



Three way comparison of Chemical Spaces using BCUTs (and avoiding structural exchange)



Jens Loesel
CompSci CoE Sandwich



Outline

Part I

- Motivation
- Background / Theory

Part II

- Examples
- Critical discussion

Part III

- Summary

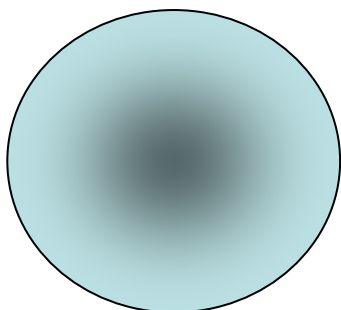


Motivation

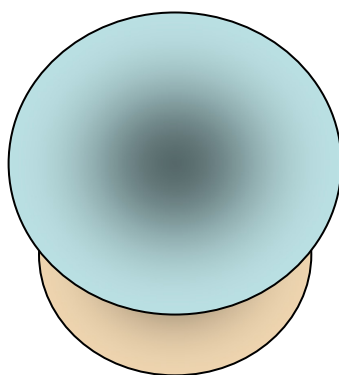
- Common questions asked by management
 - Can you compare the external set X with our own file
 - Can you advice how difference it is to our own file
 - Can you evaluate the quality
 - But please don't spend to much time on it / we need it tomorrow
- Recurring unease
 - Different is easy to define – but does it really mean good
 - In some cases it's not practical or even possible to get individual structure information – what to do in these cases

Chemical space simplified

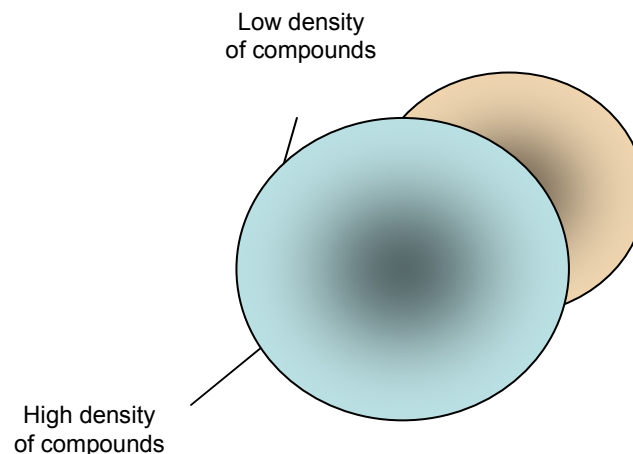
- Three typical cases of comparing chemical spaces
 - Set A (blue) is the reference, Set B (yellow) an unknown (Test) Set



B is subset of A
random, no difference
in Chem Space



B overlaps A
'Small' difference
in Chem Space



B is distinct from A
'Huge' difference
in Chem Space



Normalize Chemical Space

5 steps to a single score

1. The Background
 - Define a set of compounds to represent all drug-like space
2. Binning
 - Divide the chemical space into bins (using BCUT)
3. Normalize the Reference Set
 - Assign a value between 0 and 1 for each bin based on reference and background (higher means better)
4. Map the Test Set into the same space
 - Assign scores from Step 3 to each compound
5. One single score between 0 and 1 to describe them all
 - Average all individual scores from step 4



In case you prefer math

Scaling factor

$$k = \frac{N}{A}$$

Score for
Individual bin

$$score_i = \frac{k \times A_i}{k \times A_i + N_i}$$

Score for the
Test Set

$$score = \frac{1}{B} \times \sum_{i=1}^{i=B} score_i \times B_i$$

with

- N = total number of compounds in background Set
- N_i = number of compounds of background Set in cell I
- A = total number of compounds in Reference Set
- A_i = number of compounds of Reference in cell I
- B = total number of compounds in unknown Set
- B_i = number of compounds of unknown in cell i



Yes - but

- What do you use as background
 - The Pfizer corporate screening file
 - it's big, it's drug like, it's diverse
 - Be pragmatic – don't try to be perfect
- What constitutes a reference set
 - Any set of compounds really
 - Ro5 (non)compliant, big/small (MW), lipophilic (logP), (un)stable (RLM, HLM), soluble, active against target x, active against any target, calculated or measured
 - Natural products would make a great additional set
- How big does the reference and test set need to be
 - 100K+ for reference, low hundreds for test set



No need to exchange structure

- The scoring is structure less
 - Scoring solely based on occupancy numbers
- Data exchange
 - Send out BCUT space definition and bin boundaries
 - Receive back bin occupancy numbers
- The exchanged information is safe
 - BCUT cells can't be turned back into structures
 - Advantage towards fingerprints. A genetic algorithm can turn them back into structures which are identical or at least close to original compound



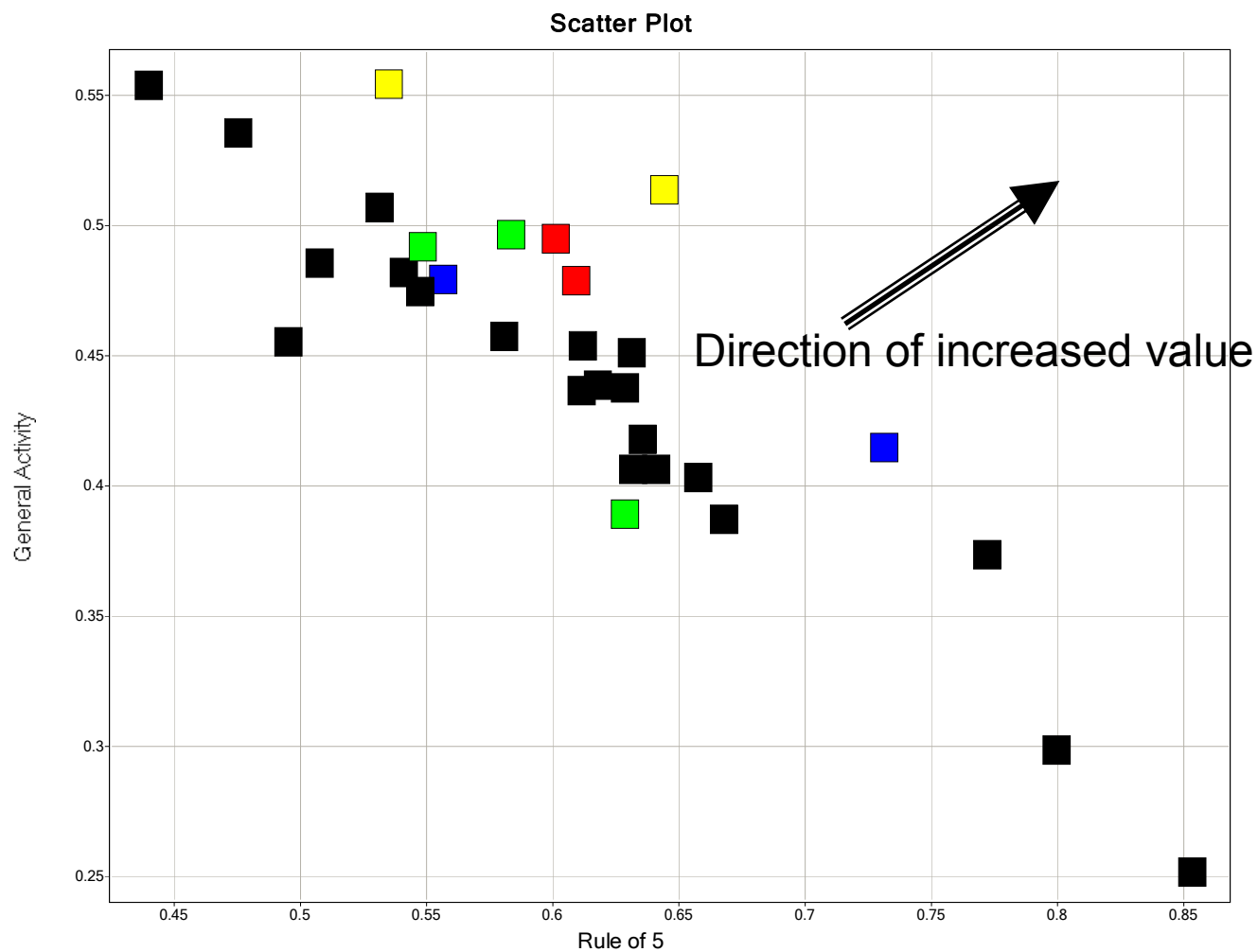
Case study I

1. Focus on external vendors / screening sets
2. Focus on simple scores
 - Known activity against ANY target, Ro5, logP, MW, solubility,
3. For simplicity ensure high score (> 0.5) is always 'good' and low score (< 0.5) is 'bad'
 - If necessary invert score
 - i.e logP displayed = $1 - \text{score}$ for high logP reference set

 - Public Vendors black
 - Pfizer green
 - External sets under investigation red
 - Launched Drugs (all, Ro5) yellow

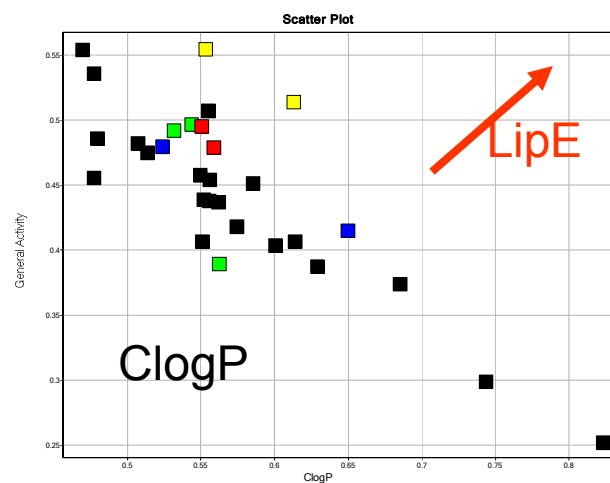
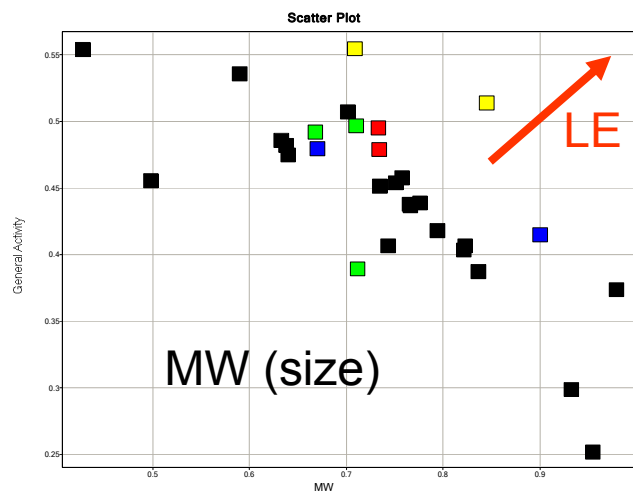
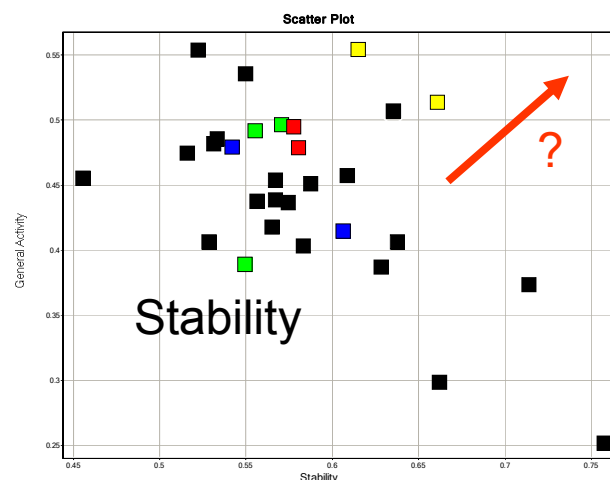
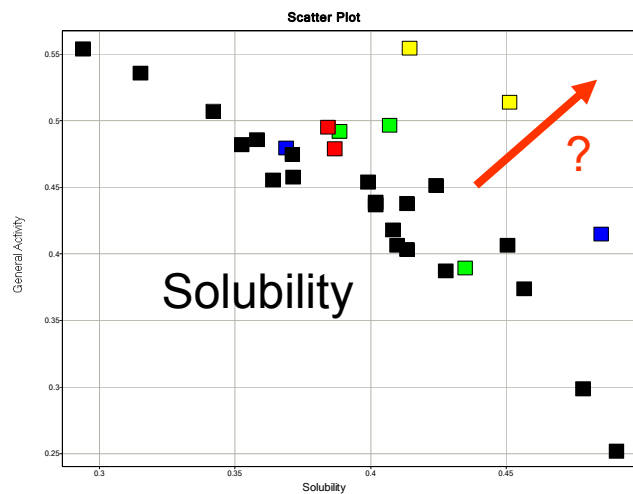


Ro5 vs general activity





Other relationships





What have we learned

- Near linear trend between all activity space and important properties for oral drugs
 - Activity space and good oral property space are separate
 - Finding the rare overlap in space is the daily work of MedChem
- Launched Drugs – the constant outliers
 - It is possible to break out of the trend line
 - But most suppliers don't manage
 - Even 100% compliant Ro5 drug set scores 'low' on Ro5 space !!
- The easiest way to be different is to be extreme
 - New compounds should be different – but in a 'good' way which seems quite elusive



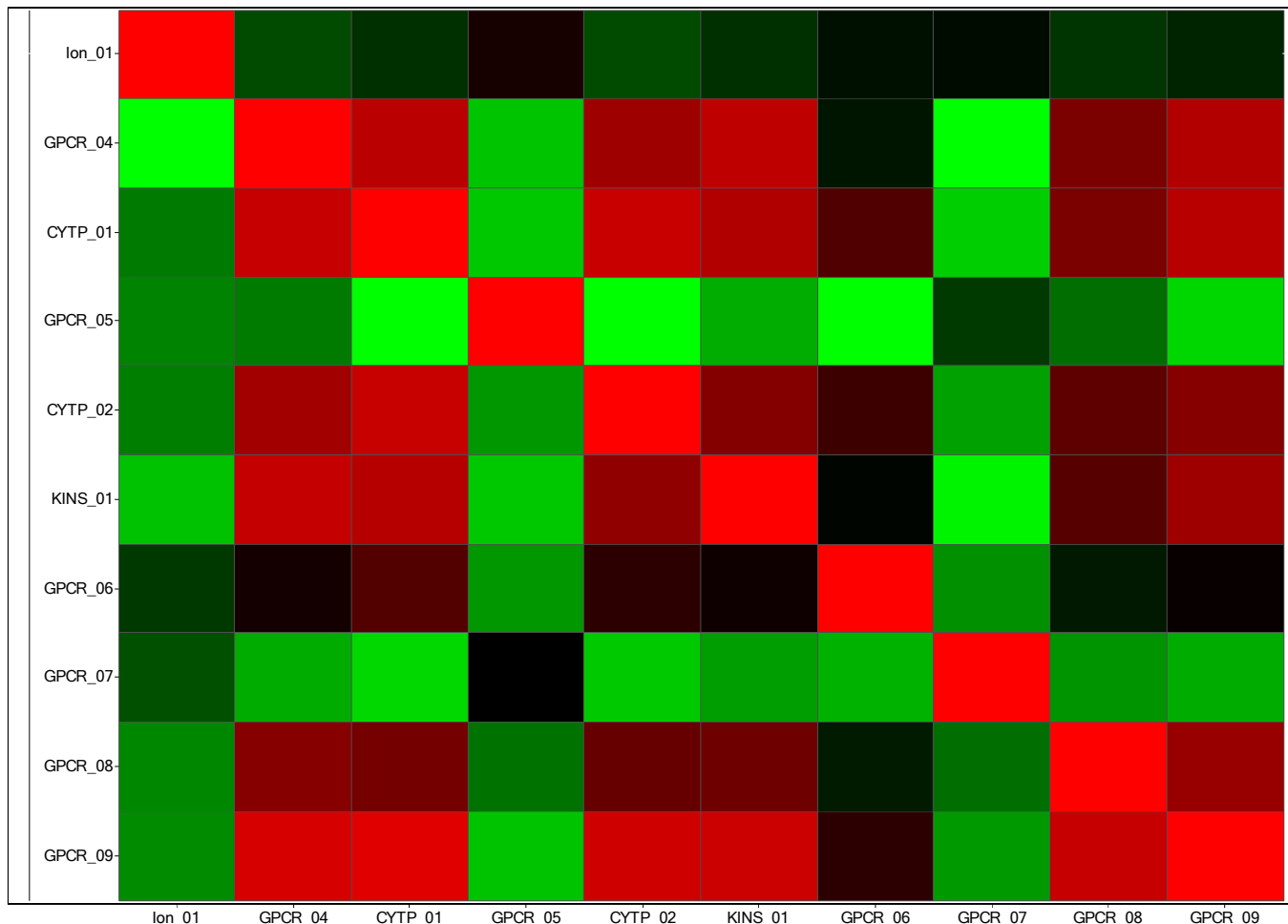
Case study II

- Mapping gene target scores against each other
- Example to use a reference set as test set and vice versa
- Work in progress



Partial Heat map of 500*500 array of target spaces

Heat Map





Poly Pharmacology

- High scores are in agreement with known close targets
- Scores do allow to map across target families to find targets with known related structures
- The methodology allows to generate a full matrix – including low scores for unrelated targets

- Still a lot more to do



Summary

- Using a three way comparison of chemical spaces puts 'difference' into context
- These comparisons can be quantified in a single score
- Activity and good properties are in different areas of chemical space making the identification of new drugs a hard business
- The methodology can be applied to gene targets to identify relationships between their chemical spaces



Acknowledgements

- Alex Alex, Tony Wood, Jeremy Everett who started the work asking me to quantify quality an external data set
- The Scripps Institute for early discussions
- Colleagues in CompChem and the CompSci CoE for input along the way
- All my listeners here