Fast similarity searching – making the virtual real

Stephen Pickett, GSK

Introduction

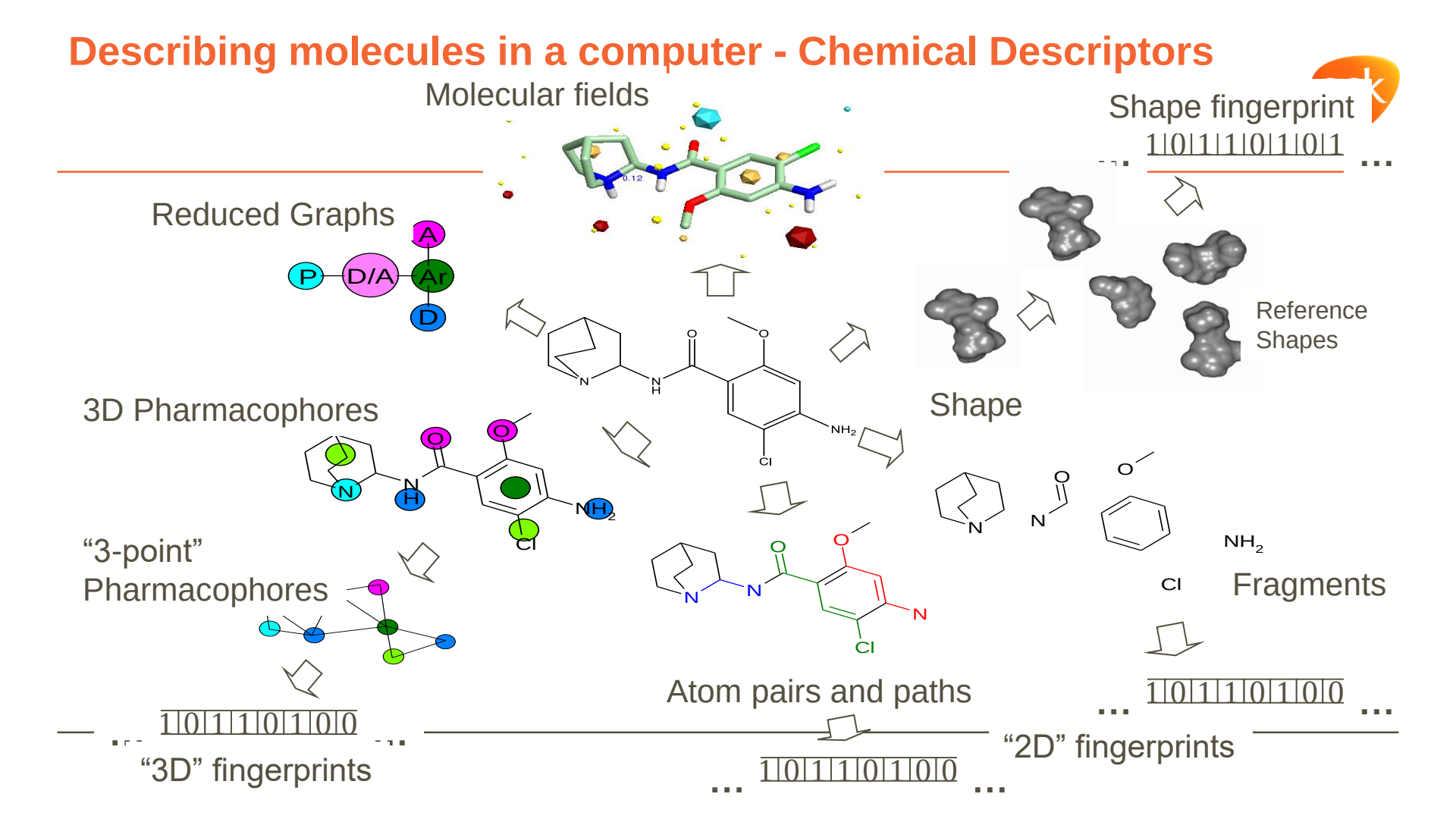
- Introduction to similarity searching
- Use cases
- Why is speed so crucial?
- Why MadFast?
- Some performance stats
- Implementation in LiveDesign
- Where next?

Similarity Searching

- 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

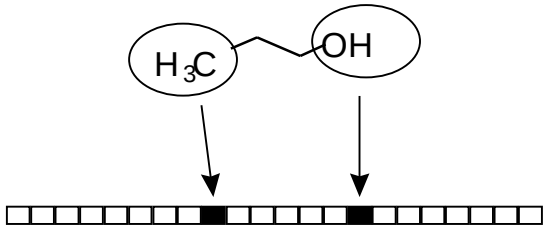
Describing molecules in a computer - Chemical Descriptors

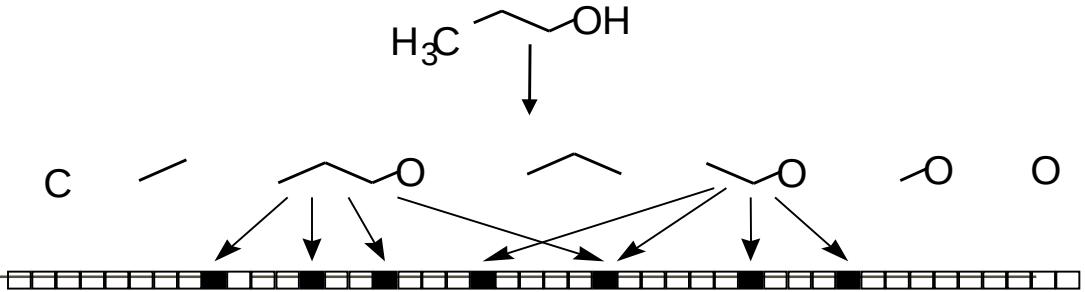


Chemical Fingerprints

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

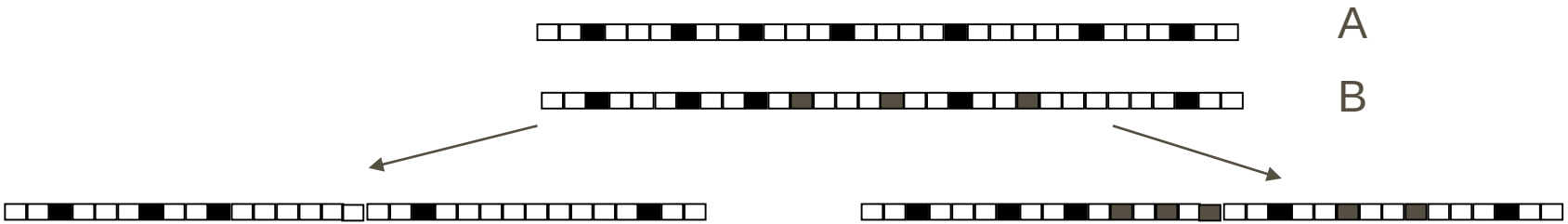




Fingerprint Similarity

Metrics

– Many metrics
  – Willett et al. J. Chem. Inf. Comput. Sci. **1998**, 38, 983
  – Tanimoto most common



$$Tanimoto = \frac{A \bigcap B}{A \bigcup B} = \frac{c}{a+b-c}$$

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

MadFast

- 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

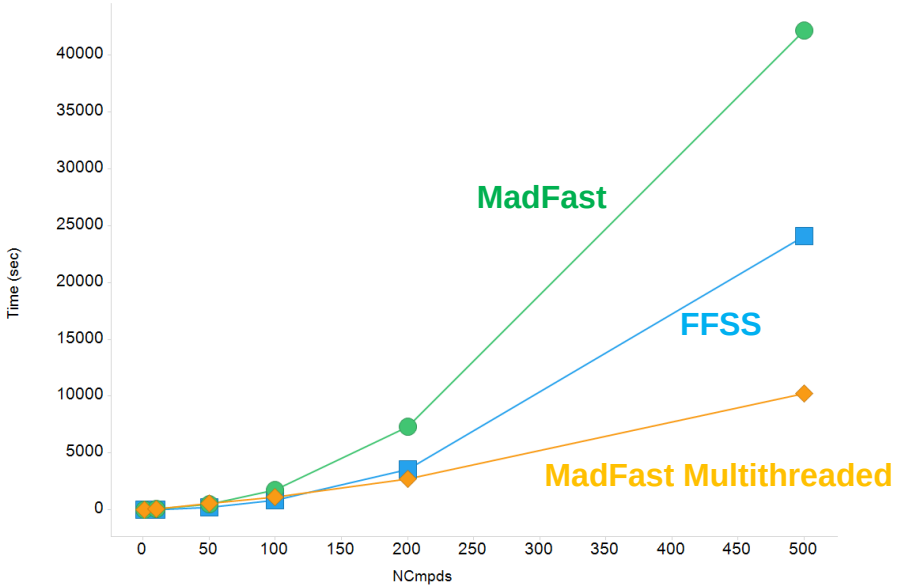
MadFast Performance

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

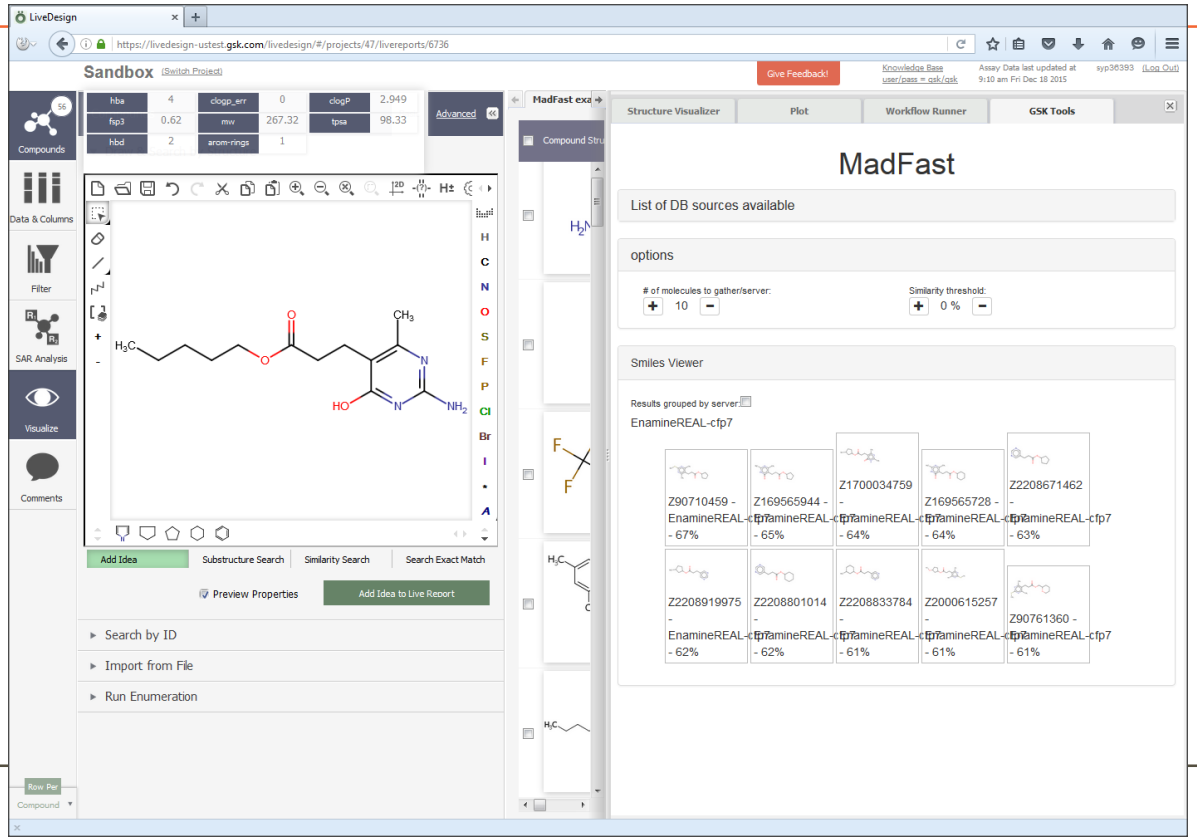


- Typical use case
  - Sphere exclusion clustering
- MadFast
  - Need to specify both maxdissim and count
  - Split query file chunks 2.5K and run in linear batches
    - Memory overhead
- Requires more optimisation for this use case

Integration with LiveDesign

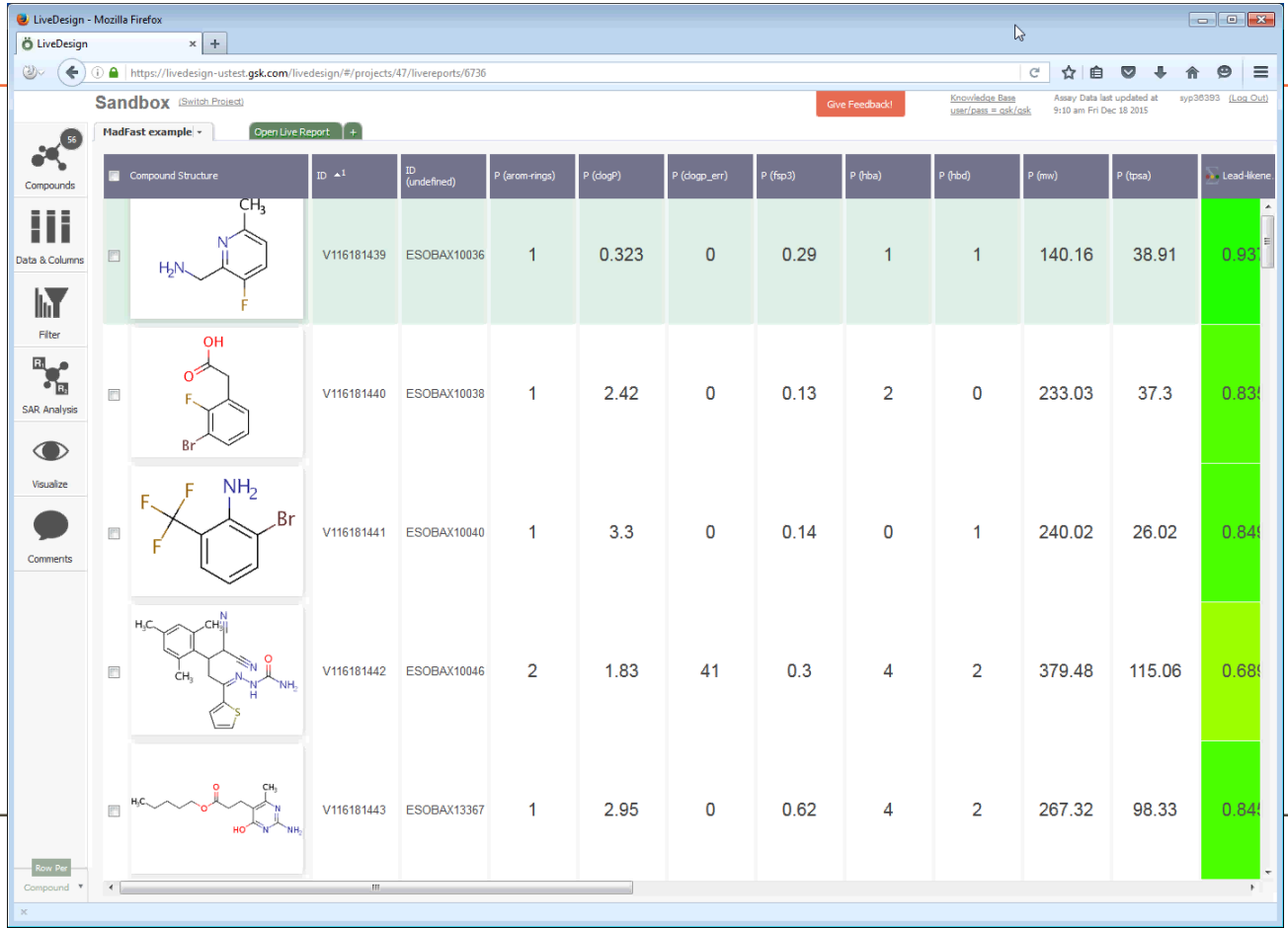
Realtime searching of >30M compounds

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



Integration with LiveDesign

Video



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