



Successes and pitfalls in the mining of HTS data

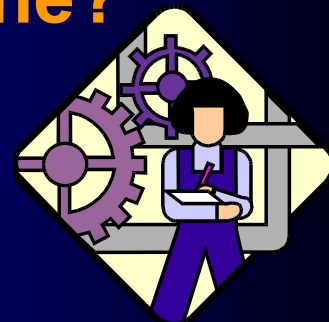
**Gavin Harper
Cheminformatics
GSK**

Overview

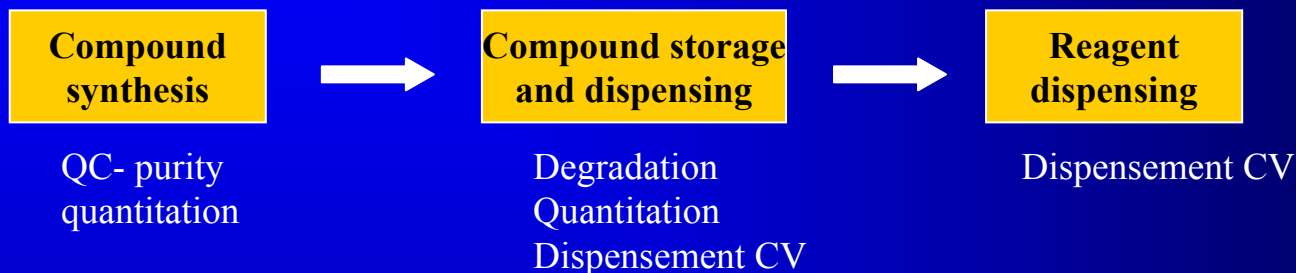
- Understanding the HTS process
- Objectives of HTS analysis
- Screening the right compounds
- Where are the hits?
- Automated analysis methods
- SIV - a chemist's tool for analysing HTS data
- Case histories



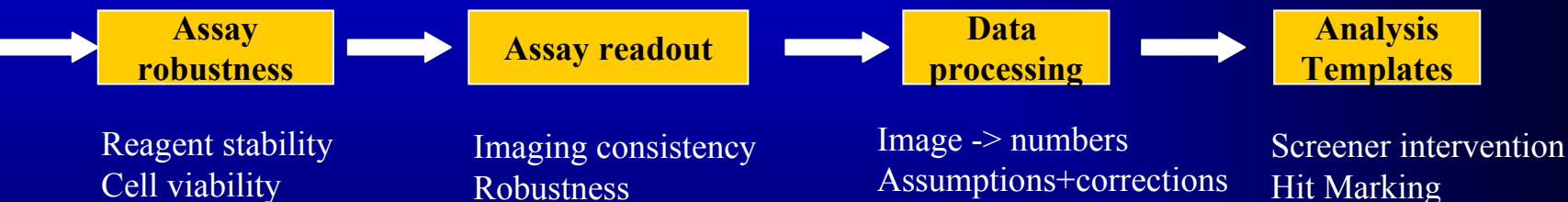
The HTS process - Welcome to the Machine?



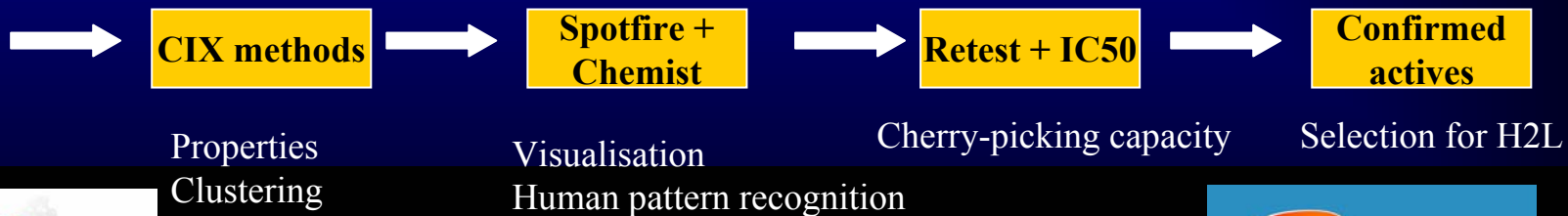
Compound Handling



Screening



Analysis



Issues affecting HTS success

- **Compound issues**

- Screening the right compounds
- Is the compound what it says on the label?
- Interfering compounds
- Promiscuous inhibitors

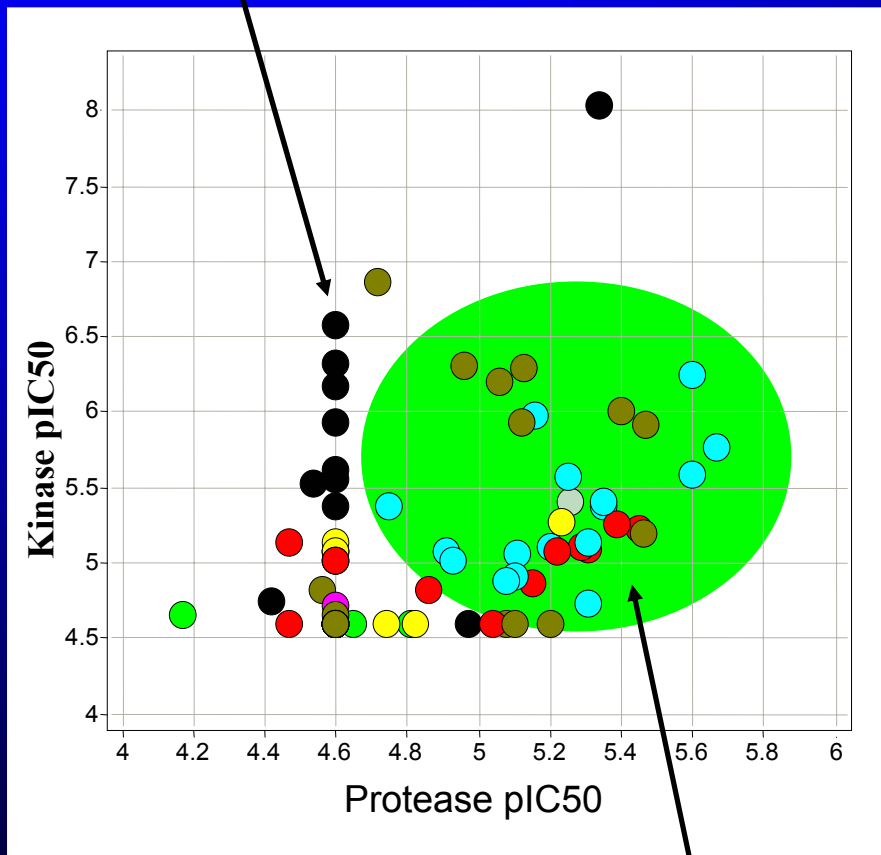
- **Screening issues**

- Hit identification
- Robustness to compounds
- Consistency through run
- Automation errors



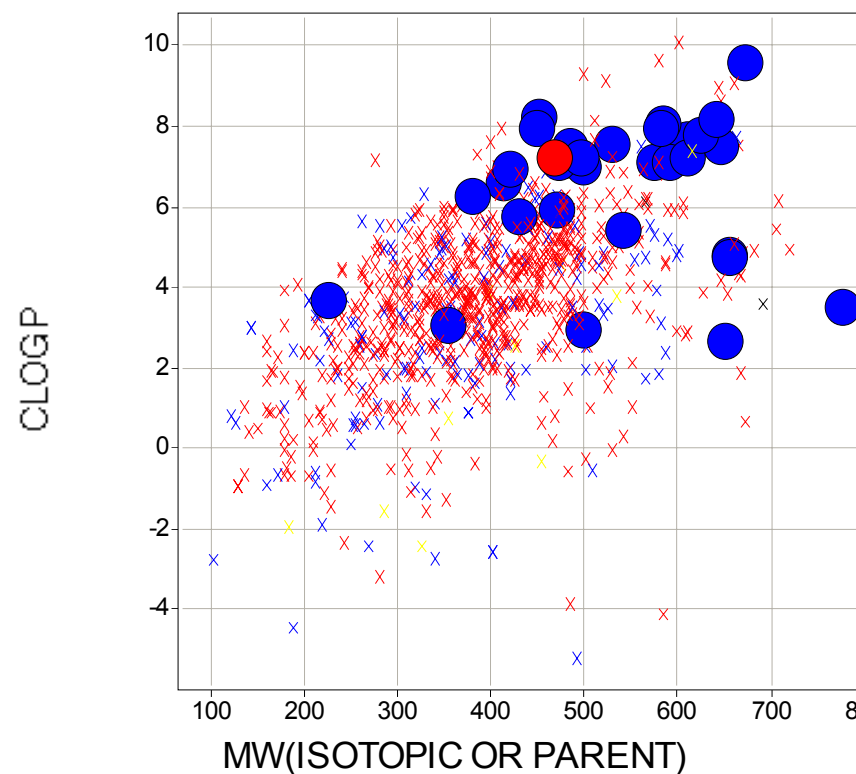
Promiscuous Inhibitors

Known Kinase compounds



These compounds (large blobs) are lipophilic high Mw and negatively charged (blue). Many lack a conserved ATP-site H-bond acceptor.

Scatter Plot



Non-kinase compounds

McGovern et al. J Med Chem 45, pp 1712-1722 (2002)



Assay Data Analysis

- Improve quality of compounds to be screened
- Define “active” in a meaningful way
- Can we use statistics / pattern recognition to find hits automatically?
- Can project chemists look at compounds individually using their expert knowledge?



Improving quality of compounds to be screened

For a sample to form part of the collection

- It has to be of a minimum purity
 - to be determined by the QA project
- It has to pass a set of agreed *in silico* filters
 - good starting points
 - developability
- Multiple lead series per screen
 - Multiple chemotypes => 2D representation
 - Collection model provides rationale and design guidelines
- Leads for all targets
 - 3D Pharmacophore coverage
 - *The Biophore Concept*. S.D. Pickett in Protein-Ligand Interactions: From Molecular Recognition to Drug Design, Volume 19 (Series: Methods and Principles in Medicinal Chemistry. Series Editors: R. Mannhold, H. Kubinyi, G. Folkers). Eds. H.-J. Böhm and G. Schneider (2003). John Wiley & Sons.



QA Project

- Merging fSB and fGW collections provided opportunity to analyse all historic samples for purity and identity.
- The new GSK sample collection populates 3 ALS systems to support μ HTS globally.
 - > 1 million compounds.
 - “Pure” and “Sure”
- The QA project required 3,775 microtitre plates, each containing approx. 324 samples.



After: new GSK screening collection

Openlynx Browser - [PR000004.rpt]

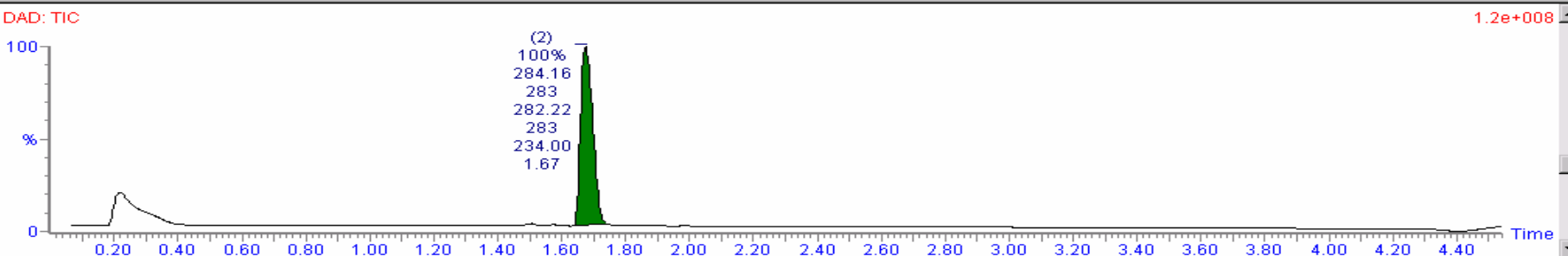
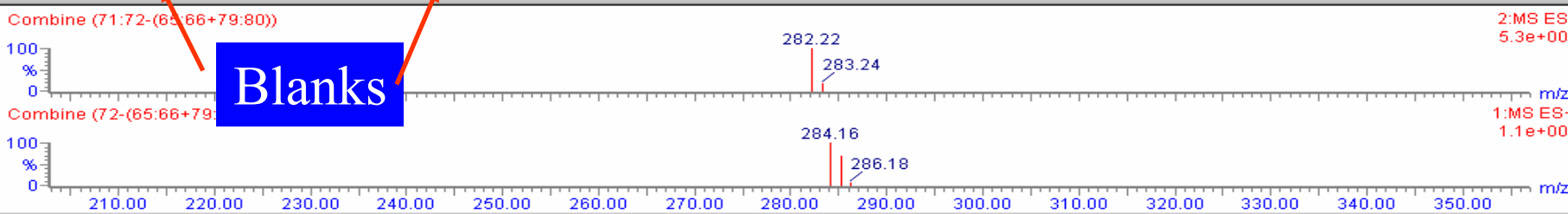
File Edit View Window Help

Plate: 1,2 Vial: 22,7

C13H12F3N3O (283.093) - (Found)

Submitter: PR000004-154
Sample: 1,2:22,7 ID: SM99840-077A46B10
Description:

Peak Number	Compound	Mass	Function	Time	
✗ 1			1:MS ES+	1.51	
✓ 2	Found	283.09(97%)	1:MS ES+	1.67	MS ES+ :284.093+301.093
✓ 2	Found	283.09(10...	2:MS ES-	1.67	MS ES+ :284.093+301.093
✗ 2			3:DAD	1.67	MS ES+ :284.093+301.093



For Help, press F1



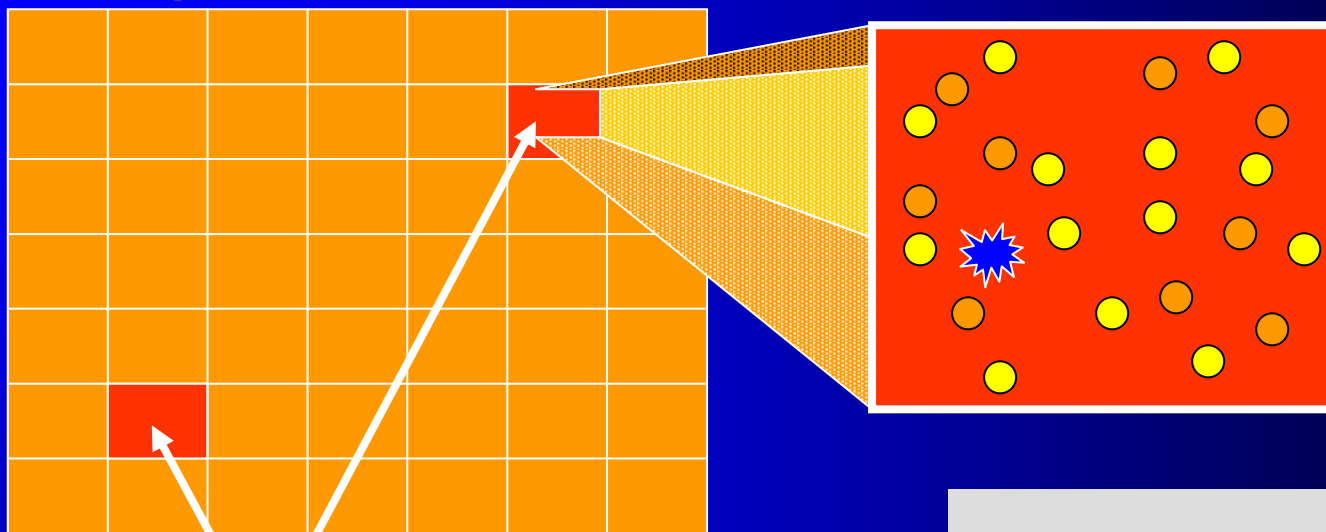
Screening Collection Model: Basic Ideas

- Relate biological similarity to chemical similarity
- Use a realistic objective
 - maximise number of lead series found in HTS
- Build a mathematical model on minimal assumptions
- How does our collection perform now in HTS?
 - relate this to our model
- Learn what we need to make/purchase for HTS to find more leads

Harper et al. *Combinatorial Chemistry & HTS*, 7(1) pp.63-70, 2004



Screening Collection Model



clusters containing
leads

- Hit
- Non-Hit
- ★ Lead

π_i

Probability that cluster
 i contains a lead
~ 1 in 100,000

α_i

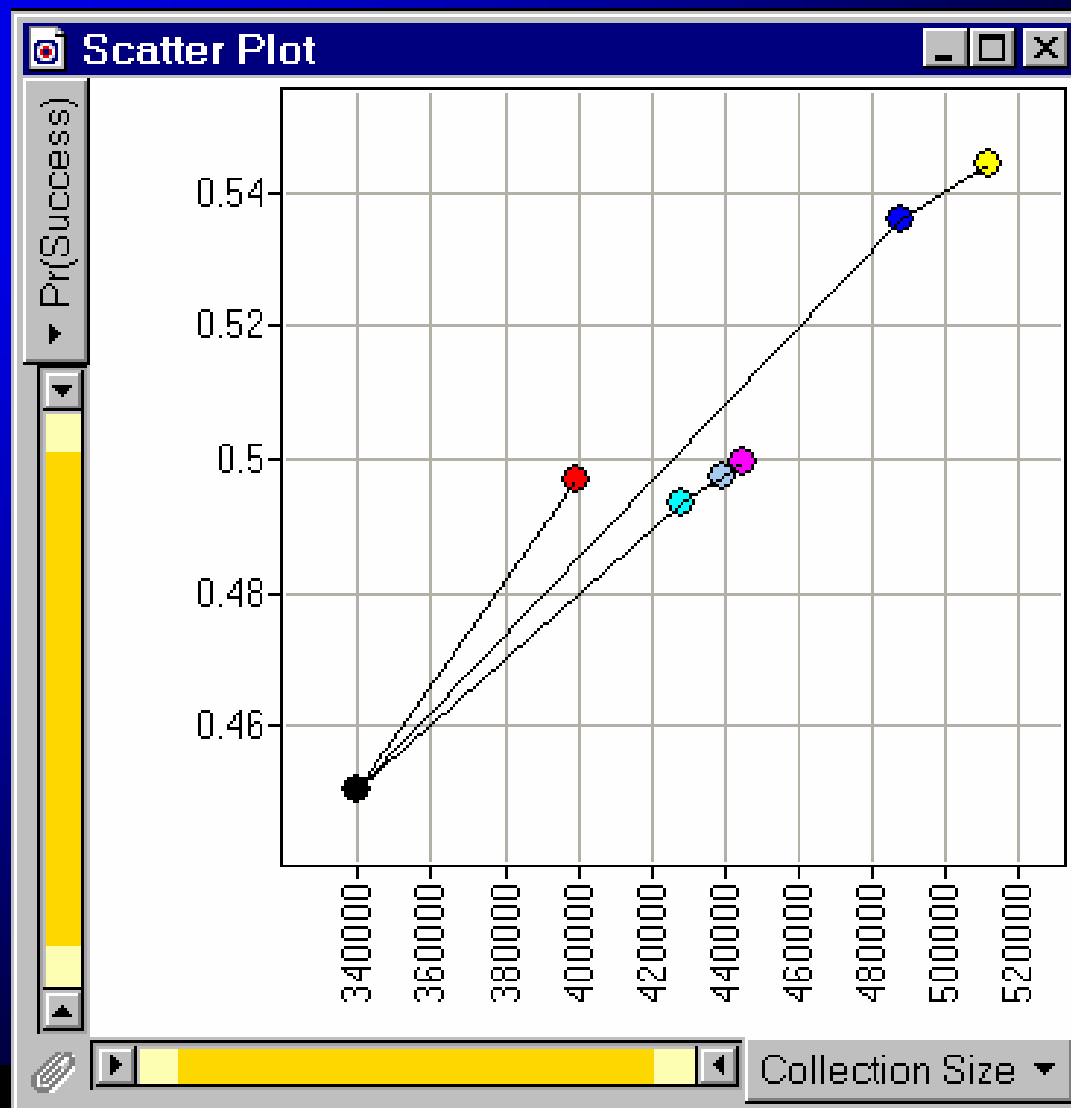
Probability that a
compound is active given
that i contains a lead

$$D(\{N_i\}_{i=1}^p) = p - \sum_{i=1}^p (1 - \alpha)^{N_i}$$



Application - Compound Purchase

More
Leads



Collection Size

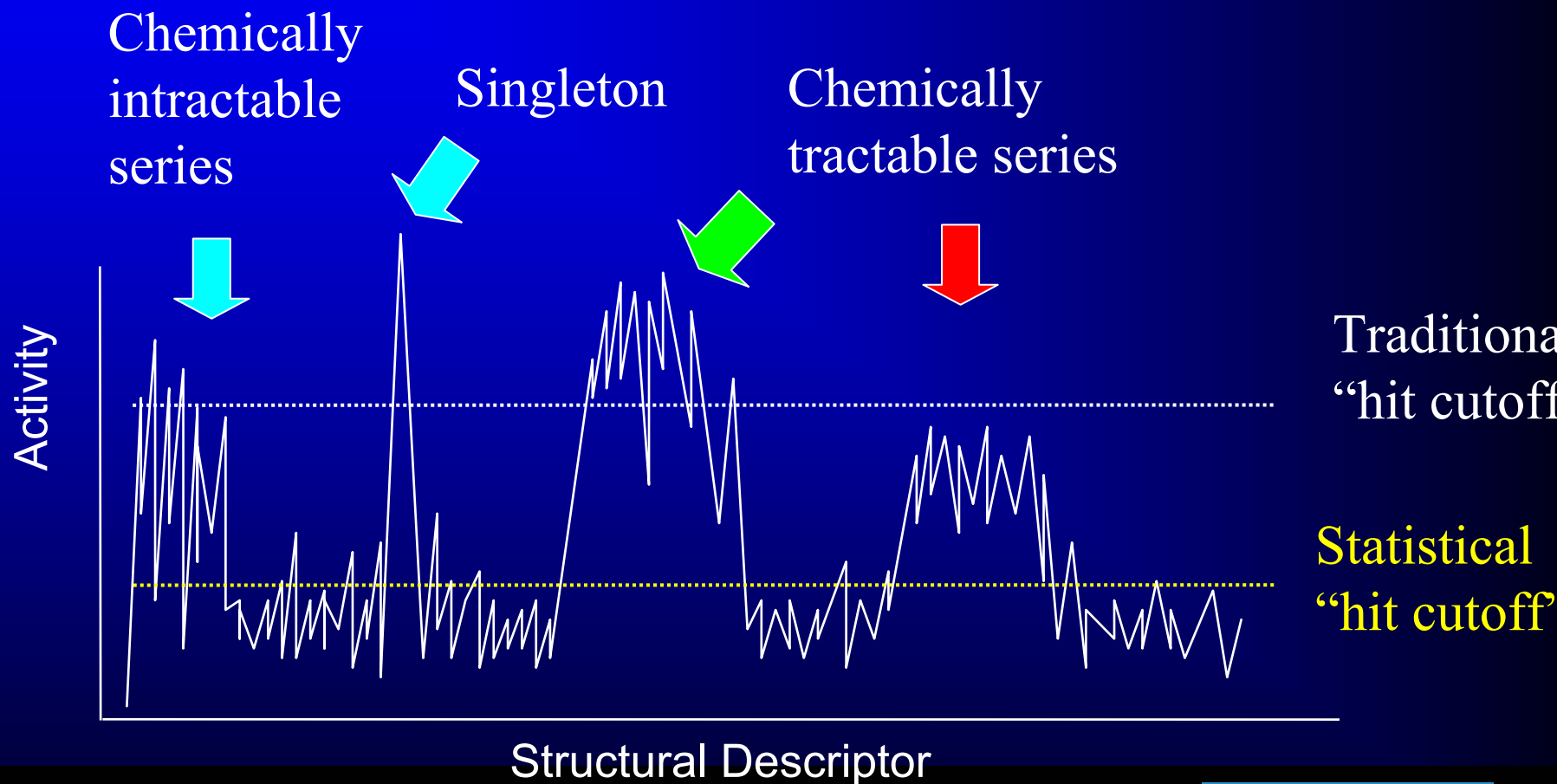


Assay Data Analysis

- Improve quality of compounds to be screened
- Define “active” as significantly different from inactive samples
- Can we use statistics / pattern recognition to find hits automatically?
- Can project chemists look at compounds individually using their expert knowledge?



HTS data analysis: schematic

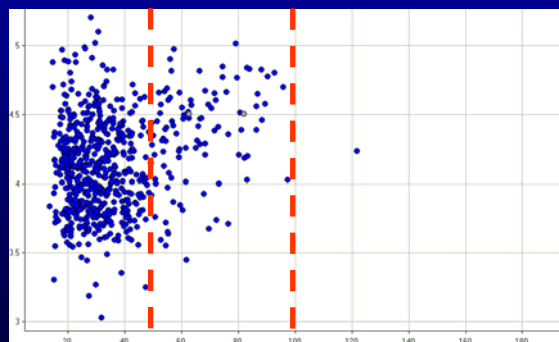
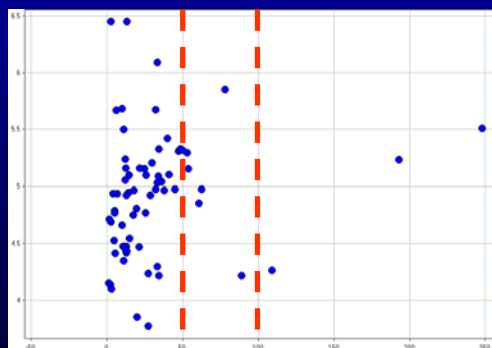
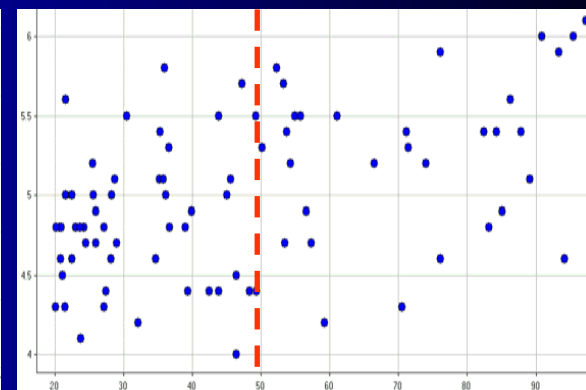
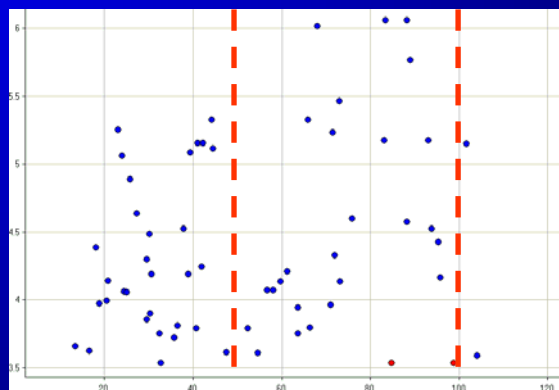
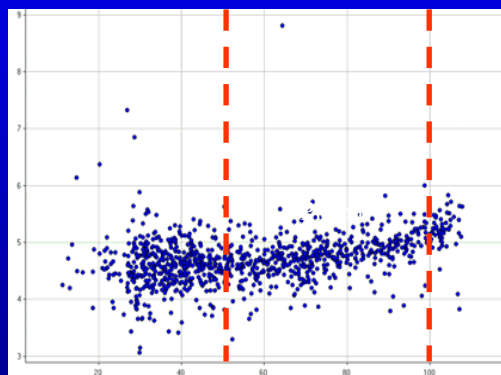


Where are the hits?

- Selecting hits based on a simple primary cutoff to fit downstream processes does not work.

pIC50

50% 100%



% I from primary

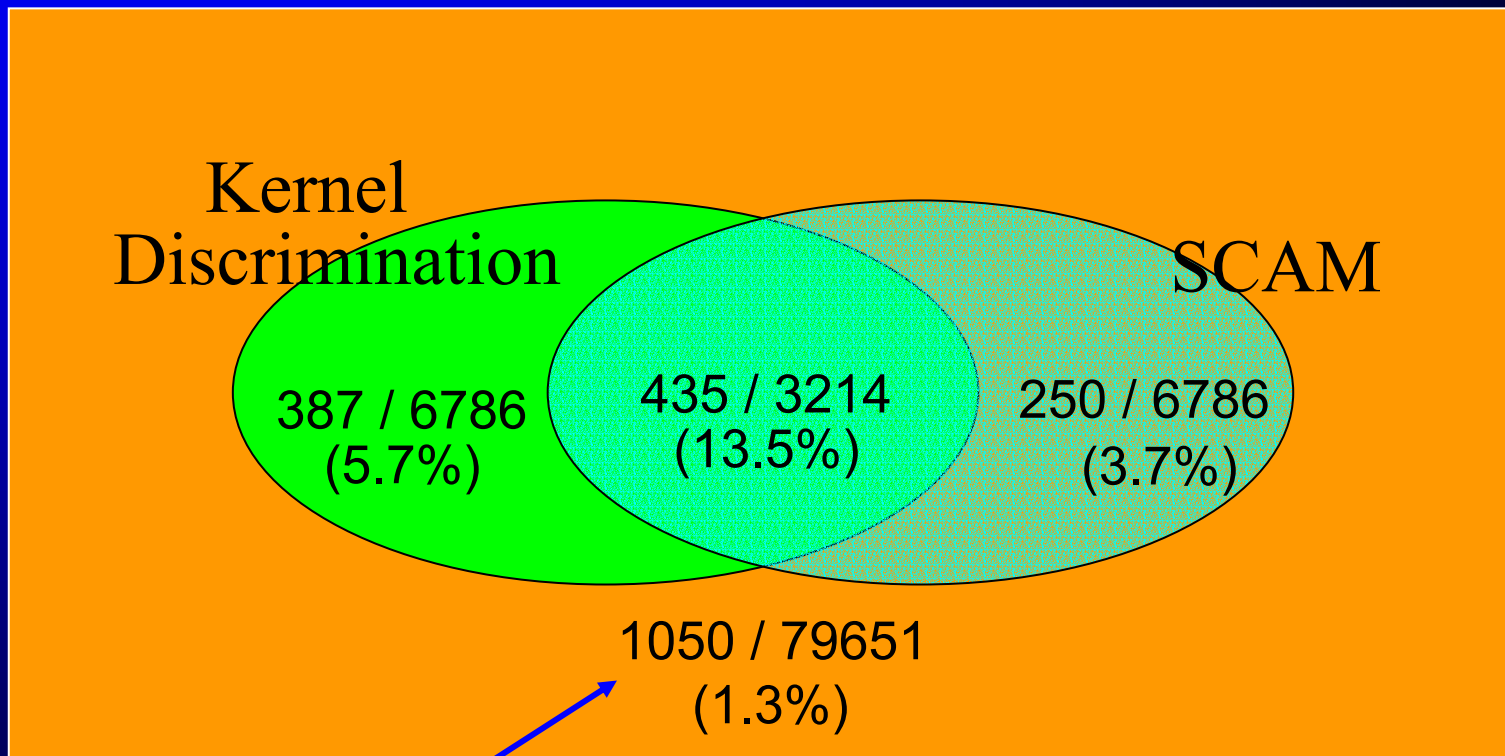
Assay Data Analysis

- Improve quality of compounds to be screened
- Define “active” as significantly different from inactive samples
- Can we use statistics / pattern recognition to find hits automatically?
 - tests with various algorithms suggest that we may still miss a lot of hits
 - progress many unsuitable compounds
- Can project chemists look at compounds individually using their expert knowledge?



Fully automatic methods miss things but can be complementary

2.1K actives / 96K inactives



Actives / # Compounds

Harper et al. *JCICS* 41(5) pp.1295-1300 (2001)

Chen et al. *JCICS* 38(6) pp.1054-1062 (1998)

Assay Data Analysis

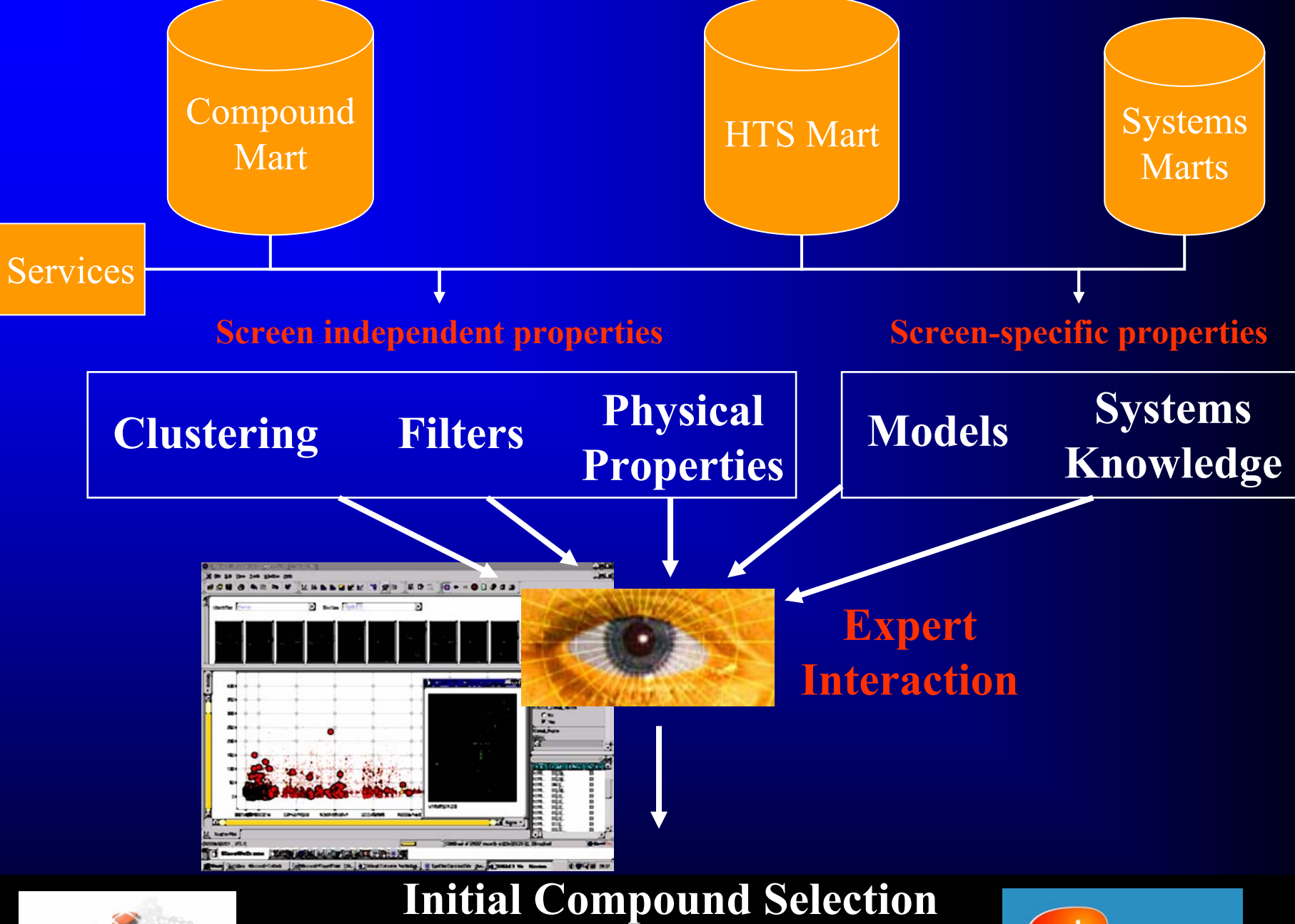
- Improve quality of compounds to be screened
- Require a measure from the screeners of what is “active”
- Can we use statistics / pattern recognition to find hits automatically?
- Can project chemists look at compounds individually using their expert knowledge?
 - If we can make it easy and intuitive.



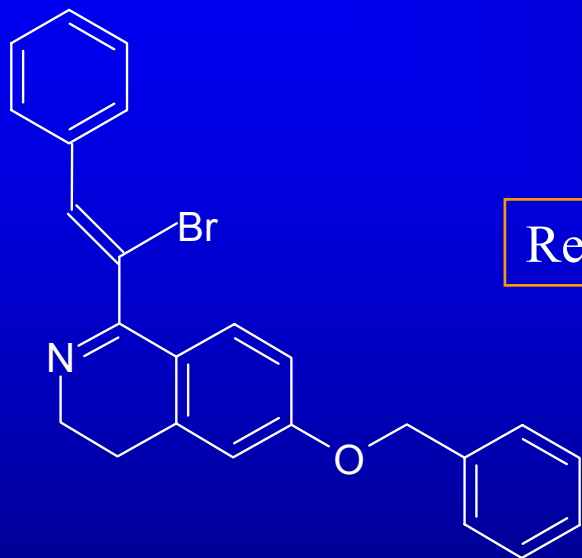
SIV - a tool for interactive analysis

- A combination of computational methods, with the combined results visualised to aid sample selection
- Visualisation is usually through Spotfire
 - GSK structure visualiser for integrated viewing of structures from SMILES in datasheet.
- Our experience is that no single method works all of the time, therefore it is normal to select several
 - e.g. clustering (various flavours), physical properties, reactivity filters, 3D pharmacophores , Data-Driven analysis etc.
 - Most techniques just look at the “actives” (though there may be many thousands of these!), but others use all of the data
- “Actives” cut-off is defined statistically from the data - not implicitly via a capacity constraint

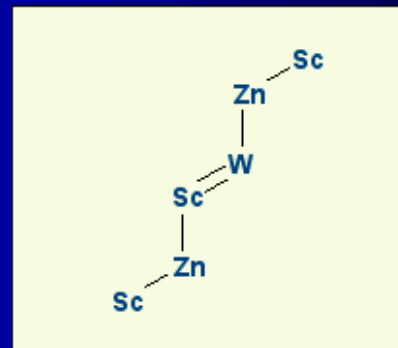




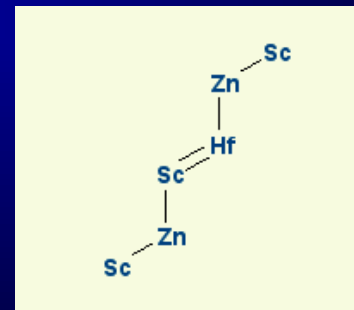
REDUCED GRAPHS



Reduced graph



Neighbour



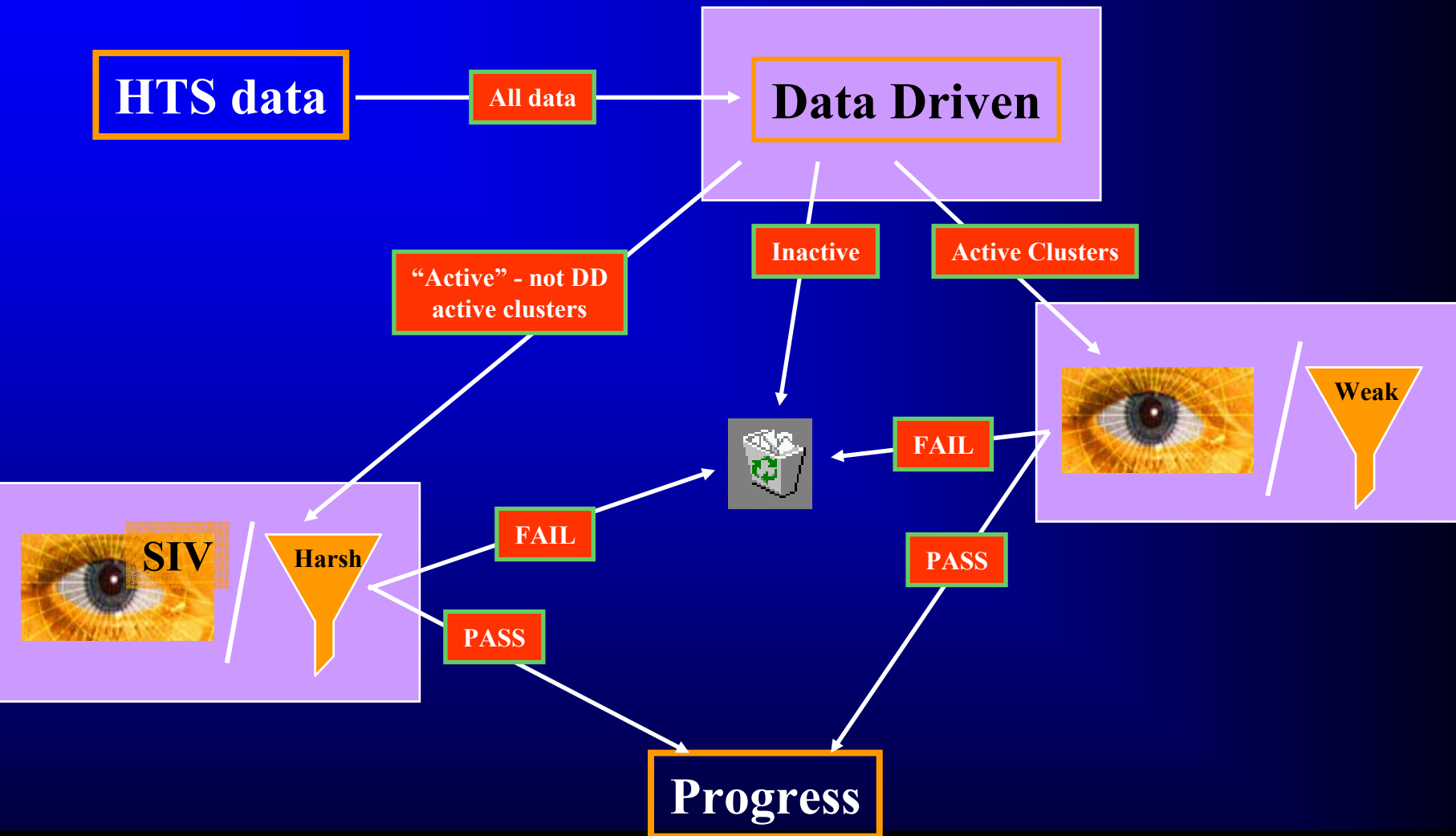
Includes acids, bases, donors, acceptors, aromatics, rings etc.

Outline of Data-Driven Algorithm

- **Sort all motifs by scoring function**
 - Prioritises clusters on activity
 - Rewarding large clusters
- **Choose top scoring motif**
 - a cluster is formed from all matching molecules
- **Repeat process with remaining molecules**
- **The user makes decisions on interesting stuff (while they're still awake!)**
- **Add “grey data” hits to what a traditional automatic method would progress.**
- **It won't deal with all those singletons!**



Possible Use

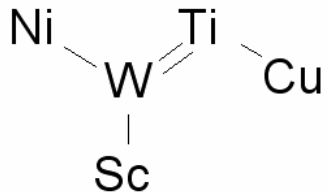


NCI Aids Data Set

Spotfire DecisionSite 7.0 - sivfile.csv - [Scatter Plot]

File Edit View Visualization Tools Window Help

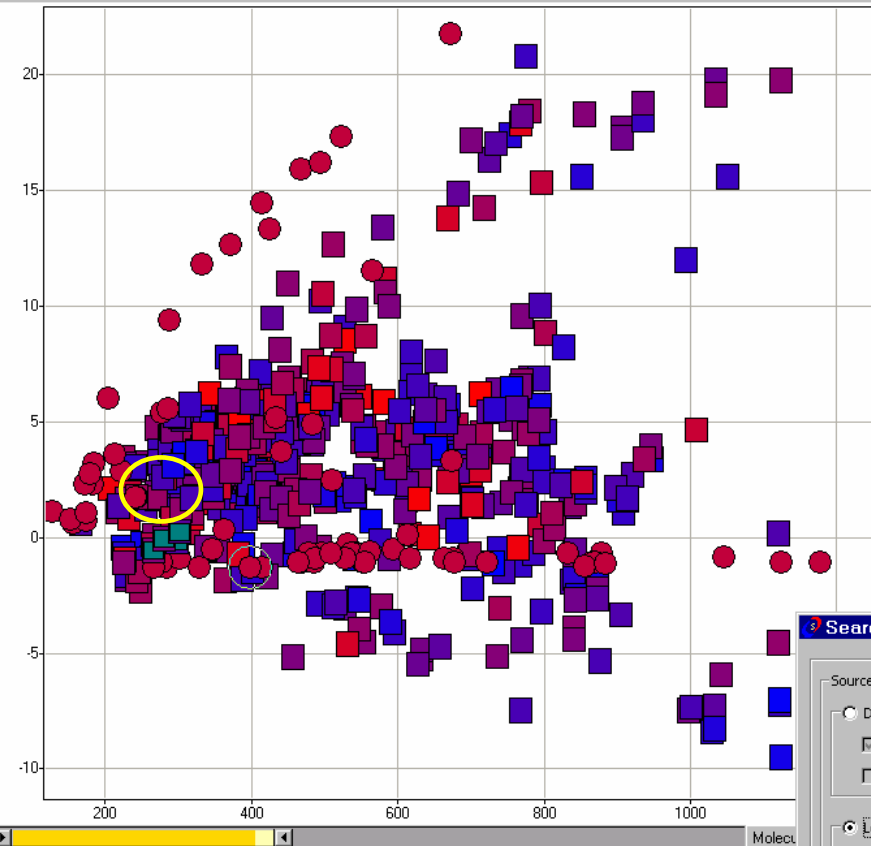
CLIX Dual Structure 1
Highlighted: RG_50



Ni — W = Ti — Cu
|
Sc

CLIX Search

[Click here to begin a new search](#)



Scatter Plot

Table

Query Devices

reduced graph

Complete_Link_Clus 1 162

N_Complete_Link 1 31

Graph_Frame 1 38

N_Graph_Frame 1 132

Change NRInns: GC

Legend

Scatter Plot

Color by Activity

-38.835213 109.94326

Shape by Motif

0 1

Search

Source

Database

Registry ACD

GULB WDI

Local Data Set

All Records

Marked (4 records)

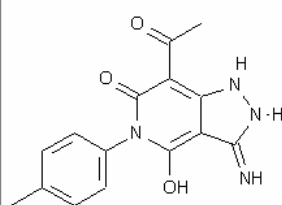
SMILES Field

SMILES

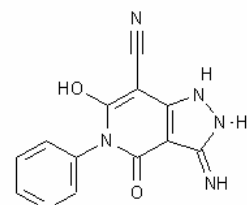
Search Type

Substructure

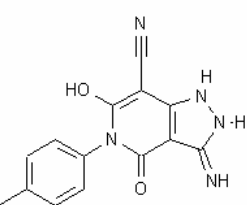
CLIX Visualizer 2



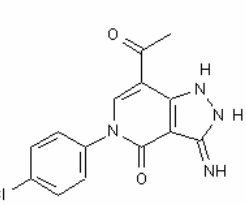
692005



692006



692031



692043

SIV

- **Good Interaction with Data Enables Excellent Expert Data-Mining**
- **Easy, Intuitive Interface**
 - “how many mouse-clicks to get where I want to be?”
- **Interactive Selection**
 - “what gets rejected if I apply this filter?”....
 - “fine, but I want to keep these 3 compounds”
- **Flexible Analysis**
 - a.k.a. “I did it MY way”



Case histories



“Typical screens”

- “Typical” screen
 - 15-30K primary hits
 - 2K progressed
 - hit/IC50 rate typically 0.25 +/- 0.25
 - small correlation to %I (0.2, 0.25, 0.3 averages)
- False positives are more of a problem than false negatives
 - we have some good methods for “rescuing” false negatives
 - false positives blur the signal, and hence the effectiveness



A high hit rate screen.

- **Primary data analysis:**

- initial 58,355 hits
- tighter 32,124 hits => used in analysis
 - fail CIX filters: 12,307 (38%)
 - remaining 19,817
 - SIV 1,712 (8.6%; 5.3% of unfiltered)

- **IC50s (four replicates: interference)**

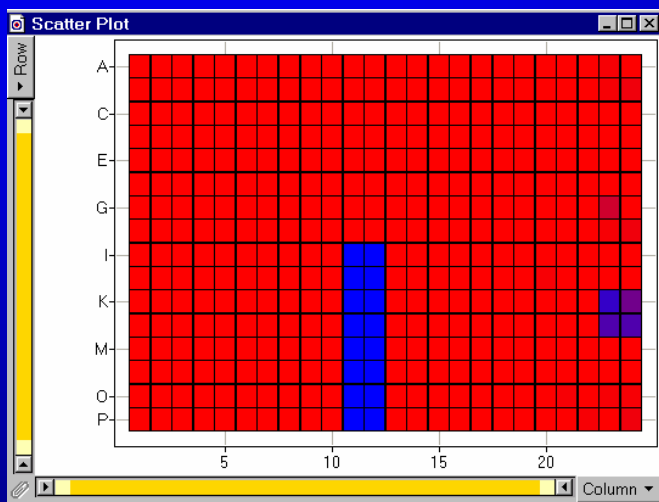
- Active 490 (27%; 1.5% of unfiltered)
- Inactive 508 (27%)
- Interference 765 (44%)

No leads identified

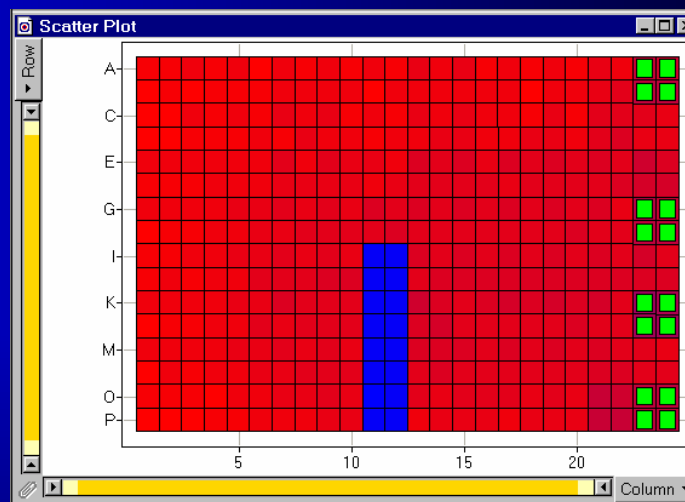
(Current hit to lead series identified by focussed screen)



Number of “hits” in each well



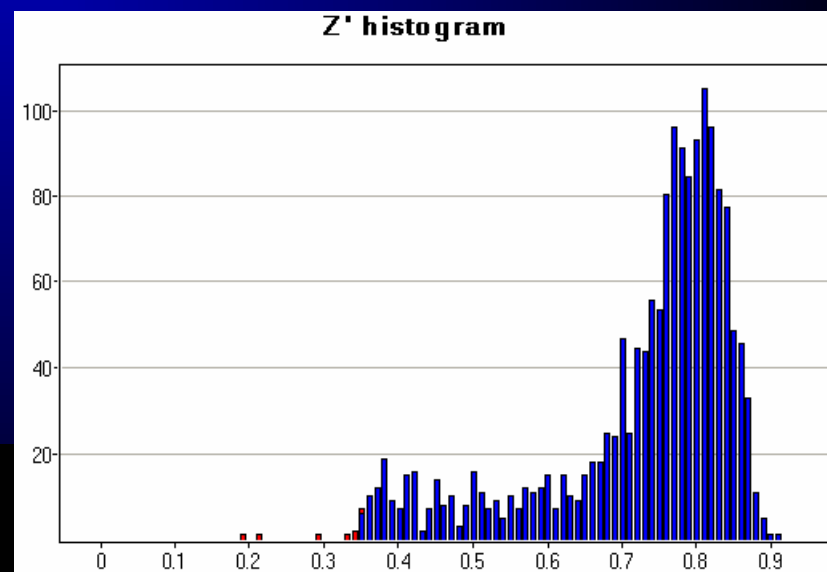
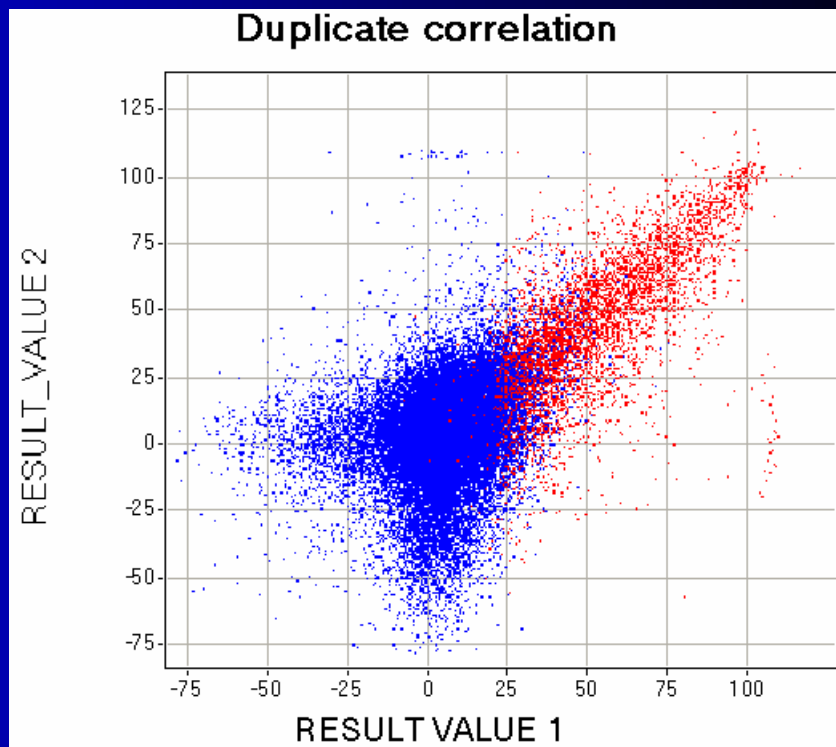
409/617 (66%) of hits
in
4/352 (1%) of wells



499/617 (81%) of hits
in
16/352 (4%) of wells

A true high hit rate screen

- **>56000 hits (12.8% hit rate)**
 - select 5000 for retest (only 2000 in many campaigns)
 - over 90% of hits never retested
- **83% retest rate!**
 - choose 1600 for IC50 from these
- **93% give a d-r curve!**
 - selected 289 for solid IC50 determination
 - 70 compounds still of interest after solid testing, all with $IC_{50} < 2\mu M$
- **We will not be able to pursue the majority of these series**



Summary?

No such thing as the average screen - this is no machine

- The HTS process comprises many steps, all of which are prone to error
- Much data is lost as we go through the process, until, ultimately, chemists see one number per sample
- For some screens, the process does work well
- For many screens, intervention is required
 - Specialist intervention can add real value
- There are many opportunities for projects to improve our processes. We should look at the enterprise as a whole, and the goals for HTS, before choosing which processes to target.



Acknowledgements

- **Cheminformatics**

- Stephen Pickett, Darren Green, Andy Whittington
- Harkamal Tumber, Sunny Hung

- **CASS**

- Andrew Leach, Giampa Bravi
- Steve Lane, Zoe Blaxill

- **DR Chemistry**

- **Molecular Screening**



Example SIV

27297

structures, >10% activity at least once in testing

21085

unique OIs

13782

>10mg solid

screen-specific

8839

Substructural filter

8823

active by statistical method

6750

after applying reactivity filters

1600

after selection by chemists

Second look (different clustering algorithm)
with ALL FILTERS OFF except solid availability.
Looked twice at anything with high potency.

1785



