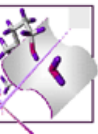


# Chemoinformatics @ UCB

**Will Pitt**

**UCB**

**Granta Park  
Cambridge  
UK**



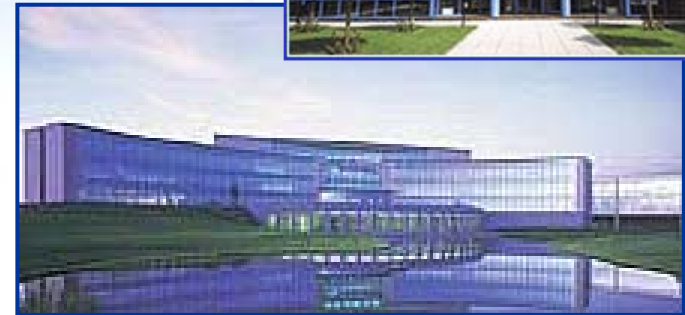
UKQSAR and  
Chemoinformatics Group  
Autumn Meeting 2005



No part of this presentation is to be reproduced without prior permission of the author

# UCB

- A global biopharmaceutical leader
- Headquarters in Brussels, Belgium
- Specialising in the fields of
  - central nervous system disorders
  - allergy and respiratory disease
  - immune and inflammatory disorders
  - oncology.
- Sales of €3 billion in 2003
- Key products include
  - Keppra (antiepileptic),
  - Xyzal and Zyrtec (antiallergics)
- R&D Sites in Cambridge, Slough and Brussels
- 1,700 people in R&D



# Chemoinformatics @ UCB

- Chemoinformatics
- Chemical Space
  - What is it ?
  - How big is it ?
- VEHICLE: heteroaromatic Space
  - Construction
  - Analysis
  - Synthetically accessible VEHICLE
  - Uses
- Recombinatorix: lead series space
- Summary

# Chemoinformatics

The application of statistical and mathematical techniques to problems in chemoinformatics

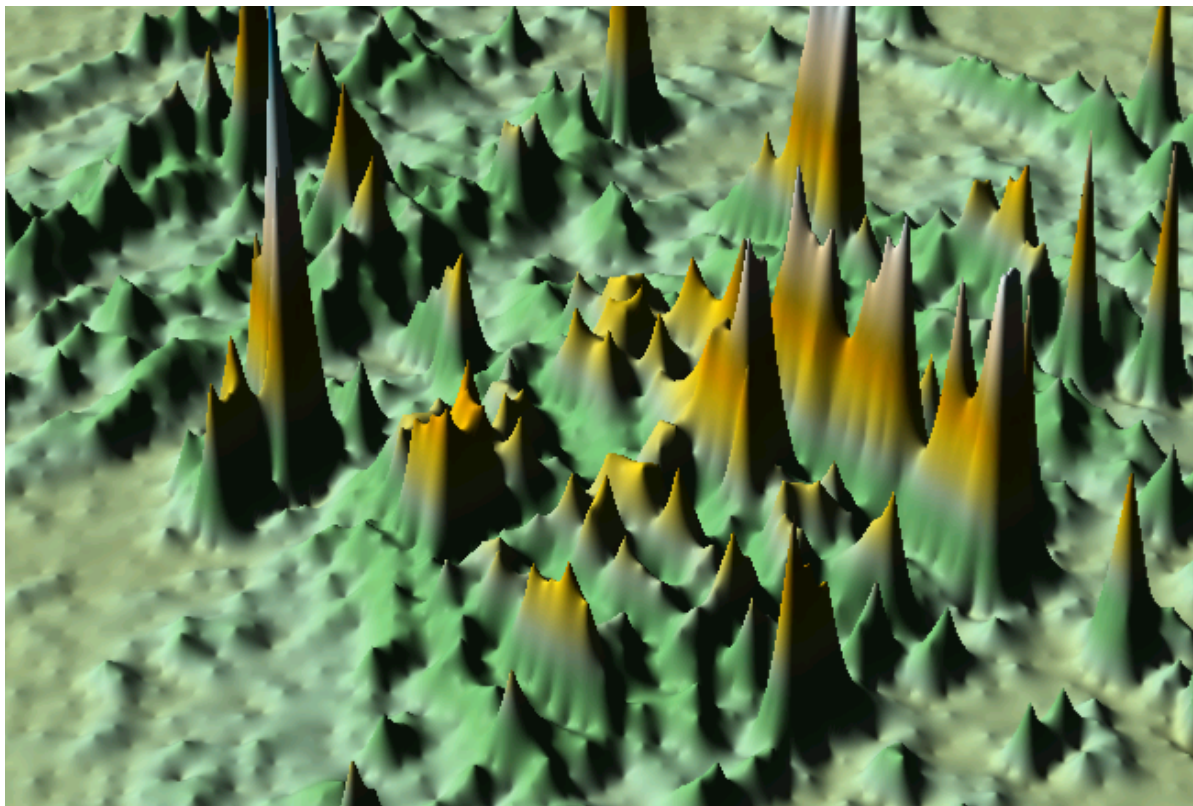
- This talk: chemical space analysis

# What is Chemical Space ?

- A complete collection of chemical structures
- Real or virtual compounds
- Can be fragments
- Viewed / analysed in 2 or more dimensions
  - Near neighbours and distant relatives
  - Similarity and diversity
  - Density, coverage, holes
  - Numbers of compounds: maximum possible, maximum available ...
- Axes : principal components, BCUTS etc.
- Often purely abstract concept

# Chemical Space : Examples

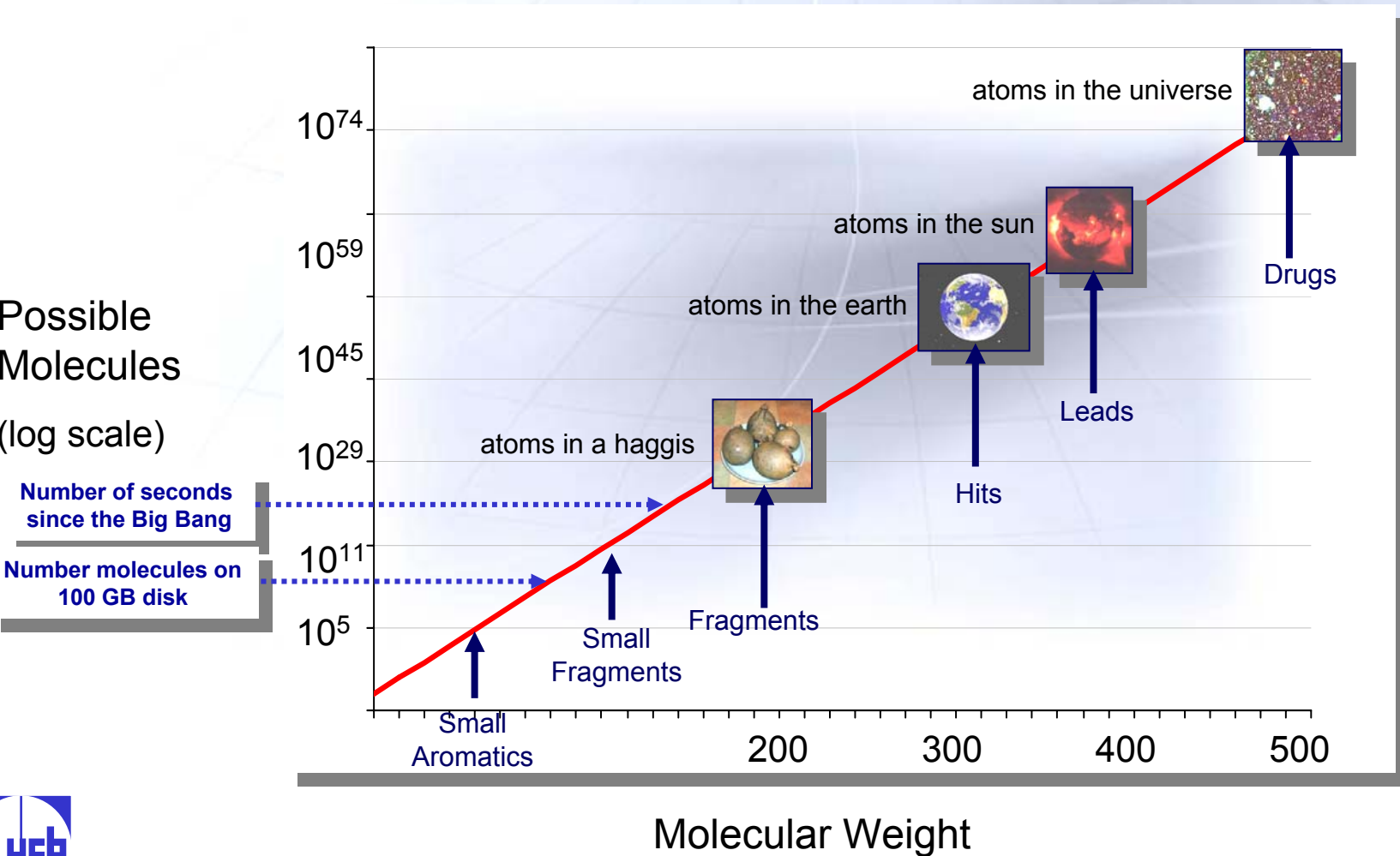
Detail of UCB screening deck in BCUT plane



# How Big is Chemical Space ?

- **Cayley 1875** *Cayley, A., Ber. Deutsch. Chem. Ges., 8 (1875), 1056-1059*
  - Enumeration of  $C_nH_{2n+2}$  alkanes e.g. 799 if  $n=13$
- **Bohacek 1996** *Bohacek R.S. et. al., Med.Res.Rev., 1996, 16, 3-50*
  - $10^{60}$  in universe of organic molecules: max. 30 atoms, 4 rings, 10 branches
- **Weininger 1998** *Weininger, D., Enc. Comp. Chem. 1998, 1, 425-430*
  - $10^{29}$  Small n-Hexane derivatives, 150 substituents
- **Ertl 2003** *Ertl, P., J. Chem. Inf. Comput. Sci., 2003, 43, 374-380*
  - $10^{20} - 10^{24}$  estimate from fragmentation of available molecules
- **Leach 2004** *A. Leach, Practical Introduction to Chemoinformatics, Univ. Sheffield, 2004*
  - $10^8$  monoamides,  $10^7$  secondary amines,  $10^{11}$  diamides from reagents in ACDC
- **Fink 2005** *Fink, T. et. al., Angew.Chem.Int.Ed., 2005, 44, 2-6*
  - 14 M compounds, max 11 atoms
- **UCB 2005**
  - Super approximation is  $14^{(n \times 1.72)}$  where n is number of atoms
  - Estimated by growing smiles strings with simples rules

# How big is chemical space?

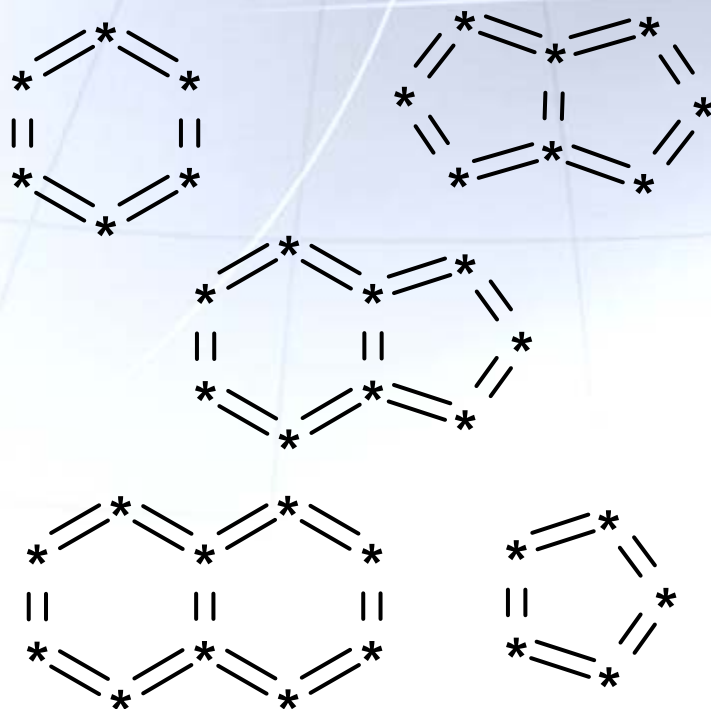


# Chemoinformatics @ UCB

- Chemoinformatics
- Chemical space
  - What is it ?
  - How big is it ?
- **VEHICLE: heteroaromatic space**
  - Construction
  - Analysis
  - Synthetically accessible VEHICLE
  - Uses
- Recombinatorix: lead series space
- Summary

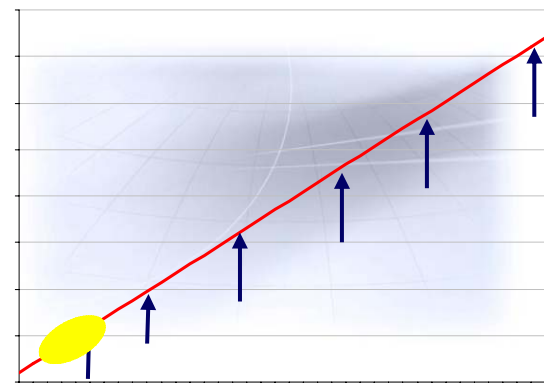
# Virtual Exploratory Heterocyclic Library (VEHICLE)

- Complete enumeration of all aromatic monocycles and bicycles



# Why Planar Heteroaromatics ?

- Tractable Size
- Pharmaceutical considerations
  - Compounds get larger during 'optimisation'
  - Smaller compounds more likely to show activity
  - Smaller compounds are more 'drug-like'
  - May be more 'novelisable'
  - Scaffold of molecule important  
source of novelty of chemical series
- Planar heteroaromatics are key to a medicinal chemists thinking

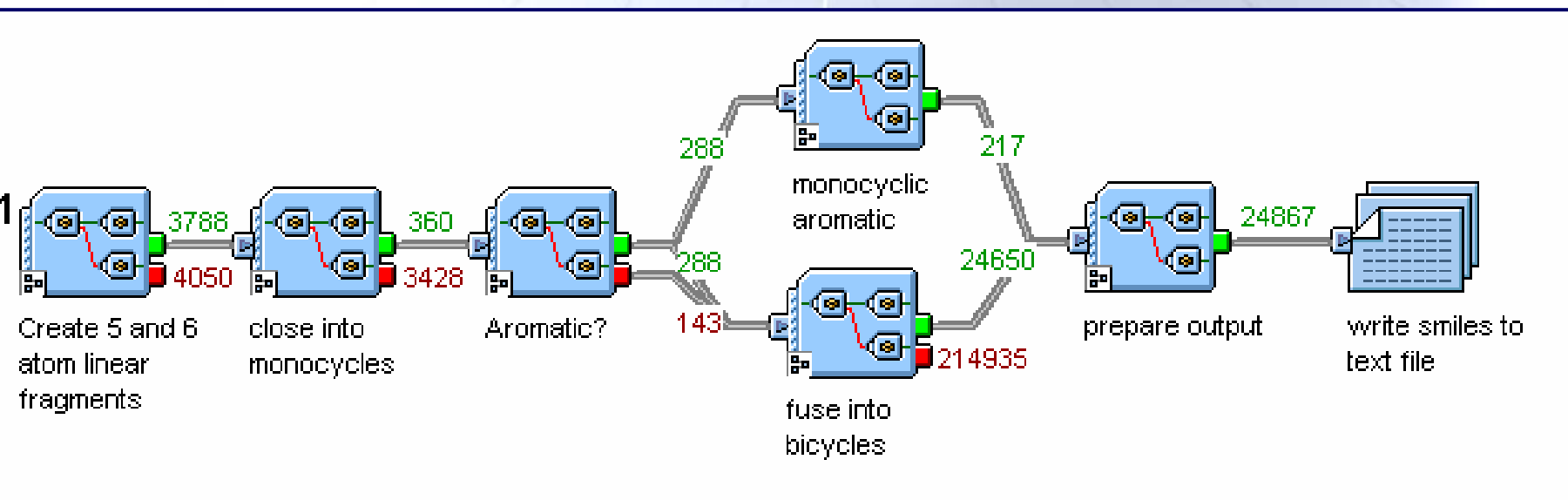


# VEHICLE

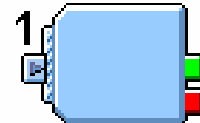
## boundaries

- 5 and 6 membered rings
- C,N,S,O
- Neutral
- No tautomers
- Obey Hückel's rule
- Only exocyclic carbonyls and always as cyclic amides or ketones

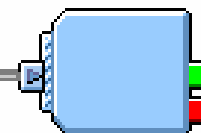
# VEHICLE Construction Pipeline



Elapsed Time: 2 Minutes 23 Seconds



Enumerate  
Combinatorial  
Reaction

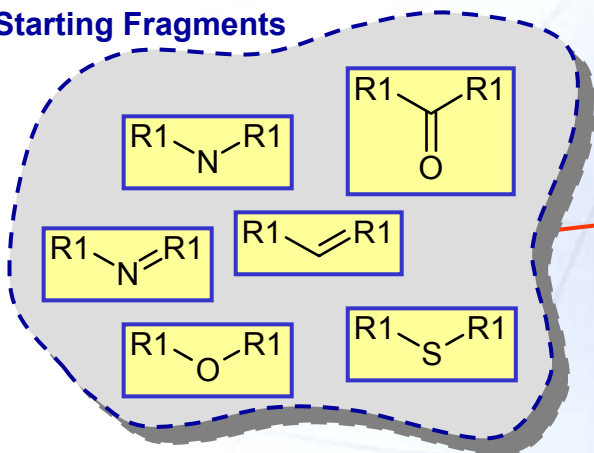


Perform  
Reaction on  
each Molecule

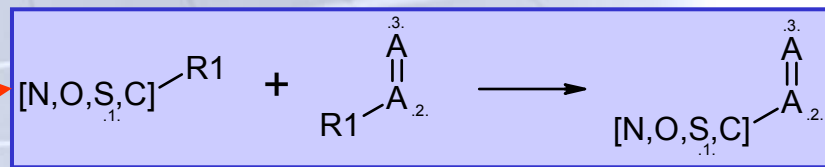
# VEHICLE Construction

## Example "Reactions"

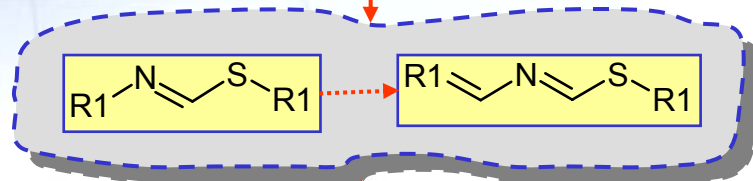
### Starting Fragments



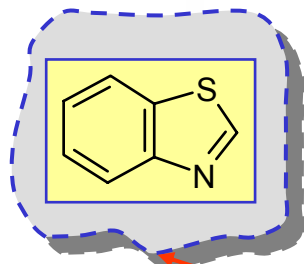
### Chain Formation



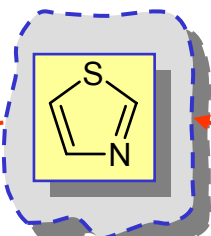
### Chains



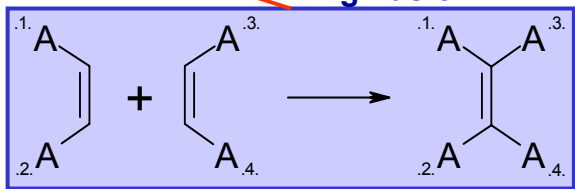
### Bicycles



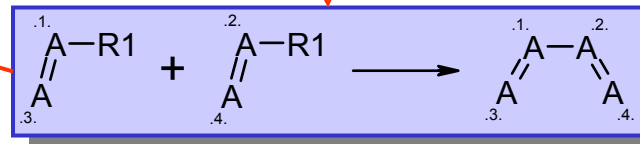
### Monocycles



### Ring Fusion



### Ring Closure

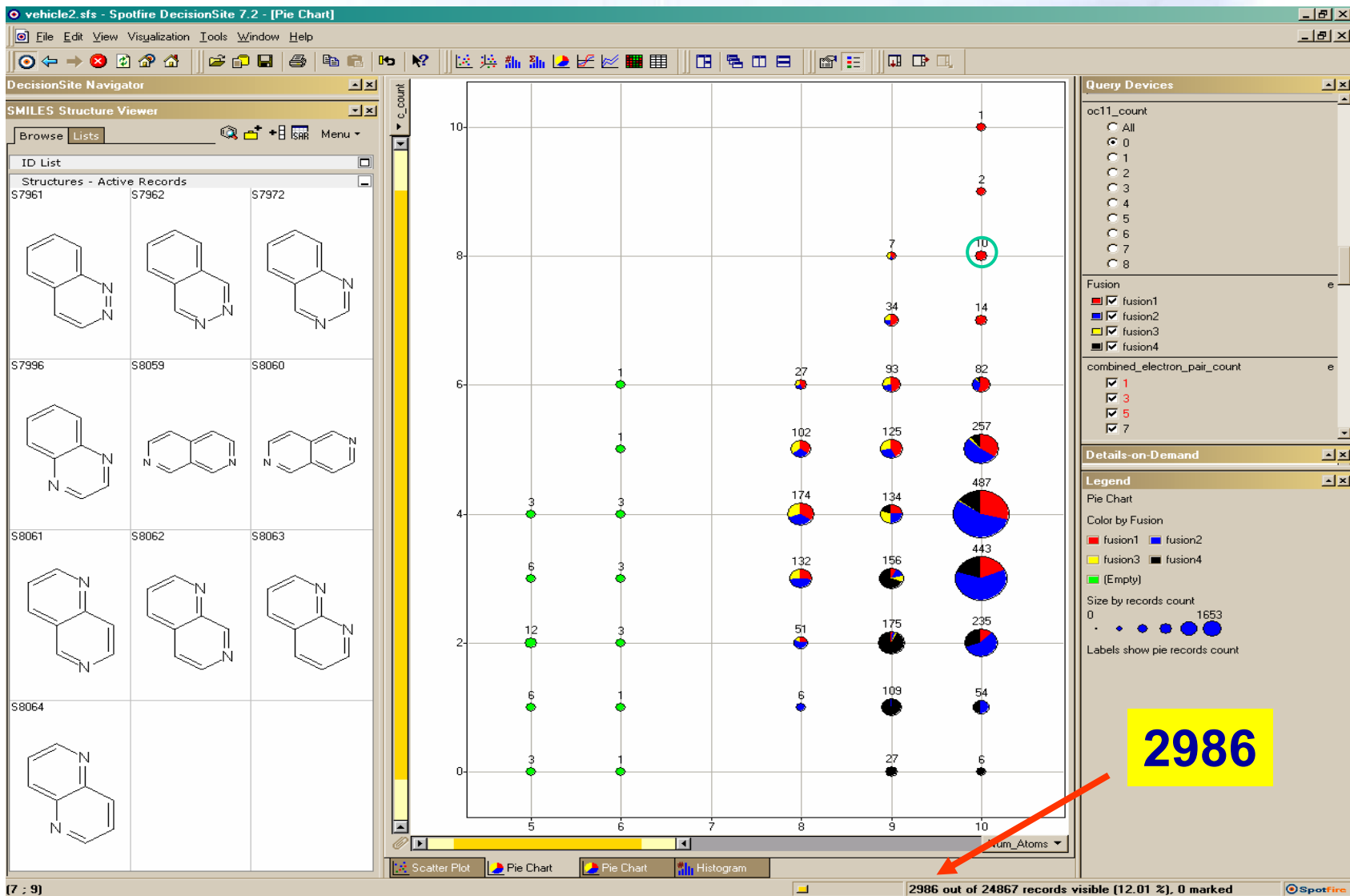


# Chemoinformatics @ UCB

- Chemoinformatics
- Chemical Space
  - What is it ?
  - How big is it ?
- VEHICLE: heteroaromatic Space
  - Construction
  - Analysis
  - Synthetically accessible VEHICLE
  - Uses
- Recombinatorix: lead series space
- Summary

# VEHICLE Analysis

## Mono and Bicycle Space

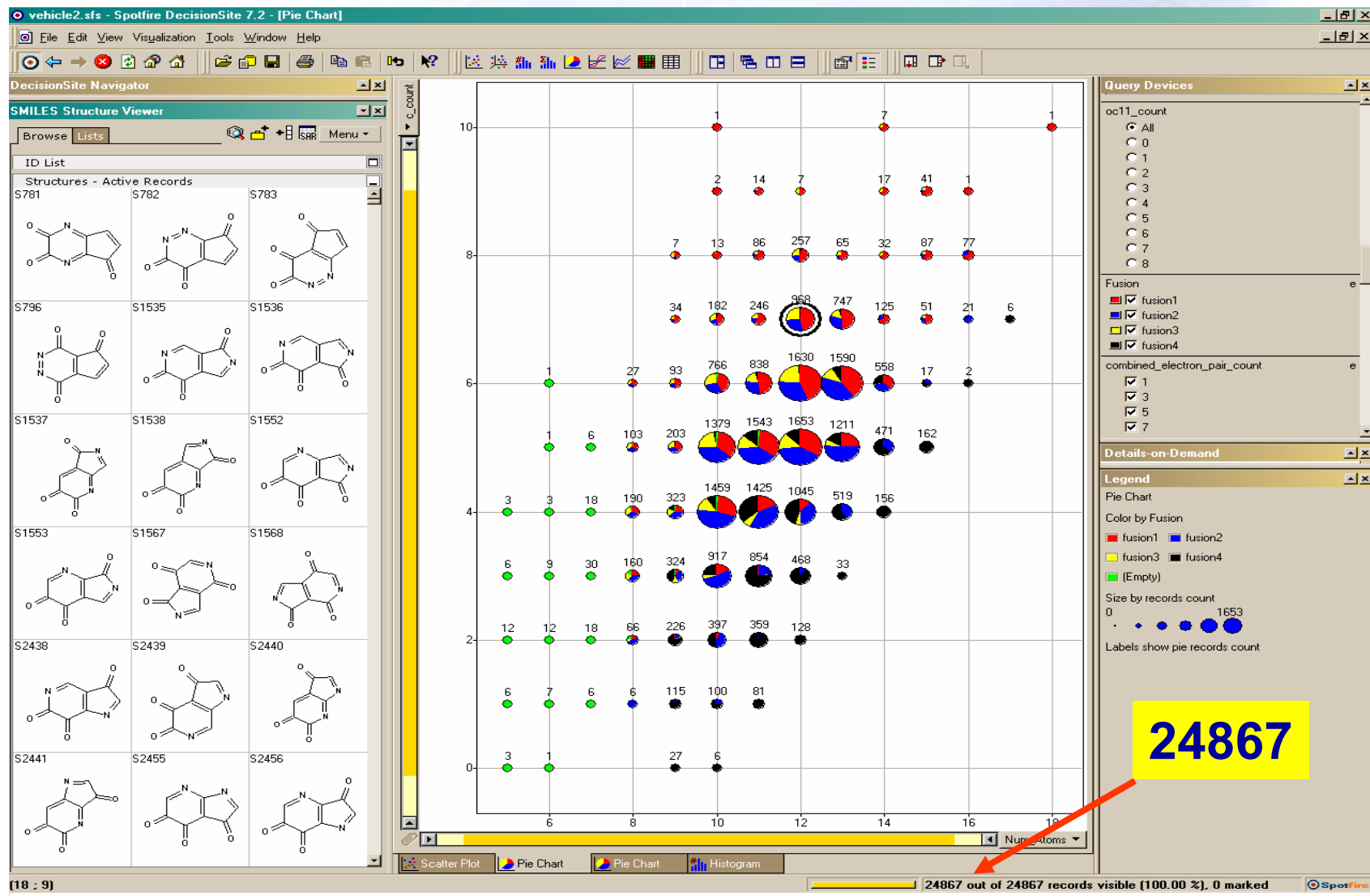


**2986**



# VEHICLE Analysis

## + Exocyclic Carbonyls

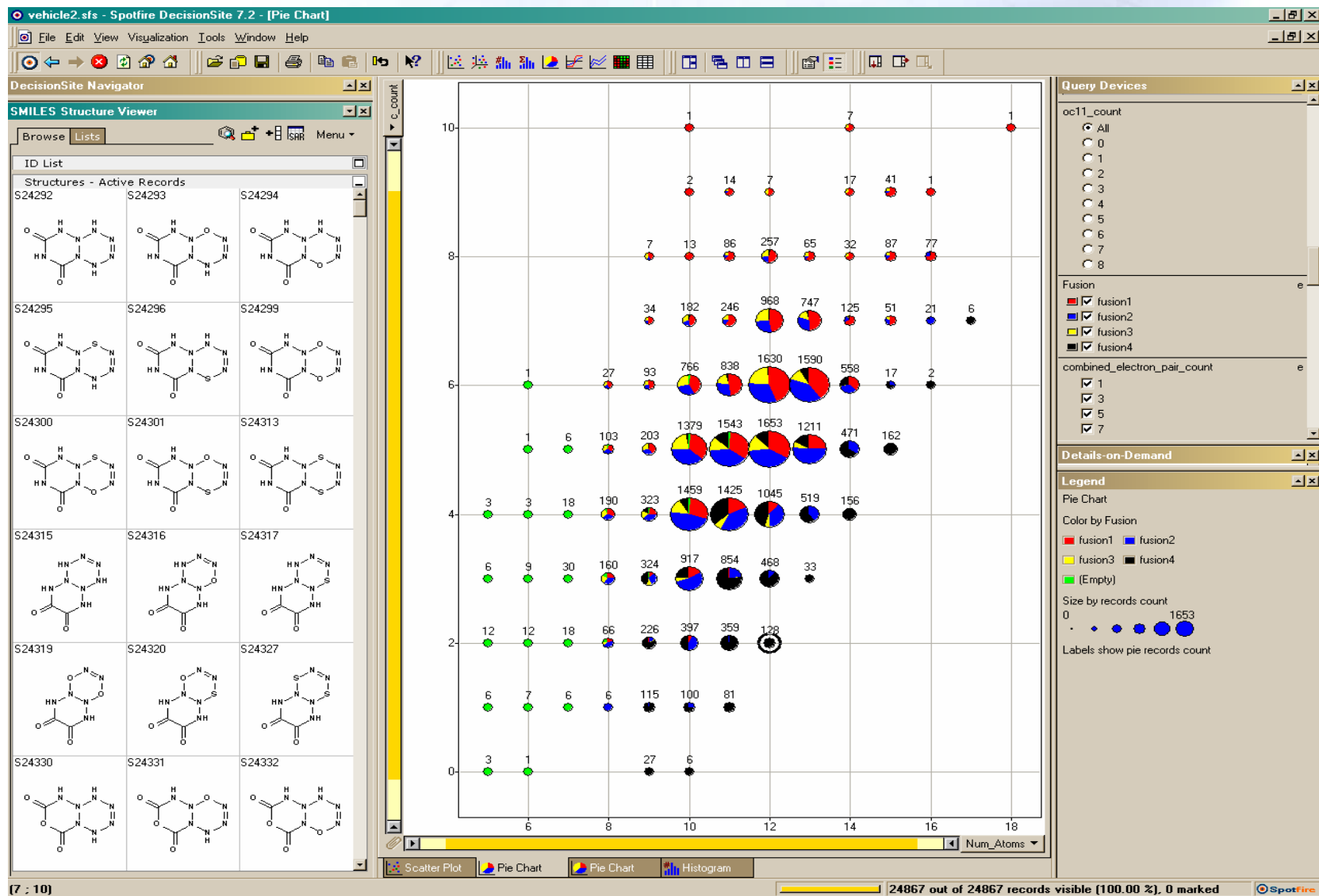


24867



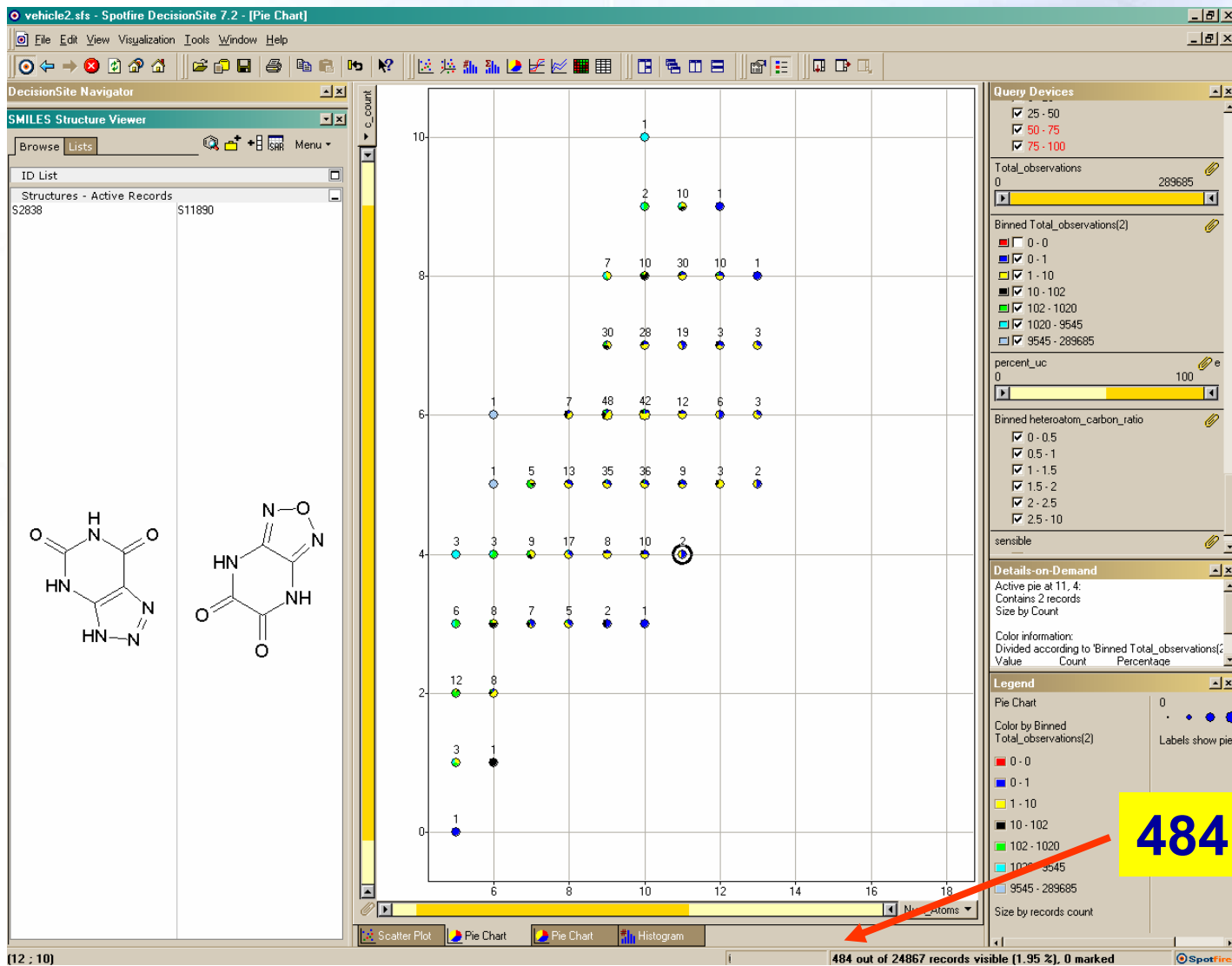
# VEHICLE Analysis

## The Outer Reaches of Aromatic Space



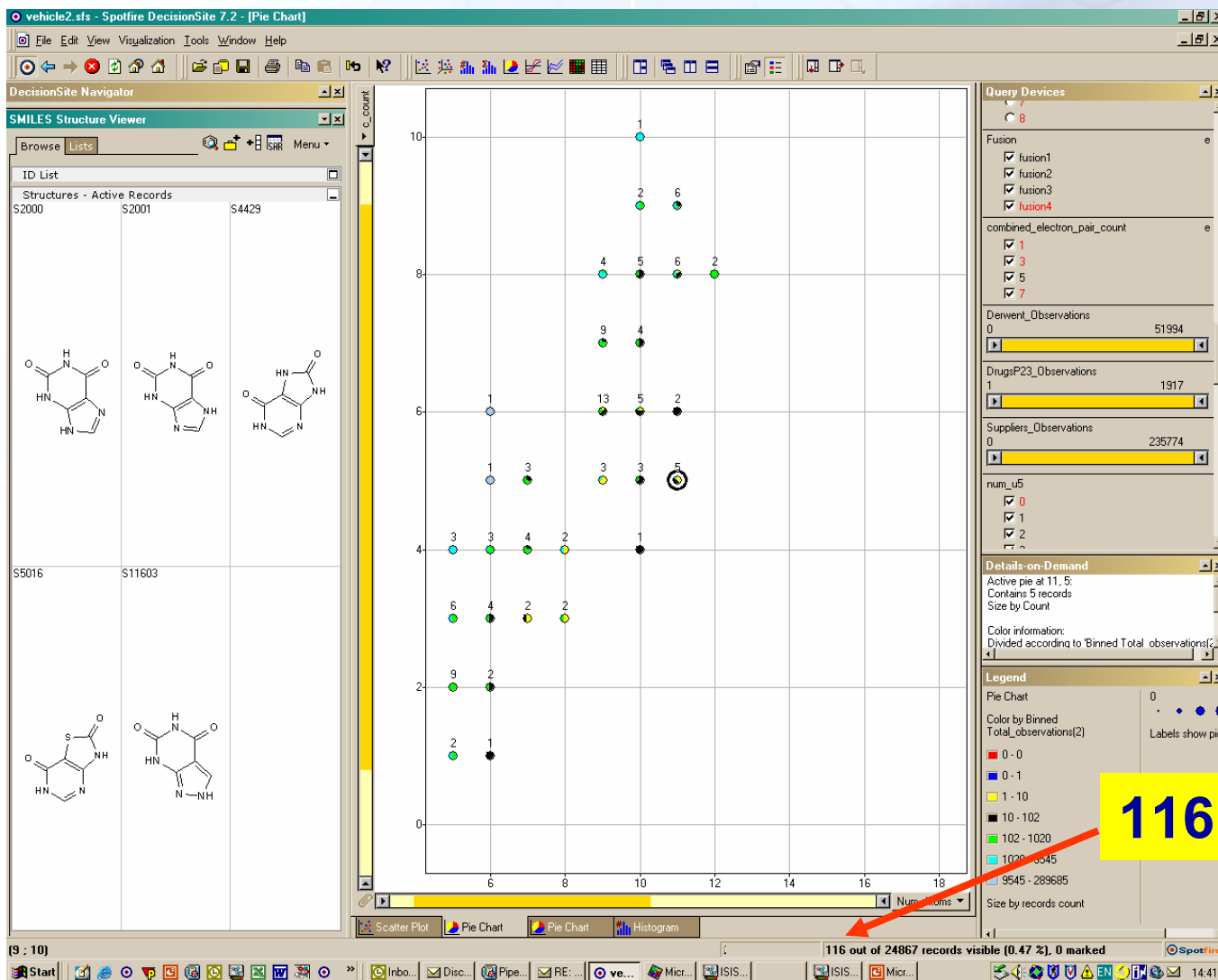
# VEHICLE Analysis

## Available Heterocycles



484 (2%)

# VEHICLE Analysis in Drugs



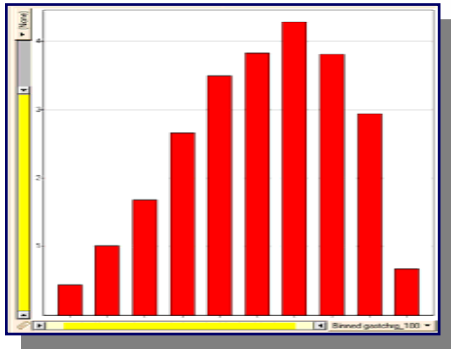
**116 (0.5%)**



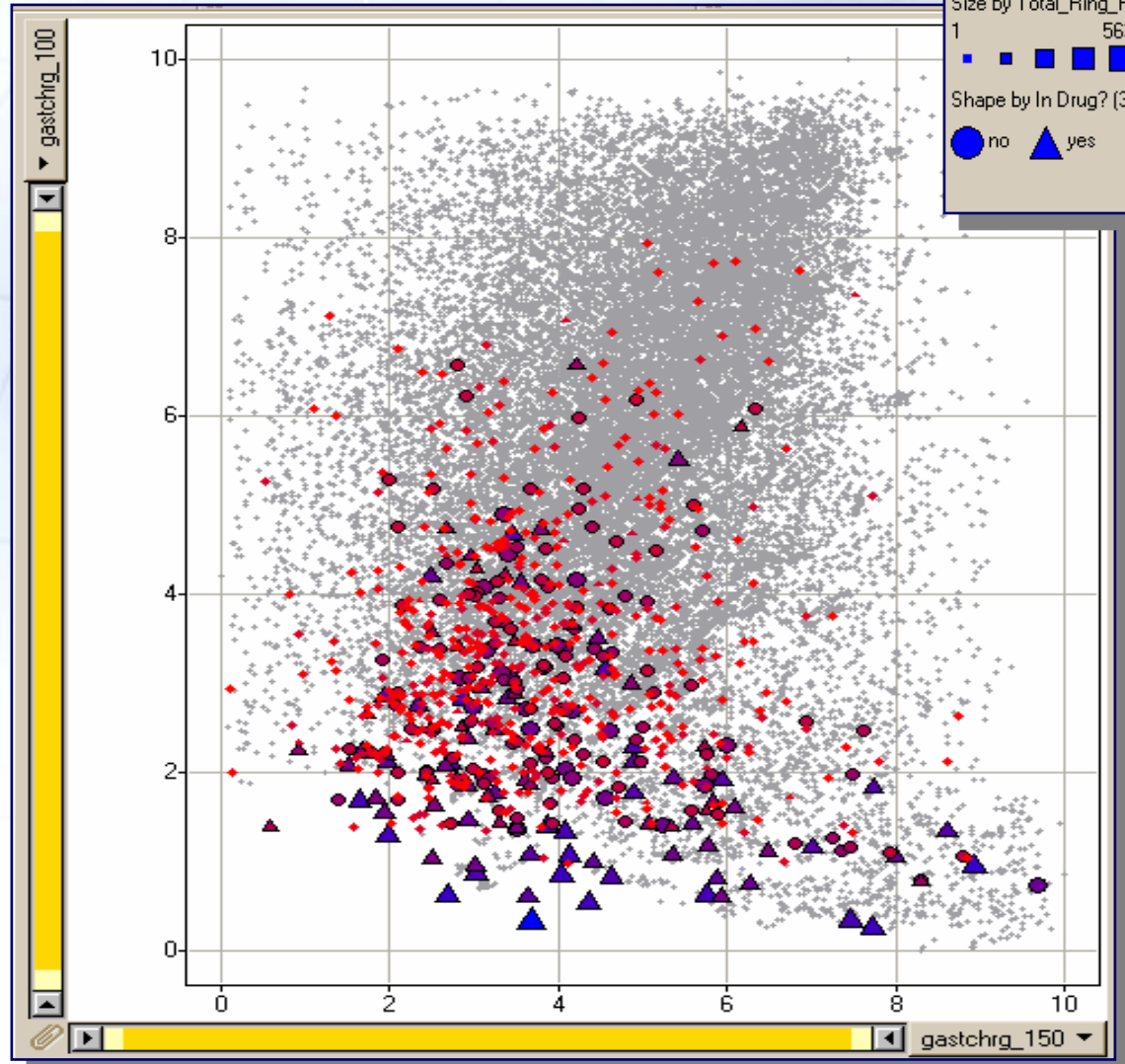
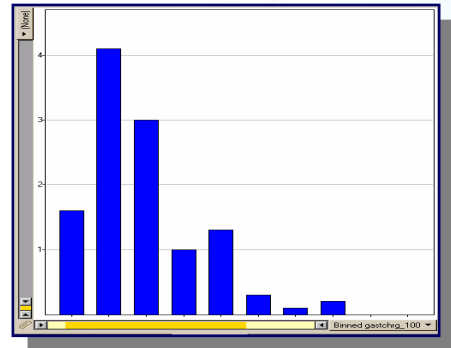
# VEHICLE Analysis

## BCUT space

All VEHICLE



Present in Drugs



# Chemoinformatics @ UCB

- Chemoinformatics
- Chemical Space
  - What is it ?
  - How big is it ?
- VEHICLE: heteroaromatic Space
  - Construction
  - Analysis
    - Synthetically accessible VEHICLE
  - Uses
- Recombinatorix: lead series space
- Summary

# Synthetically Accessible VEHICLE

Quantitative Structure Synthesisability Relationships (QSSR)

- Predict which compounds/fragments are synthetically feasible
- Train using known compounds
- Develop method to separate
  - synthesised and synthesisable: **good**
  - difficult or impossible to synthesise: **bad**
- Assume most compounds not made already are difficult or impossible to make (in near future)
- Use completely enumerated sets, such as VEHICLE
- Use whole compound set for training
- Use unseen data to gauge power of prediction

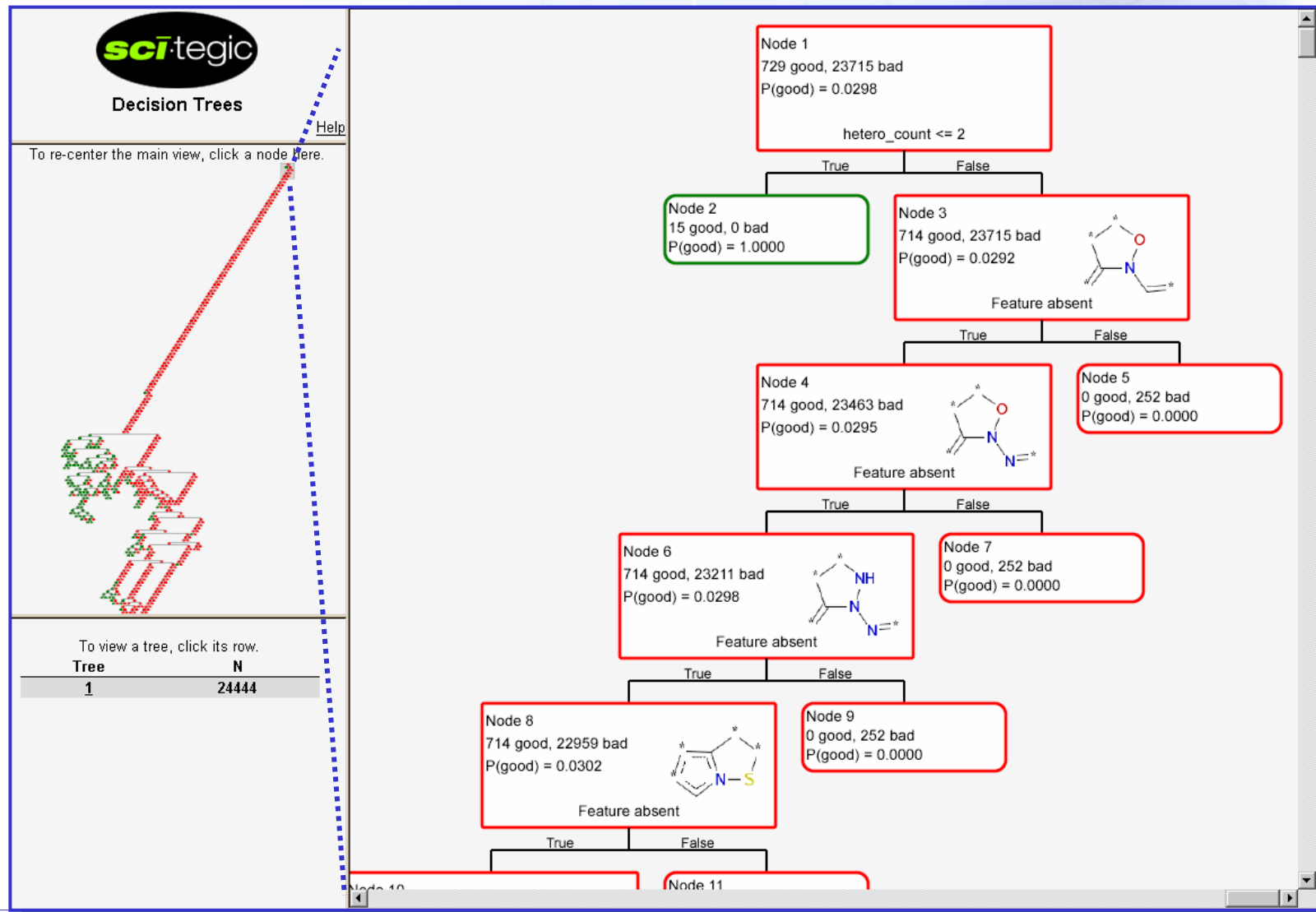
# Synthetically Accessible VEHICLE

## Decision Tree Classification

- Buja pure bucket split method
- R implementation via Scitegic PipelinePilot interface
- Descriptors: ECFP\_4, Num of heteroatoms etc.
- Classify into Good or Bad
- Single tree with all VEHICLE included

# Synthetically Accessible VEHICLE

## Decision Tree Classification



# Synthetically Accessible VEHICLE

## Decision Tree Training Results

	Good	Bad
Predicted Good	729	1894
Predicted Bad	0	21821

# Synthetically Accessible VEHICLE

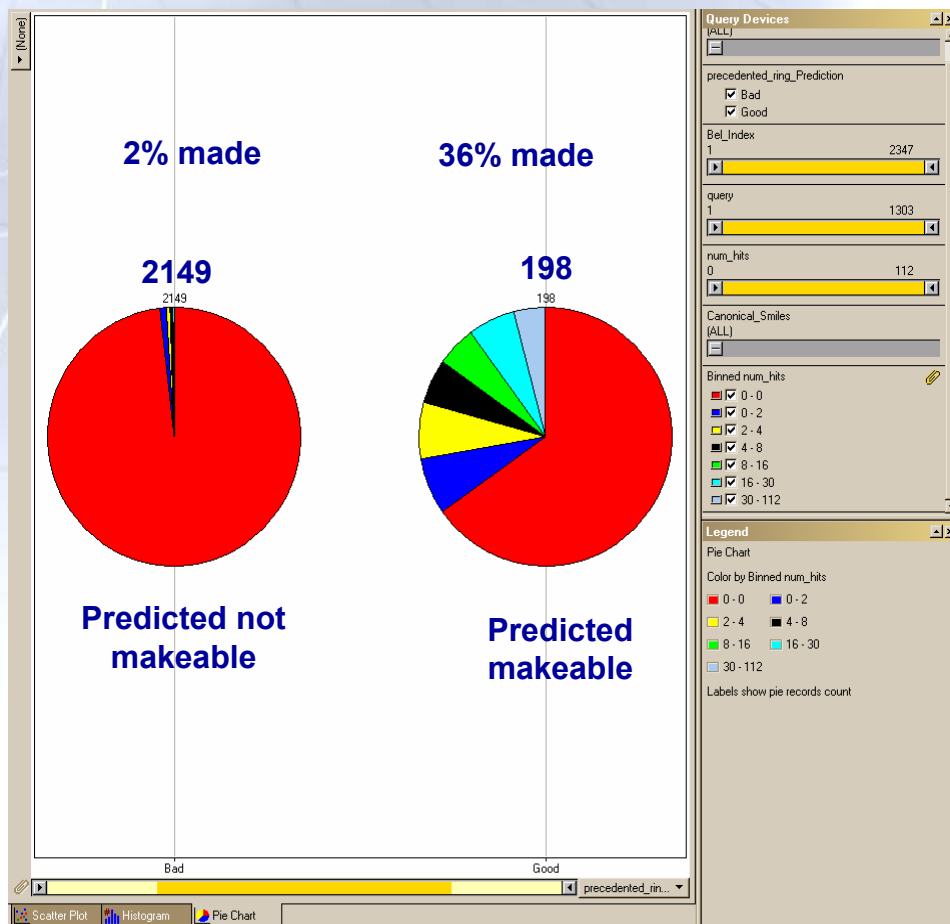
## Decision Tree Validation 1

- 36 heteroaromatic rings from UCB collection not in available compounds training set
- Tanimoto similarities to closest in training set 1.0 – 0.68
  - 12 regioisomers
  - 10 N insertions
  - 6 tautomers
  - 2 misc
  - 2 two differences
  - 3 not similar
- All predicted synthesisable

# Synthetically Accessible VEHICLE

## Decision Tree Validation 2

- Take random 10% of predicted 'good' and 'bad'
- Search Beilstein reaction literature database
- Have any compounds predicted bad been made?



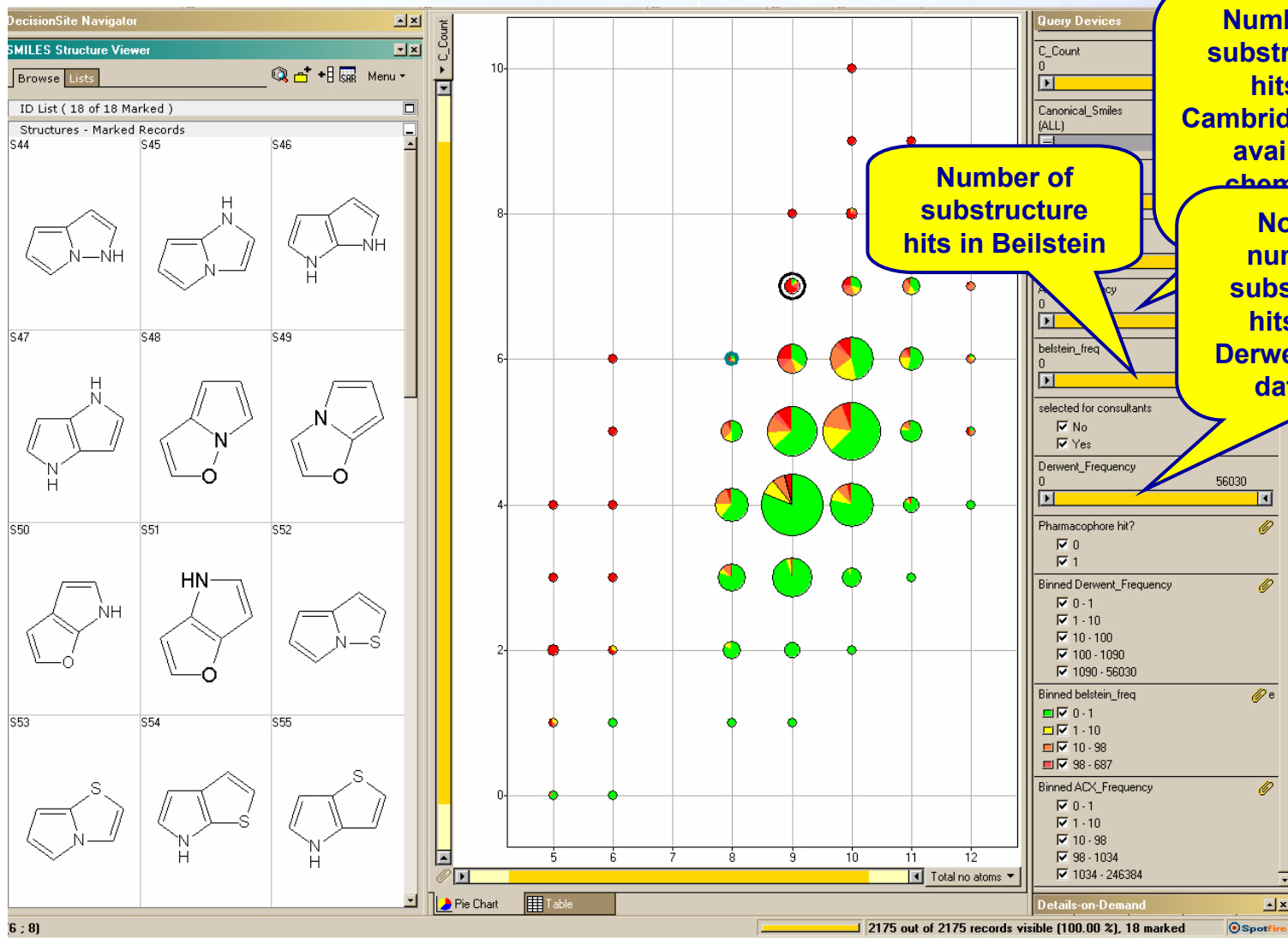
# Synthetically Accessible VEHICLE summary

- Interpolation, not extrapolation
- Cannot over train
- Estimated ring systems in the literature: 1847
  - $729 + (1894 \times 0.36) + (21821 * 0.02)$
- Estimated not in literature but possibly makeable : 1212
  - $1894 * 0.64 = 1212$
- Maybe a third of these are med. chem. interesting : 400
- **Potentially ~ 400 completely novel and interesting VEHICLE heteroaromatics available**

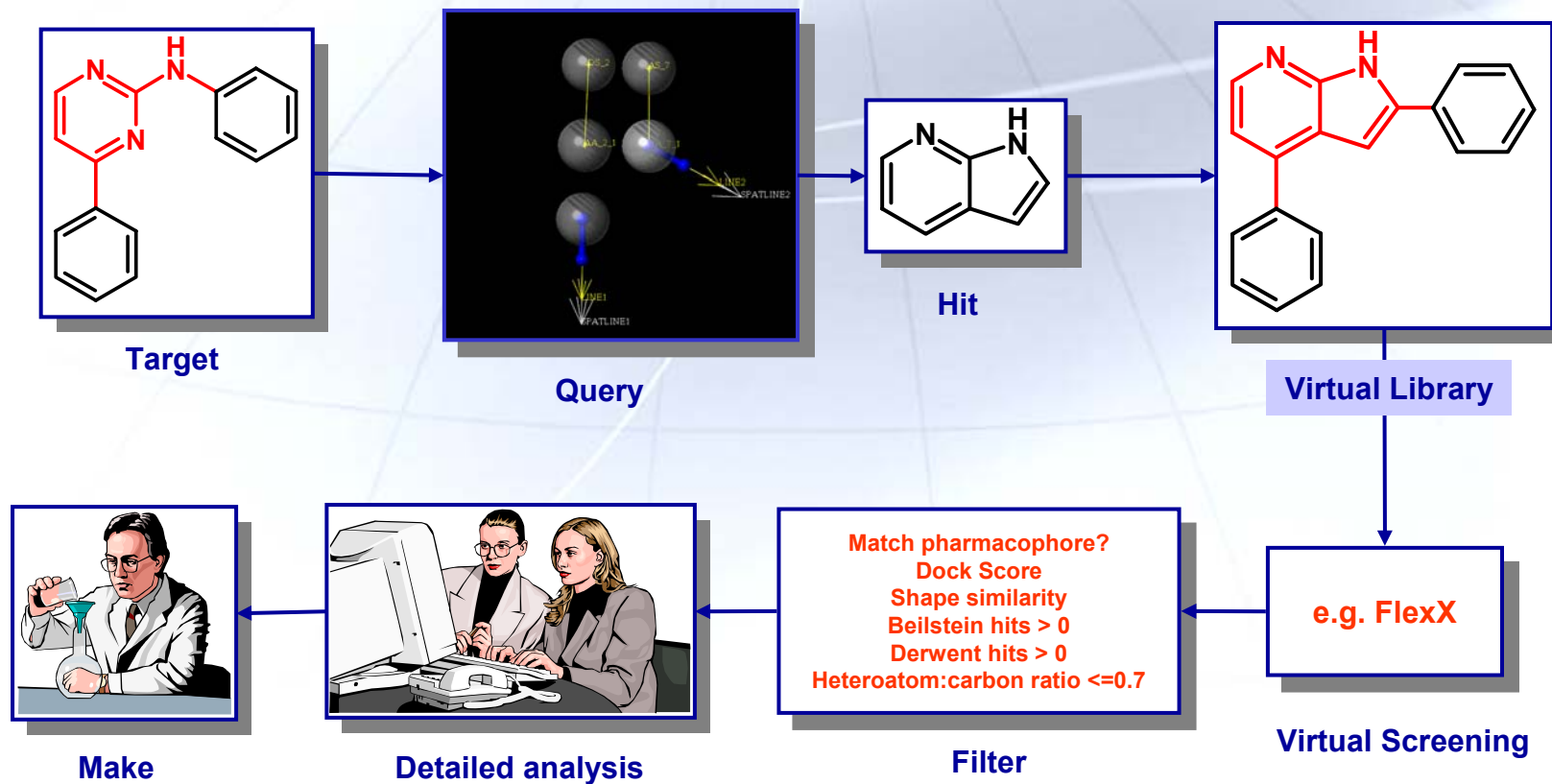
# Chemoinformatics @ UCB

- Chemoinformatics
- Chemical Space
  - What is it ?
  - How big is it ?
- VEHICLE: heteroaromatic Space
  - Construction
  - Analysis
  - Synthetically accessible VEHICLE
  - Uses
- Recombinatorix: lead series space
- Summary

# VEHICLE



# Unity Search of VEHICLE Scaffold Hopping



# VEHICLE Conclusions

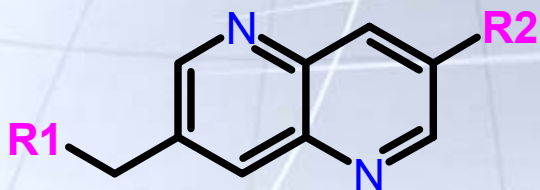
- Get the size of chemical space you are working in
- Structures yet to be made (novelty)
- Provides complete set for substructure searches – e.g. availability, novelty, synthetic routes
- Substructure replacement : complete set of options
- Novel method of predicting synthetic tractability (QSSR)

# Chemoinformatics @ UCB

- Chemoinformatics
- Chemical Space
  - What is it ?
  - How big is it ?
- VEHICLE: heteroaromatic Space
  - Construction
  - Analysis
  - Synthetically accessible VEHICLE
  - Uses
- Recombinatorix: lead series space
- Summary

# Molecular Recombinatorix

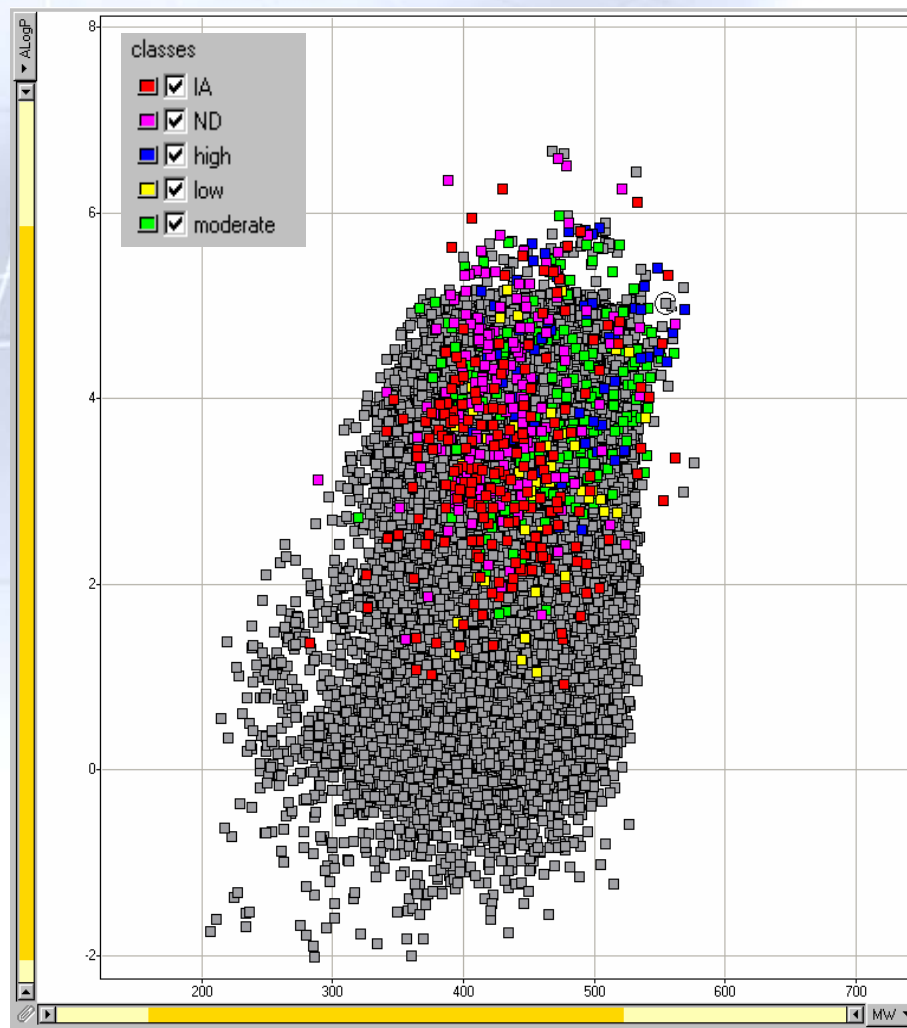
- Identify core structure



- Rip off all R groups used, retaining attachment point
- Generate all possible structures by recombining all R groups onto the core
  - Restricting R groups to the existing positions

# Recombinatorix example

- Project X
  - 571 compounds made
  - 245 different R1s
  - 167 different R2s
- Recombinatorix Space
  - 40915 possible molecules
  - 18996 not made but with good calculated properties (MW < 450, AlogP < 4)
  - Only 3% exemplified
- Approach allows us to
  - plot made and possible compounds
  - Score them by properties, predicted activities etc



# Summary

- The chemical universe is mind bogglingly big
- But the most interesting regions are of a tractable size
- We can begin to think about total virtual coverage of some aspect
  - E.g small heteroaromatics
- Useful reminder that even mature med. chem. projects rarely sample much space

# Acknowledgements

- **Colin Groom**
- Ben Perry
- Dave Parry
- Trevor Perrior
  
- Dorica Naylor
- Alicia Higuieruelo