution	21,614	38,514

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

Complexity Analysis





1. Enumerating ring and chain patterns

[More Patterns]

2. Generating canonical names for each atom pattern

Canonical name : A **Canonical name : B** **Canonical name : C**

3. Matching canonical name against the hierarchy roots of the database

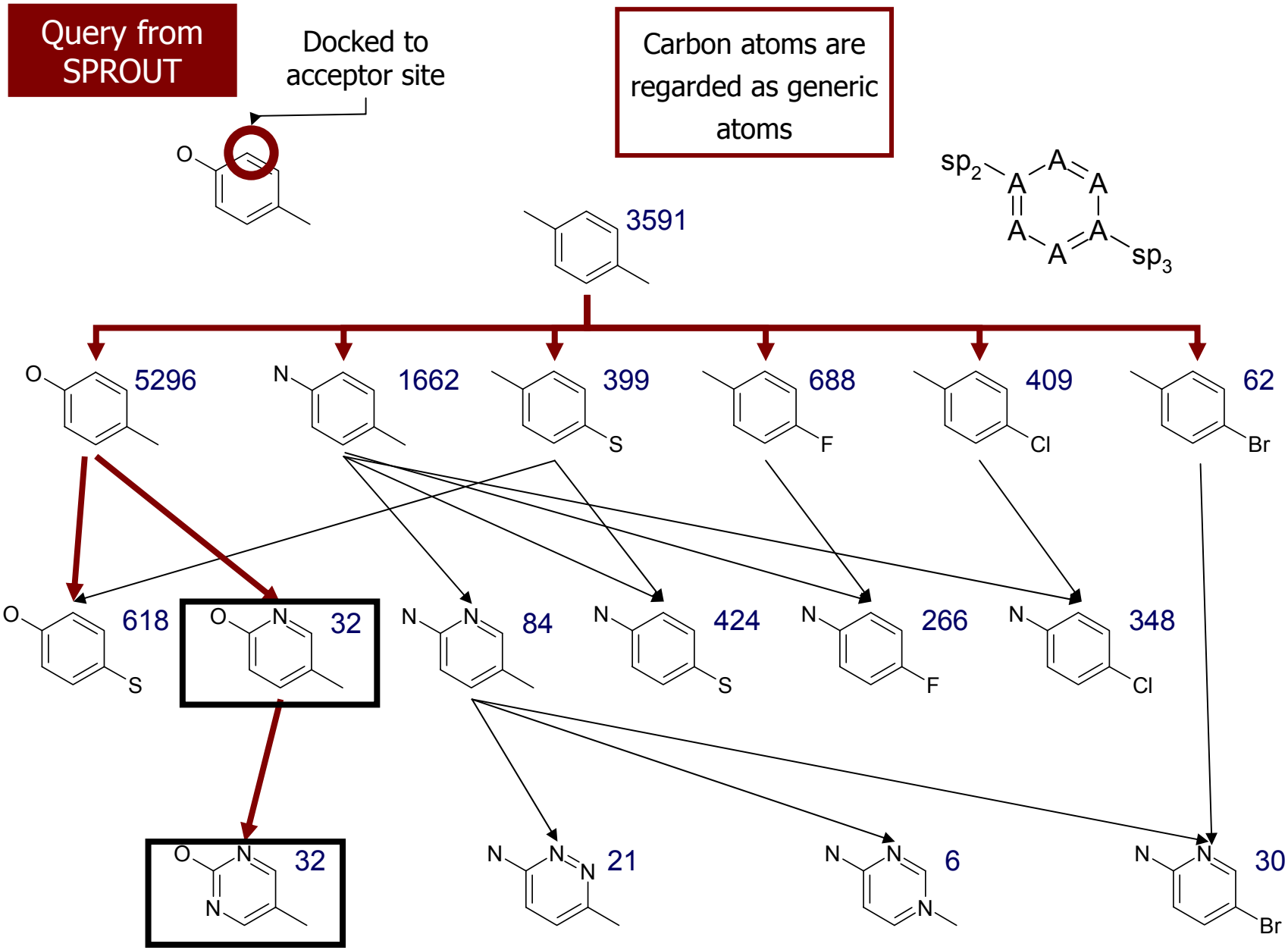
5. Rank structures by complexity score

DATABASE of hierarchies + frequency of occurrences

4. Retrieval of frequency of occurrences → Calculate score

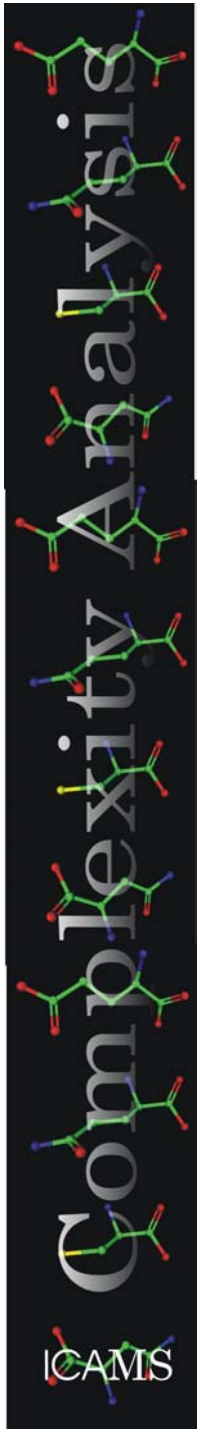
Speed of Complexity Analysis ~ 1000-1200 structures / minute on Linux PC (3GHz)

Query Matching in Hierarchy



Complexity Analysis

ICAMS



Calculation of Complexity Score

Concept

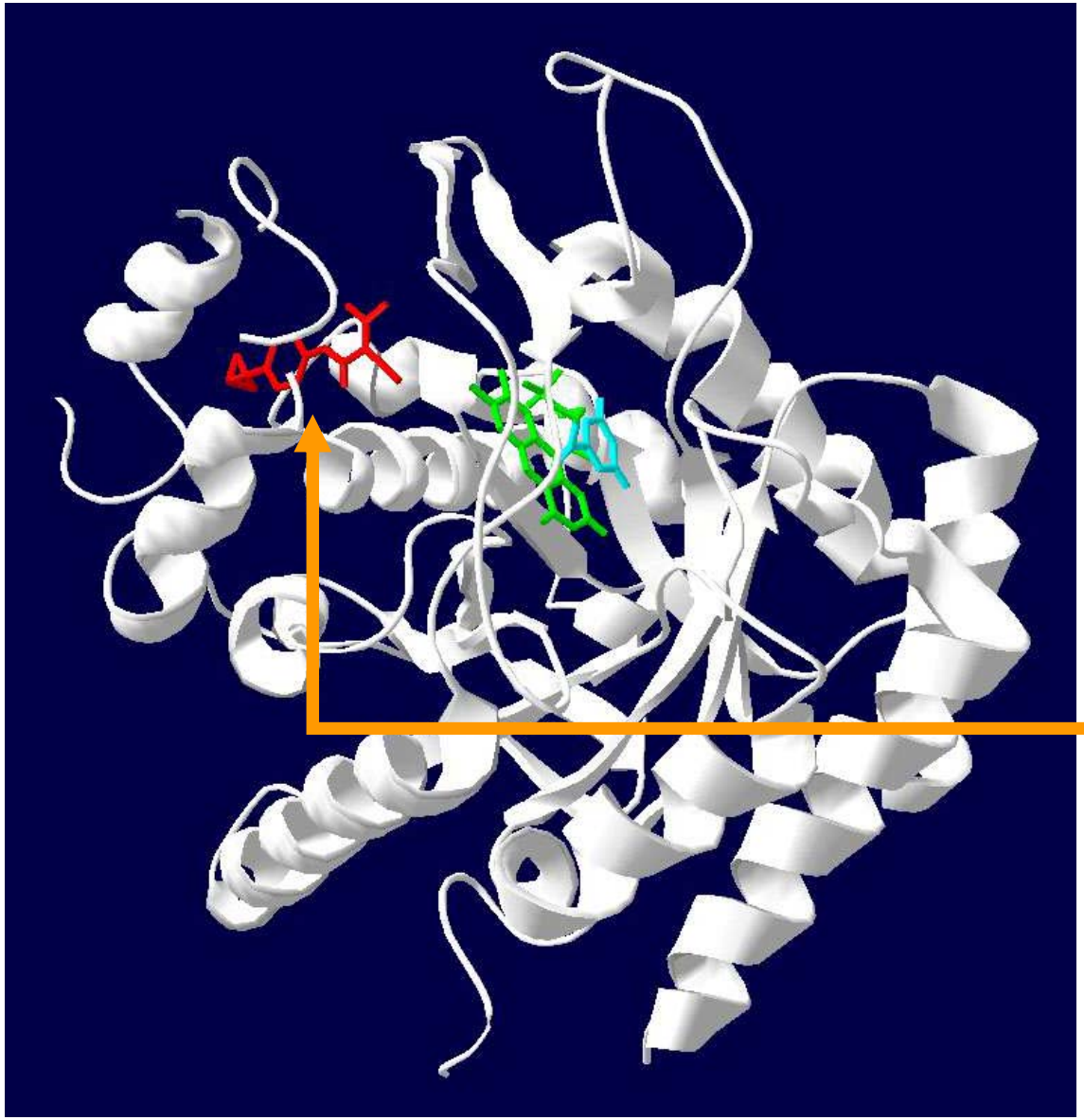
Penalise atom patterns which are infrequent or not present in the complexity database.

$$\text{Total}_{\text{COMPLEXITY SCORE}} = \frac{\sum \text{Pattern}_{\text{COMPLEXITY SCORE}}}{\text{Num of Patterns}}$$

$$\text{Pattern}_{\text{COMPLEXITY SCORE}} = \begin{cases} \left(1 - \dfrac{\ln(\text{Matched Occurrences})}{\ln(\text{Max Occurrences})}\right) * \textbf{Penalty} & \text{, if pattern exists} \\ 2 * \textbf{Penalty} & \text{, if pattern does not exist} \end{cases}$$

The complexity analysis is followed by ranking the putative ligands according to their evaluated complexity score.

Case Study (Dihydroorotare Dehydrogenase)



Attractive target enzyme for the development on new anti-malarial agents [1]

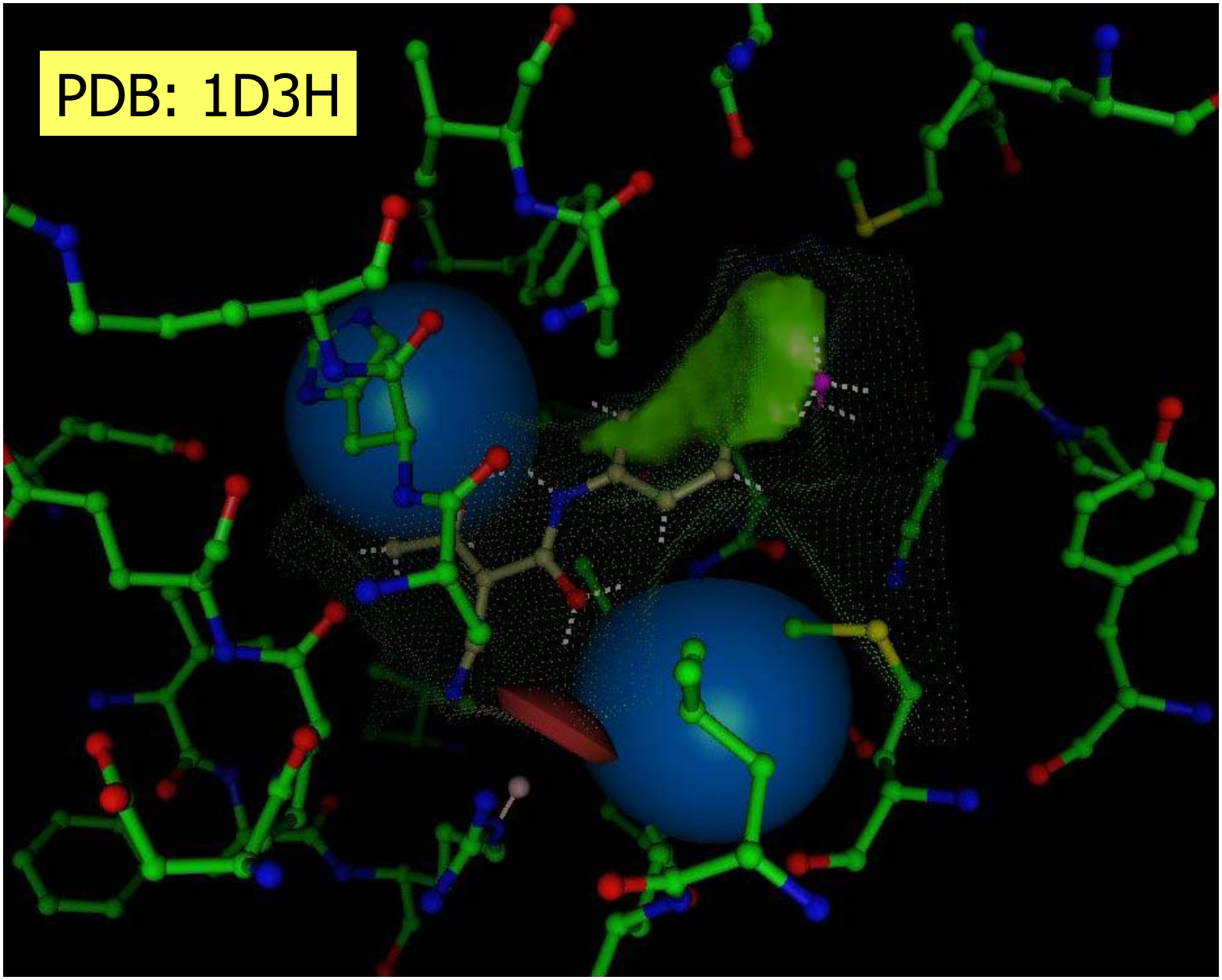
Pocket occupied by A77 1726 is targeted to propose new inhibitors

[1] Jeffrey Baldwin, Azizeh M. Farajallah: The Journal of Biological Chemistry 2002 (No.44) pp 41827-41834

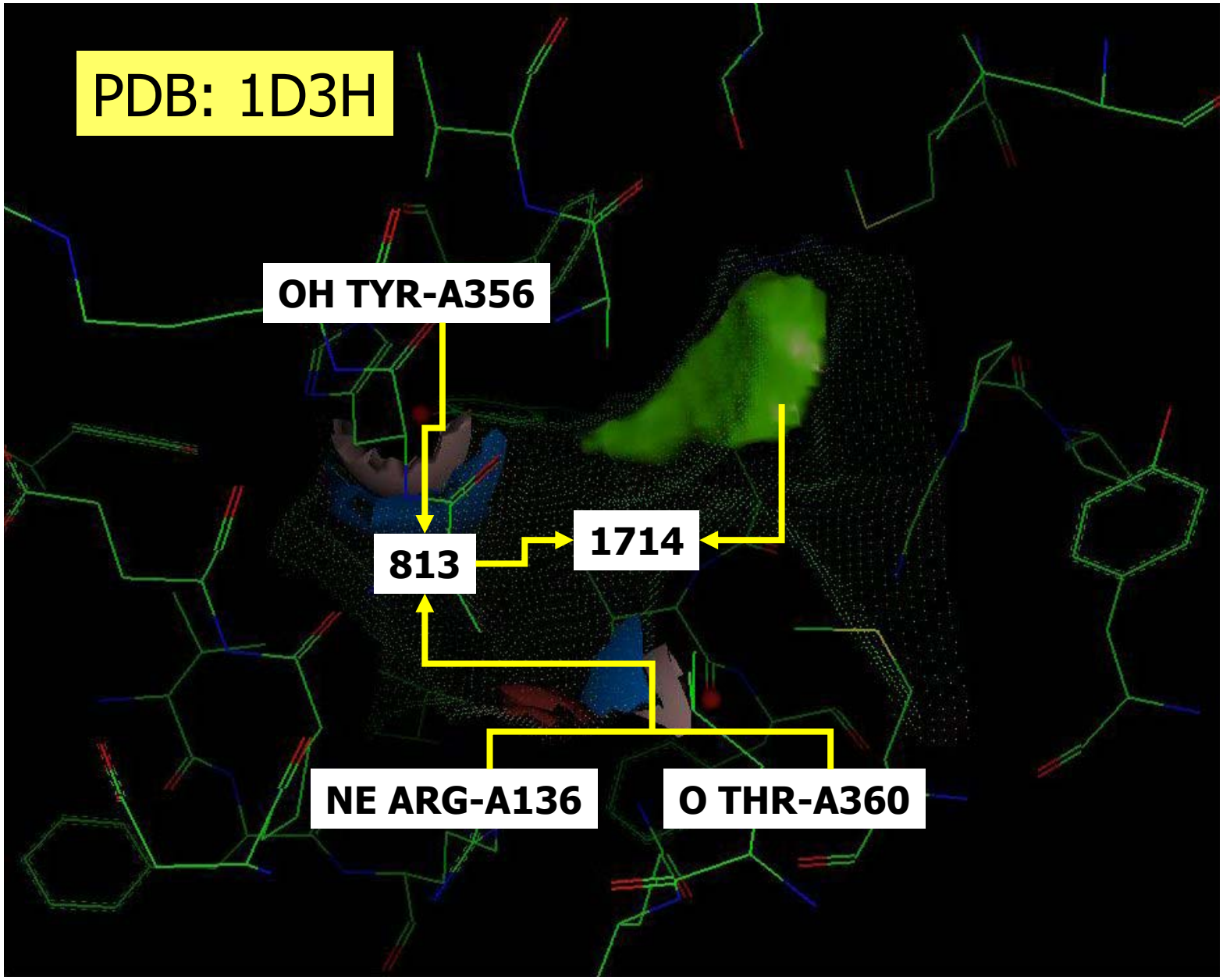
Complexity Analysis

ICAMS

Case Study (Structure Generation)



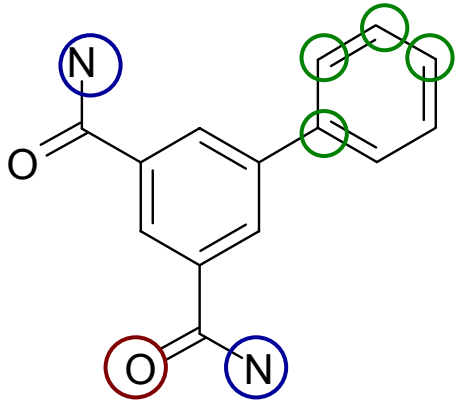
Case Study (Structure Generation)



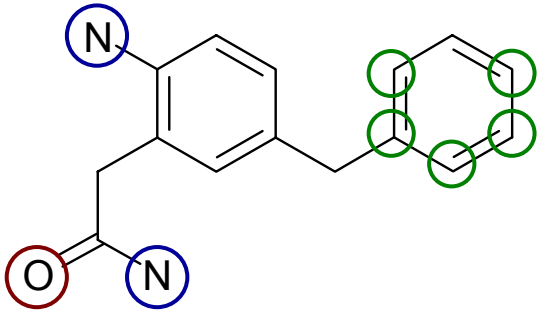
Complexity Analysis

ICAMS

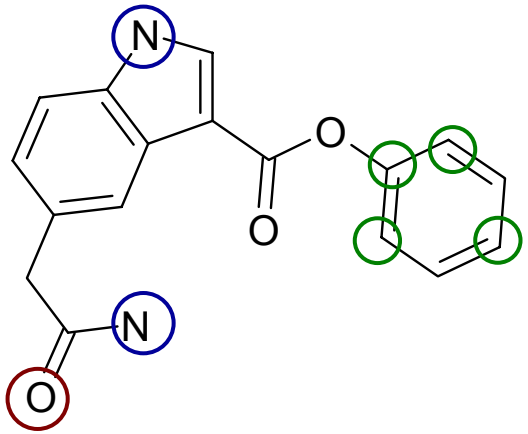
Case Study ("Simple" Structures)



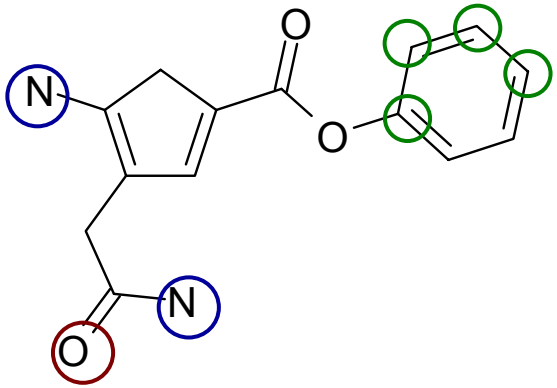
Complexity Score :	**14.29**
Binding Score :	**-8.42**



Complexity Score :	**14.63**
Binding Score :	**-7.53**

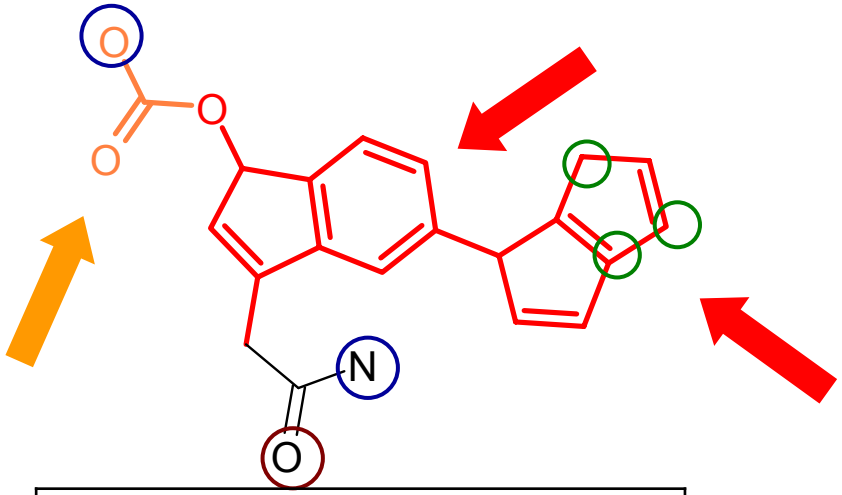


Complexity Score :	**18.64**
Binding Score :	**-9.35**

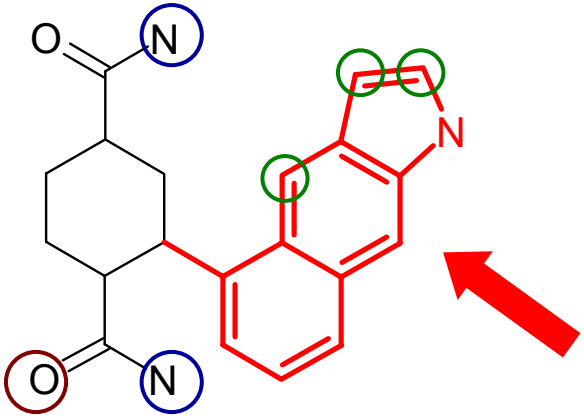


Complexity Score :	**19.92**
Binding Score :	**-7.79**

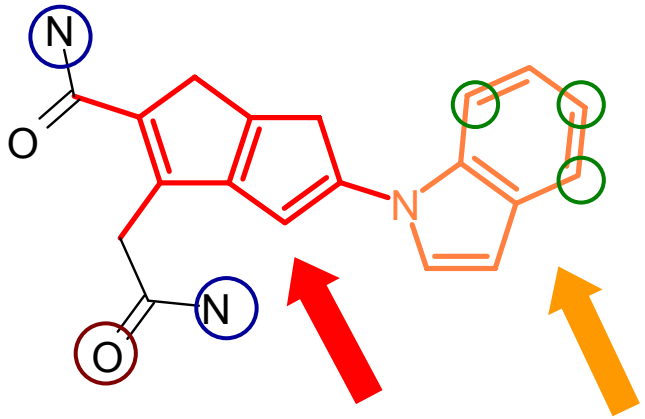
Case Study (Complex Structures)



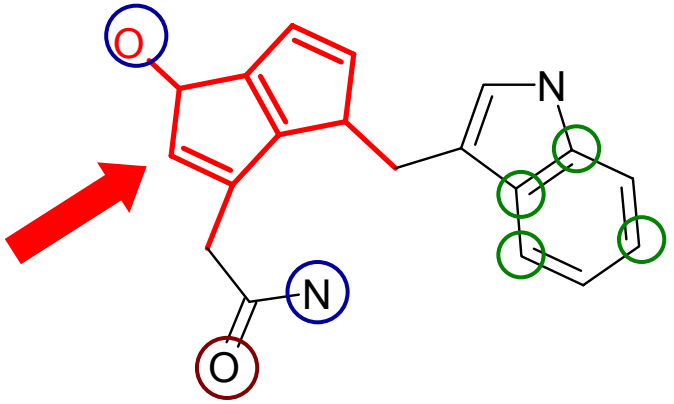
Complexity Score :	**54.77**
Binding Score :	**-10.07**



Complexity Score :	**52.37**
Binding Score :	**-10.80**



Complexity Score :	**48.47**
Binding Score :	**-8.04**

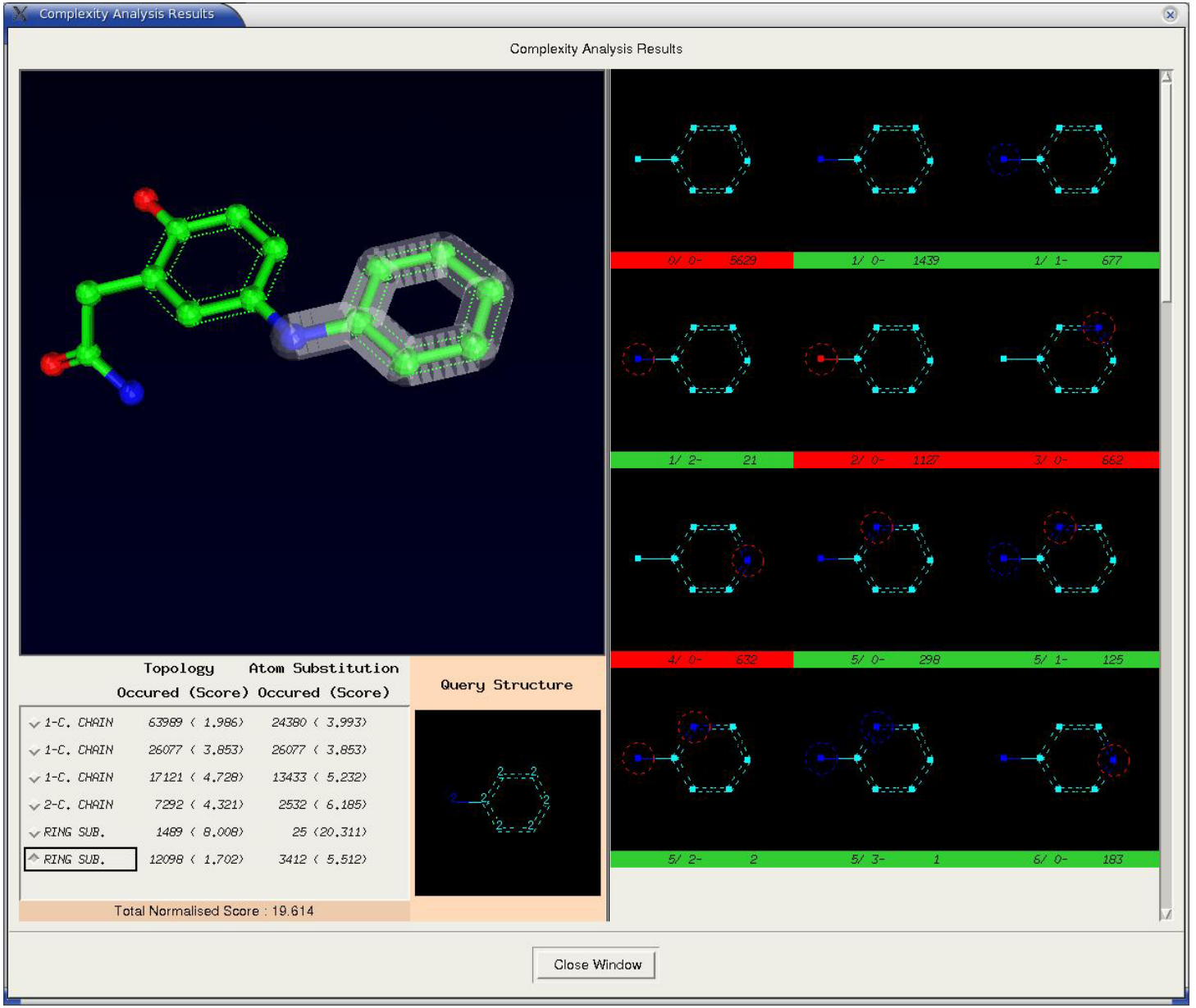


Complexity Score :	**35.42**
Binding Score :	**-9.92**

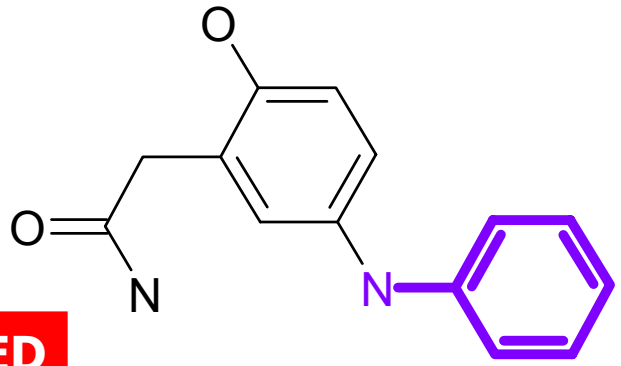
Complexity Analysis

ICAMS

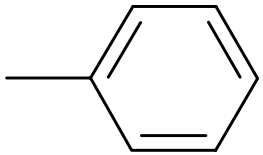
Atom Substitution Patterns

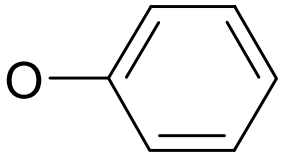


Atom Substitution Patterns

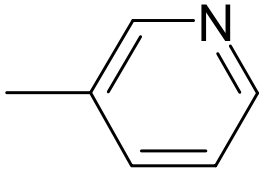


NOT MATCHED

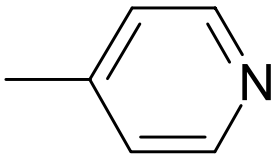

5629



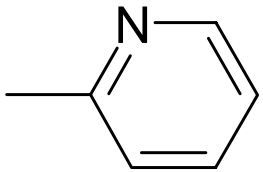
1129



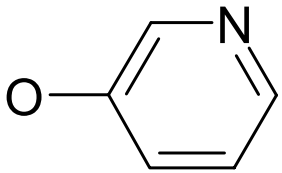
662



632

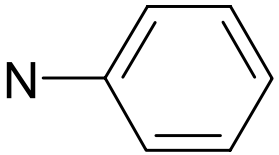


357

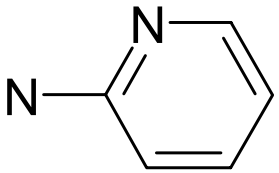


56

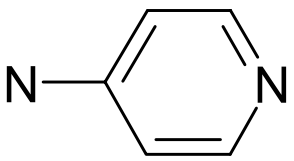
MATCHED



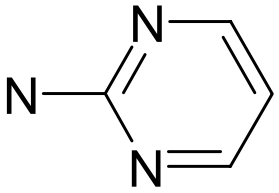
2137



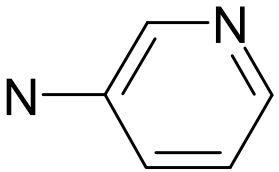
426



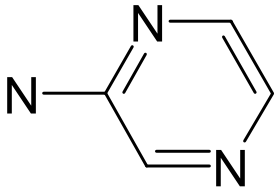
384



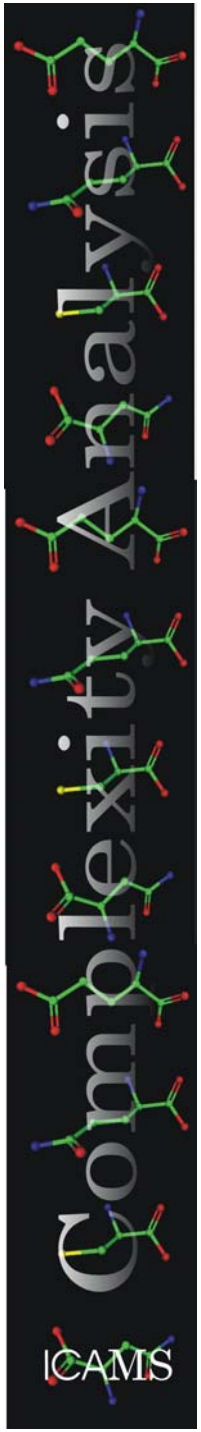
261



146



24



Conclusion

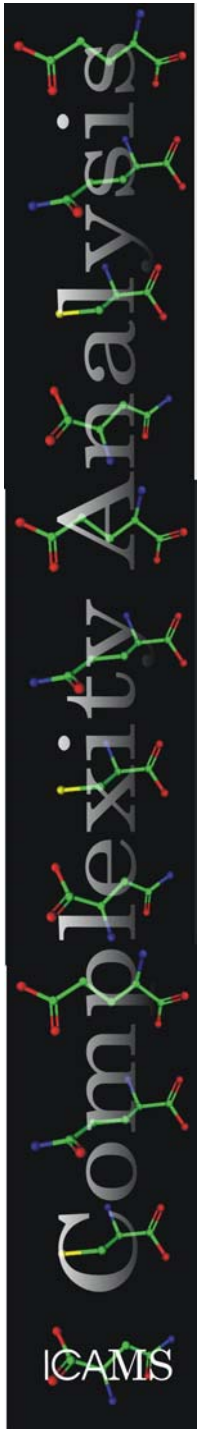
Complexity analysis based on structural motifs of existing drugs and compounds provides a fast and effective method to rank structures and eliminate complex structures prior to the computationally more expensive estimation of binding affinity.

Warning

This approach is based on characteristics of existing drugs and compounds

↓

Structures with novel structural features may get incorrectly penalised for being complex



Future work

Hetero atom substitution

Current method

Currently, only region of polar hydrogen target sites are substituted, leaving all other atoms intact.

Proposed method

The distribution of the substitution patterns can be utilised to drive hetero atom substitution.

Acknowledgement

- Prof. A.P. Johnson
- All past and present members of ICAMS
- MDL for providing the MDDR database