The University Of Sheffield.

Similarity searching using multiple starting points

Peter Willett, University of Sheffield, UK

The University Of Sheffield.

Overview of talk

- Introduction
- Similarity searching when multiple bioactive reference structures are available
- Turbo similarity searching, based on using nearest-neighbours
- Conclusions

The University Of Sheffield.

Similarity searching: I

- Use of a similarity measure to determine the relatedness of an active *target*, or *reference*, structure to each structure in a database
- The *similar property principle* means that high-ranked structures are likely to have similar activity to that of the target
- Similarity searching hence provides an obvious way of following-up on an initial active

The University Of Sheffield.

Similarity searching: II

- Similarity searching using a single target structure now a common feature in chemoinformatics software systems
- How to search with multiple, structurally unrelated target structures, e.g.,
  - Diverse hits from HTS
  - Compounds from a public database (e.g., *MDL Drug Data Report* and the *World Drugs Index*)
  - Competitor compounds

The University Of Sheffield.

Comparison of search techniques: I

- Given a set of active molecules, how can a database be similarity-ranked in order of decreasing probability of activity?
- Extensive simulated virtual screening experiments on the *MDL Drug Data Report* (MDDR) database, using
  - Molecules represented by 2D fingerprints (UNITY fingerprints in the initial experiments)
  - Inter-fingerprint similarity calculated using the Tanimoto Coefficient

The University Of Sheffield.

Comparison of search techniques: II

- Several different techniques were tested
  - Hert, J. *et al*., "Comparison of fingerprint-based methods for virtual screening using multiple bioactive reference structures." *J. Chem. Inf. Comput. Sci.*, **44**, 2004, 1177.
- Best results obtained by
  - Combining the rankings resulting from separate searches using data fusion
  - Approximation of the binary kernel discrimination method for machine learning

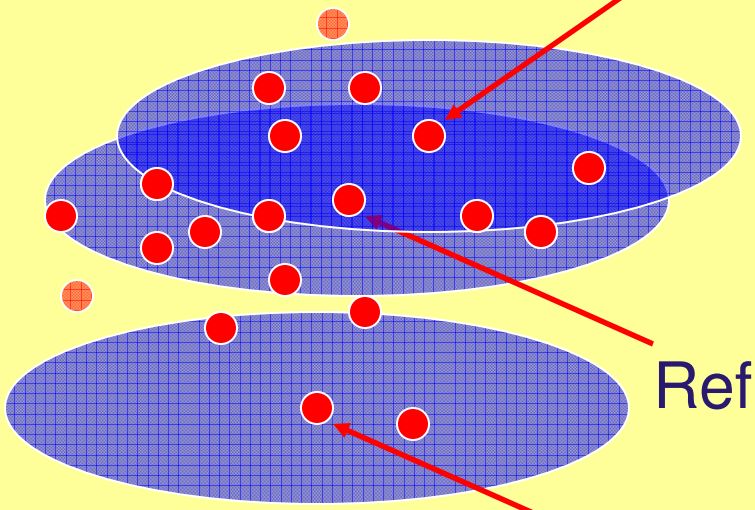
The University Of Sheffield.

Comparison of search techniques: III

- Here, focus on data fusion, where combine different rankings of the same sets of molecules
- Two basic approaches
  - Generate rankings from the same molecule using different similarity measures (*similarity fusion*)
  - Generate rankings from different molecules using the same similarity measure but different molecules (*group fusion*)

The University Of Sheffield.

Group fusion

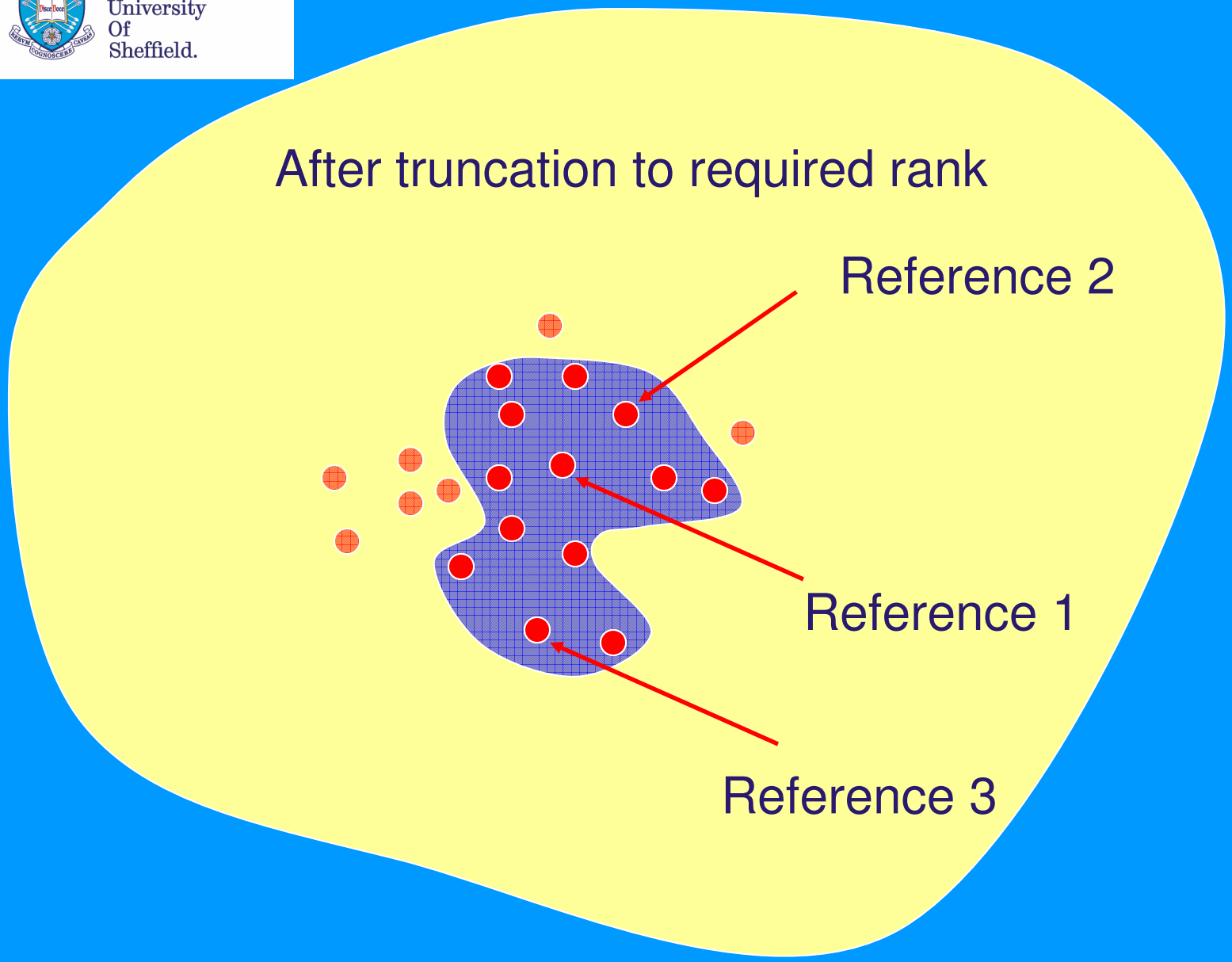


Reference 2

Reference 1

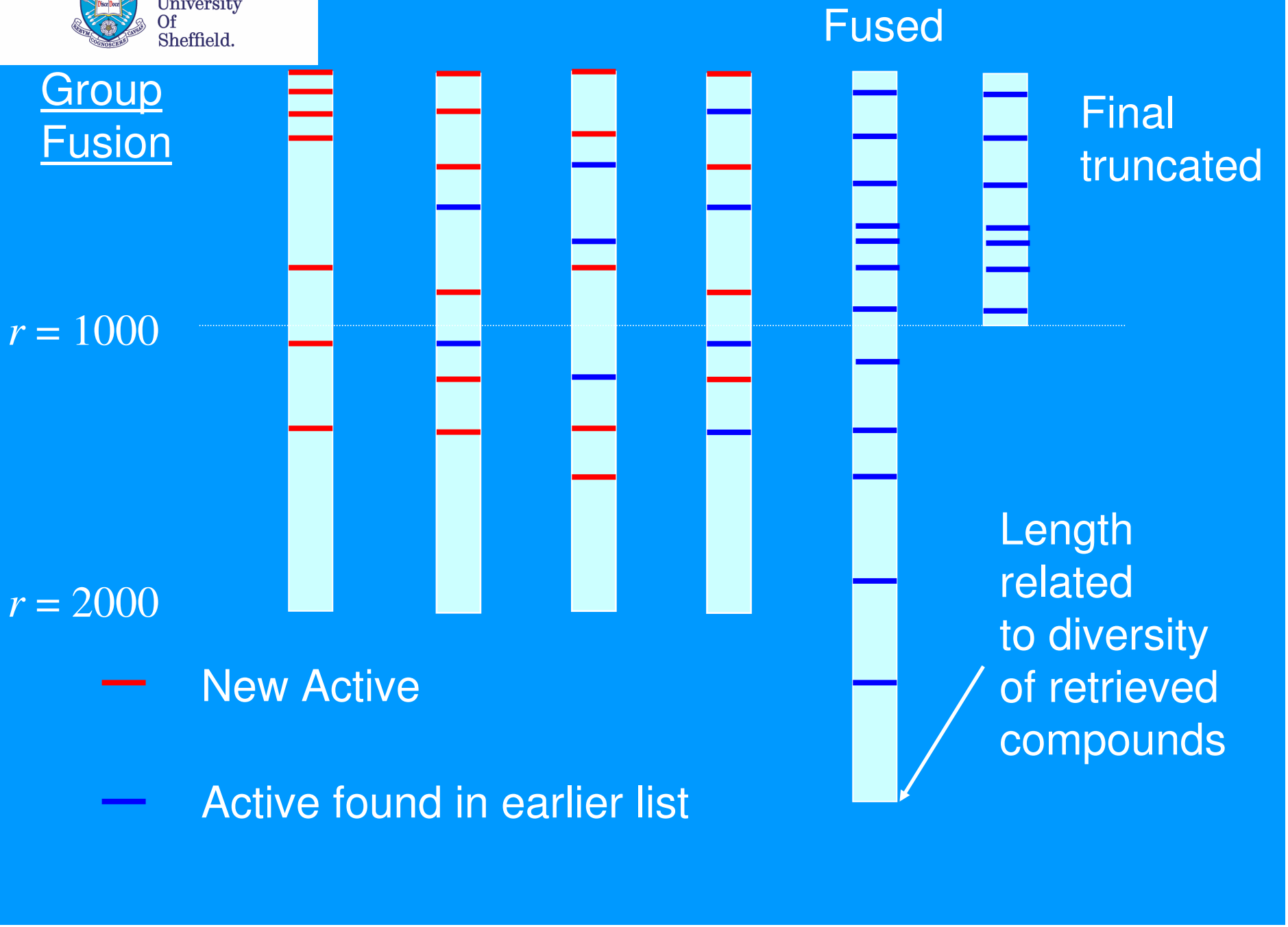
Reference 3

The University Of Sheffield.



The University Of Sheffield.

Group Fusion



The University Of Sheffield.

Group fusion rules

- Fusion of scores or fusion of ranks (normal in similarity fusion)
- SUM rule : add the scores (ranks) from the similarity lists for some database molecule and then re-rank the resulting sums
- MAX rule : re-rank using the maximum score (minimum rank) attained in any of the lists

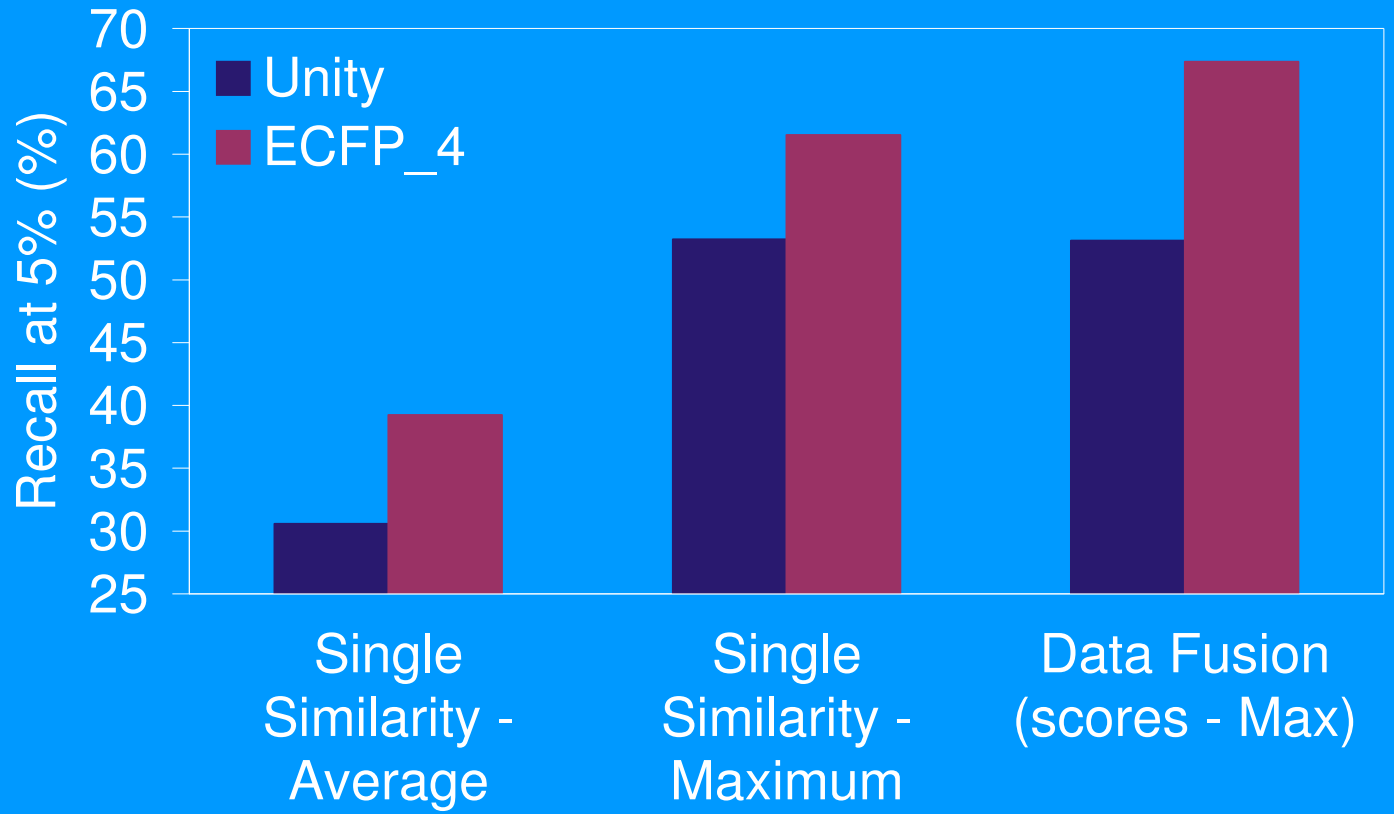
The University Of Sheffield.

Experimental details

- MDDR with ca. 102K molecules
  - 11 activity classes
  - 10 sets of 10 randomly chosen compounds from each activity class
  - All similarities calculated using the Tanimoto Coefficient
- Best group-fusion results obtained using combination of scores and the MAX rule
  - Comparison with average and best single-molecule searches

The University Of Sheffield.

Use of multiple reference structures



The University Of Sheffield.

Comparison of 2D similarity measures

- Extensive comparative experiments
  - Scitegic ECFP_4 best of the 14 types of 2D fingerprint
  - Tanimoto best of the 12 types of similarity coefficients
  - Whittle, M. *et al.*, "Enhancing the effectiveness of virtual screening by fusing nearest-neighbour lists: A comparison of similarity coefficients" *J. Chem. Inf. Comput. Sci.*, **44**, 2004, 1840.
  - Hert, J. *et al.*, Topological descriptors for similarity-based virtual screening using multiple bioactive reference structures." *Org. Biomol. Chem.*, **2**, 2004, 3256

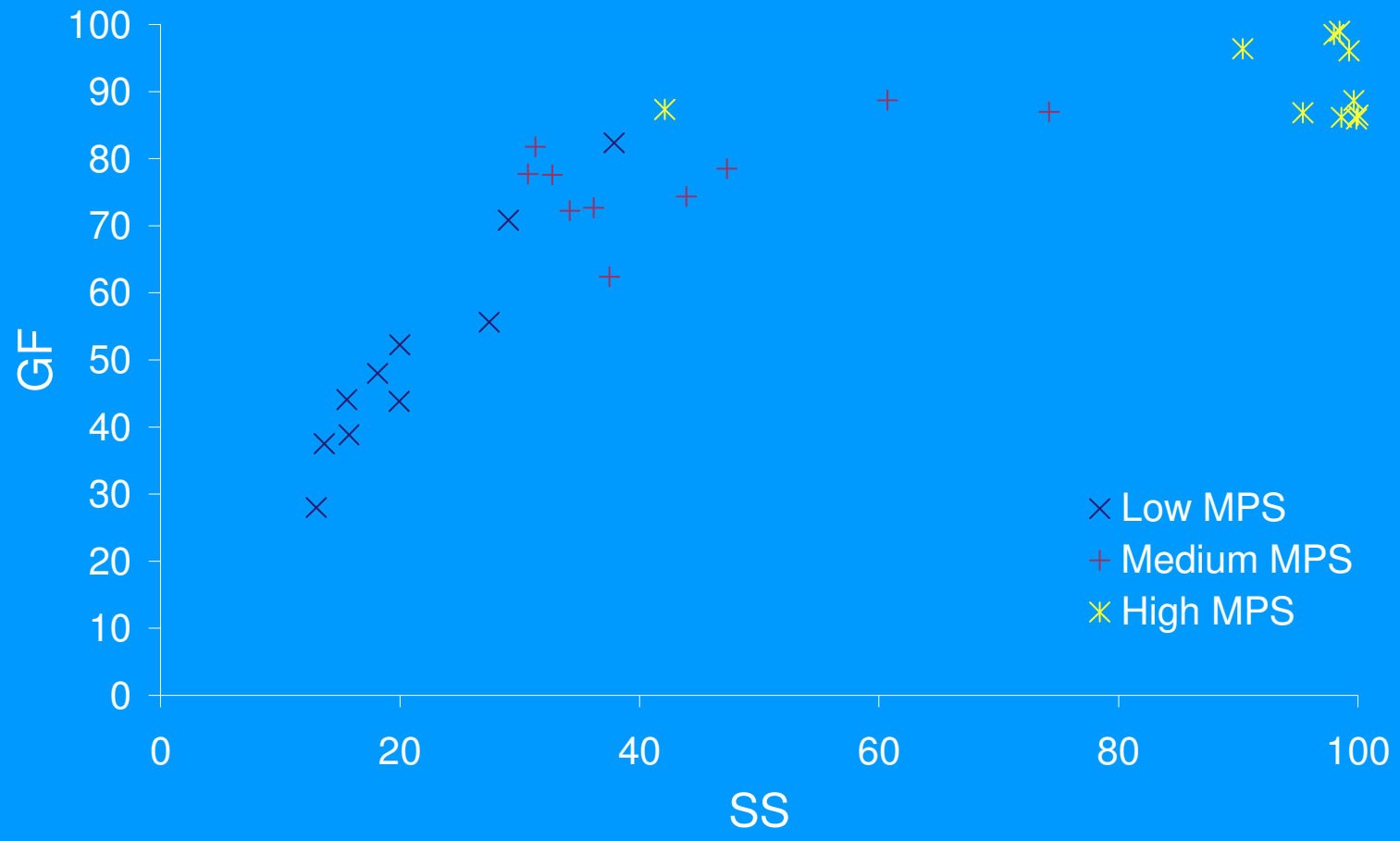
The University Of Sheffield.

Effect of structural diversity

- Some evidence to suggest that the enhancement was greatest with the most diverse sets of actives.
- More detailed experiments where chose 10 MDDR activity classes that
  - Contained at least 50 molecules
  - Had the smallest, or the largest or the median mean pair-wise Tanimoto similarity (similar results if use numbers of scaffolds)

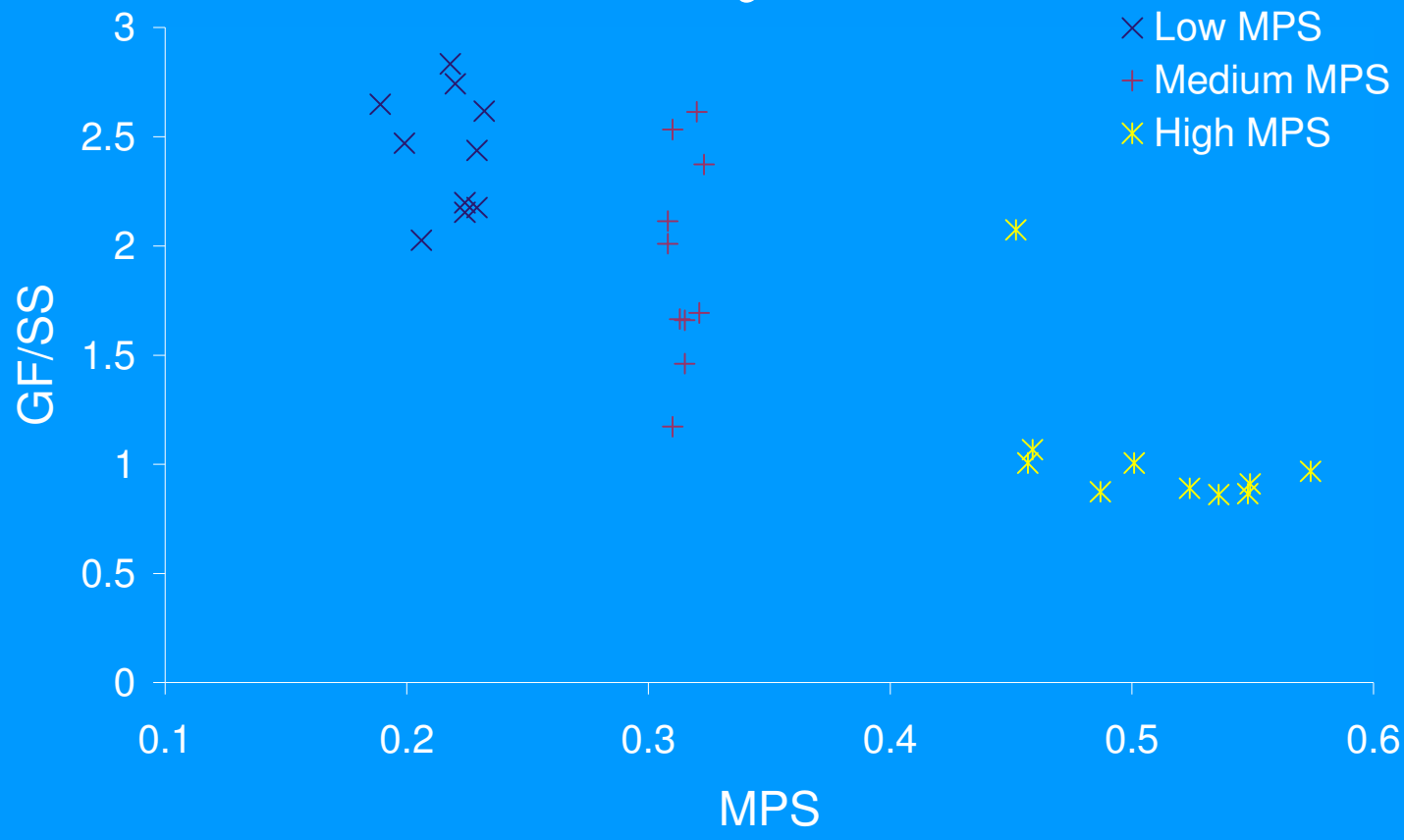
The University Of Sheffield.

Recall for group fusion and similarity searching



The University Of Sheffield.

Variation in relative recall with mean pair-wise similarity



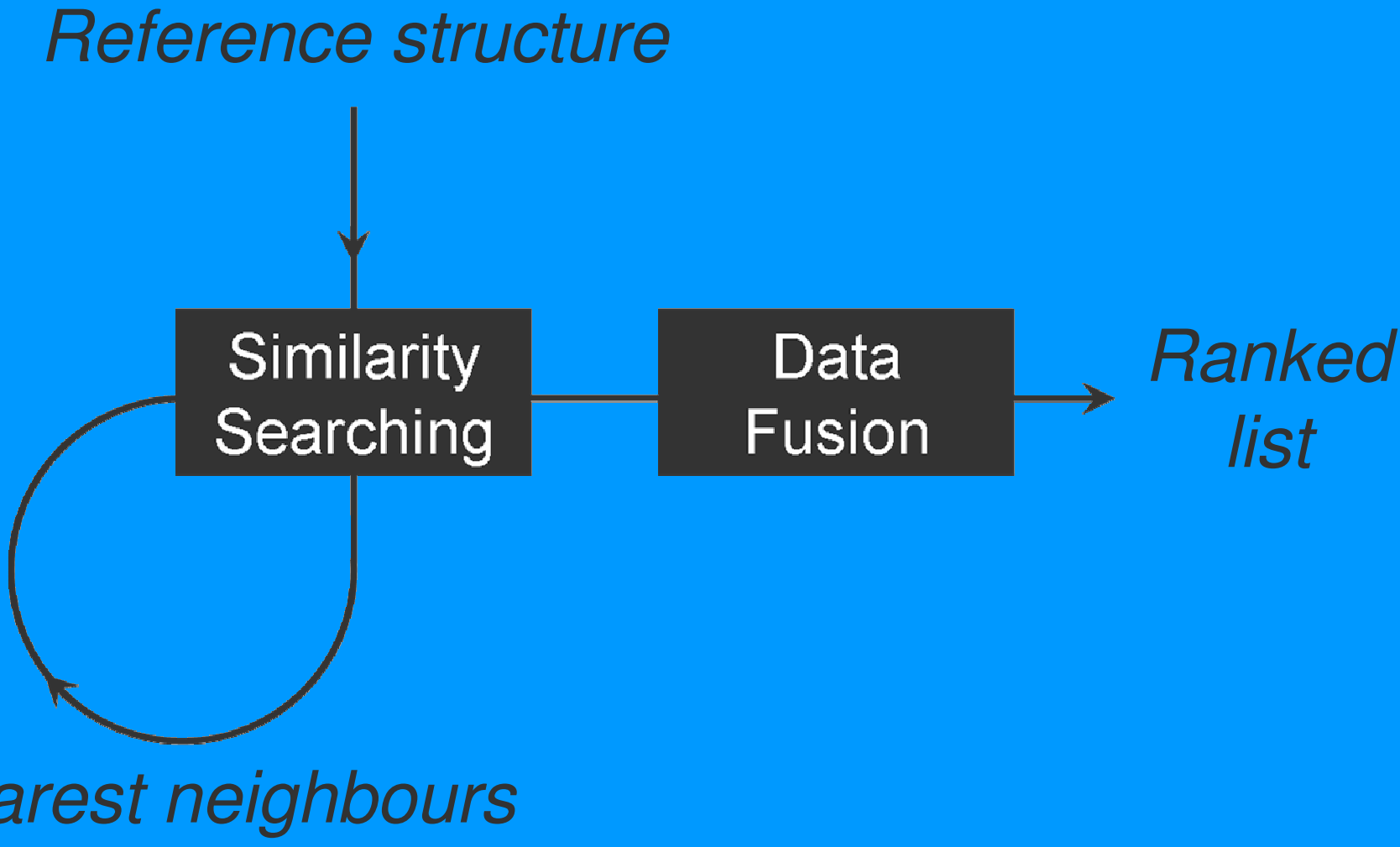
The University Of Sheffield.

Turbo similarity searching: I

- Similar property principle: nearest neighbours are likely to exhibit the same activity as the reference structure
- Group fusion improves the identification of active compounds
- Potential for further enhancements by group fusion of rankings from the reference structure and from its *assumed active* nearest neighbours

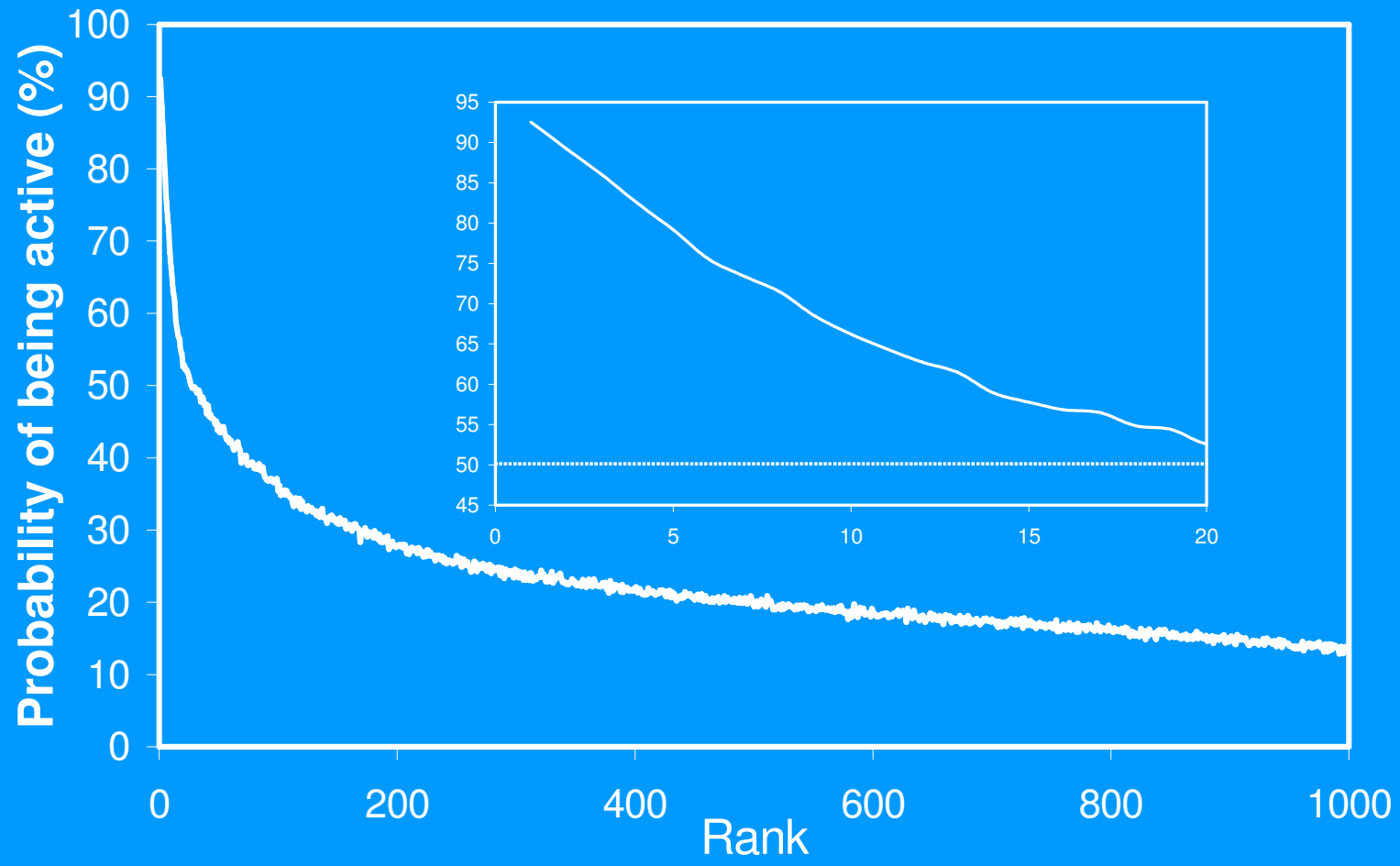
The University Of Sheffield.

Turbo similarity searching: II



The University Of Sheffield.

Probability of activity for nearest neighbours



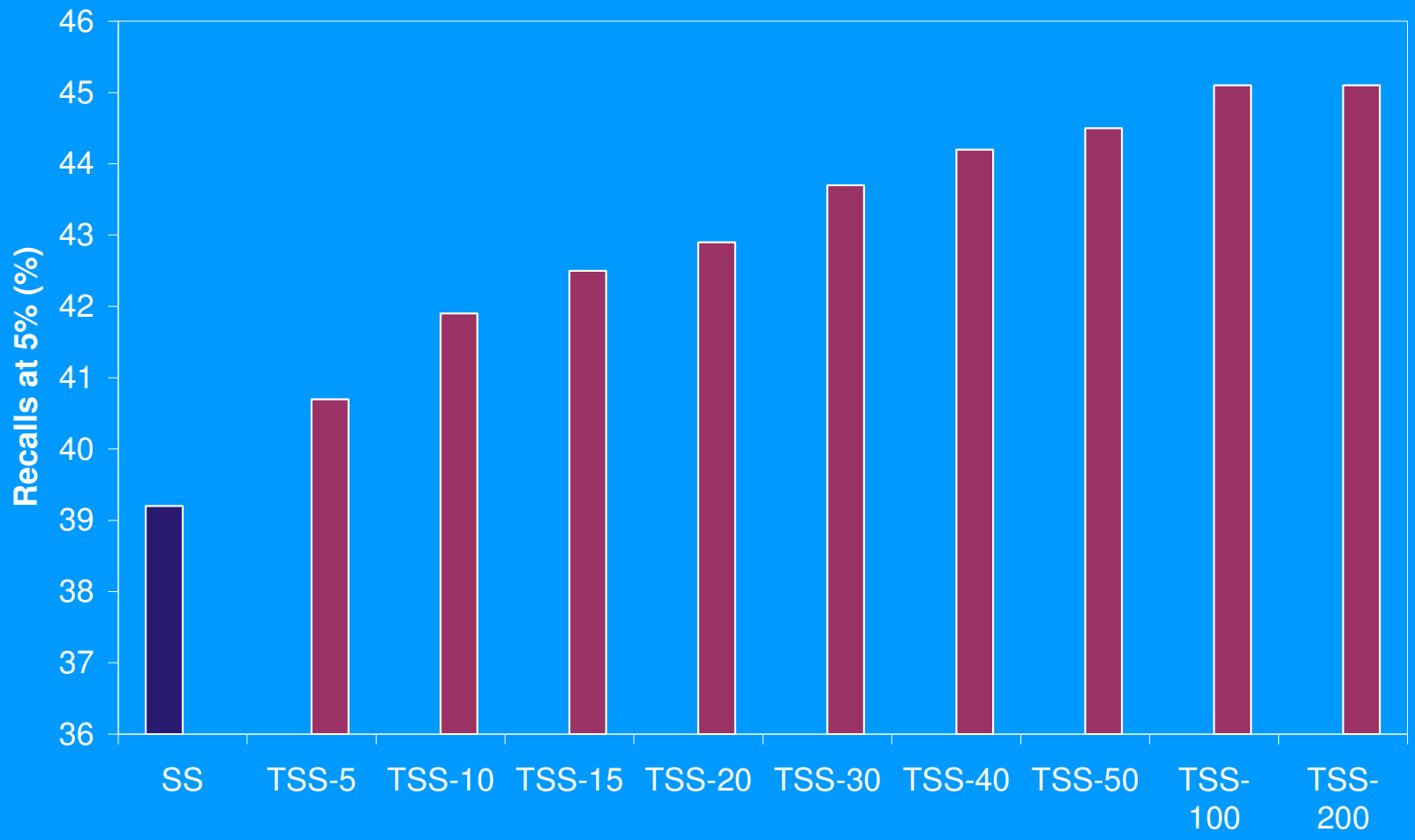
The University Of Sheffield.

Experimental details

- MDDR data set of 11 activity classes and 102K structures as used previously
  - In all, 8294 actives in the 11 classes, with (turbo) similarity searches being carried out using each of these as the reference structure
  - ECFP_4 fingerprints/Tanimoto coefficient
  - MAX group fusion on similarity scores
  - Increasing numbers of nearest neighbours

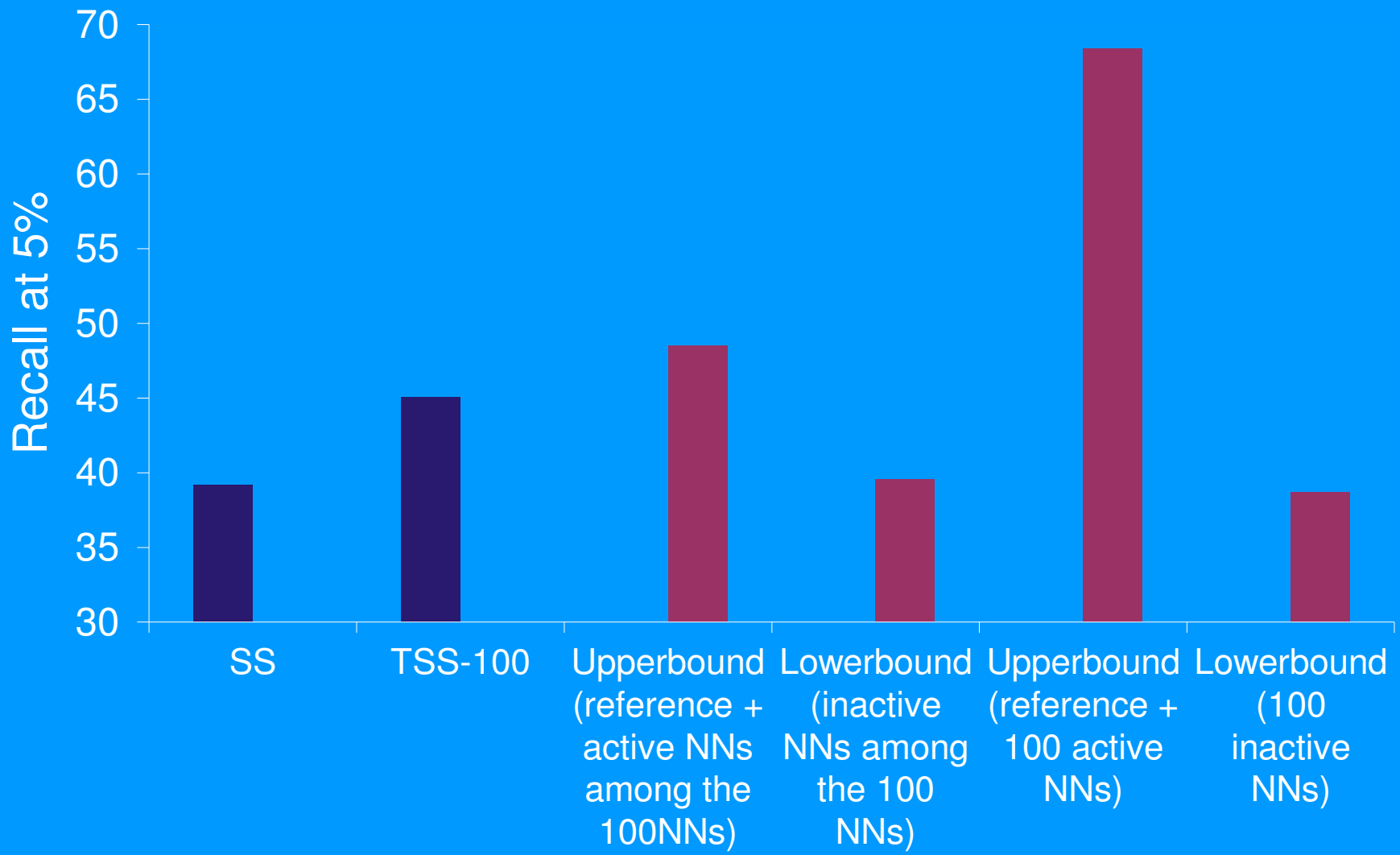
The University Of Sheffield.

Numbers of nearest neighbours



The University Of Sheffield.

Upper and lower bound experiments



The University Of Sheffield.

Rationale for upper bound results

- The true actives in the set of assumed actives yield significant enhancements in performance
- The true inactives in the set of assumed actives have little effect on performance
- Taken together, the two groups of compounds yield the observed net enhancement
- Hert, J. *et al*., "Enhancing the effectiveness of similarity-based virtual screening using nearest-neighbour information." *J. Med. Chem.*, in the press.

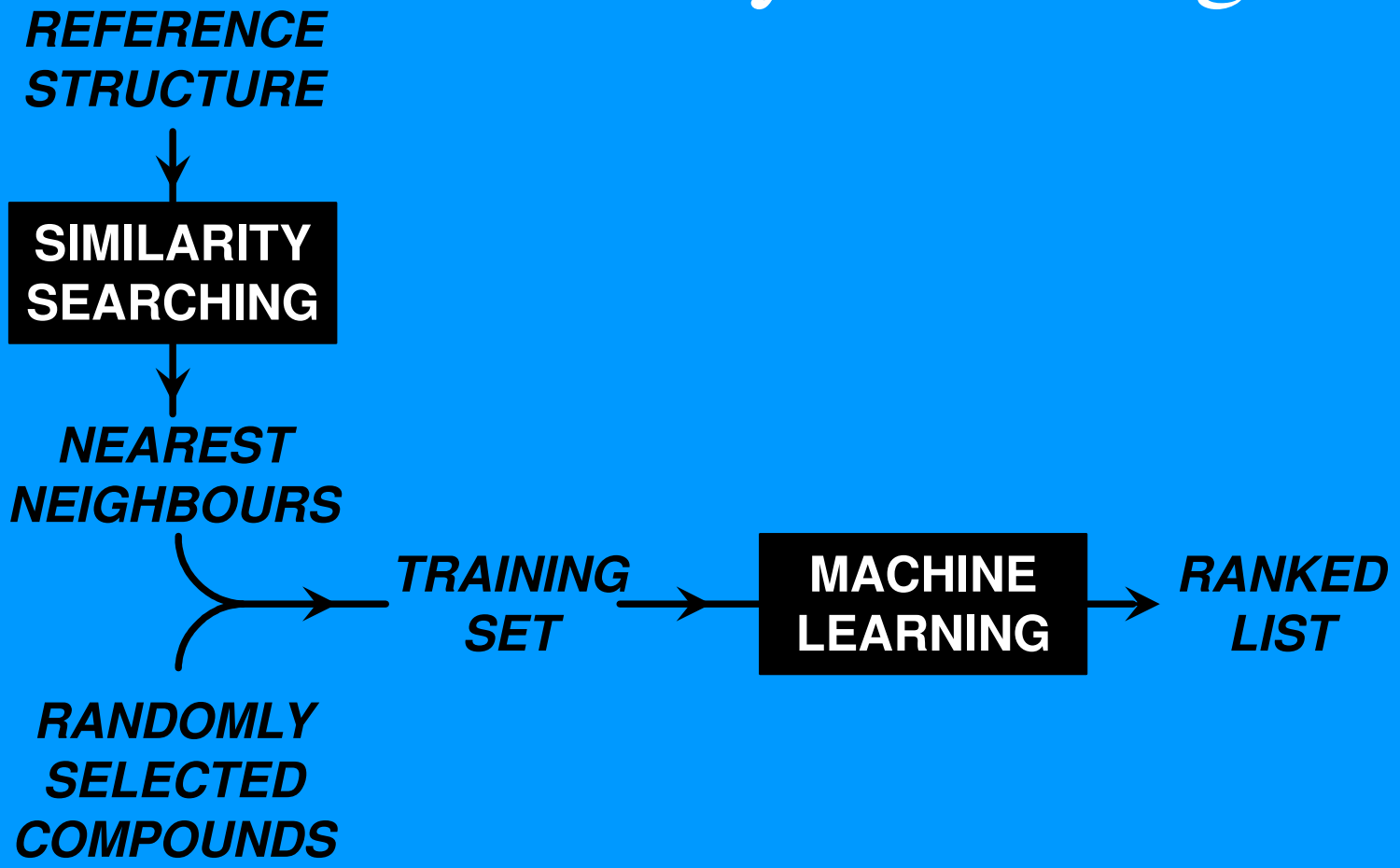
The University Of Sheffield.

Use of machine-learning methods for similarity searching: I

- Turbo similarity searching uses group fusion to enhance conventional similarity searching
- Machine learning is a more powerful virtual screening tool than similarity searching
  - But requires a training-set containing known actives and inactives
- Given an active reference structure, a training-set can be generated from
  - Using the *k* nearest neighbours of the reference structure as the actives
  - Using *k* randomly chosen, low-similarity compounds as the inactives

The University Of Sheffield.

Use of machine-learning methods for similarity searching: II

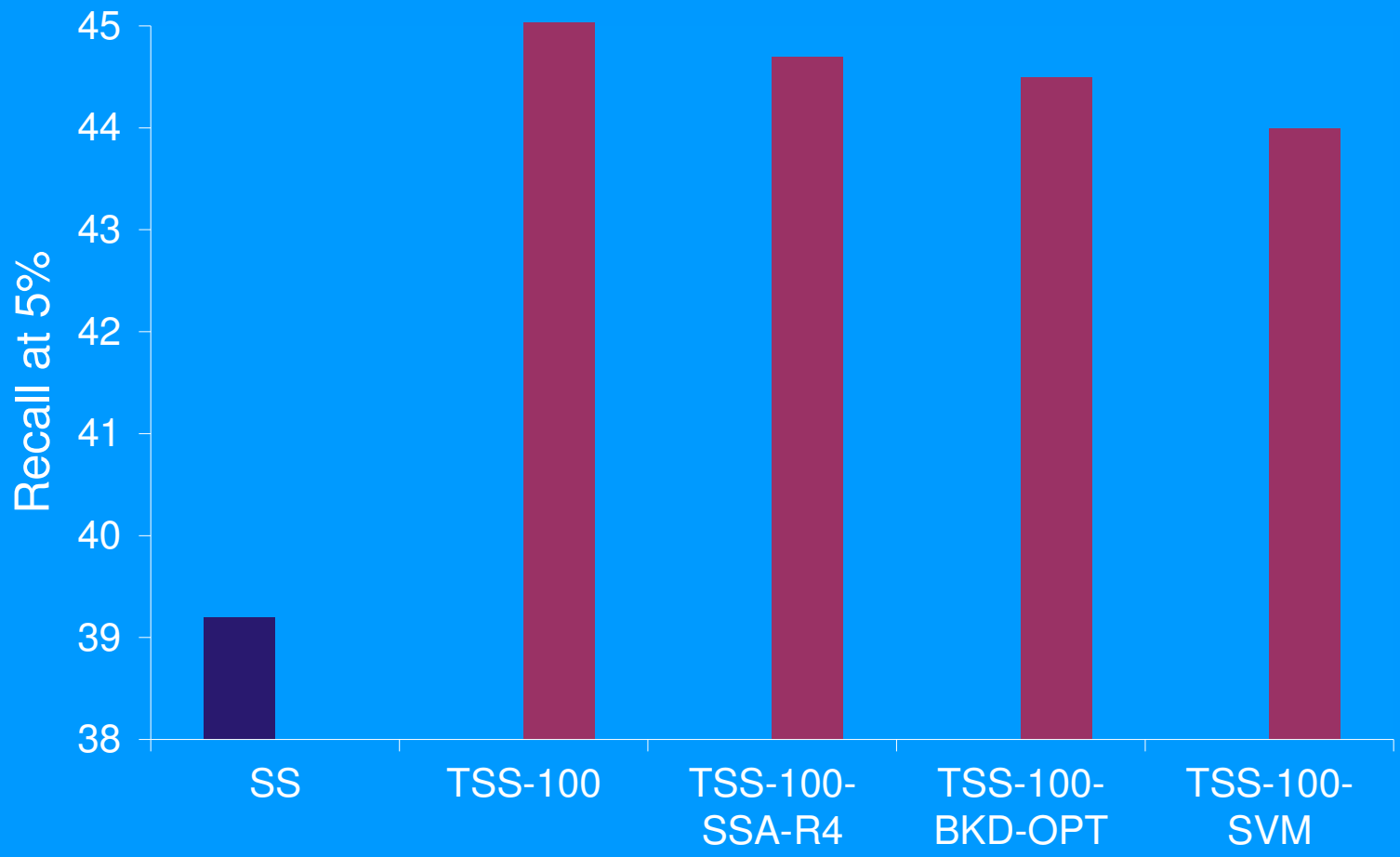


Initial experiments: I

- Three machine-learning techniques in the second stage
  - Substructural analysis
    Best results with the R4 probabilistic weight
  - Binary kernel discrimination
  - Support vector machine
- MDDR dataset as used previously, with 100-molecule training-sets

The University Of Sheffield.

Initial experiments: II



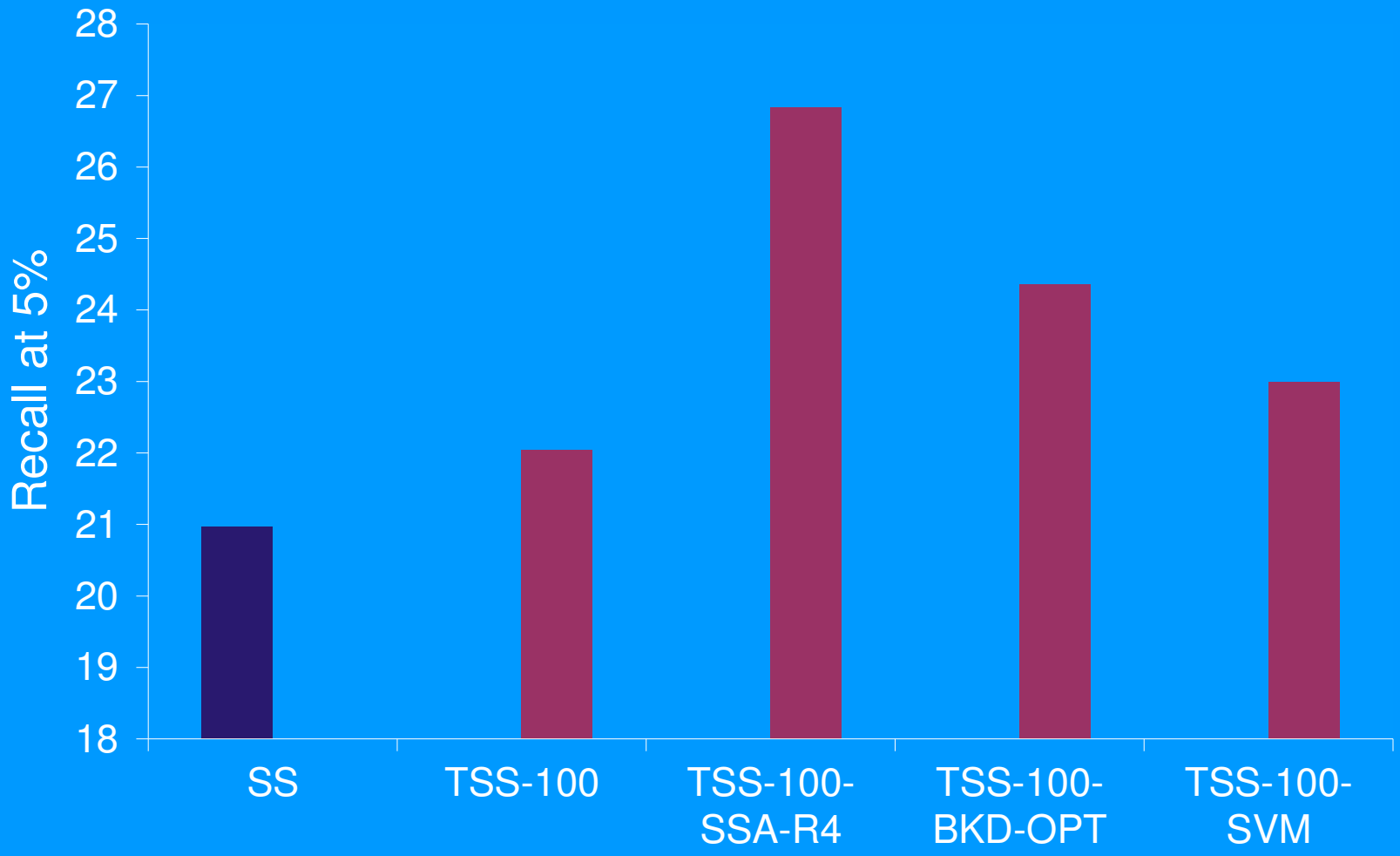
The University Of Sheffield.

Additional experiments: I

- Initial results rather disappointing, but some improvements noted with the most diverse datasets
- Further experiments with the set of 10 MDDR activity classes with the lowest mean pair-wise Tanimoto similarity

The University Of Sheffield.

Additional experiments: II



The University Of Sheffield.

Conclusions: I

- Fingerprint-based similarity searching using a known reference structure is long-established in chemoinformatics
- When small numbers of actives are available, group fusion will enhance performance when the sought actives are structurally heterogeneous

The University Of Sheffield.

Conclusions: II

- Can also enhance conventional similarity search, even if there is just a single active, by assuming that the nearest neighbours are also active
- Can be effected in two ways
  - Use of group fusion to combine similarity rankings (overall best approach)
  - Use of substructural analysis to compute fragment weights (best with highly heterogeneous sets of actives)

Acknowledgements

- Collaborators
  - Jerome Hert, Martin Whittle and David Wilton
  - Pierre Acklin, Kamal Azzaoui, Edgar Jacoby and Ansgar Schuffenhauer
  - Alexander Alex, Jens Loesel and Jonathan Mason
- Funding, software and data support
  - Barnard Chemical Information, Daylight Chemical Information Systems, MDL Information Systems, Novartis Institutes for BioMedical Research, Pfizer Global Research and Development, Royal Society, Scitegic, Tripos, and the Wolfson Foundation