

# Fast similarity searching – making the virtual real

*Stephen Pickett, GSK*

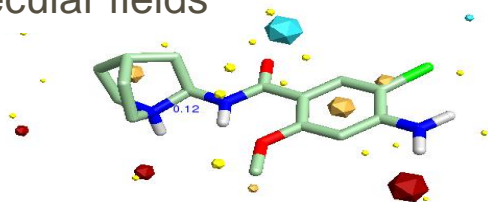
- Introduction to similarity searching
- Use cases
- Why is speed so crucial?
- Why MadFast?
- Some performance stats
- Implementation in LiveDesign
- Where next?

- Similarity property principle
  - If two compounds are similar they should have similar properties.
    - A. M. Johnson, G. M. Maggiora (1990). *Concepts and Applications of Molecular Similarity*. New York: John Willey & Sons
- If this did not apply “on the whole” then SAR would not exist and lead optimisation would be “random”
- How do we measure similarity
- What do we mean by “similar property”?
  - Use case dependent

# Describing molecules in a computer - Chemical Descriptors



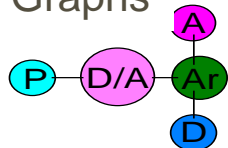
Molecular fields



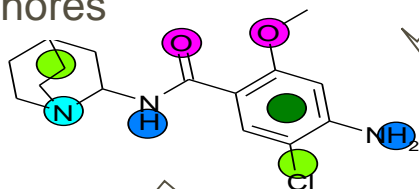
Shape fingerprint

110111011011

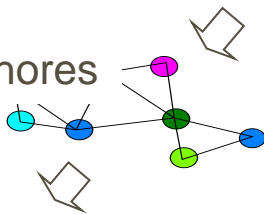
Reduced Graphs



3D Pharmacophores



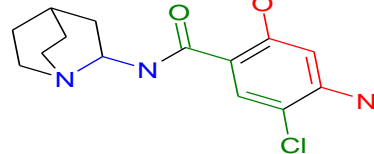
"3-point"  
Pharmacophores



110111011011

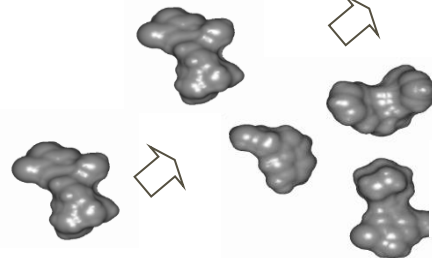
"3D" fingerprints

Atom pairs and paths

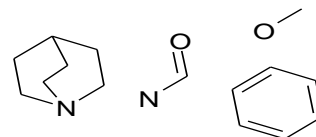


110111011011

Shape



Reference Shapes



NH<sub>2</sub>

Cl

Fragments

110111011011

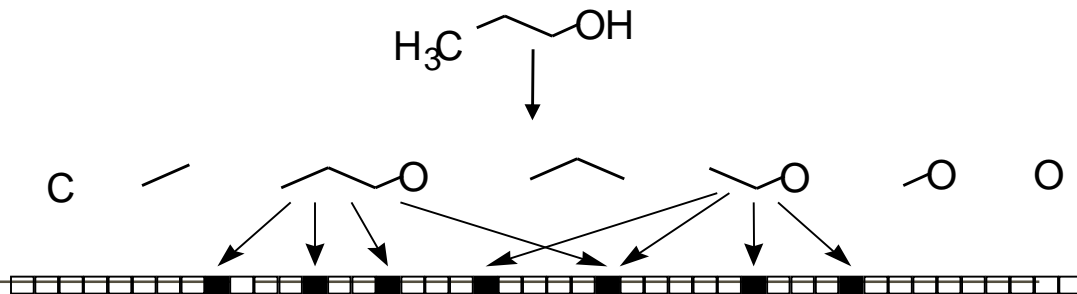
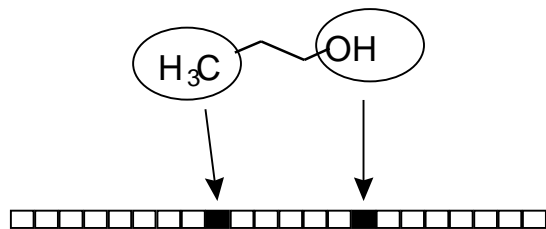
"2D" fingerprints

# Chemical Fingerprints



*Several types*

- “Fingerprints” or “bit-strings”
  - 00010000100010
- MACCS keys
  - Presence or absence of pre-defined fragments
- Daylight-like (ChemAxon Cfp)
  - All paths within system-defined ranges (bonds)
  - Includes element type information
  - “Hash” into fingerprint of predefined length
  - Each fragment sets multiple bits so cannot go back



# Fingerprint Similarity

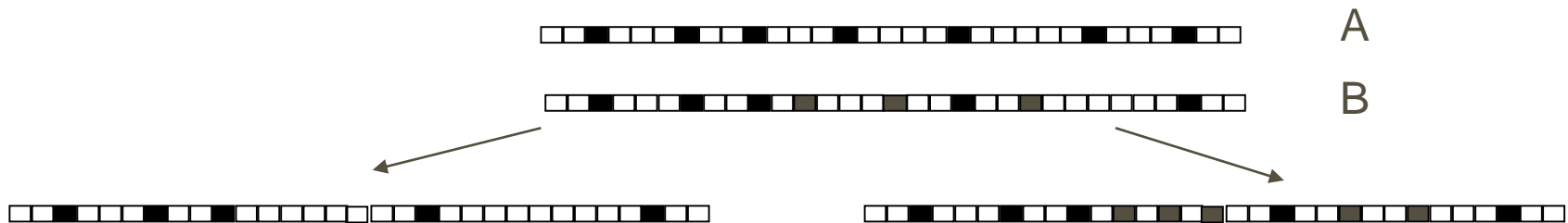


## Metrics

### – Many metrics

– Willett et al. J. Chem. Inf. Comput. Sci. **1998**, 38, 983

### – Tanimoto most common



$A \cap B = c$

$A \cup B$

$$Tanimoto = \frac{A \cap B}{A \cup B} = \frac{c}{a + b - c}$$

# Fingerprint Similarity

*How to define “similar”*

---



- Recommended cutoffs
  - dependent on fingerprint
    - Papadatos et al. J. Chem. Inf. Model. **2009**, 49, 195
  - and use case
    - clustering, analogue searching, collection design

# Why performance matters



- At GSK we cluster the whole screening collection (>2M) every weekend
  - Typical naïve N x N matrix generation can take many hours / days for large data sets
  - Highly parallelisable
  
- Similarity and clustering rate limiting step in compound acquisition process
  
- Requirement to search in **real time** very large libraries of available compounds
  - ZINC > 100M
  - eMolecules ~ 5M
  - EnamineREAL >30M tangible compounds



# Making similarity go faster



- Form of Tanimoto coefficient allows speed up of calculations
  - Presorting of fps
  - Comparing query on-bits to target on-bits with threshold
  - Lookups for similarity computation
  - Optimising code at machine level
- GSK ffss (Sunny Hung)
  - Comparator used here
- Chemfp (Andrew Dalke)
  - Fast Tanimoto searching
  - Efficient storage and extensive code optimisation
  - Command line tools
- SmallWorld (NextMove)
  - Graph edit distance based
  - Efficient storage and retrieval

- 
- ChemAxon development project
    - GSK has provided input and testing
  
  - Efficient precomputation of target fingerprint storage
  - Very fast querying via command-line
  - Web server to provide REST API
  - Demo UI
  
  - Fits with our current infrastructure
    - Potential to link to cartridge for management

# MadFast Performance



*Most similar to 790K set*

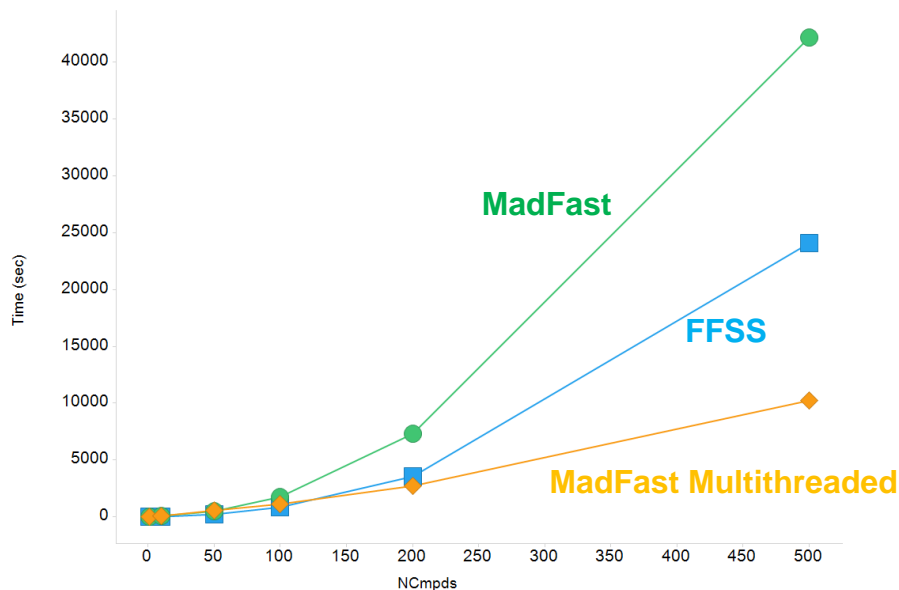


- Typical use
  - Closest cluster centroid
- MadFast
  - Single thread mode
  - ~3 fold faster on large sets.
- HP Z820
  - Intel Xeon E5-2667 @ 2.9 GHz
  - 24 processors

# MadFast Performance



*All vs All Tanimoto > 0.85*



- Typical use case
  - Sphere exclusion clustering
- MadFast
  - Need to specify both maxdissim and count
  - Split query file chunks 2.5K and run in linear batches
    - Memory overhead
- Requires more optimisation for this use case

# Integration with LiveDesign

Realtime searching of >30M compounds

– Reviewer Note: all structures on this and the next slide are vendor compounds, available on-line and not GSK compounds.



The screenshot displays the LiveDesign web application interface. On the left, a 'Sandbox' panel shows a table of molecular properties:

Property	Value	Property	Value
hba	4	clgp_err	0
clgp	2.949		
fp3	0.62	rmv	267.32
		tpsa	98.33
hbd	2	arom-rings	1

The central workspace shows a chemical structure of a complex molecule with a hydroxyl group, a methyl group, and an amine group. Below the structure are search options: 'Add Idea', 'Substructure Search', 'Similarity Search', and 'Search Exact Match'. A 'Preview Properties' button and an 'Add Idea to Live Report' button are also visible.

On the right, the 'MadFast' panel is active, displaying a 'List of DB sources available' and search options. The 'options' section includes a '# of molecules to gather/server:' set to 10 and a 'Similarity threshold:' set to 0%. The 'Smiles Viewer' section shows results grouped by server, with the following data:

Server	Compound ID	Similarity
EnamineREAL-ctp7	Z90710459	-67%
	Z169565944	-65%
	Z1700034759	-64%
	Z169565728	-64%
	Z2208671462	-63%
EnamineREAL-ctp7	Z2208919975	-62%
	Z2208801014	-62%
	Z2208833784	-61%
	Z2000615257	-61%
	Z90761360	-61%

# Integration with LiveDesign

Video



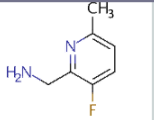
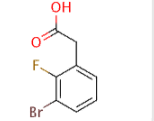
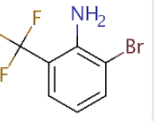
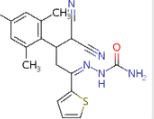
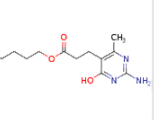
LiveDesign - Mozilla Firefox

LiveDesign

https://livedesign-ustest.gsk.com/livedesign/#/projects/47/livereports/6736

Sandbox [Switch Project] Give Feedback! Knowledge Base user/pass = gsk/gsk Assay Data last updated at 9:10 am Fri Dec 18 2015 syp38393 (Log Out)

HadFast example - Open Live Report +

Compound Structure	ID	ID (undefined)	P (arom-rings)	P (dogP)	P (dogp_err)	P (fsp3)	P (rba)	P (rbd)	P (mw)	P (tpsa)	Lead-liken.
 <chem>CNCC1=CC=C(F)N=C1</chem>	V116181439	ESOBAX10036	1	0.323	0	0.29	1	1	140.16	38.91	0.93
 <chem>O=C(O)c1cc(F)c(Br)cc1O</chem>	V116181440	ESOBAX10038	1	2.42	0	0.13	2	0	233.03	37.3	0.83
 <chem>Nc1cc(Br)ccc1C(F)(F)F</chem>	V116181441	ESOBAX10040	1	3.3	0	0.14	0	1	240.02	26.02	0.84
 <chem>Cc1ccc(C)cc1C(=O)Nc2nc3c(ncn3C)sc2</chem>	V116181442	ESOBAX10046	2	1.83	41	0.3	4	2	379.48	115.06	0.68
 <chem>Cc1nc2c(ncn2C)sc1C(=O)OCCCC</chem>	V116181443	ESOBAX13367	1	2.95	0	0.62	4	2	267.32	98.33	0.84

Row Picker  
Compound

Presentation title

14

# Conclusion



- 
- MadFast provides a great solution for interactive searching
    - Multiple data sources
    - Multiple fingerprint options
    - Multiple metric options
    - Integrated with LiveDesign
  
  - Command-line use for clustering applications requires more optimisation
  
  - Can we do the same with substructure searching?

# Acknowledgements

---



- Gabor Imre and the DISCO team at ChemAxon
  
- Sunny Hung (GSK)
- Martin Saunders (GSK)
  
- Benoit Mangili (Tessella) for the LiveDesign integration





# Section heading orange

*Supporting heading*