Use of data mining and chemoinformatics in the identification and optimization of high-throughput screening hits for NTDs

James Mills; Karl Gibson, Gavin Whitlock, Paul Glossop, Jean-Robert Ioset, Leela Pavan Tadoori, Charles Mowbray

ICOPA XIII

13 Aug 2014

james.mills@sandexis.co.uk

Two applications of chemoinformatics

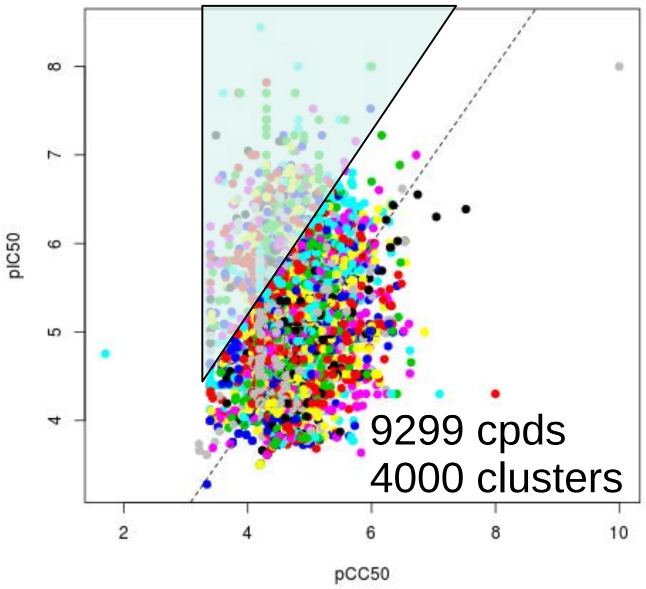
- Identification of novel series from re-triage of HTS results
  - Stage 1: identify all hit series meeting basic criteria
    - Novel and active but not toxic
  - Stage 2: prioritise series for follow-up
    - Generation of view of all related chemical matter
      - eMolecules: commercially available chemical space
      - ChEMBL: bioactivity space

- Optimization of compounds within a lead series
  - Concept of additivity in SAR
  - Apply additivity to compound design

Input data

- Clustered actives from phenotypic HTS*
  - IC_{50}: potency against *Leishmania*
    - Mouse macrophage assay
  - CC_{50}: toxicity against human cell line
  - Seek hits with >10-fold window
    - Or evidence that this can be attained

pIC_{50} vs pCC_{50} coloured by cluster



9299 cpds
4000 clusters

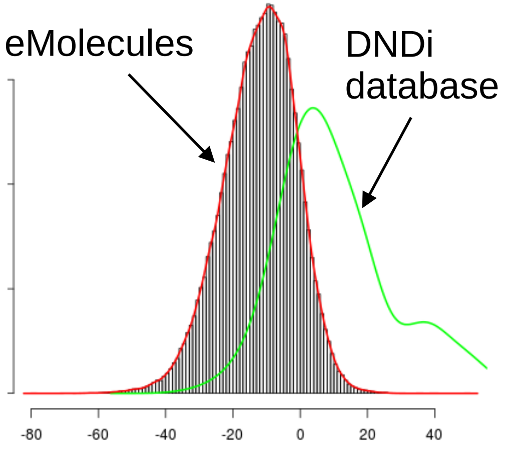
- Pre-existing DNDi dataset
  - *Leishmania* IC_{50} and CC_{50} data for 5000 cpds
    - Data collated from multiple projects and series
  - Seek to avoid this chemical space
    - *i.e.* require novel hit matter

* Note that no data for inactives were available

Stage 1: criteria to triage hit clusters

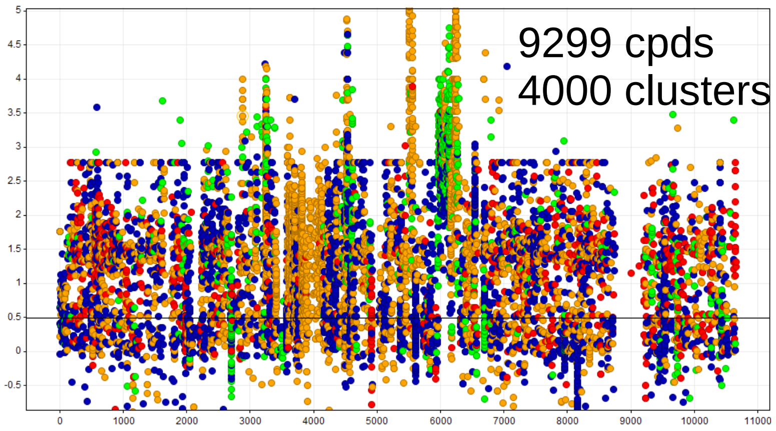
- Is cluster enriched in compounds with >5-fold window?
  - In complete dataset, 60% of hits have >5-fold window
  - Is proportion of cpds with 5-fold window better than 60%?
    - P < 0.1: proportion could have arisen by chance
- Within pre-existing DNDi chemical space?
  - Built Bayesian model to score cpds
    - High: contains features common in DNDi dataset
  - Favour compounds with low scores
- Structural alerts based on toxicity literature
  - Traffic-light system
- Drug-like properties (MWt <500, clogP < 6…)

Histogram of Bayesian scores



Hit series triage

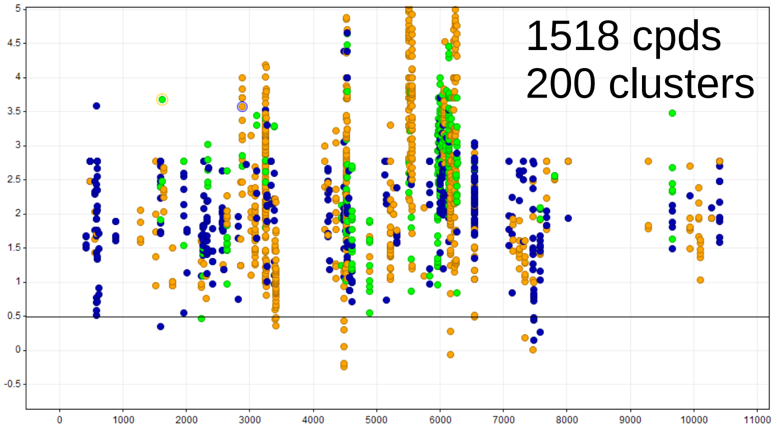
Potency/Toxicity window vs *cluster, coloured by alerts (blue = clean)*



9299 cpds
4000 clusters

Automated filters

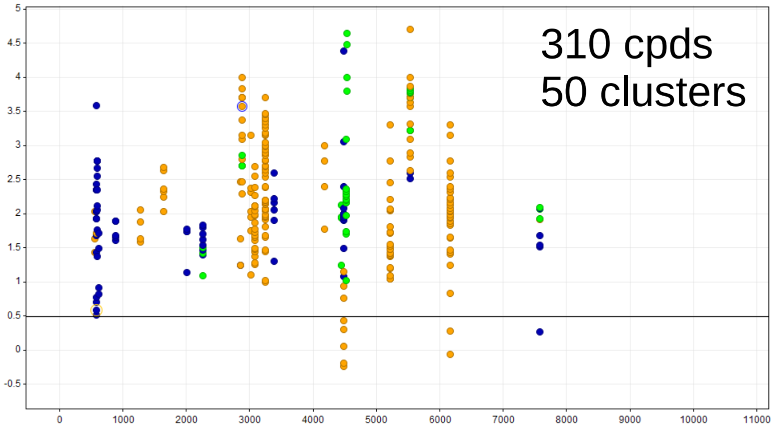
1) Properties
2) Alerts
3) DNDi-like
4) P < 0.1



1518 cpds
200 clusters

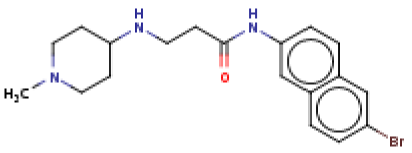
Manual selection

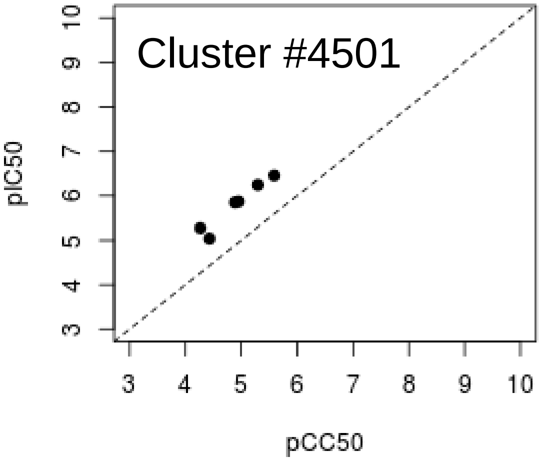
1) Known series
2) Synthesis
3) Developability



310 cpds
50 clusters

e.g.

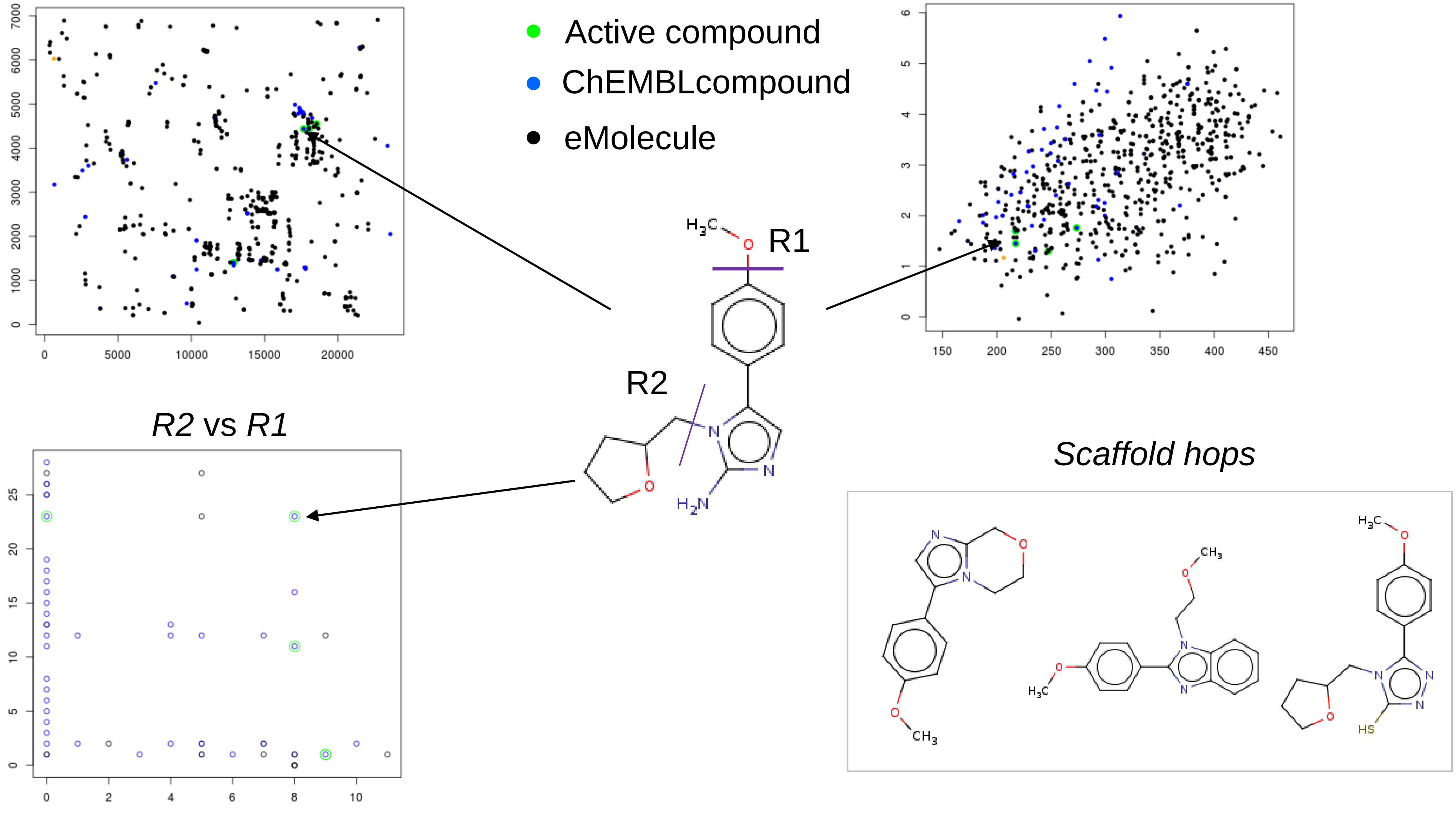




Stage 2: prioritisation of clusters

- Which of the 50 clusters should we follow up on?
- Sent data package to panel of NTD med chem experts to assess:
  - Probability of compound optimisation to drug
    - Potency *vs* toxicity
    - Scope for modification
      - Are there compounds to order in and screen?
      - Rapidly test local SAR of core and substituents
    - Potential off-target activity
    - Likely ADMET properties (metabolic stability etc.)
  - Precedence in neglected diseases
- Each cluster tagged as high/medium/low priority

Characterisation of local chemical space



Property index vs structure index

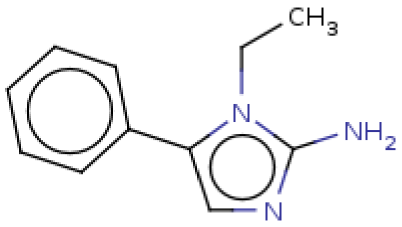
MWt vs clogP

- 🟩 Active compound
- 🟦 ChEMBLcompound
- ● eMolecule

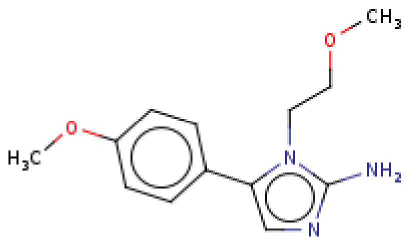
R2 vs R1

Scaffold hops

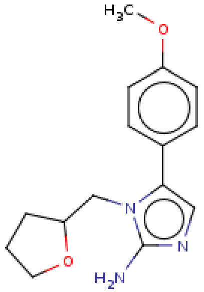
Identification of local bioactivity space

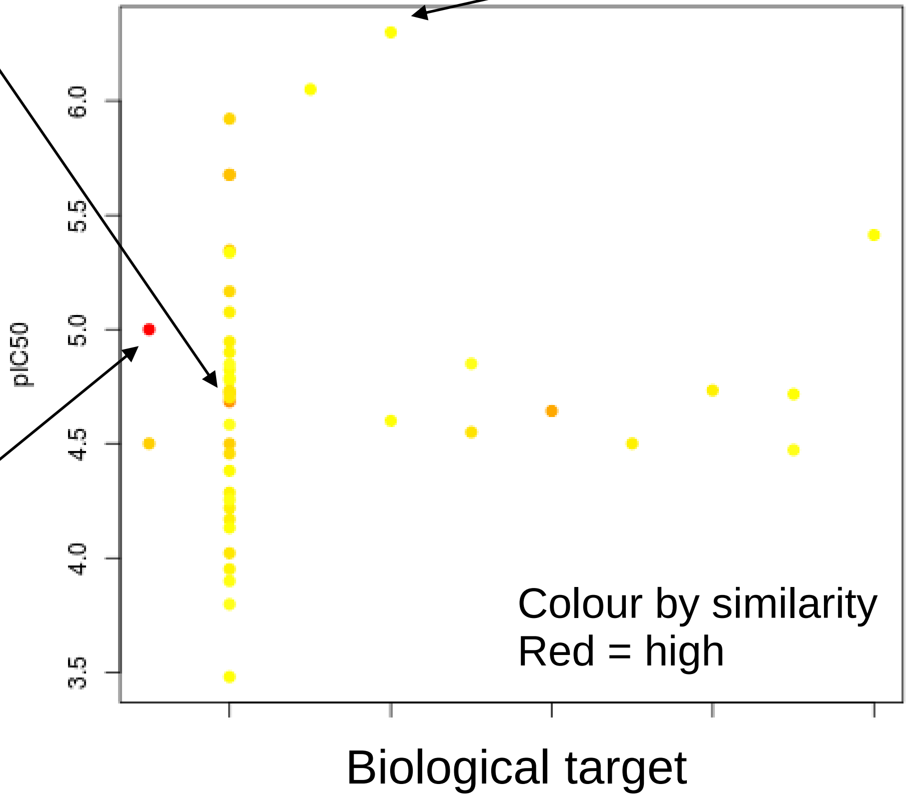


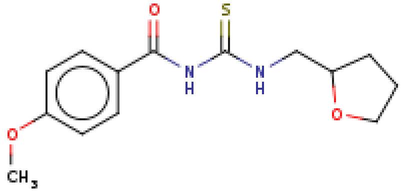
CHEMBL1649788
Similarity = 0.72
Pseudomonas 21µM



CHEMBL1561575
Similarity = 1.00
Arachidonate lipase 10µM

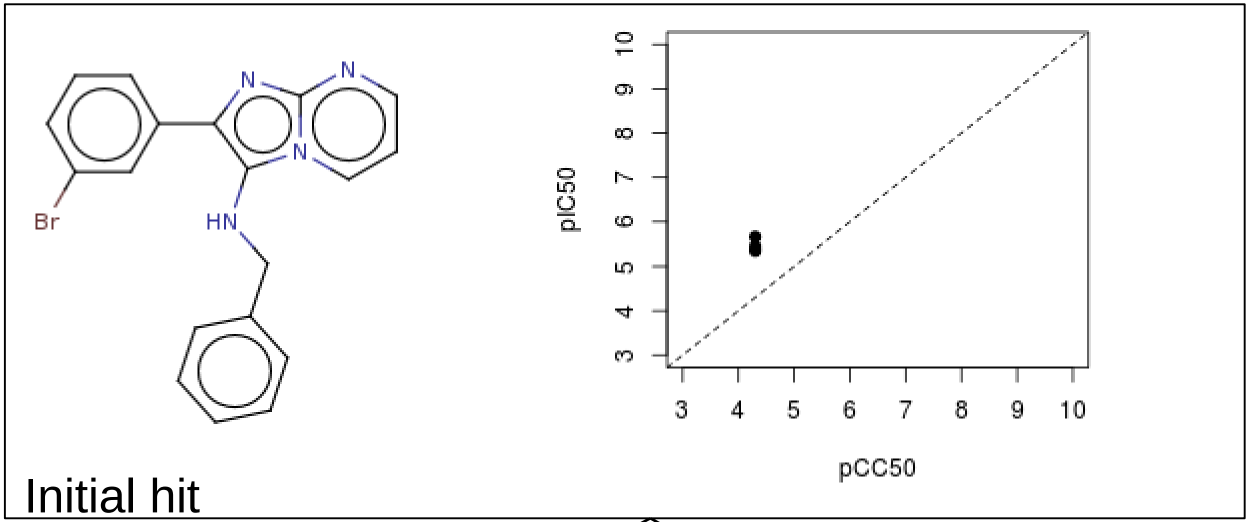




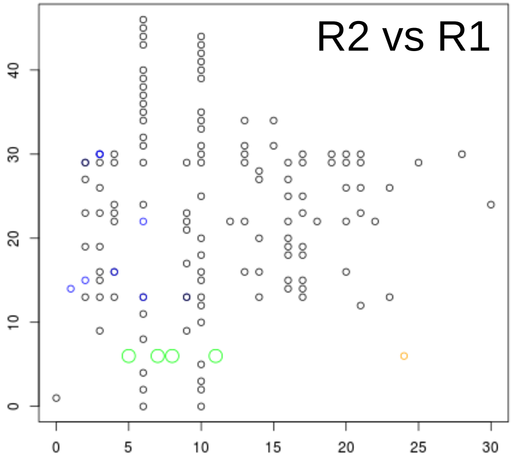


CHEMBL1448630
Similarity = 0.48
DNA helicase Q1 890nM

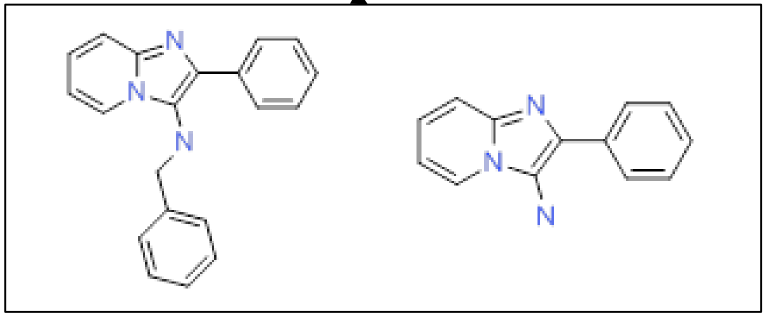
Result: identification of novel series



Initial hit



Confirmed activity within series



Actives from scaffold hop

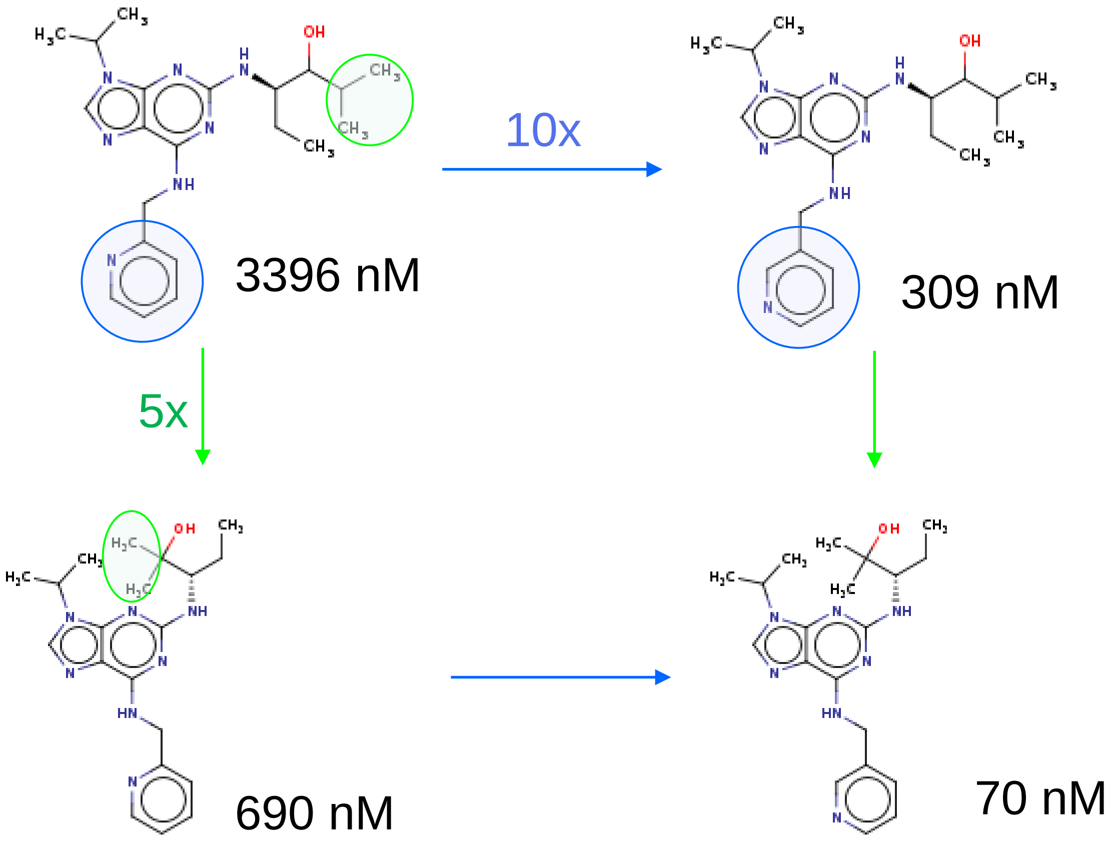
Hit re-triage summary

- Revisiting and consolidating legacy data can prove useful
  - Identified novel series with anti-*Leishmania* activity

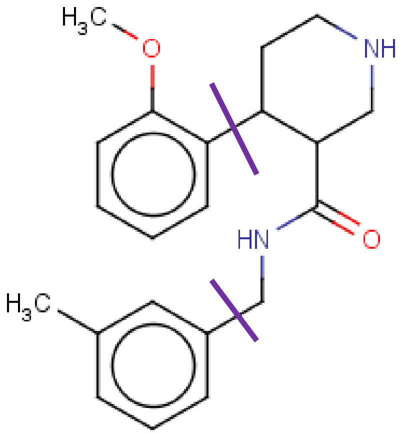
- Largely automated HTS triage to identify 50 chemotypes
  - Series enriched in actives with a window over toxicity
  - Singletons with a window over toxicity
- Deeper dive to prioritise the 50 selected chemotypes
  - Assessed local chemical space for precedented compounds
    - Evidence that synthesis is possible
    - Select compounds for immediate screening
  - Assessed bioactivity data to suggest mode of action/toxicity

Lead optimization: SAR additivity

- Double-mutant analysis
  - *e.g.* CDK2 compounds from ChEMBL: additive SAR



T. brucei piperidine series



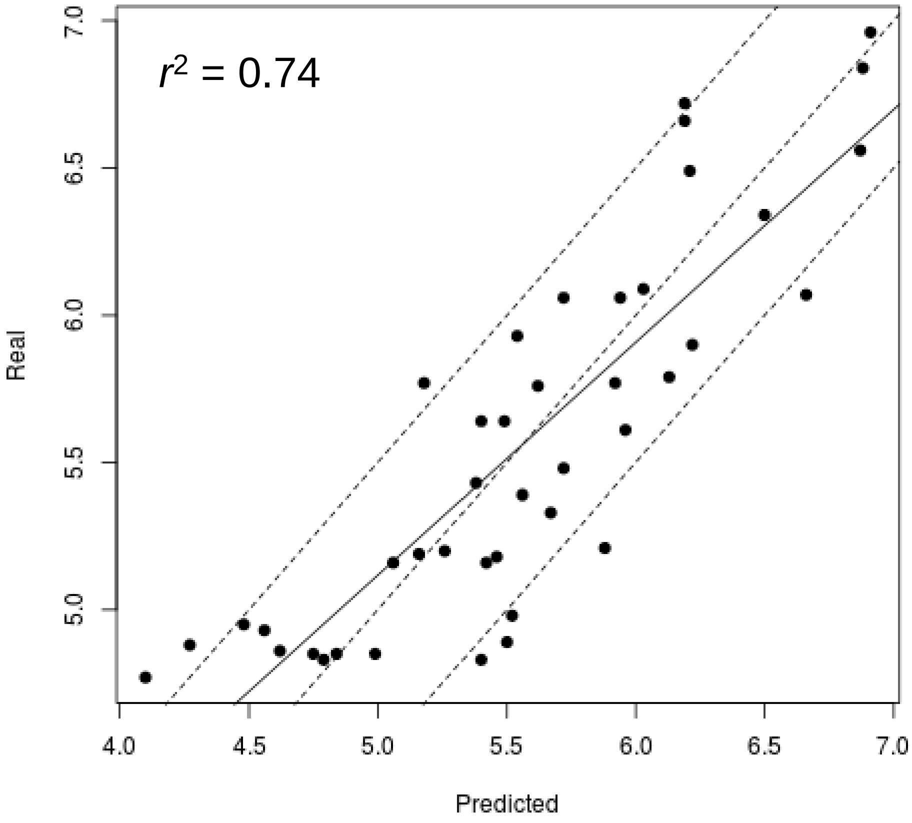
IC_{50} 458nM at *T. brucei*
CC_{50} 44µM

- Assess additivity of series
- Apply additivity to prediction of more potent compounds

Assessment of additivity

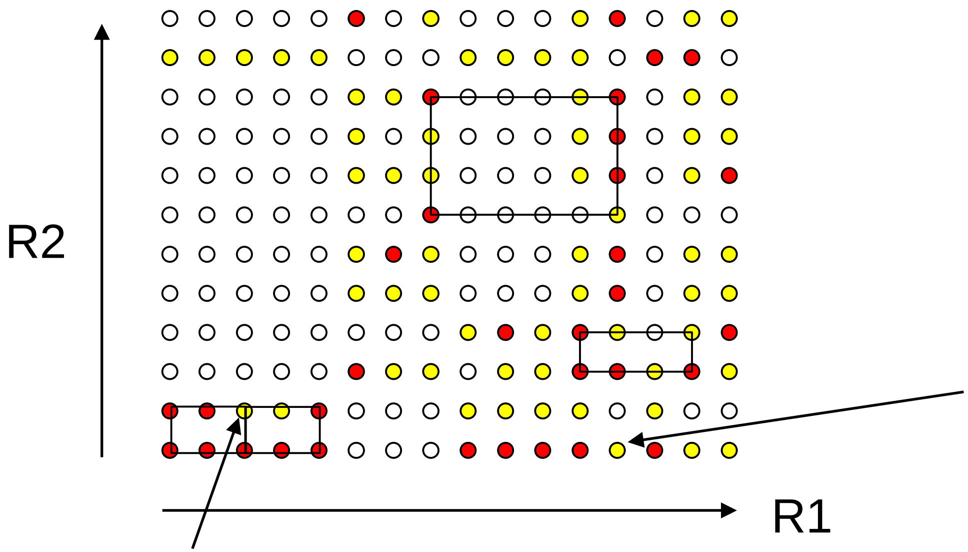
- For each square, predict potency of 4th compound from other 3
- Deviations from prediction
  - <10-fold: within experimental error
  - >10-fold: non-additivity?
    - or submit for retest
- This series shows additive SAR
  - Use squarewise analysis to predict
  - Expect accuracy within 3-fold

Real vs predicted pIC50



Application of additivity

- Fill gaps in chemical space
- Predict potency for 4th corner of all possible squares



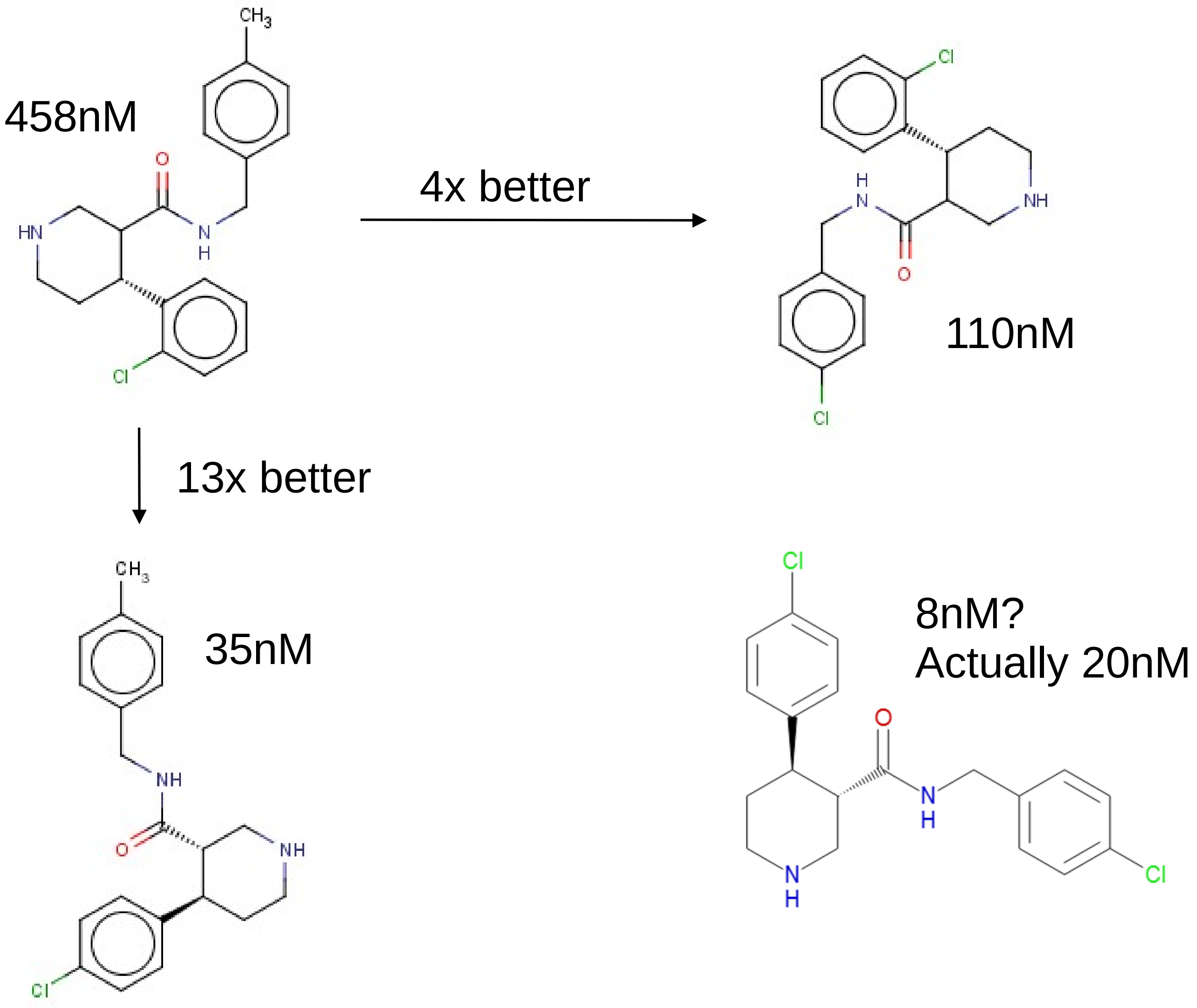
- 🔴 Already made
- 🟡 Activity predicted
- ○ Activity not predicted

Making this compound would enable large number of predictions

Use mean prediction if more than one square predicts

- Also use to suggest informative compounds

Results from squarewise analysis



458nM

4x better

110nM

13x better

35nM

8nM?
Actually 20nM

Summary

- Extract maximal information from accumulated data
  - In particular public datasets (ChEMBL)

- Identification of novel series by re-triage of HTSs
  - Reduce dataset to series of interest
  - Extract all salient information for these series
  - Apply med. chem expertise to interpret information

- Optimize potency within a given series
  - Assess additivity of all data
  - Apply additivity to low-risk prediction of unmade compounds

Acknowledgements

- iThemba: synthesis on *T. brucei* project
- EBI: ChEMBL database
  - https://www.ebi.ac.uk/chembl/
- LMPH Antwerp: testing on *Leishmania* project
- Scynexis: synthesis and testing on *T. brucei* project
- Pfizer: *T. brucei* chemical matter

Squarewise *vs* Free-Wilson

- Free-Wilson assigns weights to each functional group
  - Potency is sum of weights for each group

Free-Wilson	Squarewise analysis
Assumes additivity	Assumes additivity
Predicts full *n* x *n* matrix	Predicts incomplete *n* x *n* matrix
Fits variables to data	No fitting of data
All Rgp occurrences contribute to prediction *i.e.* global model	2 Rgp occurrences contribute to prediction *i.e.* local model