High-throughput medicinal chemistry: A comparison of theoretical methods

Colin Edge & Alfonso Pozzan

Computational, Analytical & Structural Sciences

GlaxoSmithKline

Today's Talk

Computational Methods for HTC

- Examples of GSK methods
  - 1D
  - 2D
  - 3D
- Current limitations
  - Data
  - Models
  - Experiments
  - Computers
  - (Knowledge?)
- GSK's strategy
  - Good in-house data
  - Better models
  - Distributed, automated system (NetHTS)

What methods are available?

Short answer : any method - for any purpose!

Methods in reality (i.e. what I shall discuss…)

- Fragmentation (1D/2D)
- ADMET prediction/property profiling (2D++)
- Fingerprints (2D/3D)
- Shape comparison (3D)
- Docking/Virtual Screening (3D++)
- Optimisation (1D/2D/3D++)

Purposes:

- Ligand-based array design
- Structure-based array design
- Early lead array design
- Prospective chemistry array design

Difficult/impractical:

- High-resource methodology (time, money, people)

Fragmentation/classification methodology

Automated hierarchy

- Adapted from Proasis2, GSK protein structure db
- Fragmentation via:
  - CLOGP
  - SMARTS
- Used for :
  - Observed common (‘privileged’?) fragments
  - R-group analysis
  - “Thing-like” scoring

ADMET predictors/filters for high-throughput chemistry

- Need to be fast(ish)
- Need to be generic ‘global’ models
  - Derived from large data set
  - Applicable to many problems
- …or carefully labelled as local models
  - Derived from specific data set, e.g.
    - Target class
    - Compound class
  - Limited transferability

Typical ADMET filter rules of thumb*

- Absorption (rules for low Abs):
  - PSA < 140
- CNS (rules for low CNS)
  - PSA < 90, MW <450, logD 1-3
- Bioavailability (for low bioavailability)
  - MW < 500, ClogP<5, N(N,O)<10, N(NH,OH)<5
- CYP450
  - Structural alerts (imidazoles, pyridine, etc)

Useful for HTC? Global vs local. More GSK examples later…

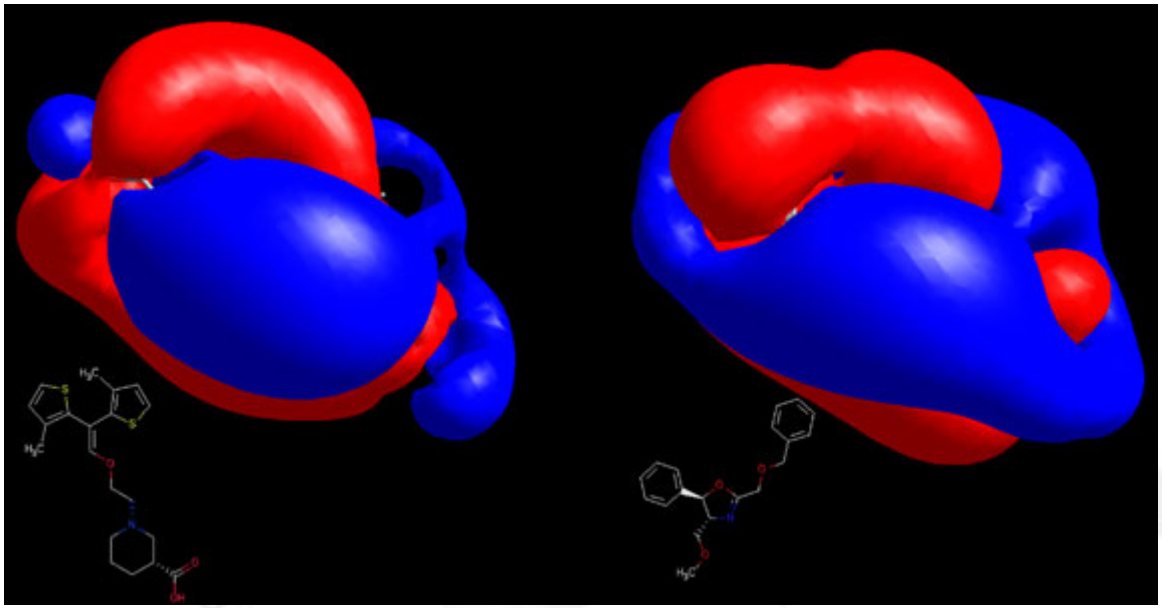
* N.B. These are literature rules, not GSK rules!
e.g. **Clark,** DDT ***8***(03)927, **Lipinski et al.,** AdvDrugDelivRev***23***(97)3

2-D and/or 3-D: Fingerprints*

- Fingerprint = simple list of features
  - Concise
  - Portable
  - Comparable
- Examples
  - Daylight – atoms and bonds (2-D)
  - 3DMill – catalyst pharmacophores (3-D)

*Review: Willett et al "Chemical Similarity Searching" JCICS ***38*** (1998) 983

Beyond 2D:
Shape comparison/bioisosteres*



Tshape > 0.75,
Telectrostatic > 0.3

- Omega (fast conformation generation)
- ROCS (shape comparison by Gaussian overlap)
- EON (electrostatic comparison)

Useful for HTC? Yes, for medium-sized libraries

*** Thornber**, "Isosterism and Molecular Modification in Drug Design" Chem. Soc. Rev. ***8***(1979)563
Open Eyes: http://www.eyesopen.com

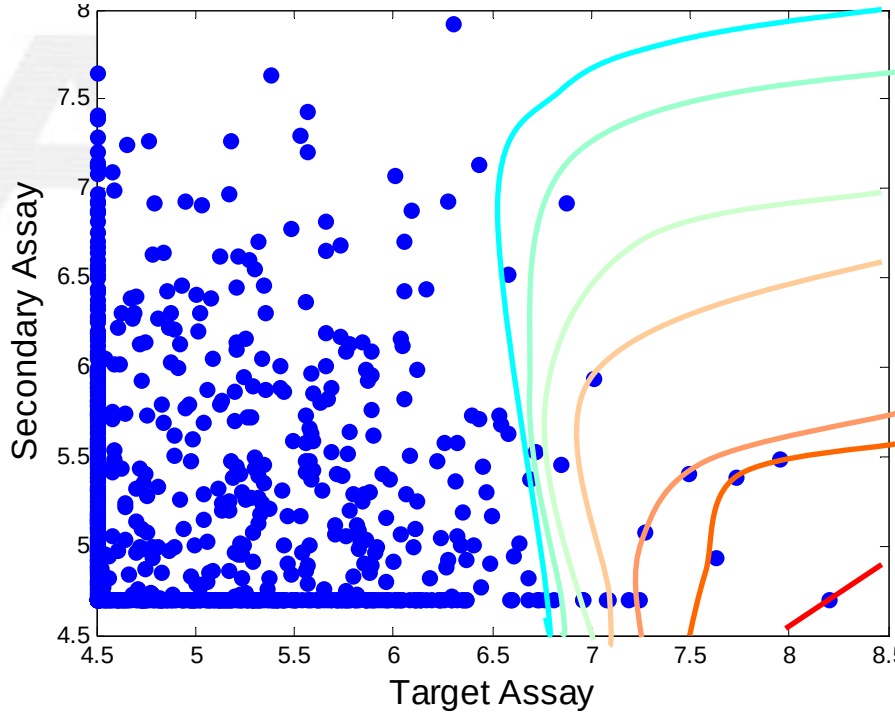
Beyond 3D: Docking/Scoring

- GSK study*
  - Used 1303 molecules versus 8 proteins
  - 10 docking programs tested
  - 37 scoring functions tested
  - ‘Answers’ not in public domain!
- The current state of the art:
  - Often suggest the right pose amongst a list of many wrong ones
  - Pretty hopeless at ranking the list
  - Definitely no good for binding affinity prediction!

* **Warren et al.,** “A Critical Assessment of Docking Programs and Scoring Functions” JMedChem ASAP 10.1021

Optimisation methods

- ADEPT*
  - GSK Daylight-based ‘work-horse’
  - Profiling
  - Diversity
- MOGA**
  - Multi-objective optimisation
  - Pareto ranking
  - Problems with large numbers of objectives



*Leach & Hann, “The in silico world of virtual libraries” DDT ***5***(00)326

Gillet et al., “Combinatorial library design using a multiobjective genetic algorithm” JCICS ***42*(02)375

Nate Woody

First step in building models

GET THE DATA!

SOBAX

- GSK collection of '*tangible*' molecules

- CODS
  - GSK inventory
- Various commercial databases…

In other words, good predictors need good data!

… and the only way to get good data is do it yourself…

The GSK screening collection*

- While other companies were chasing the millions…
- …GSK reduced its screening collection from about 1.4M to 0.6M during QA exercise
- BUT…
  - we know that they are ‘pure and sure’!
  - fewer false positives in screening
  - and thus, a good basis for models
- AND… our total has since recovered.

*** Lane et al.,** “Defining and maintaining a high quality screening collection – the GSK experience” DDT ***11*** (2006) 267

Example: Bioavailability beyond Lipinski…

- Continuous variables
  - MW and ClogP
- Discrete variables
  - Numbers of donors/acceptors
- Where's solubility? (pKa and others…)

- We measure/calculate the easy/quick things
- Our hypothesis is a function of these

Examples of changes

- Change the data
  - Measure other properties
- Change the model
  - Calculate H-bond strengths
- Change the attitude
  - Faster simulation somehow?
  - Slower experiment is acceptable?
  - Accept model limitations
  - Devise high-throughput bioavailability measure

What's wrong with the future?

- Many calculations cannot be done at all
  - There is a theoretical limit (and a much lower practical limit) to what can be calculated
- Most calculations cannot be done in time
- Computers will never be fast enough and are already too fast!
  - Colin's Law of Computer Work – an Inverse Parkinson's Law
    - "Computer Work Shrinks To Fill The Time Available"
- Human bias affects required models:
  - Rules of Thumb treated as LAWS – "its CLOGP is 5.1"
  - Gut Feeling – "can't tell you why, but I don't like the look of that"
  - Tyranny of the Urgent – "we'll have to make it anyway"
  - Fetish of Potency – "but the lead's 0.3 nanomolar!"

So what does the future really look like?

An ideal automated library design platform

- Generic needs
  - Minimum input required (e.g. SMILES)
  - No human intervention during the process
  - Suitable for high throughput
  - Modular, flexible, and other admirable buzzwords
- GSK-specific needs
  - Able to highlight other opportunities (e.g. cpd acquisition)
  - Able to increase library scope beyond the original one (recycling chemistry/DRAF* proposal)
  - Suitable for cooperative work between sites
  - Suitable for GSK/DRAF* workflow

* DRAF=GSK-speak for Discovery Research Automation Facility, our integrated screen-to-lead buildings

Pros and Cons…

- Pros
  - Flat files (not Oracle)
  - Local expertise maintained
  - Good dissemination of the compound sets
  - In-house (‘tweakable’)
- Cons
  - Flat files (not Oracle)
  - Disk space (4Gb / 1M molecules)
  - Requires scripting/programming expertise
  - In-house (requires supporting)

Currently available utilities

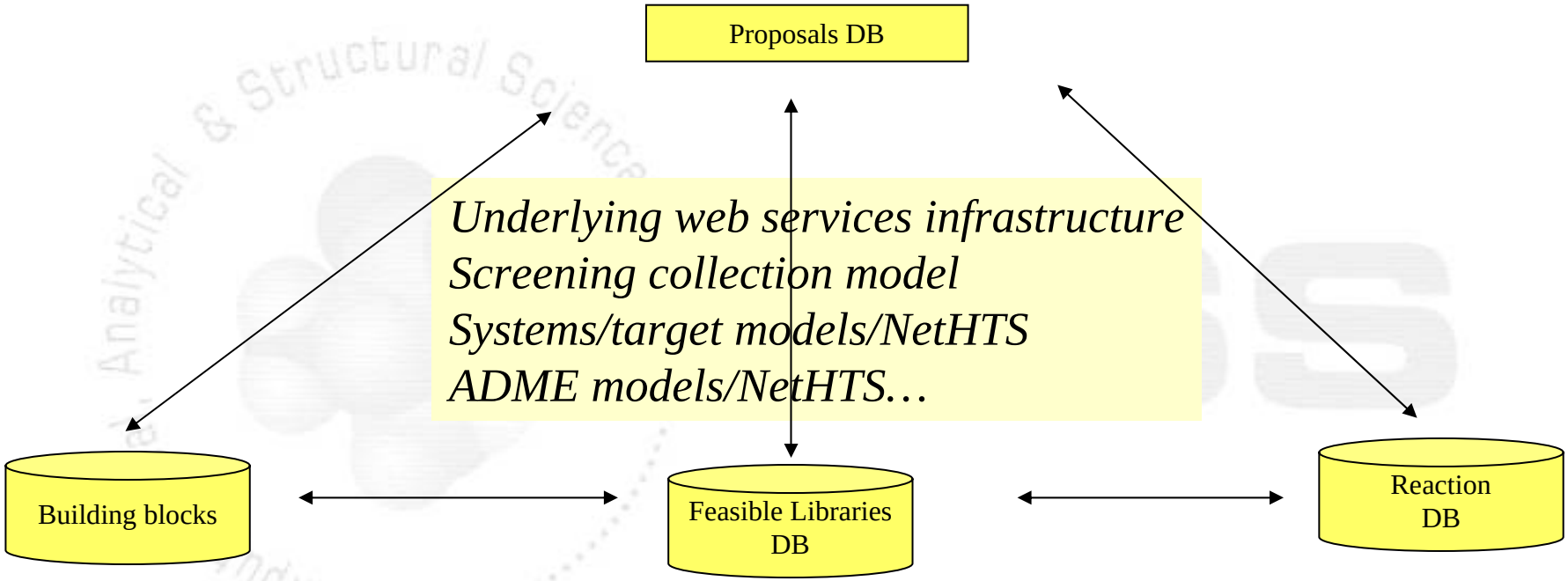
- Simple ADMET predictors
- The “similarity principle”
  - 2D (Daylight) fingerprints
  - 3D (3DMill) fingerprints
- Pattern recognition
  - Pharmacophoric fingerprints
- Systems-based

Strategy:
GSK Library Design

- **What can be made?**
  - **Feasible library** strategy to library design
- **What can be bought?**
  - **Virtual screening** of preferred suppliers
- **What can be made/bought to make what can be made?**
  - **Building block** assessment and design
- **What can be designed?**
  - **Systems*-based** prospective chemistries
- **What can be exploited further?**
  - **Cross-comparison of** virtual libraries
- **What can be recycled?**
  - **Modify/expand** appropriate planned libraries

* System = GSK-speak for target class

The Future…



Feasible libraries generated automatically whenever:

- New proposals
- New reactions
- New building blocks
- Modified reaction feasibility
- Design requires (e.g. hole identified in compound set)

Acknowledgments

Library Design Coordination Team

- Pat Brady, Jimmy Chung, Jameed Hussain, Stephen Pickett, Hideyuki Sato, Martin Saunders, Eugene Stewart, Zheng Yang,
- (Alfonso Pozzan, Colin Edge)

And…

- Tim Clark
- Francis Atkinson
- Paul Gleeson
- Nate Woody
- Bill Maclachlan