Fingerprint-Based Virtual Screening Using Multiple Bioactive Reference Structures

Jérôme Hert, Peter Willett and David J. Wilton (University of Sheffield, Sheffield, UK)

Pierre Acklin, Kamal Azzaoui, Edgar Jacoby and Ansgar Schuffenhauer (Novartis Institutes for BioMedical Research, Basel, Switzerland)

Outlines

- Introduction to similarity-based virtual screening
- Extension to multiple reference structures
- Comparison of ranking methods
- Comparison of descriptors
- Turbo similarity searching
- Conclusions

Virtual screening

- Virtual screening involves scanning databases of compounds to find molecules that may exhibit some bioactivity of interest, so as to prioritise a screening programme

Similarity searching (1)

- Use of similarity measure to determine the degree of similarity between an active reference structure and each structure in the database
- Similarity property principle means that high-ranked structures are likely to have a similar activity to that of the reference structure

Similarity searching (2)

- How can existing methods be used when several diverse structures are available?

Project overview (1)

- Given a set of molecules of known activity, how can they be used to rank a database in order of decreasing probability of their exhibiting that activity?
- The various approaches have been evaluated in simulated virtual screening experiments on the MDDR database (ca. 102K molecules) using a range of different activity classes

Project overview (2)

- Use just 2D fingerprints (of various sorts)
- Use of the Tanimoto coefficient (will not consider the effect of different similarity coefficients)

Comparison of Methods

- Three distinct approaches have been investigated
  - Single fingerprint methods
  - Data fusion methods
  - Substructural analysis methods

Single Fingerprint Methods (1)

- Two methods derived from the Stigmata approach
  - Modal approach
  - Weighted approach
- Modal / weighted fingerprint used as a query, with the Tanimoto coefficient being used to score molecules in the database

Single Fingerprint Methods (2)

Training set of actives:

mol 1: **1001011000010 10**
mol 2: **0011010000110 00**
mol 3: **1101010011111 01**
mol 4: **1011011010100 10**
mol 5: **0100111000111 01**

Modal at 40%:
1111011010111 11

Weighted:
3224153021442 22

Data Fusion Methods (1)

- Combination of different rankings of the same sets of molecules with the expectation of improving decision
- This basic idea has been used previously with considerable success (also consensus scoring) by generating different rankings from the same molecule, using different similarity measures

Data Fusion Methods (2)

- Here, the different rankings come from different molecules but use the same, Tanimoto-based similarity measure

Fusion of ranks

Fusion of scores

Data Fusion Methods (3)

- Two fusion rules investigated

$$\mathrm{SUM} : \sum_{i=1}^{N} s_i$$

$$\mathrm{MAX} : Max\left(s_i\right)$$

Substructural Analysis Methods (1)

- Classic substructural analysis (SSA):
  - Training-set containing actives and inactives
  - Weights calculated for each bit from the training-set using a weighting scheme
  - Sum of the weights for each bit present in a compound gives the score

Substructural Analysis Methods (2)

- Approximate substructural analysis without known inactives:
  - Reference structures as training-set actives
  - Approximate the training-set by the entire database
  - Use of an appropriate weighting scheme that does not make explicit use of information about the inactives

$$R1 : \log\left(\frac{A_j / N_A}{T_j / N_T}\right)$$

Substructural Analysis Methods (3)

- Binary Kernel Discrimination (BKD):

$$K_\lambda(i,j) = \lambda^{N-d_{i,j}}(1-\lambda)^{d_{i,j}}$$

  - λ is a smoothing parameter that is optimised using the training-set active and inactive compounds
  - Compounds are scored by:

$$L_A(j) = \frac{\sum_{i \in actives} K_\lambda(i,j)}{\sum_{i \in inactives} K_\lambda(i,j)}$$

Substructural Analysis Methods (4)

- Approximate BKD without known inactives
  - Reference structures as training-set actives
  - Set of 100 randomly chosen compounds from the database as training-set inactives (cf SSA approximation)

Experimental Details

- MDL Drug Data Report (MDDR) Database
- 11 activity classes selected
- 10 sets of 10 randomly chosen compounds from each activity
- 3 fingerprints investigated: 988-bit Unity, 1052-bit BCI, 2048-bit Daylight
- Results are average recalls at 5% over the 10 different trials

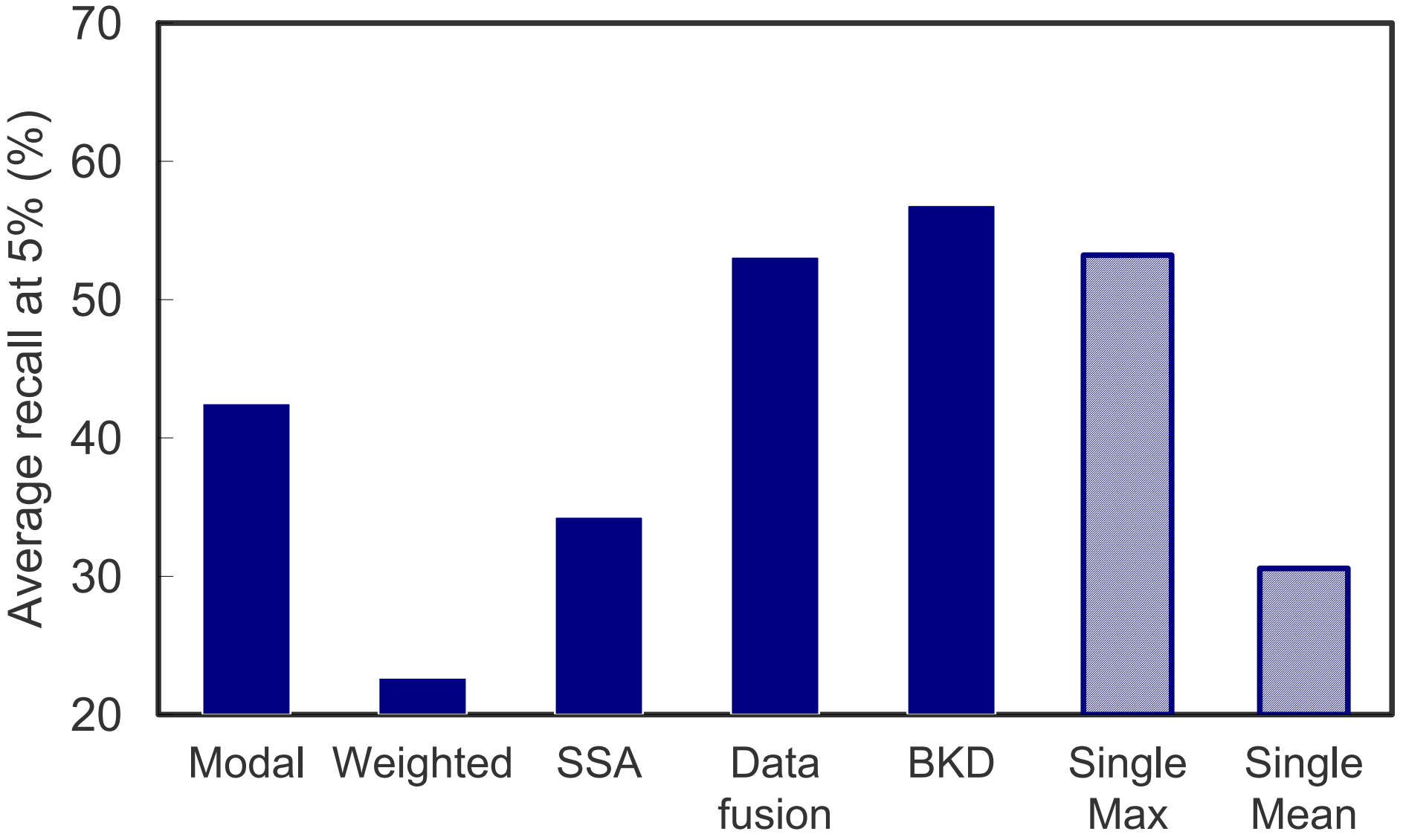
Results (1)

- Initial Experiments
  - Best threshold for modal approach is 40%
  - Best combination for data fusion is the combination of scores using the MAX fusion rule
  - The three types of fingerprint give broadly comparable results

Results (2)

- The basic results obtained with the various methods have been compared with the average and the maximum of all individual, single-molecule similarity searches

Results (3)


70
60
50
40
30
20
Average recall at 5% (%)
Modal
Weighted
SSA
Data fusion
BKD
Single Max
Single Mean

Comparison Of Descriptors

- Four different types of 2D descriptor investigated using the two best approaches from the initial part of the study
  - Structural keys
    - 1052-bit BCI fingerprints
  - Hashed fingerprints
    - 988-bit Unity fingerprints
    - 2048-bit Daylight fingerprints
    - 2048-bit Avalon fingerprints (internal Novartis system)
  - Circular substructures
    - ECFP_2, ECFP_4, FCFP_2, FCFP_4
  - Pharmacophore vectors

Pharmacophore Vectors: Similog

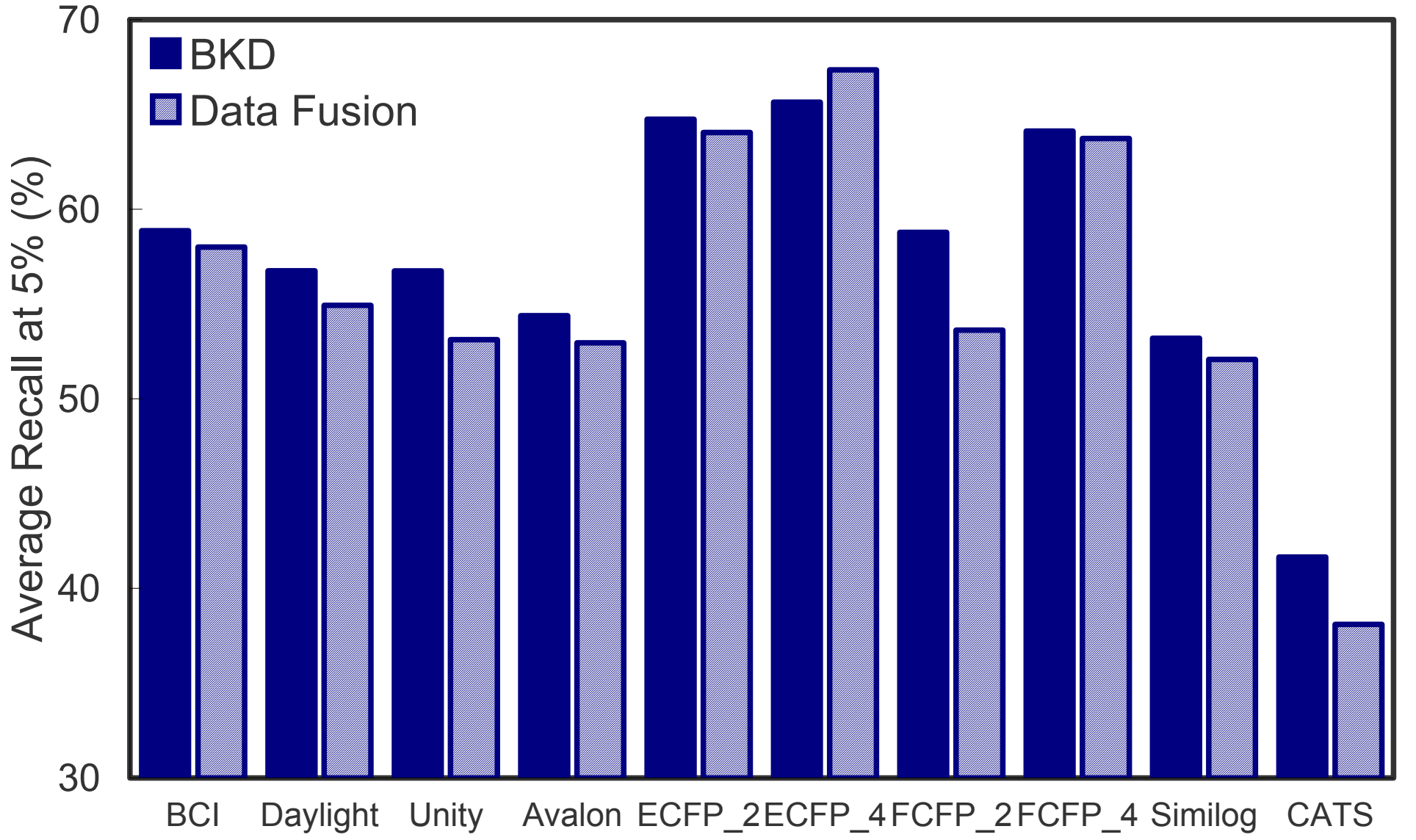
- Similog keys
  - Atom typing scheme based on four properties: hydrogen-bond donor, hydrogen-bond acceptor, bulkiness and electropositivity
  - Atom triplets of strings encoding absence and presence of properties, plus distance encoding form a DABE key
  - Vector contains a count for each of the 8031 possible DABE keys

Pharmacophore Vectors: CATS

- CATS vectors
  - Based on five atom types: hydrogen-bond acceptor, hydrogen-bond donor, positive, negative, lipophilic
  - Correlation vector representation calculated by

$$CVR_d^T = \frac{1}{A}\sum_{i=1}^{A}\sum_{j=1}^{A}\delta_{ij,d}^T$$

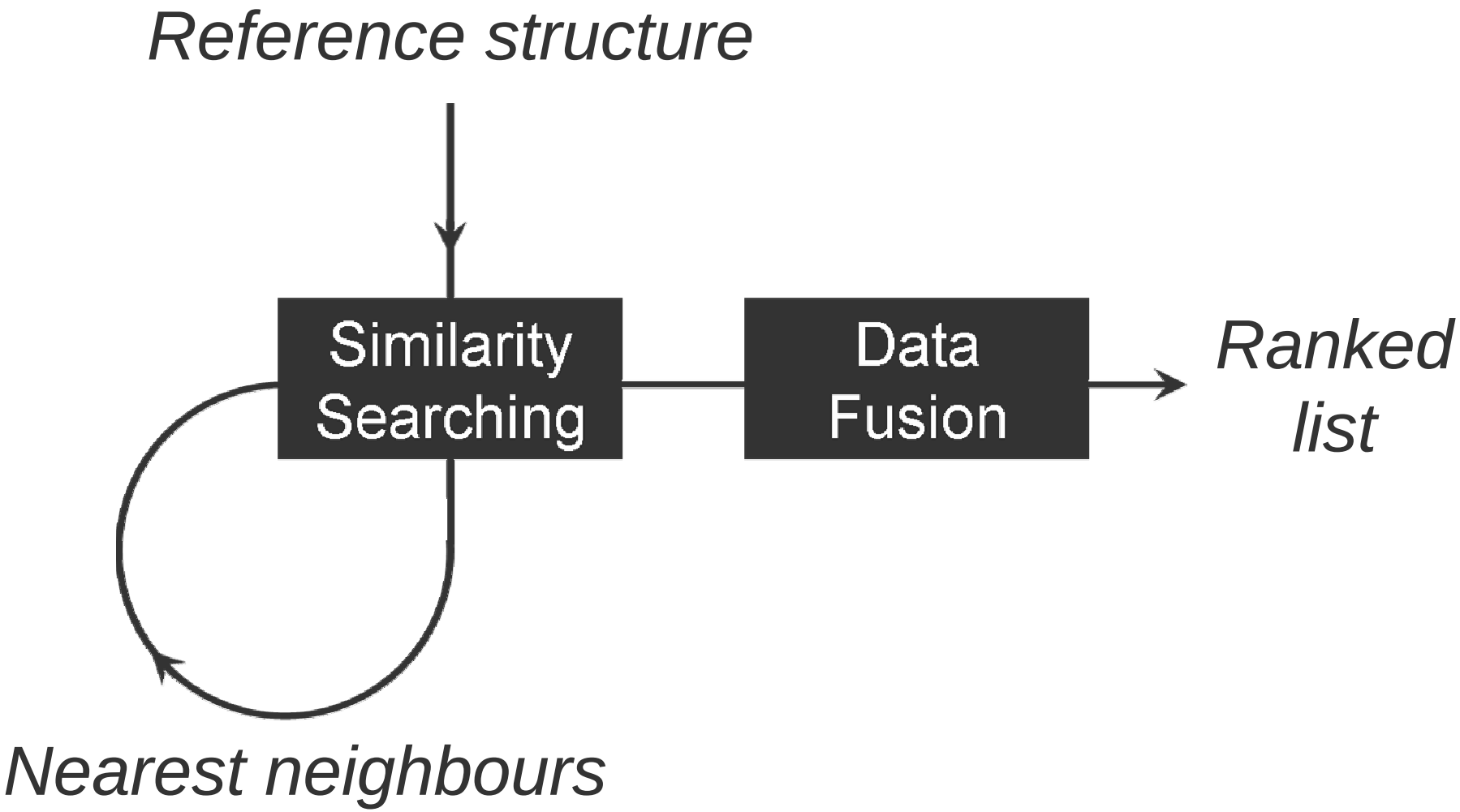
Results


BKD
Data Fusion
Average Recall at 5% (%)
70
60
50
40
30
BCI
Daylight
Unity
Avalon
ECFP_2
ECFP_4
FCFP_2
FCFP_4
Similog
CATS

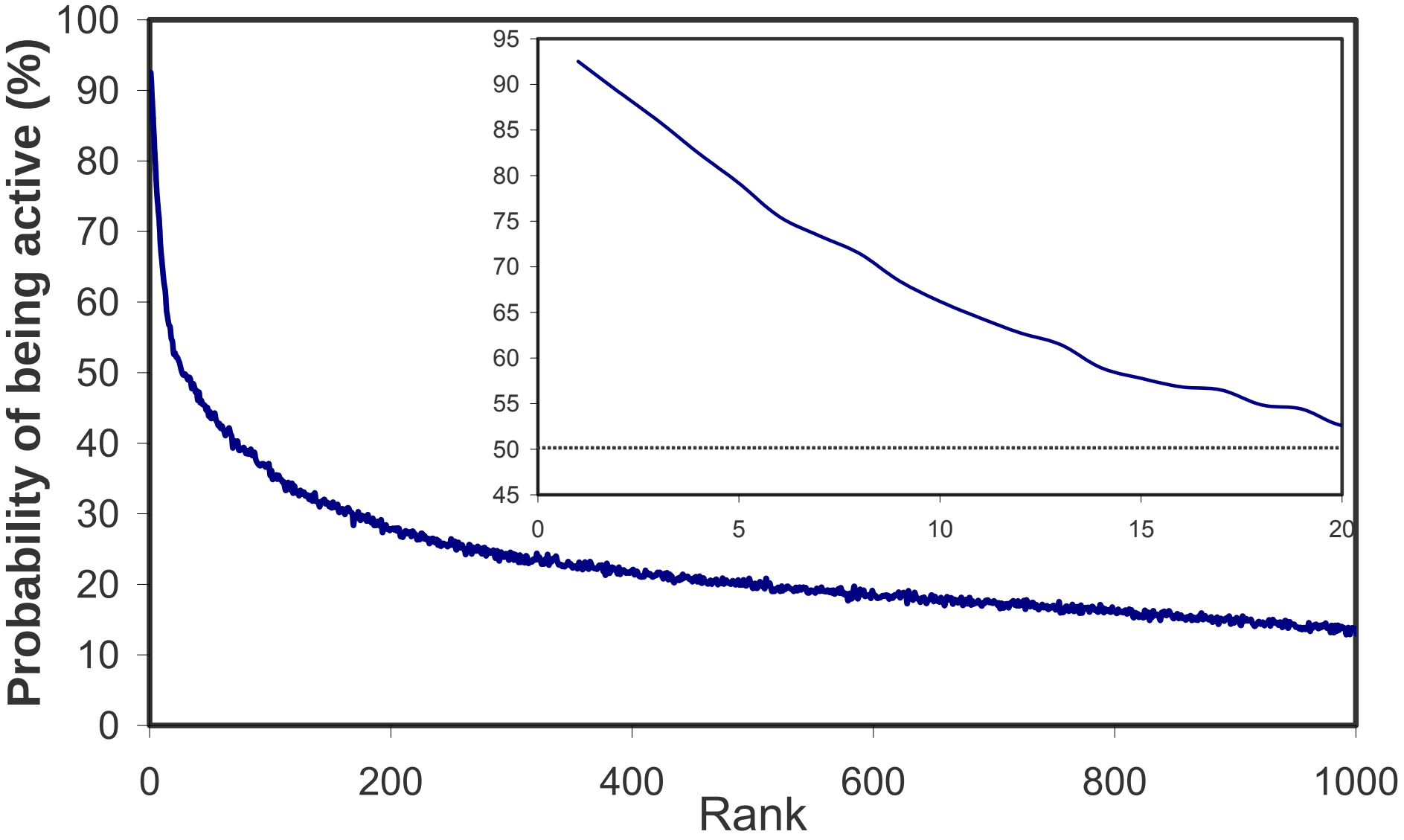
Turbo Similarity Searching (1)

- Similarity property principle: nearest neighbours are likely to exhibit the same activity as the reference structure
- Data fusion of multiple bioactive compounds is an effective way of improving the identification of active compounds

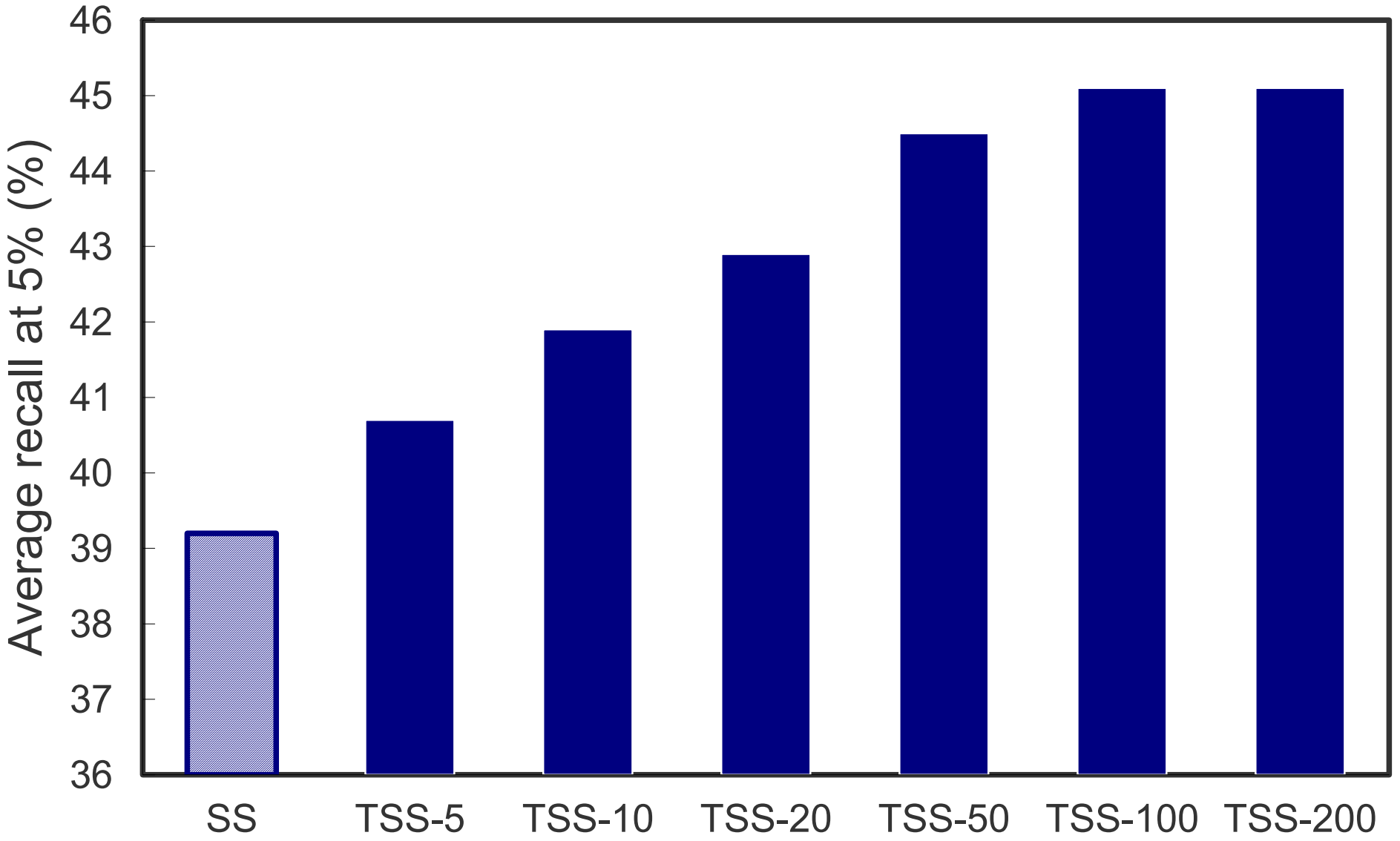
Turbo Similarity Searching (2)



Probability of being active vs Rank



Results


Average recall at 5% (%)
46
45
44
43
42
41
40
39
38
37
36
SS
TSS-5
TSS-10
TSS-20
TSS-50
TSS-100
TSS-200

General conclusions

- This work has demonstrated two effective ways of using multiple active structures in 2D similarity searching (BKD and Data fusion)
- This work has also demonstrated the general effectiveness of the circular substructure descriptors (ECFP_4 in particular)
- Turbo similarity searching is a simple way of increasing the cost effectiveness of similarity searching

Acknowledgements

- Novartis Institutes for BioMedical Research for funding
- CINF Division for funding
- MDL Information Systems Inc. for the provision of the MDDR database
- Barnard Chemical Information Ltd., Daylight Chemical Information Systems Inc., the Royal Society, Tripos Inc., Scitegic Inc. and the Wolfson Foundation for software and laboratory support.