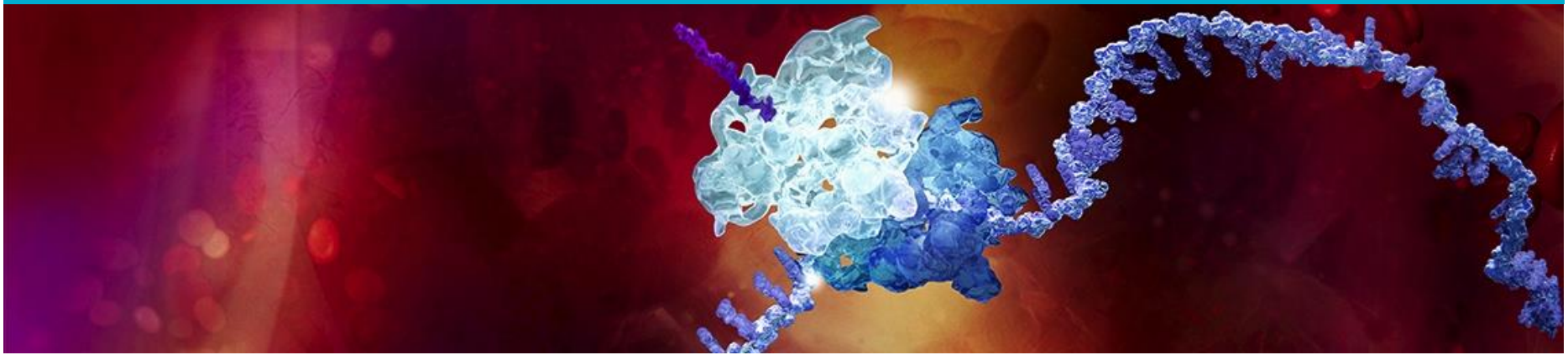
AstraZeneca
IMED Biotech Unit

Success Through Statistics in High Throughput Drug Discovery

Dr Graeme Robb

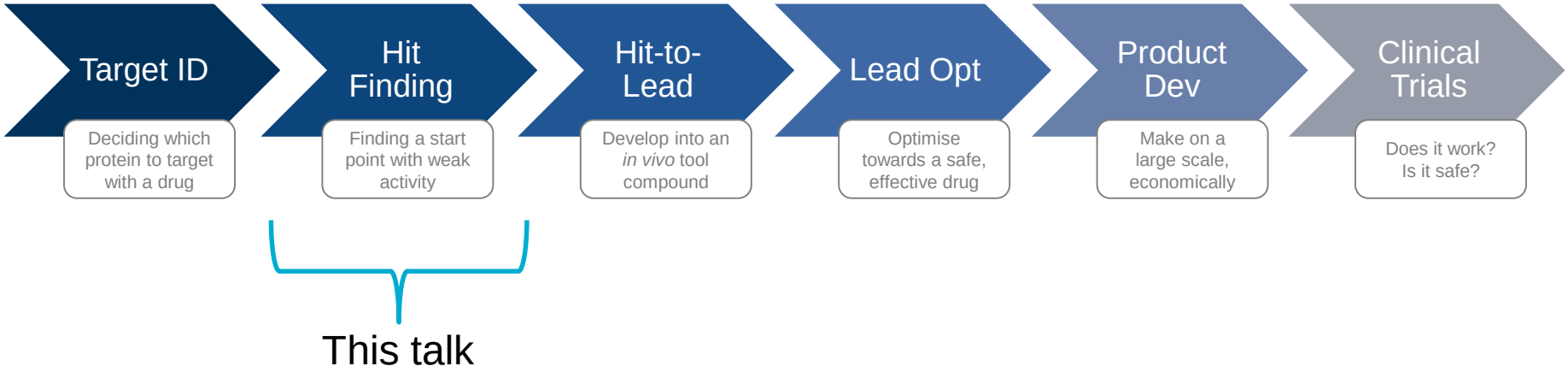
JMP User Group Meeting, UK

13th July 2017



Drug Discovery

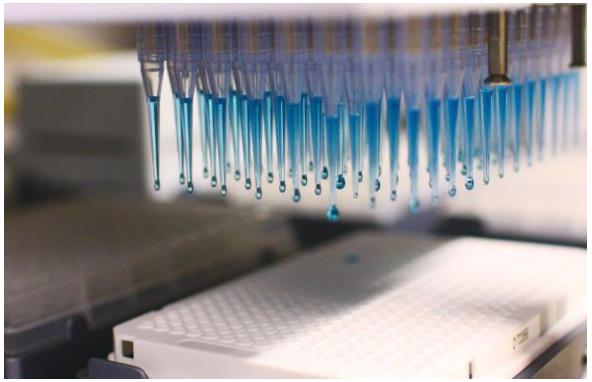
- The job of identifying a novel substance that can treat a disease or condition while remaining safe and non-toxic.



High Throughput Screening (HTS)

- Automation of assays, allowing millions of compounds to be screened

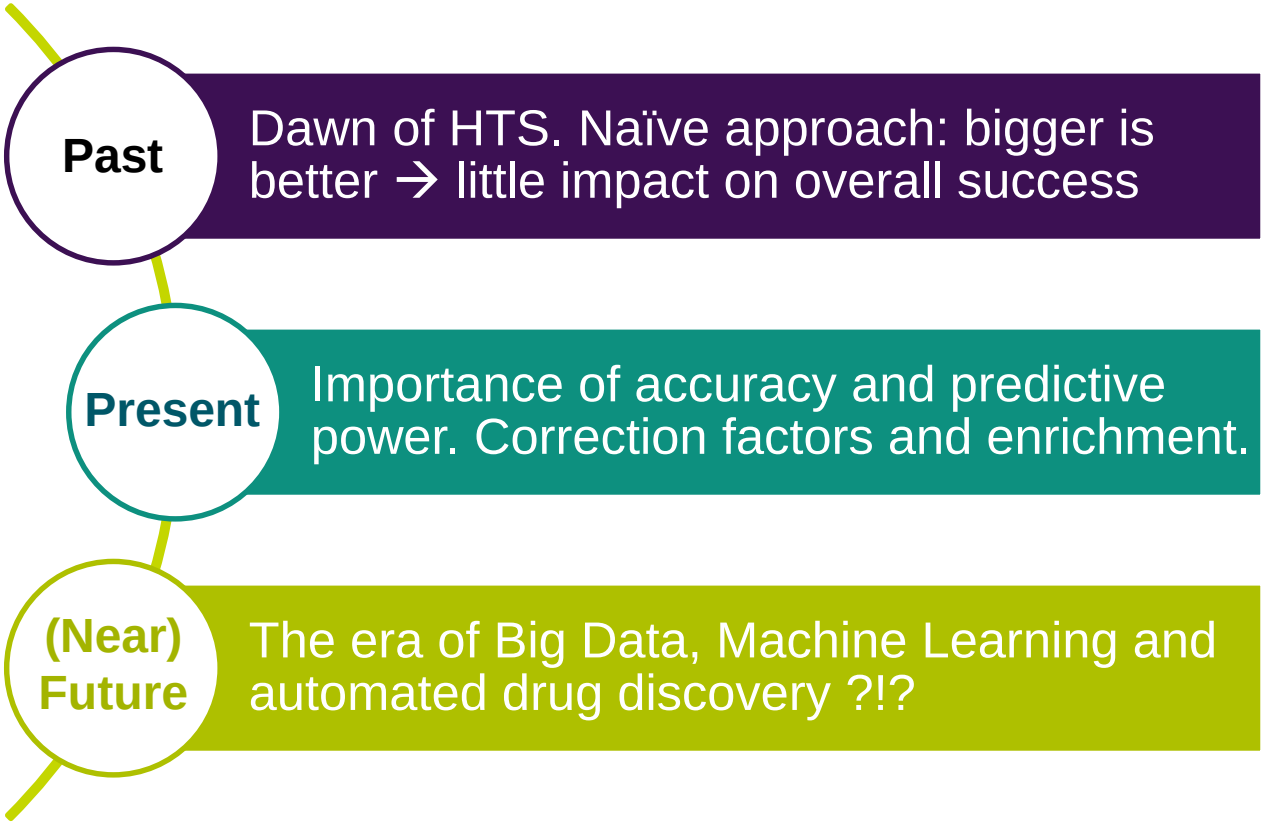
- Challenges:
  - Automation can bring systematic error
  - Assay format for HTS involves compromise
  - Challenge of signal vs noise
  - Big Data – can we use this?



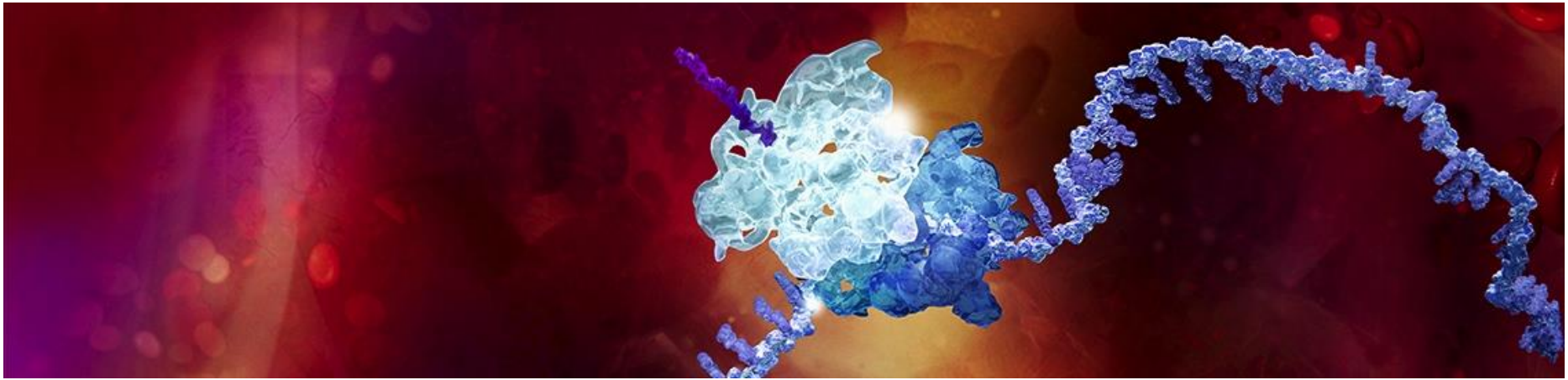
HTS Plate

HTS Robot

High Throughput Screening (HTS)



Lessons from the Past



Why hasn’t HTS delivered big rewards?

- Two theories:

The ‘low-hanging fruit’ theory

- We’ve found drugs for easy targets
- Only hard targets are left

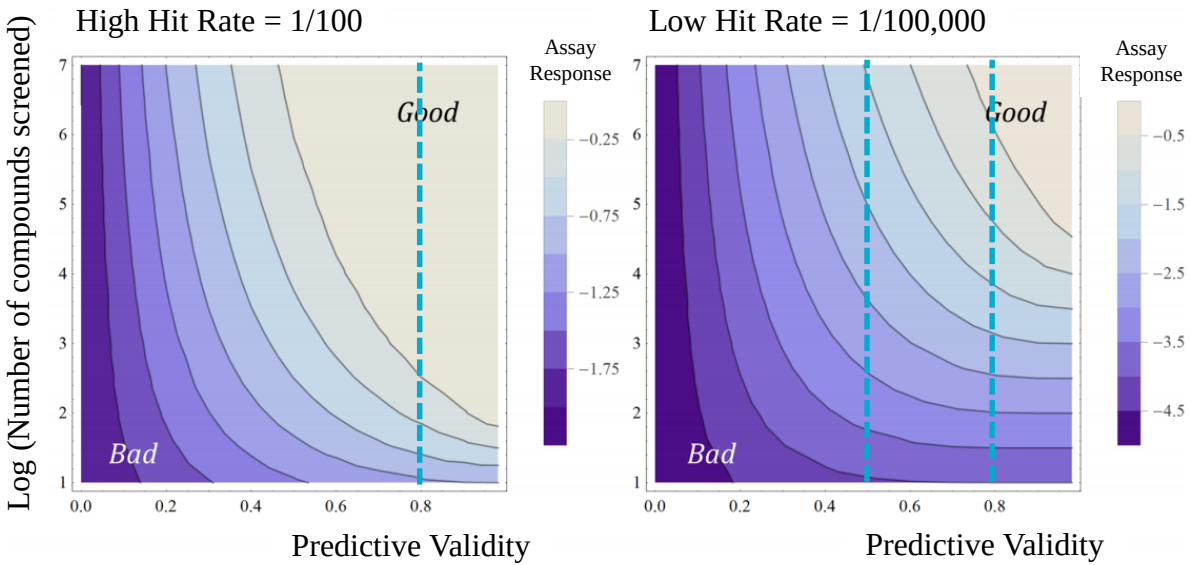
The ‘loss of predictive validity’ theory

- The compromise in HTS is that the assay is now more removed from the phenotype
- We find hits from the assay, but these may not work in the disease state

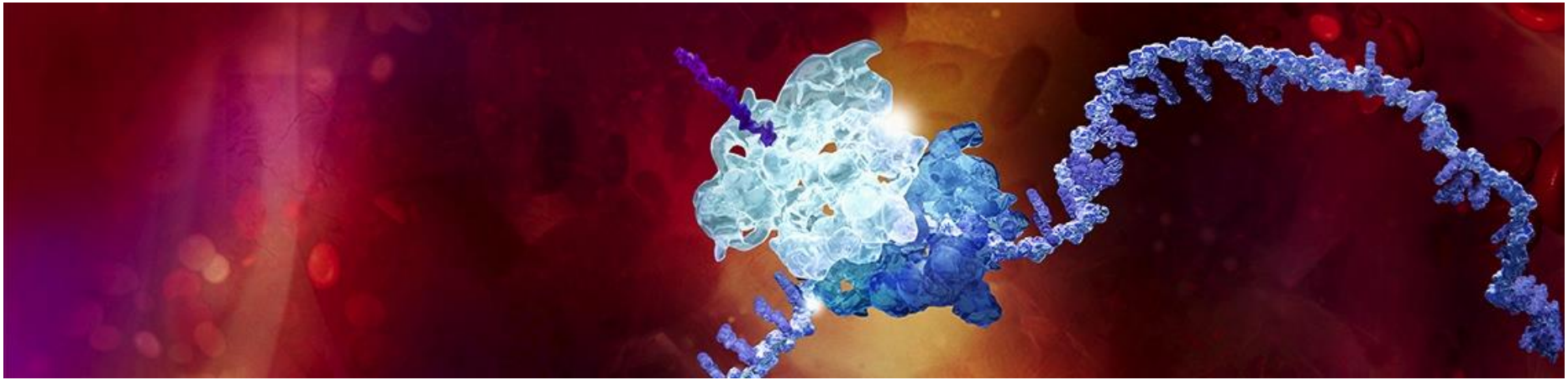
Decision Theory in Drug Discovery

When Quality Beats Quantity: Decision Theory, Drug Discovery, and the Reproducibility Crisis. Scannell JW, Bosley J, *PLoS ONE*, **2016**, 11(2): e0147215. doi:10.1371/journal.pone.0147215

- States that when looking for a 'rare positive', quality >> quantity

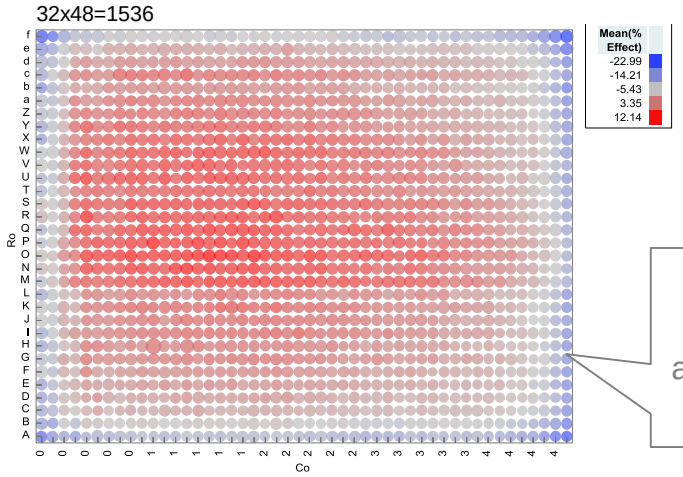


Statistics to the rescue?

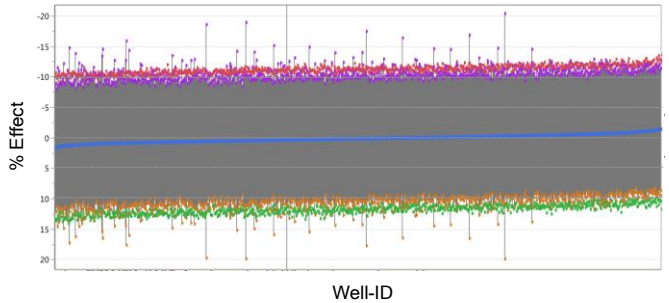


HTS Plate Correction

- Well-to-well assay variability
- Use average across all 600+ plates to correct for systematic difference
- Mean-centring
- Response Scaling
- Correct before further analysis



'Edge effects' – lower response along outer edges of the plate



Signal to Noise

- In HTS we have no replicates
  - Use Z-Score to define significant hits
  - The hits are outliers to distribution
- But standard deviation is sensitive to outliers
  - Use modified Z*-Score
  - Requires negative control on every plate
- Hits are defined:
  - *Z*-Score > N* x *StdDev*
  - *e.g. N* = 4

% effect for compound i

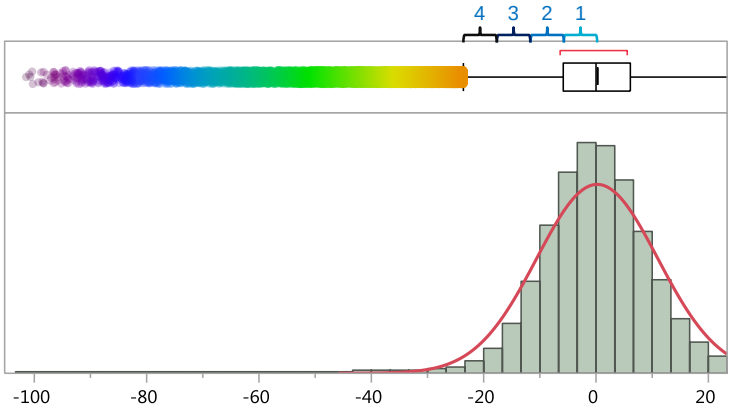
Mean % effect

$$Z\text{-}Score = \frac{X_i - \bar{X}}{S}$$

Standard Deviation

$$Z^*\text{-}Score = \frac{X_i - \bar{X}}{S_{NC}}$$

Standard Deviation of negative control



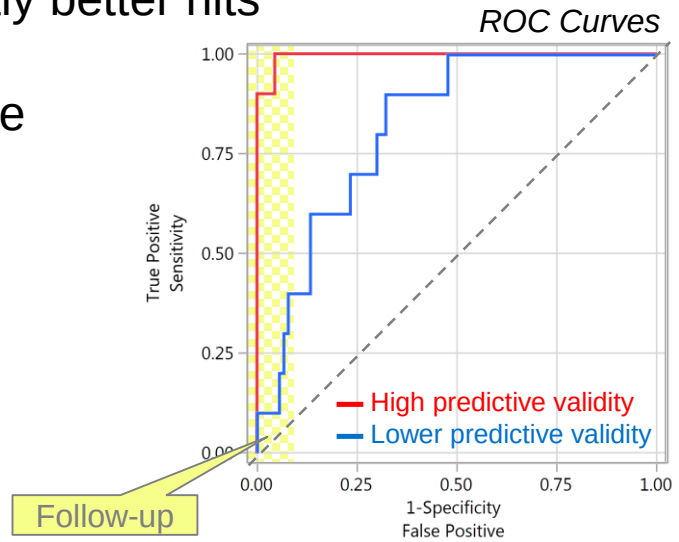
True or False

- False positives can be:
  - Systematic (technology artefact, promiscuous)
  - Random (impurity, decomposition, ‘noise’)

- False positives can ruin an HTS
  - True positives missed in a sea of apparently better hits

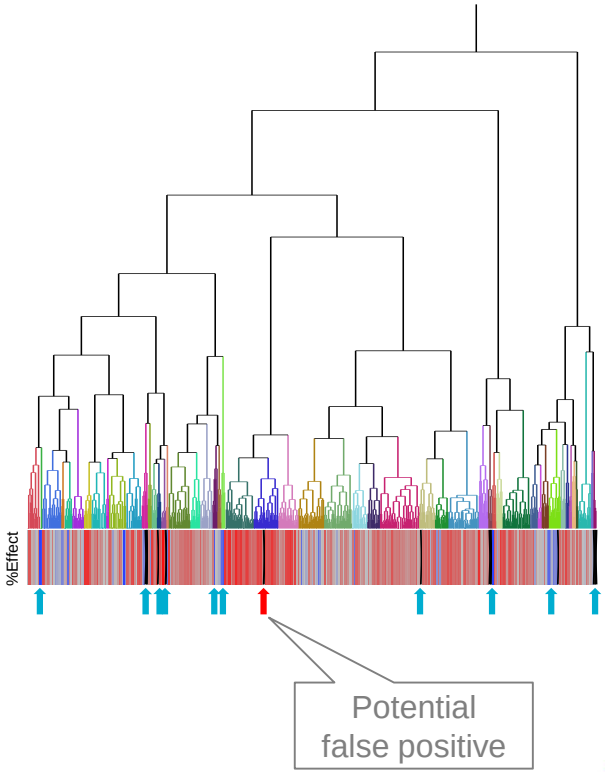
- Low predictive validity = high false positive rate
  - Can miss many true positives

- *Can use biological methods:*
  - *Artefact screen*
  - *Orthogonal endpoint assay*



Similarity Principal

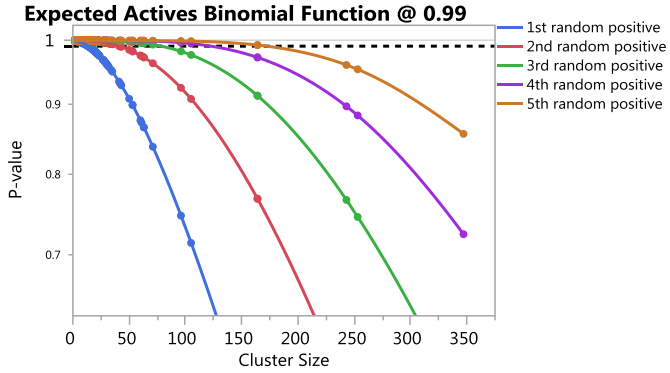
- “Similar compounds have similar biological activity”
  - Exploit with clustering (by chemical structure)
- True positive clusters contain mainly positives
- False positive clusters have only isolated hits
- Limitation:
  - Breaks down for smaller clusters
  - Chance false-positives can mislead
  - Manual step to inspect clusters

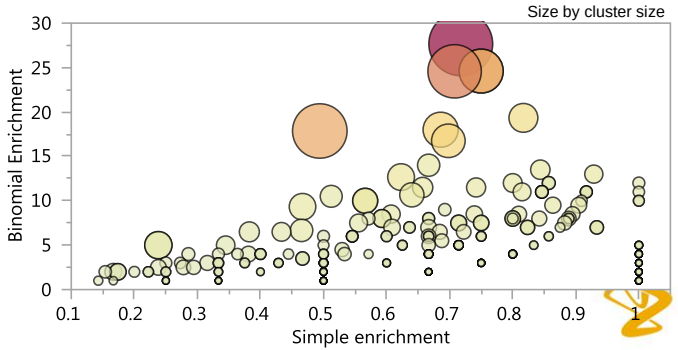


Binomial Enrichment

$$Enrichment = \frac{N^{Active}_{Observed}}{N^{Active}_{Expected}}$$

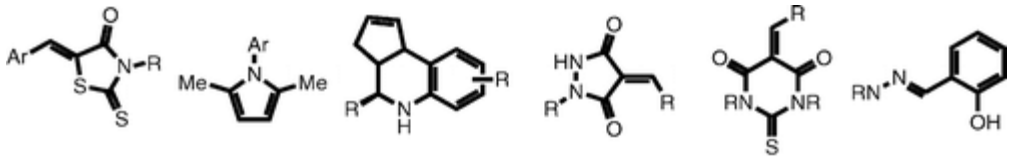
- Unbiased enrichment, accounting for cluster size
  - Expected actives from the binomial distribution
- *e.g.* for a 1% expected random false positive rate
  - If we want to be 99.9% certain of a true positive
    - N>3 clusters may have 1 false positive (33%)
    - N>17 clusters may have 2 false positives (12%)
    - N>40 clusters may have 3 false positives (7%)
    - *etc.*
- Significance is less likely in small clusters





PAINS

- Pan Assay Interference Compounds (PAINS)
  - Baell, JB, Holloway, GA, *J. Med. Chem.* **2010**, 53 (7), 2719-2740
  - e.g.



- Promiscuous compounds that either:
  - Interfere with assay technology to give a response
  - Are genuinely active at many targets (unselective)

- Enrichment analysis will show these as excellent
  - So how do we identify these as false positives?

Frequent Hitters

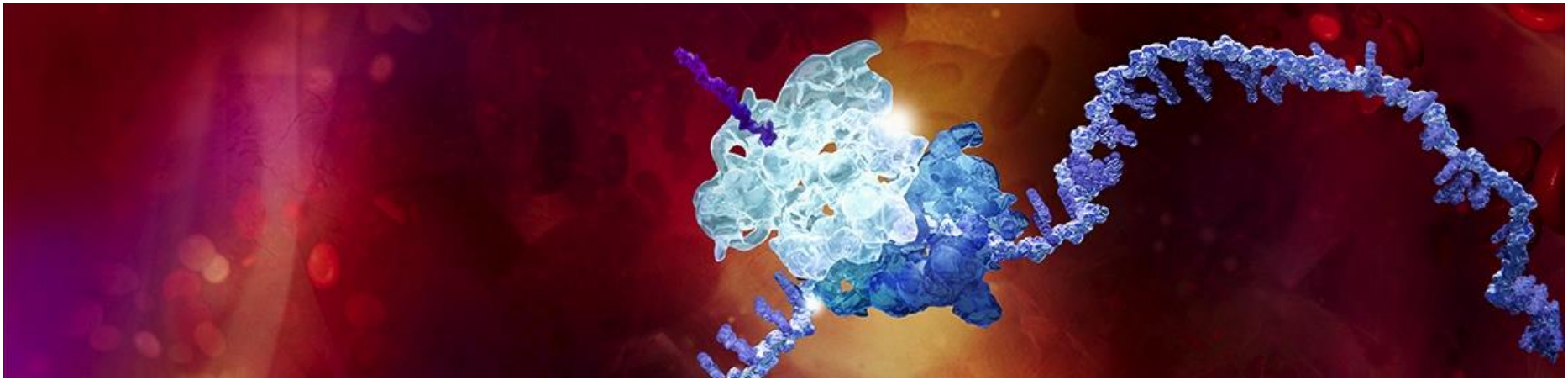
- The answer is in Big Data
  - Mine through 100s of historical HTS campaigns
  - Nissink, JWM, Blackburn, S, *Future Med. Chem.*, **2014**, 6 (10), 1113-1126 (doi:10.4155/fmc.14.72)

- Think about rolling dice
  - Chances of rolling one 6 is fairly high (1/6)
  - Chances of rolling ten 6's is much lower (1/60,466,176)

- Quantify frequent hitter behaviour:
  - A compound is active *A* times across *N* Assay
  - Expected chance of activity *p* is low, *e.g.* 5%
  - Can determine the probability of this from Binomial Survival Function (BSF)
  - If *-logBSF* = 3, then there is a 0.1% chance of this happening

$$\log BSF = {}^{10}\log\, p(a \geq A | p, N)$$

Looking to the Future

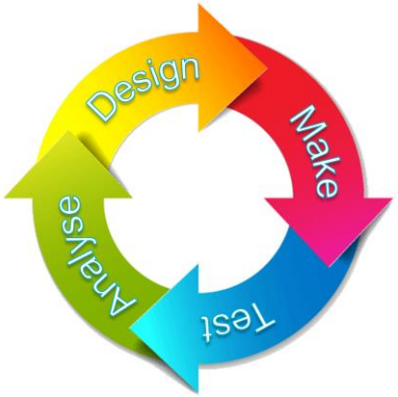


Virtual Compounds, Virtual Screens

- High Throughput Screening is limited to ~10^6 compounds
- DNA-tagged screening has pushed this to ~10^9 compounds
- Virtual Library methods are now looking to achieve ~10^{9+} compounds
  - All based on validated synthesis
- With such large sets, we need
  - Fast virtual screening. FastROCS/GPU can screen 200 per second*
  - High predictive validity (challenge!)
  - Reliable enrichment and clustering methods

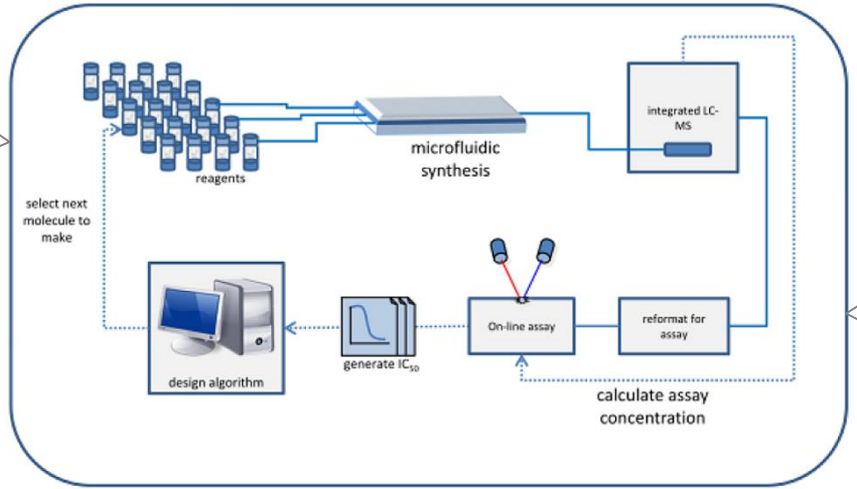
* Boström *et al.*, *Future Med. Chem.*, **2016**, 8 (14), 1739-1752

Automated Drug Discovery



- DMTA – Design, Make, Test, Analyse
  - The iterative cycle of drug discovery
- What if we could automate this?
  - Tarver, G, *et al*., *Med. Chem. Lett*., **2013**, 4 (8), 768-772

Technology is now there to automate synthesis and testing (with restrictions)

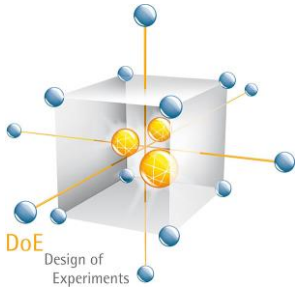


Challenge is in analysis and design. Need algorithms that optimise desired properties in non-continuous chemical space

Automated Drug Discovery

- The challenges:

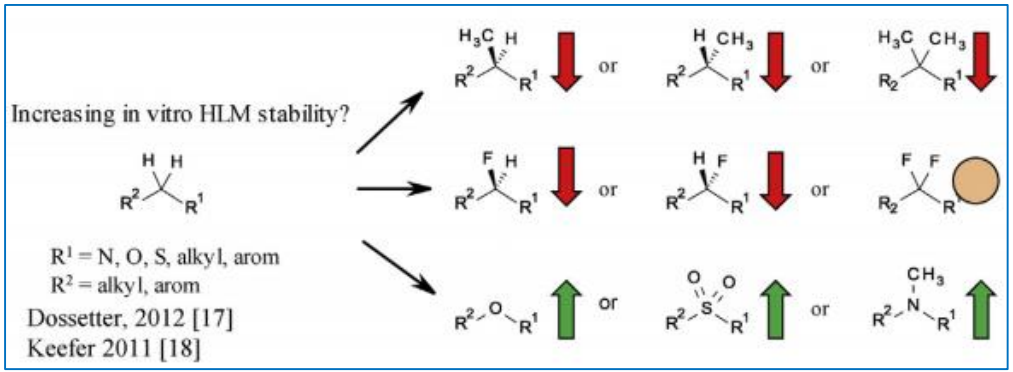
1. How to pick a starting set of compounds to begin?
   – Need some knowledge to narrow down the search space
   – Design of experiments could be used for initial test set



2. How to build a model to design the next compound?
   – Need non-linear modelling / machine-learning
   – Or use hybrid statistical / physical modelling system
   – Iteratively build, test and rebuild model
   – Harness the power of Big Data to guide the model

The Power of Big Data

- Matched-Molecular Pairs Analysis (MMPA)
  - Generate rules about how small changes affect compound properties
  - Dossetter, Griffen and Leach, *Drug Discov. Today*, **2013**, 18 (15-16), 724-731



  - Apply statistical learning to newly designed compounds
  - Optimise quicker!

Summary

Predictive Validity is important	• False positives can ruin an HTS • Quality more important than Quantity
Statistics can help identify chemistry enriched in hits	• Defining a ‘hit’ in the noise • Binomial enrichment analysis
Big Data analysis can enrich our view of an HTS	• Avoid PAINS • Learning new rules
Automated DMTA is the new challenge	• DoE to start? • Machine learning for design?

Confidentiality Notice