



GlaxoSmithKline



Use of data mining to help identify compounds that are unstable in DMSO

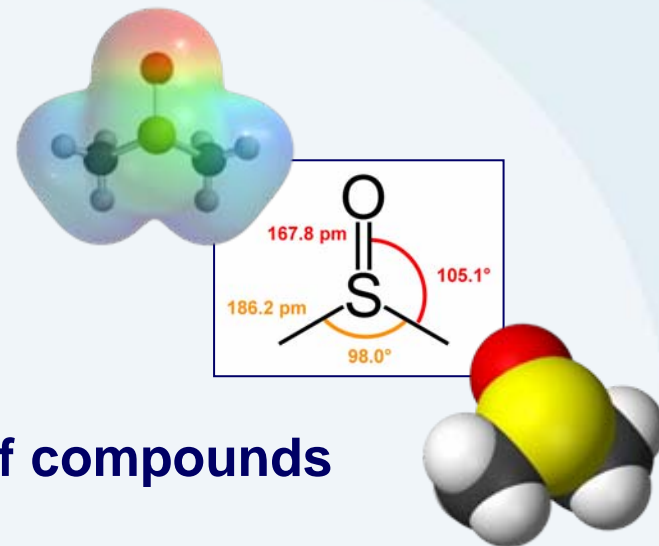
Jameed Hussain, Gavin Harper, Zoe Blaxill,
Irene Areri, Farnaz Saremi-Yarahmadi, Stephen
Pickett, Philip Sidebottom

Introduction

- Why is compound stability in DMSO important
 - Storage conditions of HTS sets
- Current methods of identifying compounds unstable in DMSO
- Identifying further substructures in GSK
 - QA data
- Data Driven Algorithm
 - Brief description
 - Descriptors
- Results of the analysis
- Future direction

Why is compound stability in DMSO important ?

- Dimethyl Sulphoxide (DMSO)
 - Relatively chemically inert
 - Relatively high melting (18.5°C) point
 - Relatively high boiling (189°C) point
- **Has the ability to dissolve a wide range of compounds**
- Ideal solvent for solution storage of large (diverse) compounds sets
 - High Throughput Screening (HTS) sets
 - >1 million cmpds
- DMSO is used to prepare all compound solutions (up to preclinical) in drug discovery



Why is compound stability in DMSO important ?

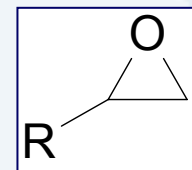
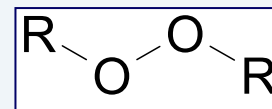
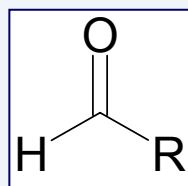
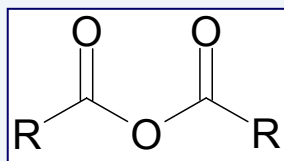
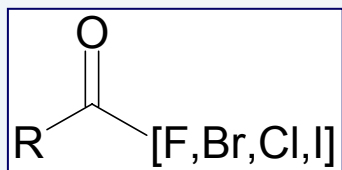
- Need to make sure compounds solutions do not change significantly over lifetime of the solutions
 - Reduce false positive and false negative results in HTS
- Avoid compound degradation and precipitation
- Typical storage conditions
 - Low temperature storage (-80°C to 4°C)
 - Reduce water uptake
 - Reduce freeze/thaw cycles (precipitation)
- Lifetime of the solutions can be a few years

Current methods to identify unstable compounds

- GSK employs a set of *in-silico* substructure filters for known unstable moieties

- Example of substructures excluded include:

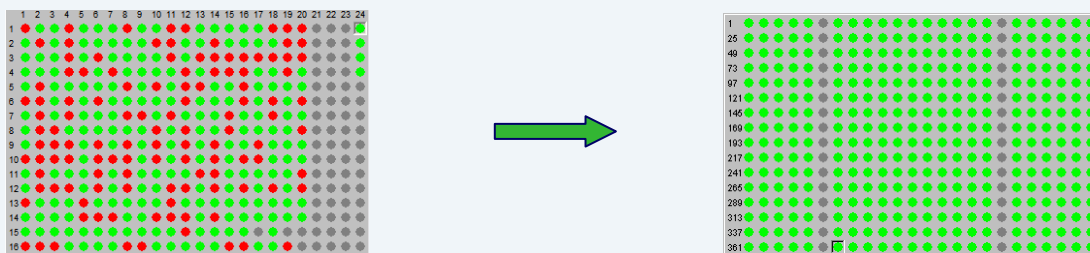
- Acid halides, anhydrides, aldehydes, peroxides, epoxides etc.



- Similar practice within the pharmaceutical industry
- Some attempt to try to model compound stability
 - COMDECOM
 - Not a trivial problem
- **The chemical diversity of HTS sets is large** (>1 million cmpds)
 - Naïve to assume we know of every unstable moiety

GSK Quality Assurance process

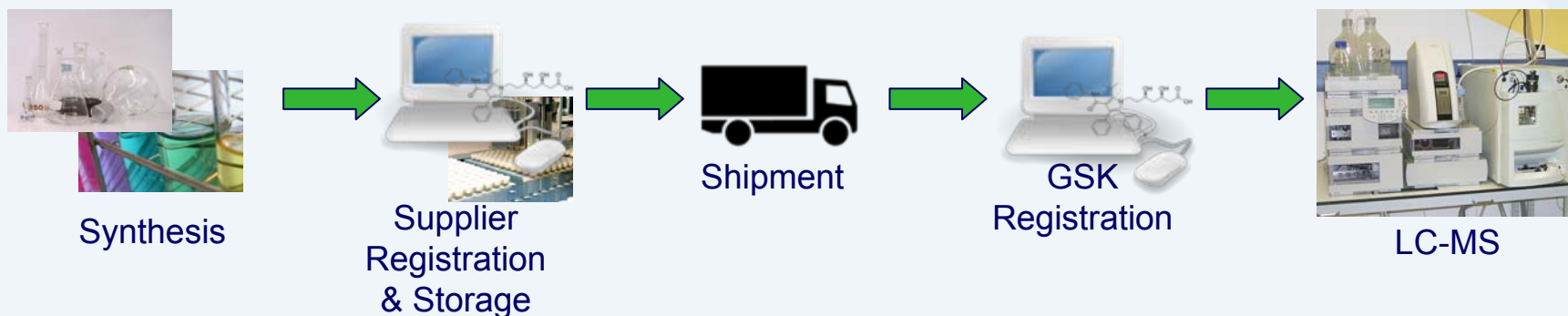
- Within GSK we have a Quality Assurance (QA) process for all compounds entering HTS collection
 - Legacy & new (in-house synthesised / purchased) compounds



- Compounds need to have a purity greater than 80%
- The process has been running for a few years
 - Amassed a lot of QA data
- Could we use this data to learn more about compound degradation ?

More about the QA data...

- The QA process tests the outcome of the whole process
 - Example: Purchased compounds..

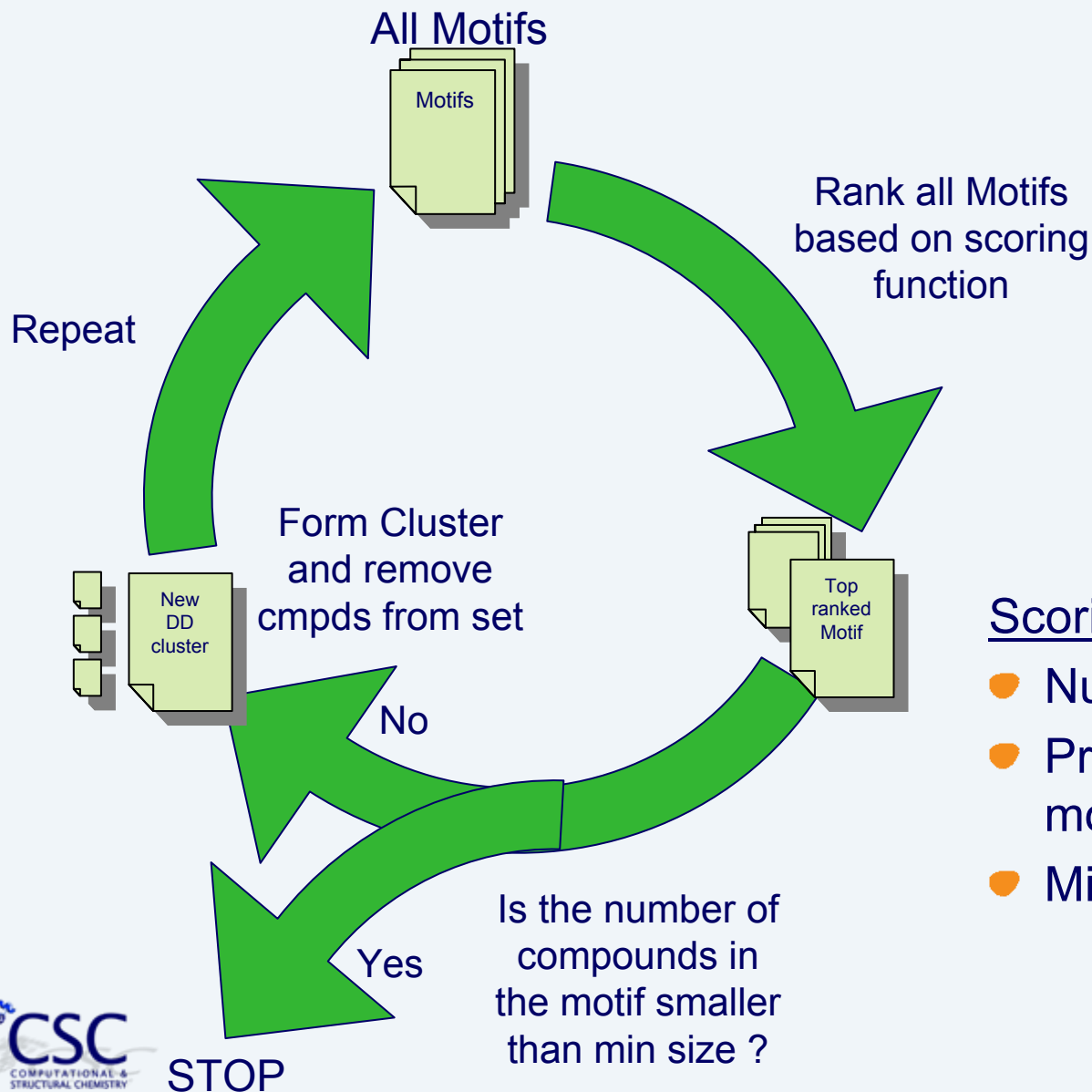


- Errors at any of these stages can lead to QA failure
 - Compound degradation is just one the possible causes
- **Data is inherently noisy**
 - Data-mining approaches

Data Driven Clustering

- Used within GSK to analyse HTS data
 - Noisy data
- Uses biological activity data to identify sizeable clusters of predominantly active compounds
- Harper *et al*, *J. Chem. Inf. Comput. Sci.* **2004**, 44, 2145-2156.
- Adapted to run against binary response data
 - QA data
 - “Active” = Compounds that failed QA process
 - “Inactive” = Compounds that passed QA process

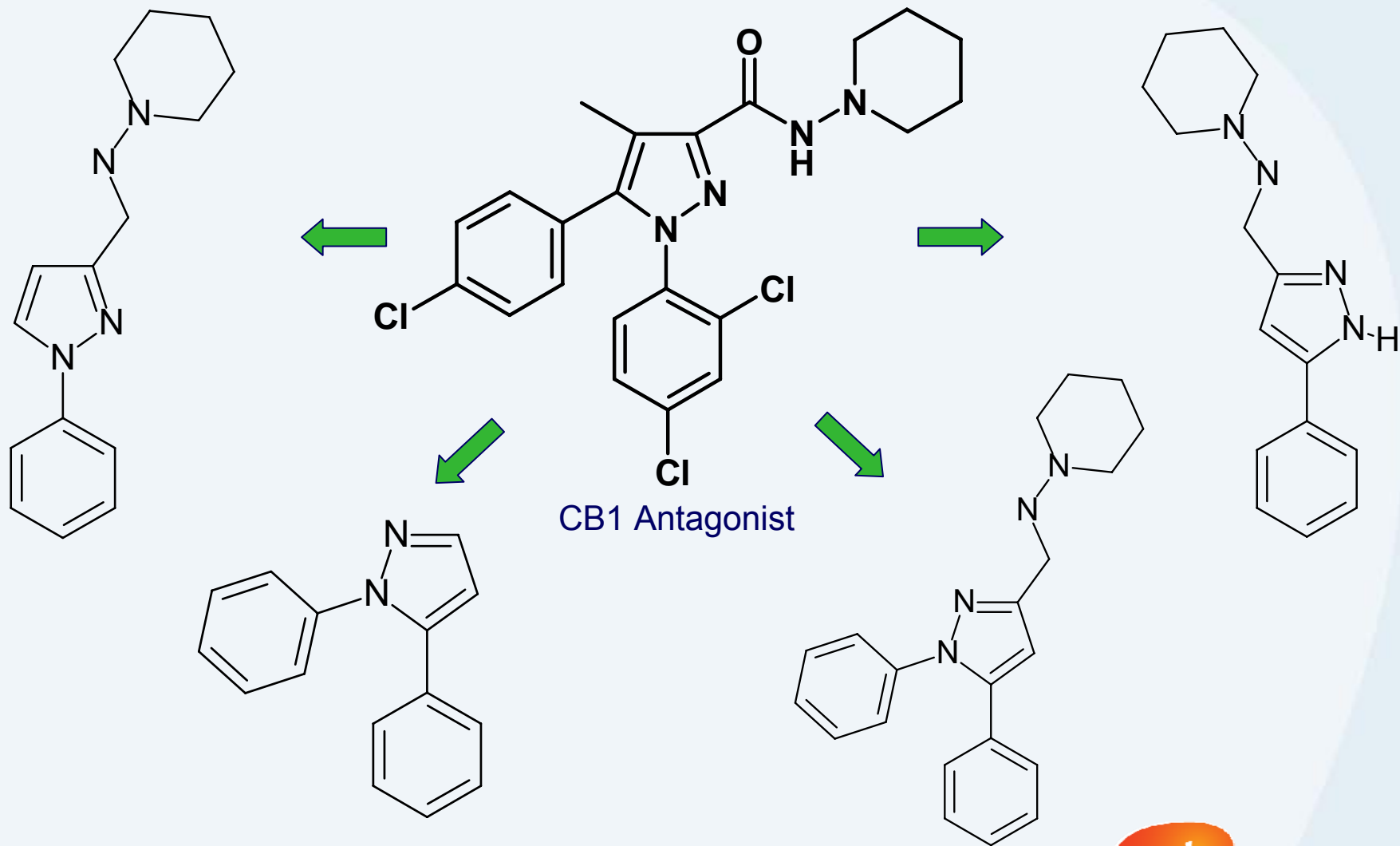
Data Driven Clustering Algorithm



Scoring Function:

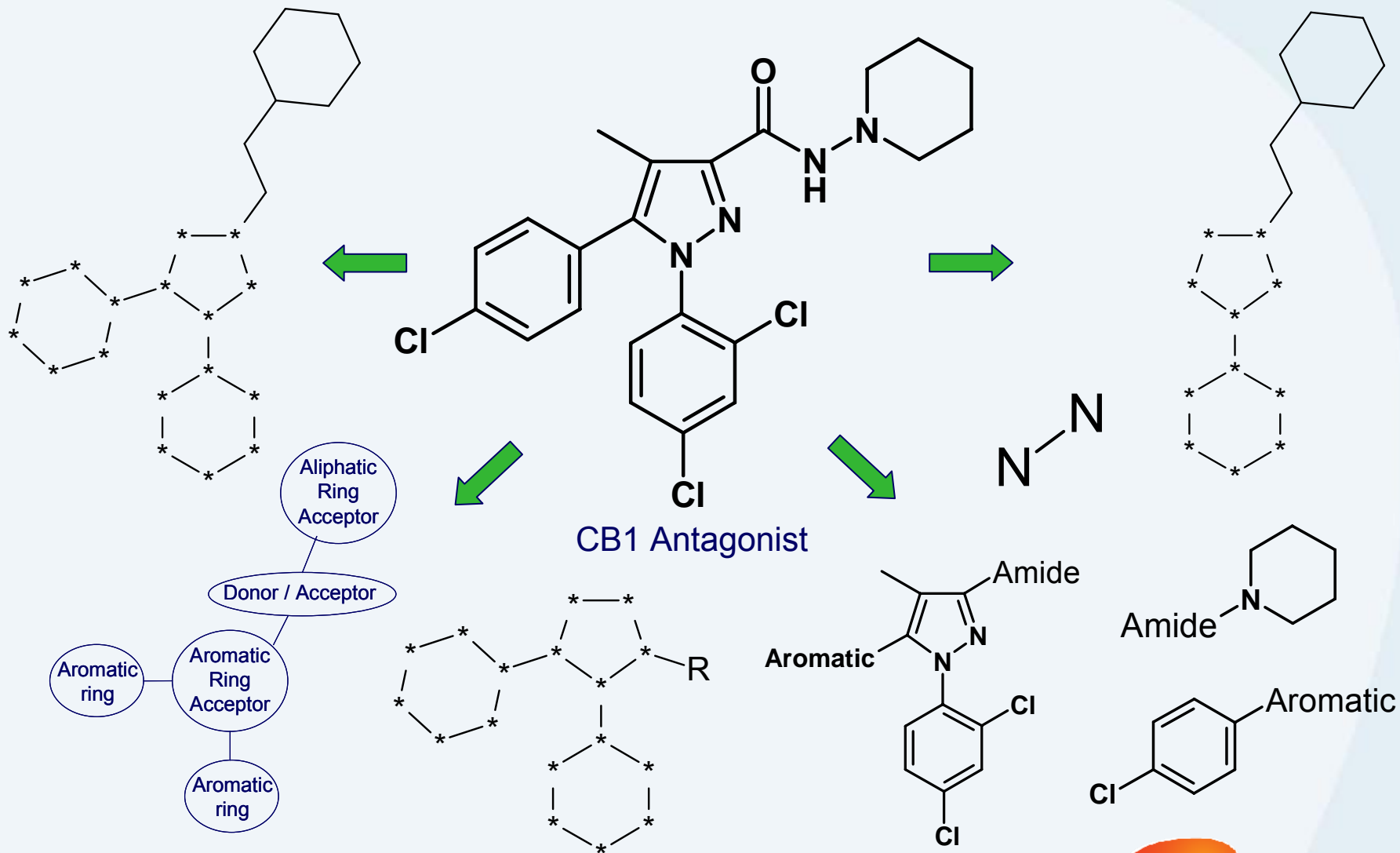
- Number of cmpds per motif
- Proportion of “actives” per motif
- Min Cluster Size 20

More Motifs..



The properties of known drugs: 1. Molecular Frameworks, Bemis and Murcko, *Journal of Medicinal Chemistry*. 1996, 39 (15), 2887-2893.

More Motifs

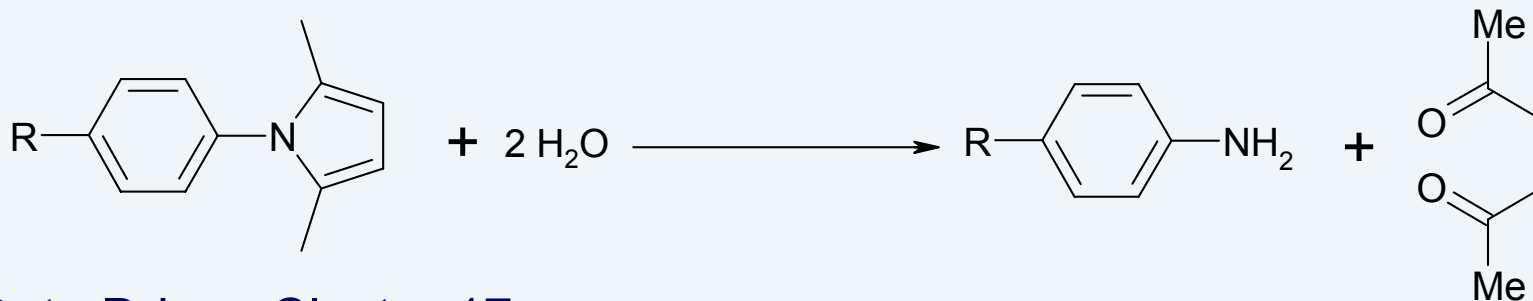


Running the Data Driven algorithm

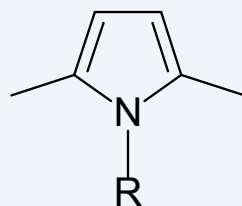
- There are several types of QA failure
 - Category 4: Pure, wrong
 - Category 6: Impure, right
 - Category 8: Impure, wrong
- Only Category 6 QA fails considered
 - ~30k compounds
 - “Active” set
- Random selection of ~900k compounds that passes QA process
 - “Inactive” set
- Run the data driven process

Results from Data Driven Analysis

- First reality check
 - Does the program identify known unstable compounds ?
- Recently learnt that certain pyrroles are unstable in DMSO



- Data Driven Cluster 17
 - Pyrrole



- RECAP fragment
- 101 cmpds
- 67.3% QA fails
- 10 different suppliers

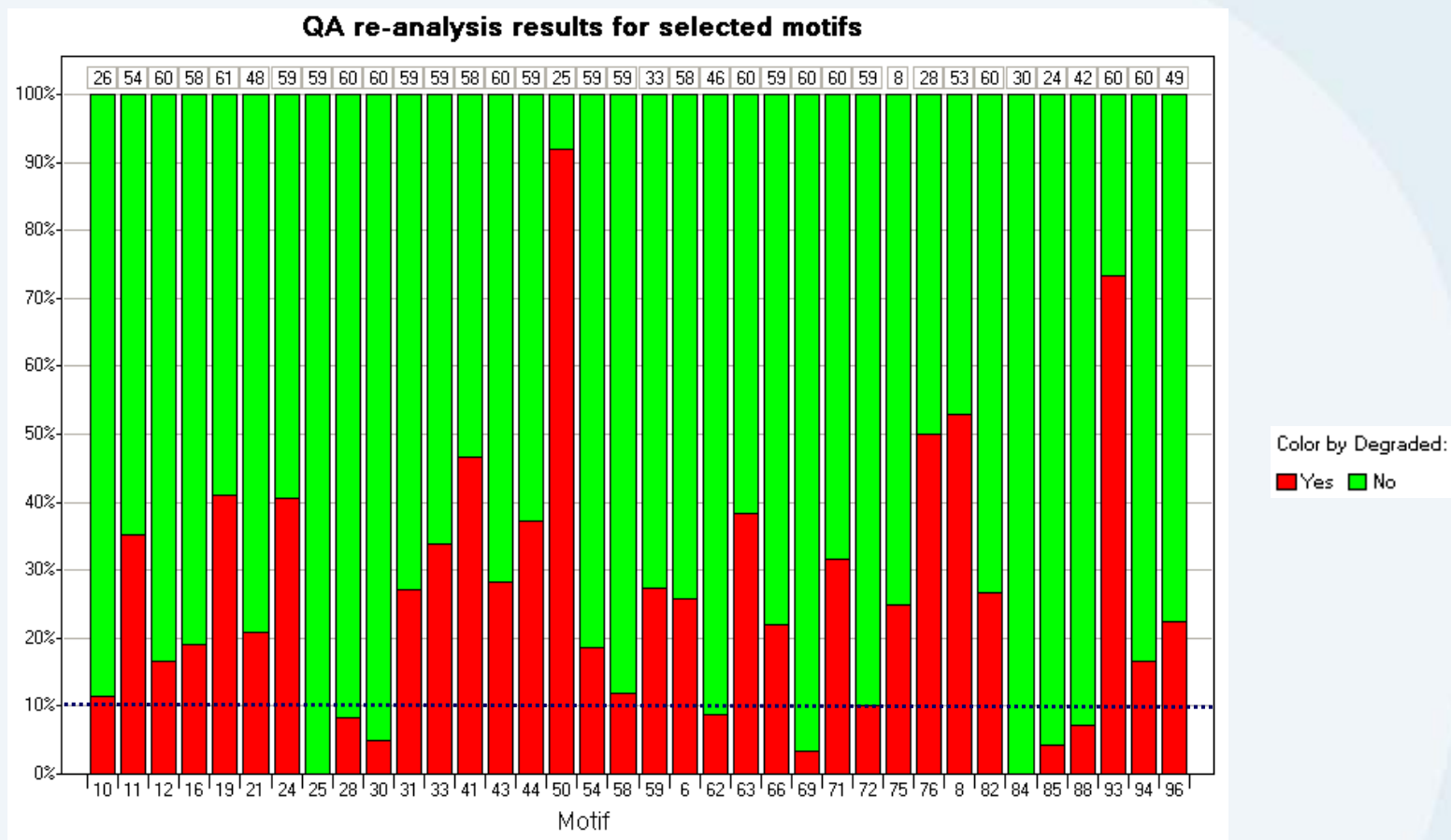
Selection of motifs for further analysis

- The first 100 Data Driven Clusters were visually analysed
- A selection of 36 Data Driven Clusters selected for further analysis
 - QA fails from a variety of suppliers / arrays
 - Minimise potential of having a library specific issue
 - Selected motifs had to be relevant to business
 - Have a number in the HTS collection
 - Some understanding of how it may degrade
 - Difficult; most cases it was non-obvious
- A selection of 1823 compounds made for re-analysis
 - For each selected DD cluster
 - Search the HTS collection
 - Randomly select up to 60 compounds for each motif

Re-analysis QA Results

- Time zero QA data available for each compound re-analysed
 - All from the HTS collection
- The compounds re-checked using same QA process
 - Compound classed as degraded if purity (determined by UV) difference greater than 10% from time zero
- Short Aside..
 - What level of degradation per motif is significant ?
 - What is the underlying rate of compounds degradation in low temperature DMSO solution storage ??
 - In GSK QA failure rate seen from XC50 hits follow-ups is ~3%
 - Decided to be conservative
 - If $\geq 10\%$ compounds degraded, motif considered unstable

Re-analysis QA Results



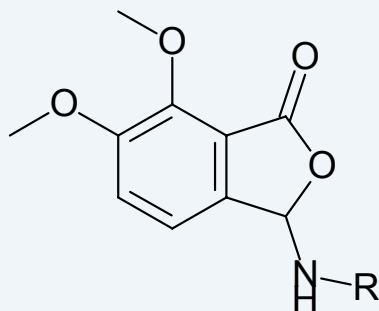
- 28 out of the 36 motifs contain $\geq 10\%$ degraded compounds
 - Degrading compounds from multiple suppliers / arrays

Results

- Results exceeded expectation
 - 28 motifs to follow up!
- Difficult to follow-up all 28 degrading motifs due to resource required
 - Restricted ourselves to follow-up study on only two motifs
- Tried to pick motifs with the “cleanest” degradation
 - Consistent patterns in LC-MS results across all degrading compounds in motif
- Results of analysis for one of the motifs will be presented
 - Analysis for other motif ongoing...

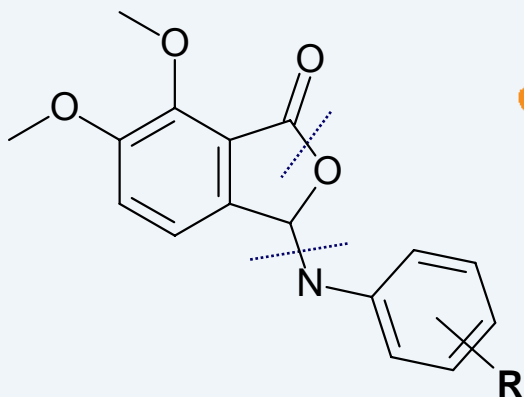
Motif 11

- Data Driven Analysis
 - RECAP fragment
 - 129 cmpds
 - 63.6% QA fails



- QA Re-analysis
 - 54 compounds tested
 - 19 (35.3%) degraded

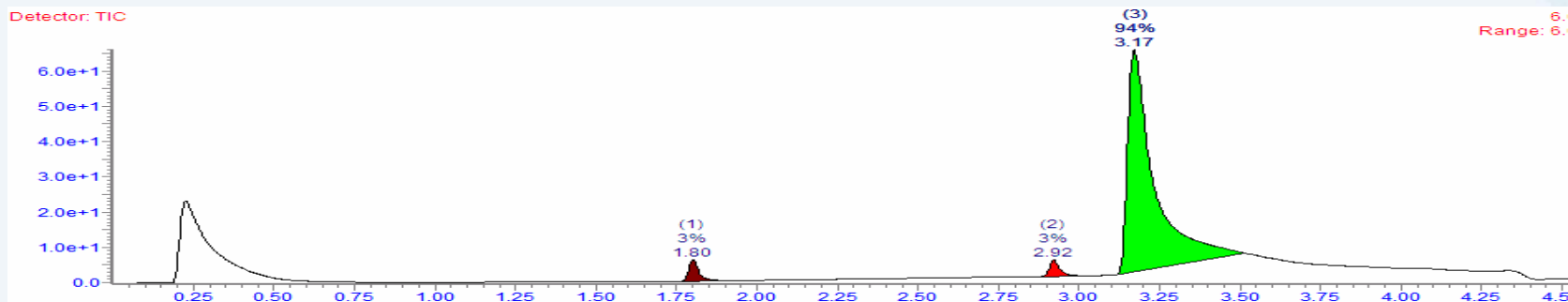
- Example degrading compound from motif



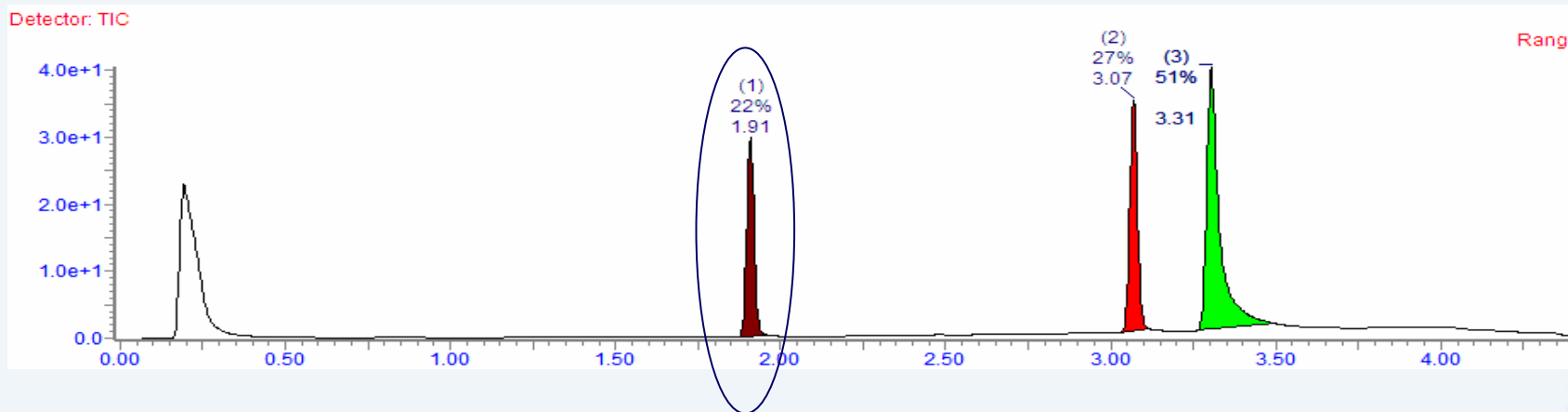
- Expected route of decomposition ?
 - Ester Hydrolysis*
 - Amine Hydrolysis

LC-MS results from QA analysis

Initial QA results (Time Zero)

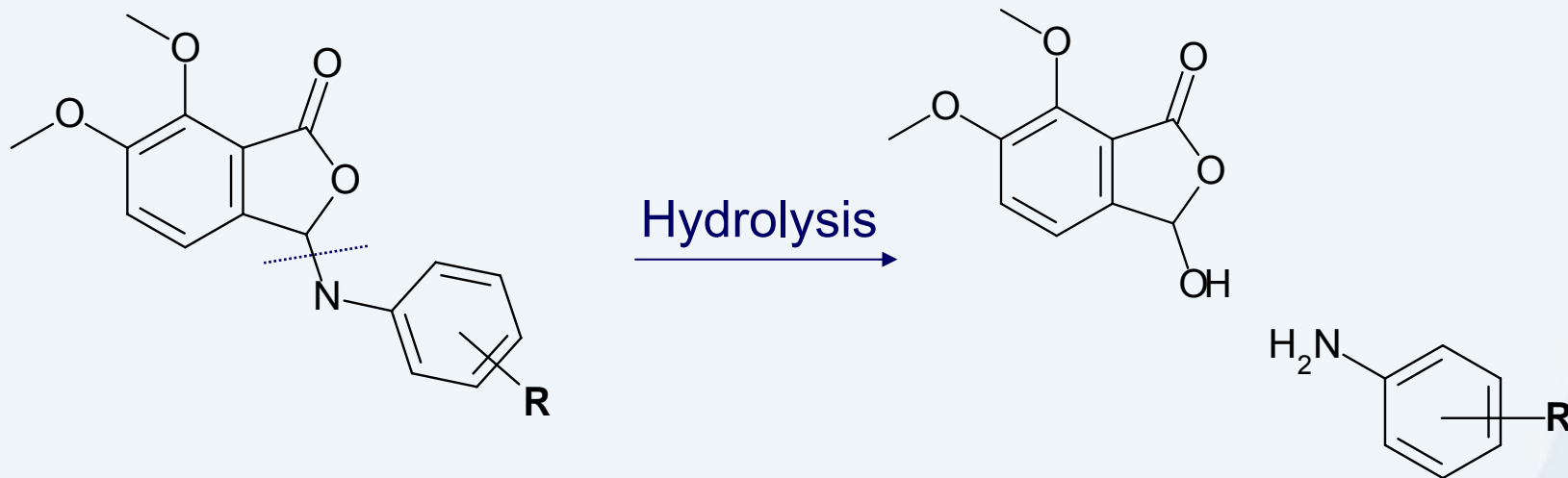


Re-analysis QA results – major impurity 192 mass difference



Motif 11

- Degrading compounds consistently show an impurity 192 mass units below compound
 - Points to amine hydrolysis
 - Confirmed by further analysis (LC-MS and NMR)
 - Ester degradant also unstable (hydrolysis of ester)

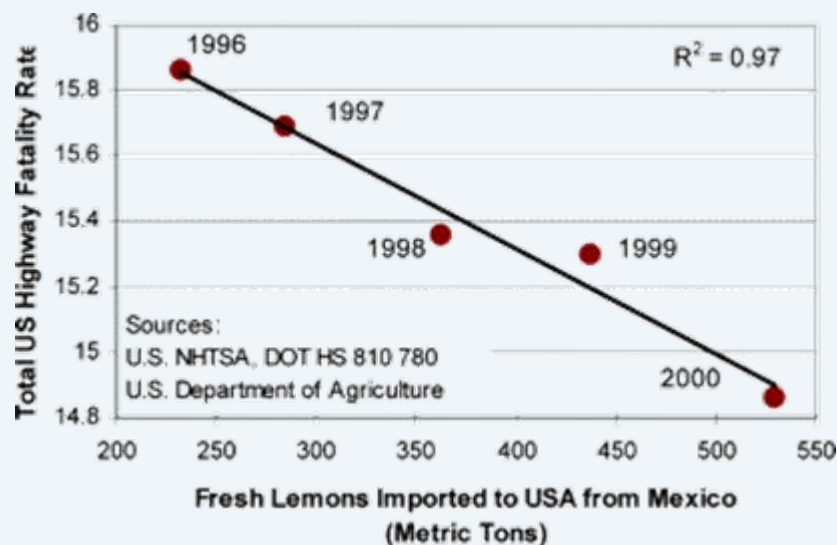


Unstable motifs

- For Motif 11
 - May have guessed that it was unstable
 - Or just post-rationalisation 😊
 - Not all compounds containing motif had degraded
 - No obvious pattern to separate the stable/unstable compounds
 - Degradation more complex than expected
- For many of the other degrading motifs it is not obvious that they are unstable in DMSO
 - Even more difficult to predict the degradation pathways
- **Compound degradation is a complex problem**

Future Direction

- All motifs identified as unstable will be added to the compound filters used within GSK
 - Used against newly synthesised and purchased compounds
 - Need to be careful (Correlation \neq Causality)



Future Direction

- Plans underway to re-check all HTS compounds that contain motifs found to be unstable
 - Removal of the impure compounds from HTS collection
- Data Driven very effective in mining QA data for DMSO unstable compounds
- Plan to run data driven algorithm periodically to identify further unstable motifs
 - Help maintain a high quality HTS collection

Summary

- Why compound stability in DMSO important
- Current methods of identifying compounds unstable in DMSO
- Identifying further substructures in GSK
 - QA data
- Data Driven Algorithm
 - Brief description of the algorithm
 - Descriptors
- Results of the analysis
- Future direction

Acknowledgements

- Silvia Bardoni
- Kathryn Jenkinson
- Alexander Wheatley
- John Hollerton
- Analytical Chemistry, MDR
- Compound Management, MDR
- Computational and Structural Chemistry, MDR

Back-up Slides

- Scoring Function

- Based on Normal approximation to Binomial
- Underlying failure rate “p” determined from full data set.
- If motif matches N molecules, X of which fail, the score is given by

- $$\frac{(X - Np)}{\sqrt{Np(1 - p)}}$$