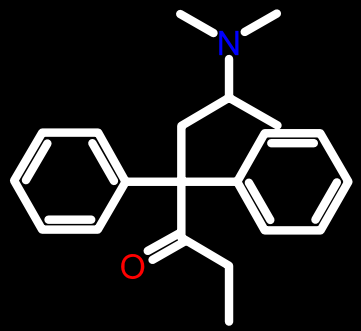
Can similarity searching be improved?

Jérôme Hert

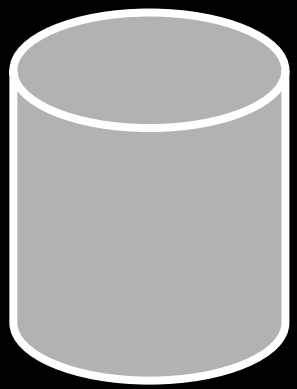
University of California, San Francisco

Similarity Searching (SS)

Reference structure



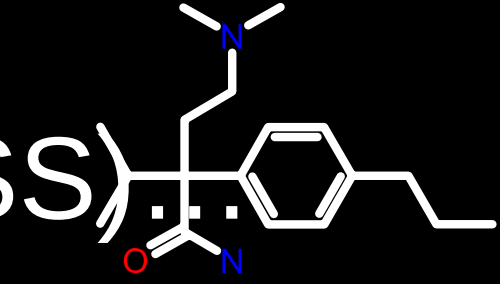
$S_{A,B}$

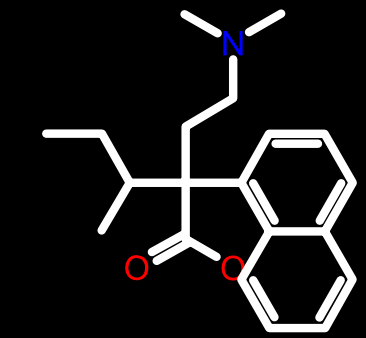


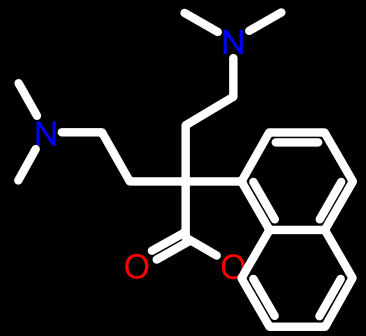
Library of candidate compounds

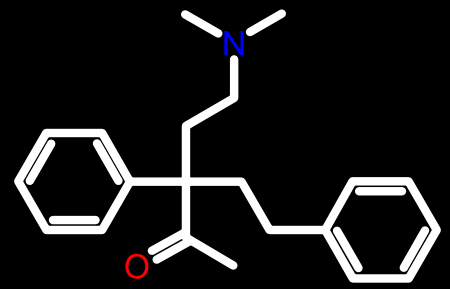
```
0011101010001

0001100000000
0000000010010
1000000111010
0000000101000
0011001000000
0001000000100
0000000000000
…
```



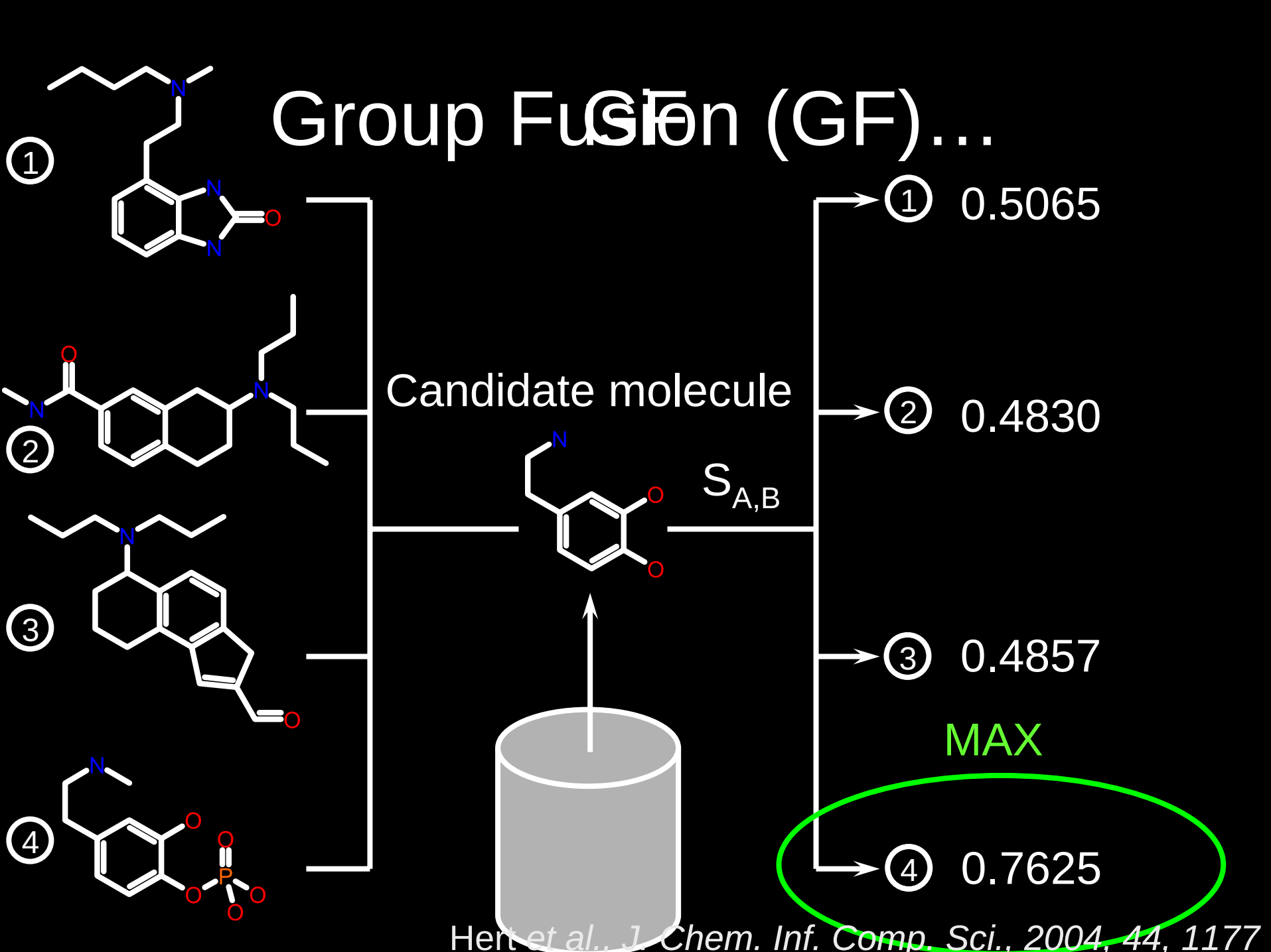






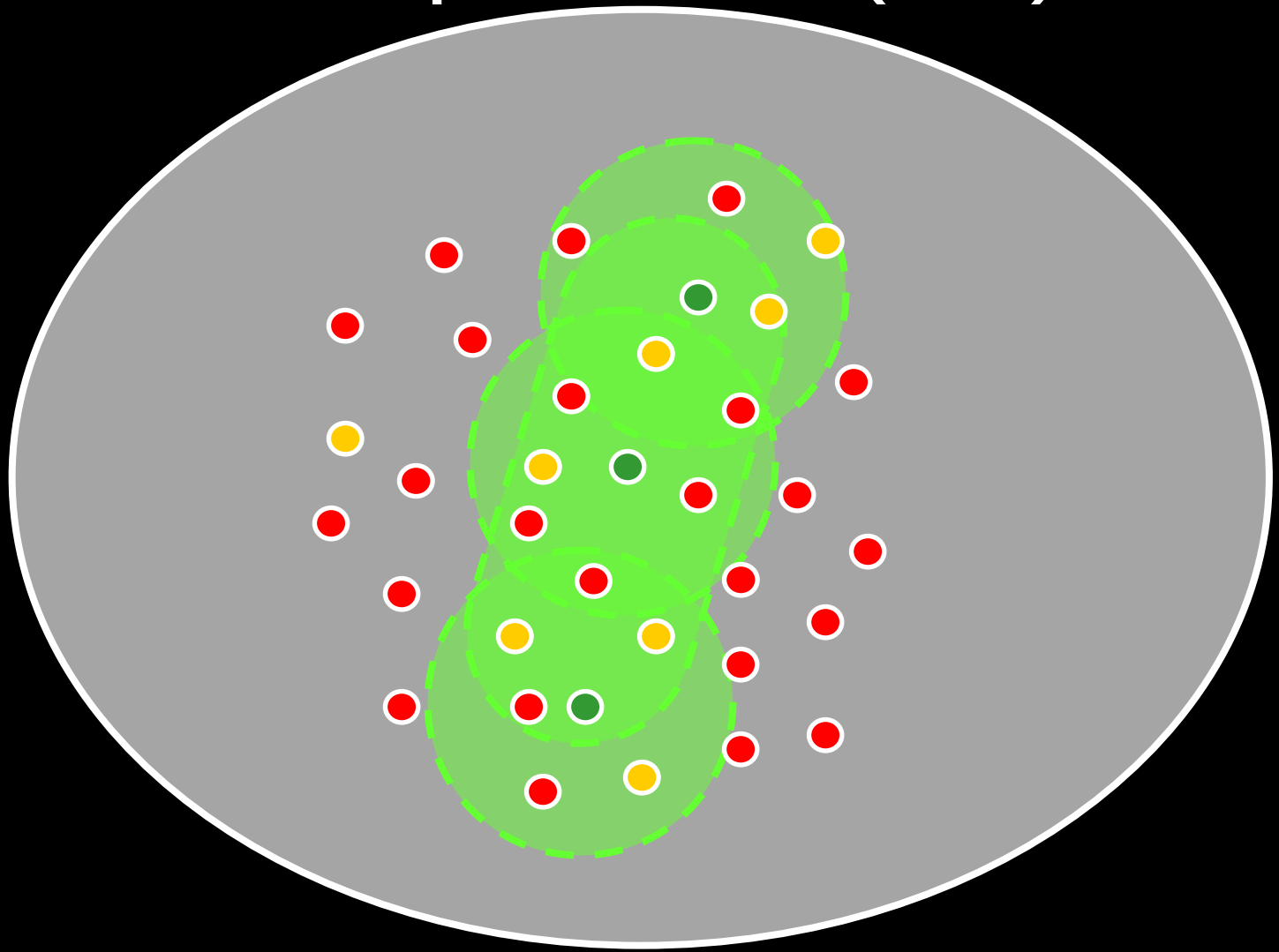
Nearest neighbors

Group Fusion (GF)…



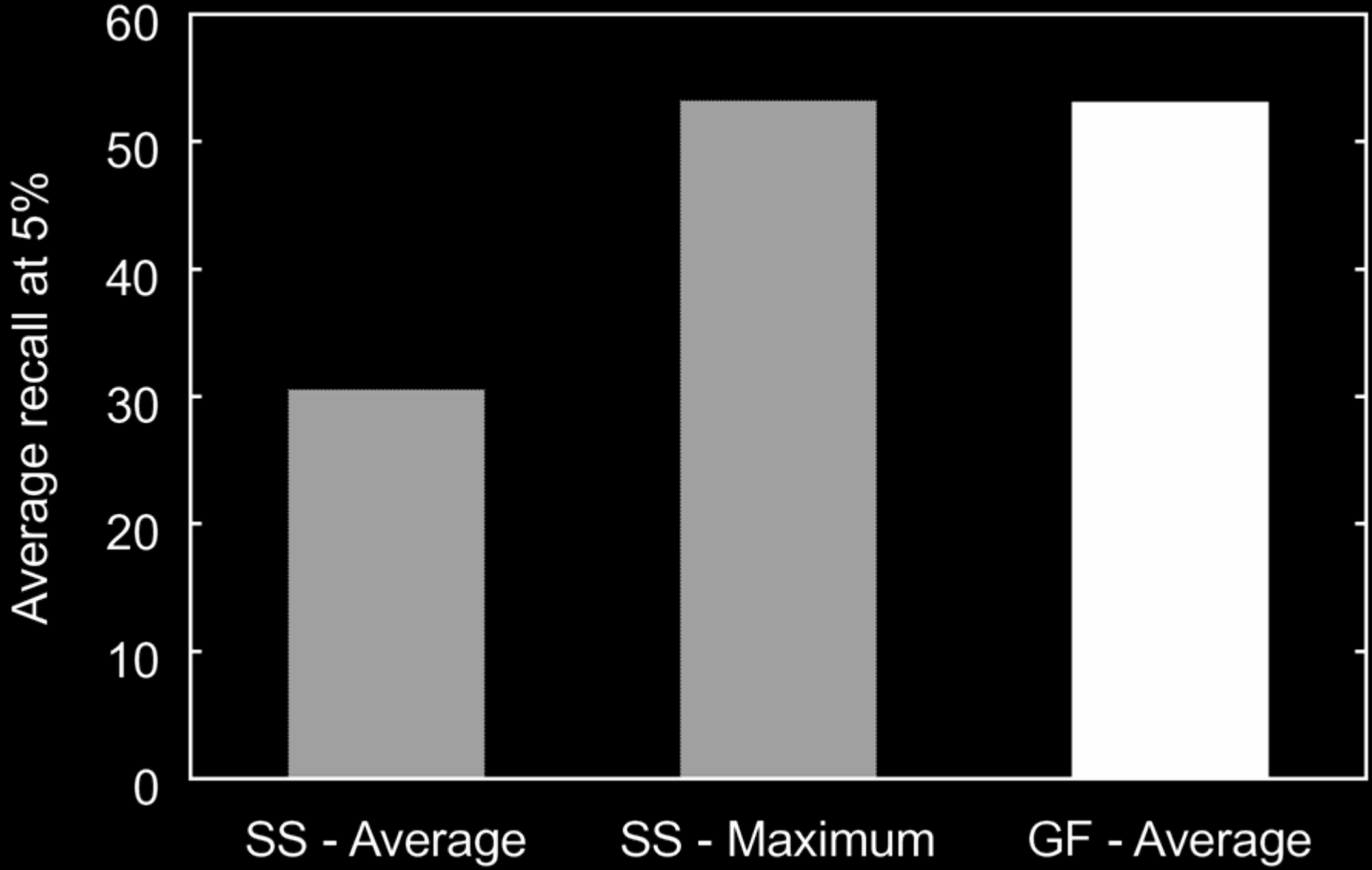
Hert *et al., J. Chem. Inf. Comp. Sci., 2004, 44, 1177*

Group Fusion (GF)…

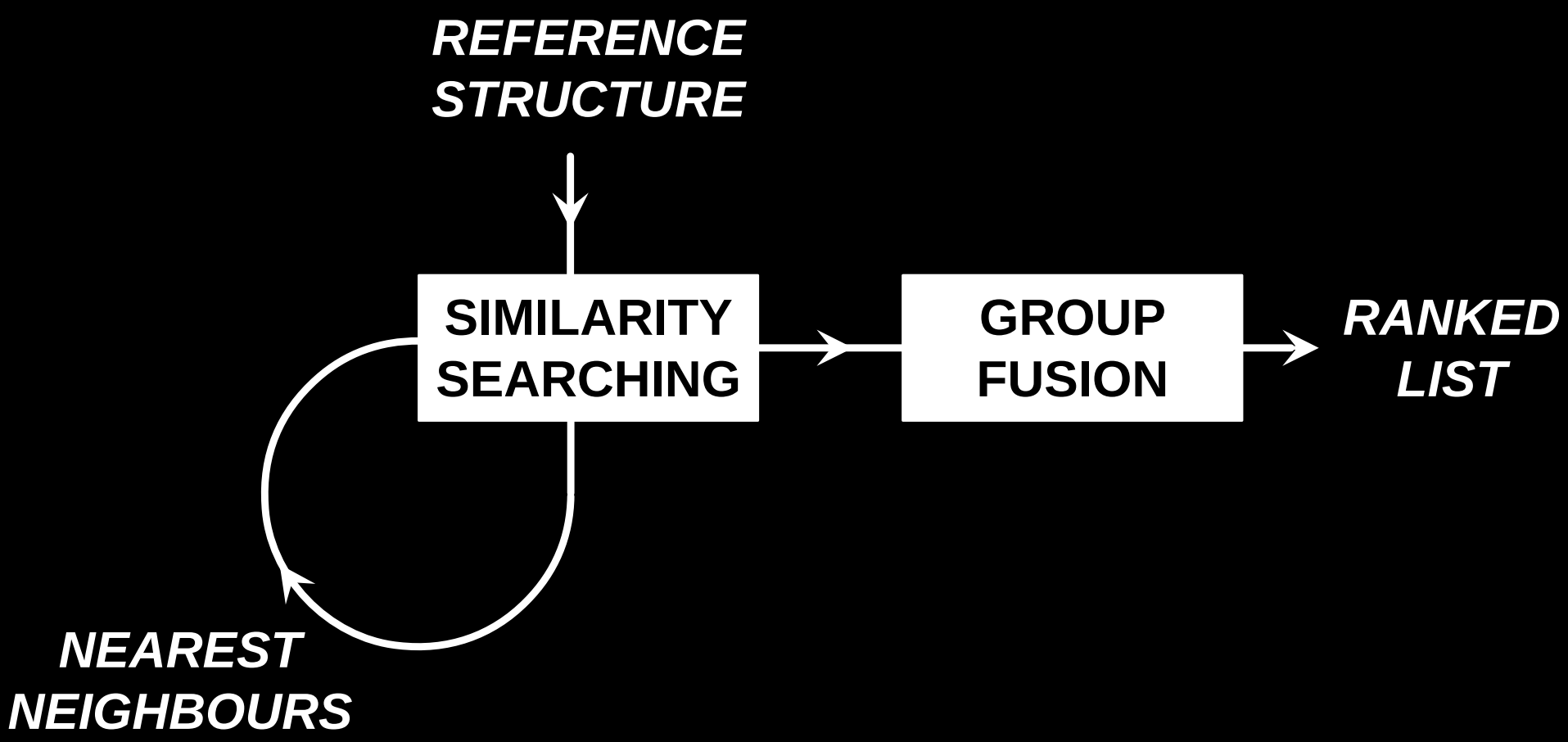


Inactive molecule Active molecule Reference molecule

SS versus GF

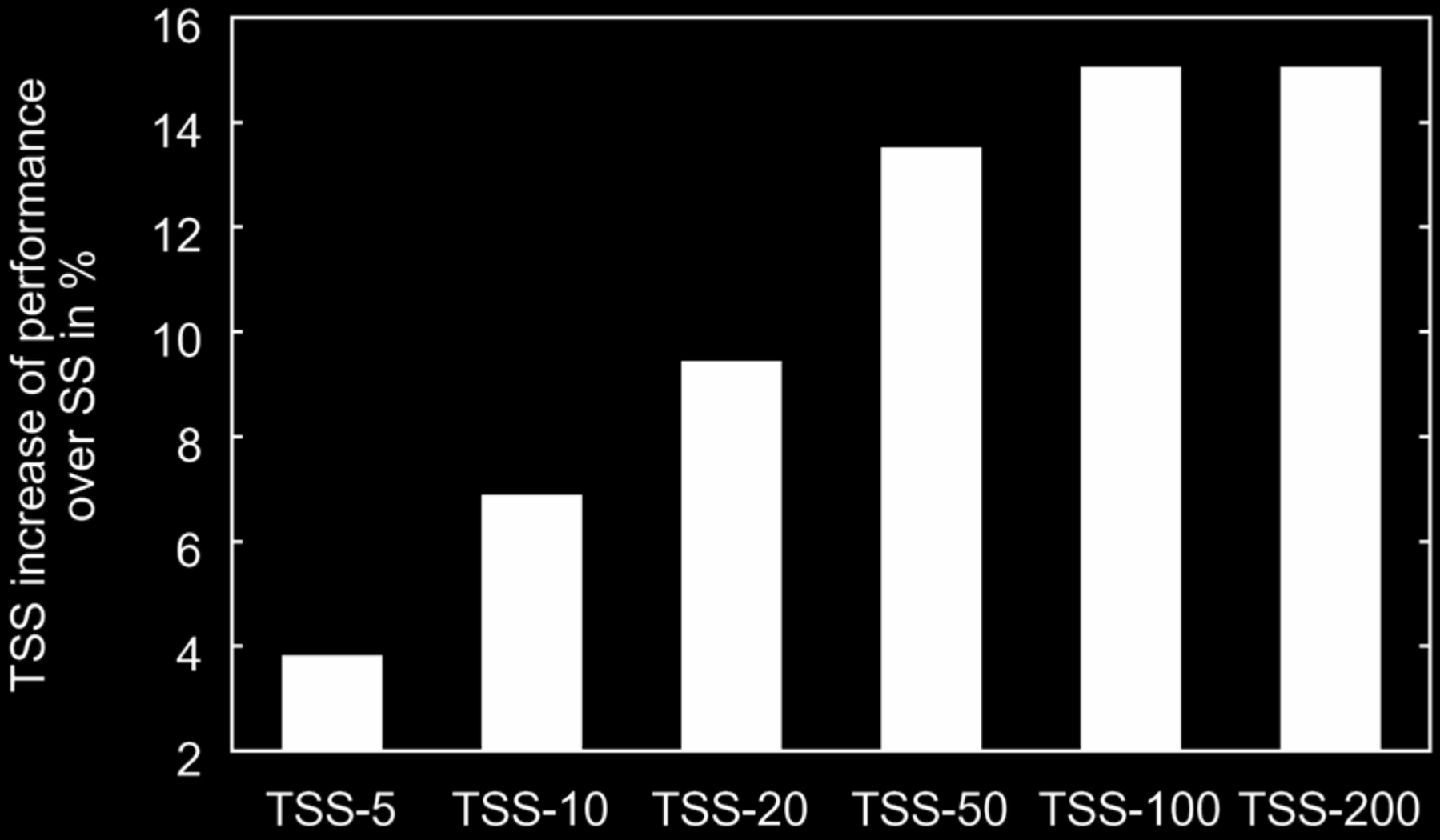

Average recall at 5%
60
50
40
30
20
10
0
SS - Average
SS - Maximum
GF - Average

Turbo Similarity Searching (TSS)…

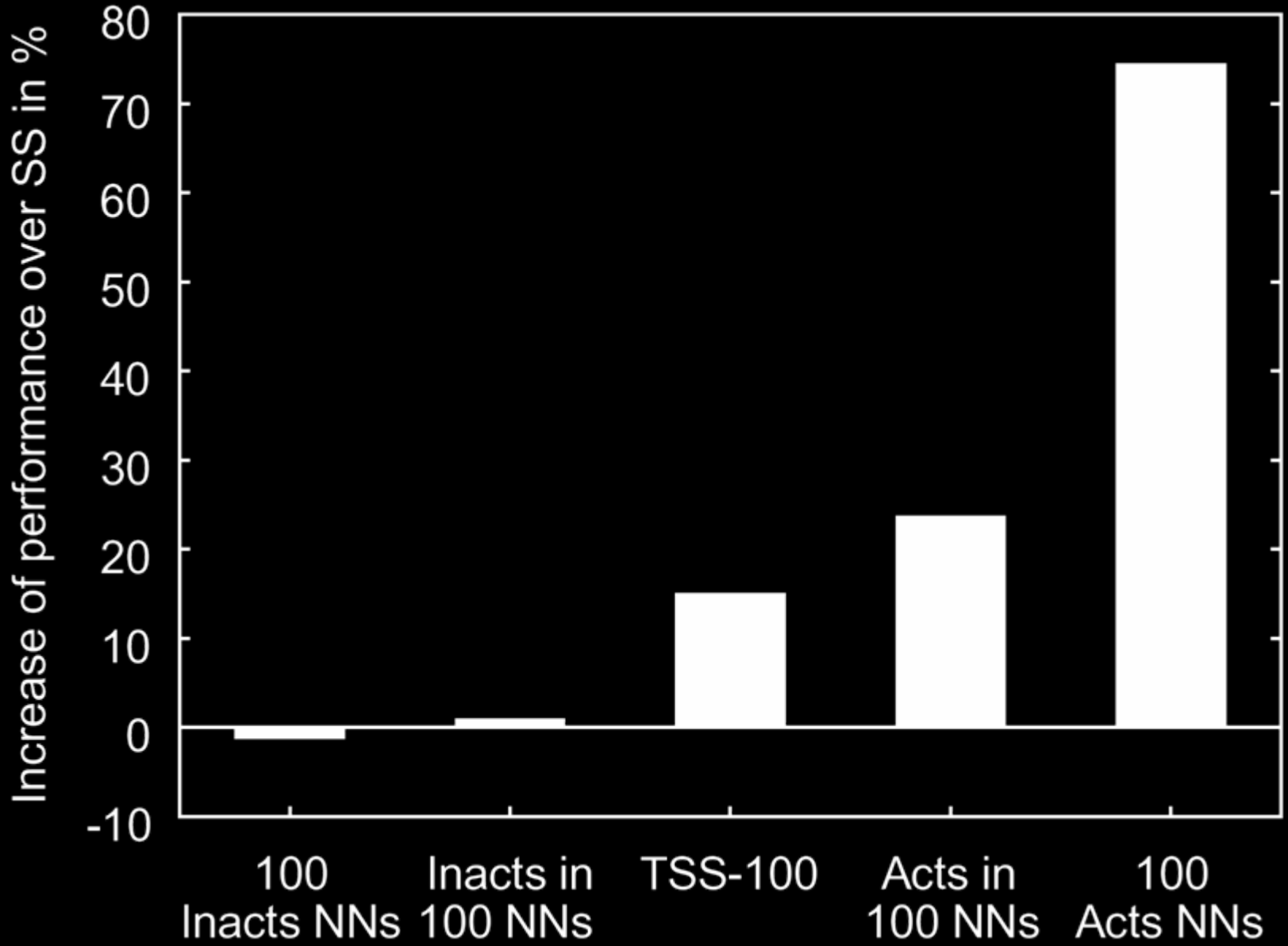


Hert *et al., J. Med. Chem., 2005, 48, 7049*

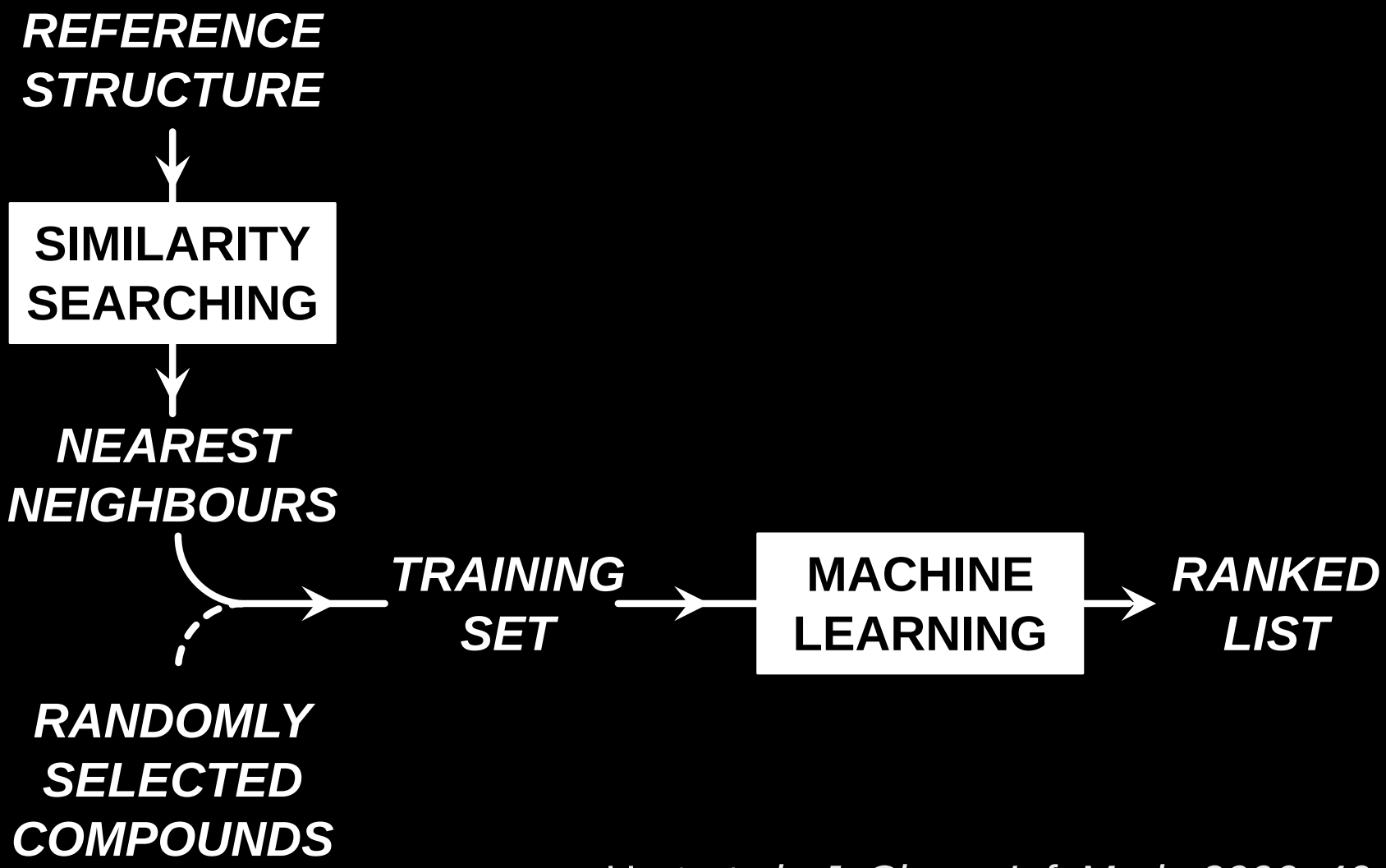
SS versus TSS



How does it work?

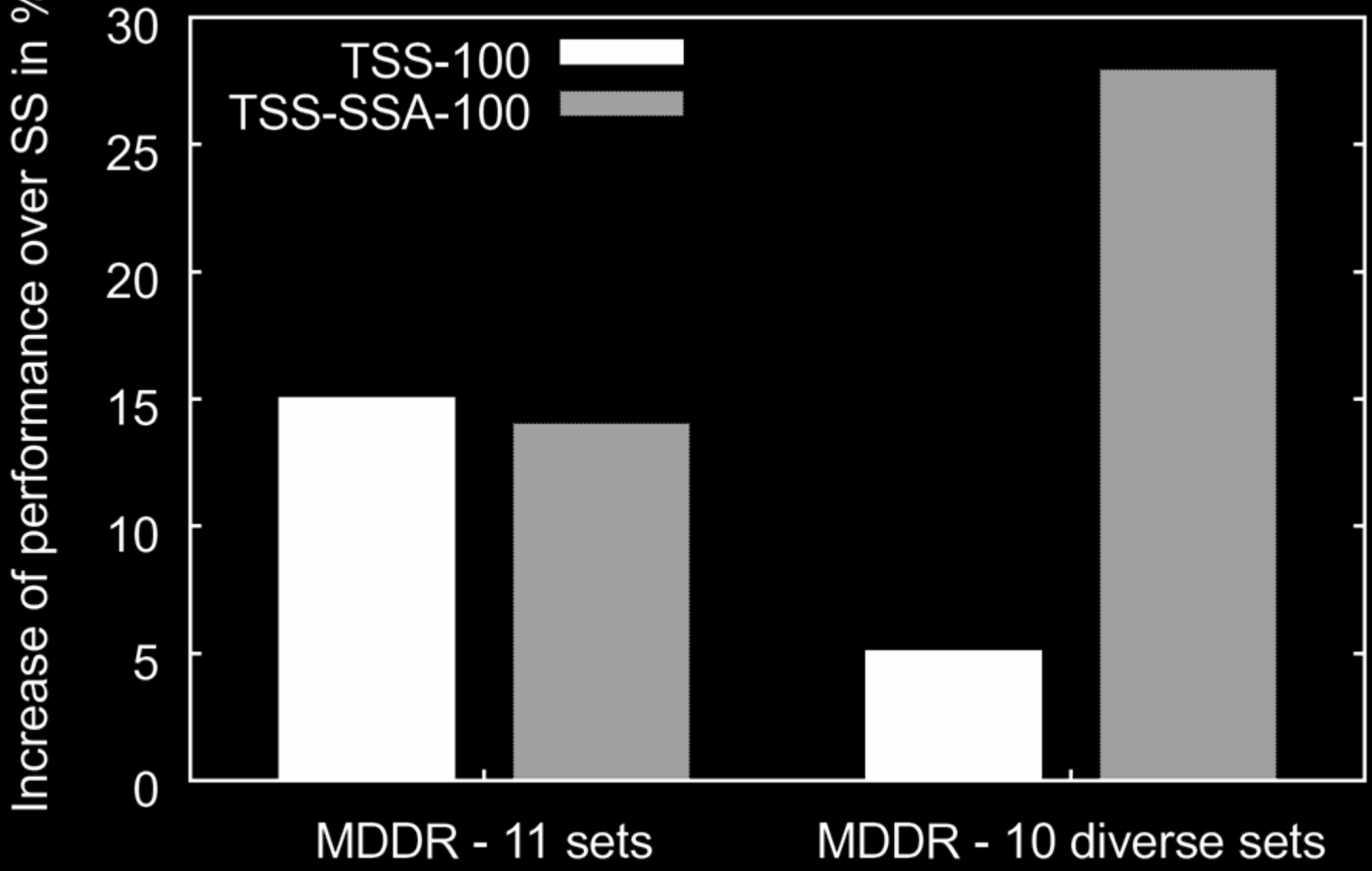


TSS Bundled



Hert *et al., J. Chem. Inf. Mod., 2006, 46, 462*

SS versus TSS versus TSS-SSA



Question

- Is the use of nearest neighbors as a mean to enhance similarity searching generally applicable?

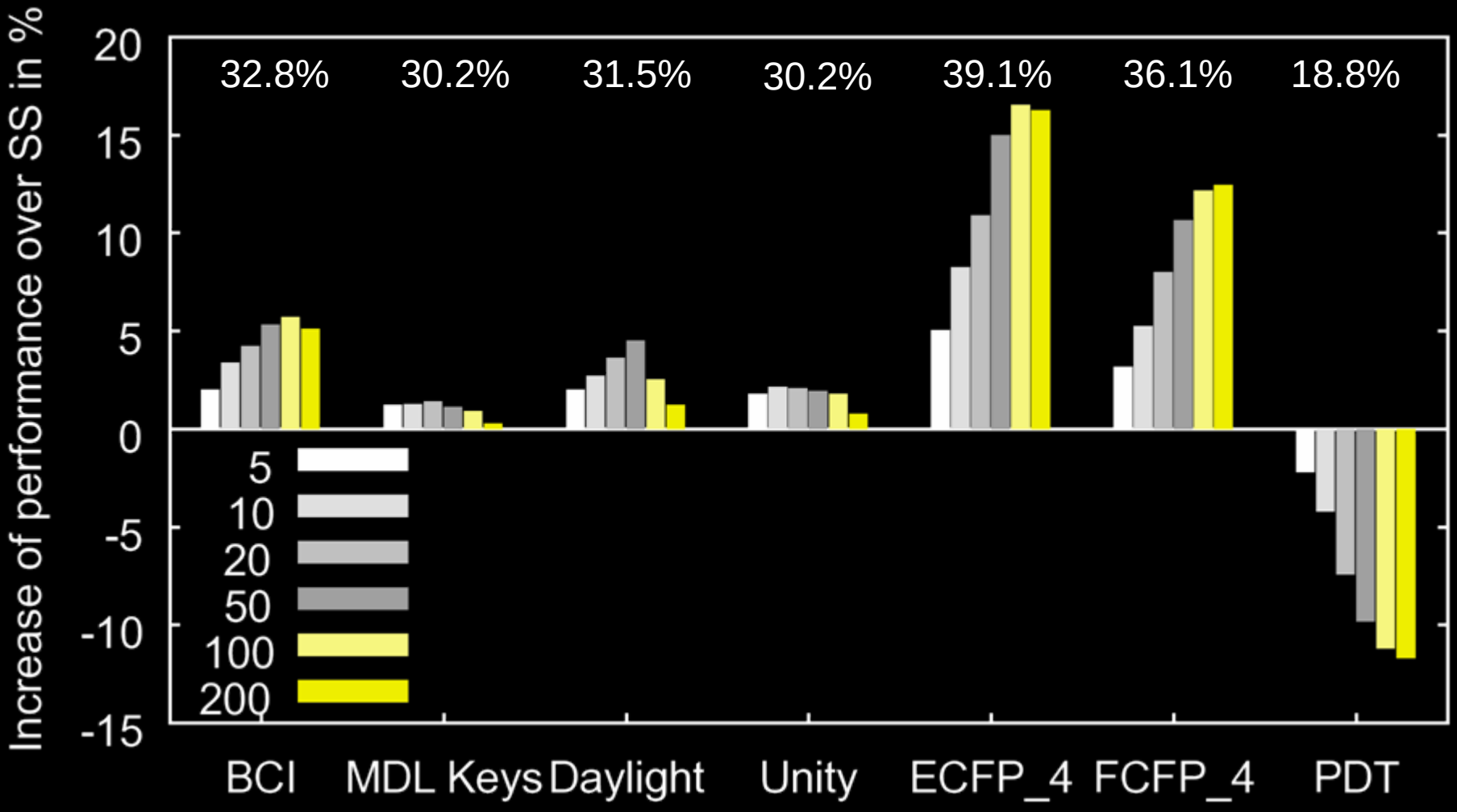
Datasets

- MDL Drug Data Repository (MDDR):
  - 100K molecules
  - 1 group of 11 sets
  - 1 group of 10 “diverse” sets (based on MPS)
- NCI AIDS dataset
  - 40K molecules
  - 393 confirmed actives to retrieve 1430 confirmed high or moderate actives.
- ACS datasets

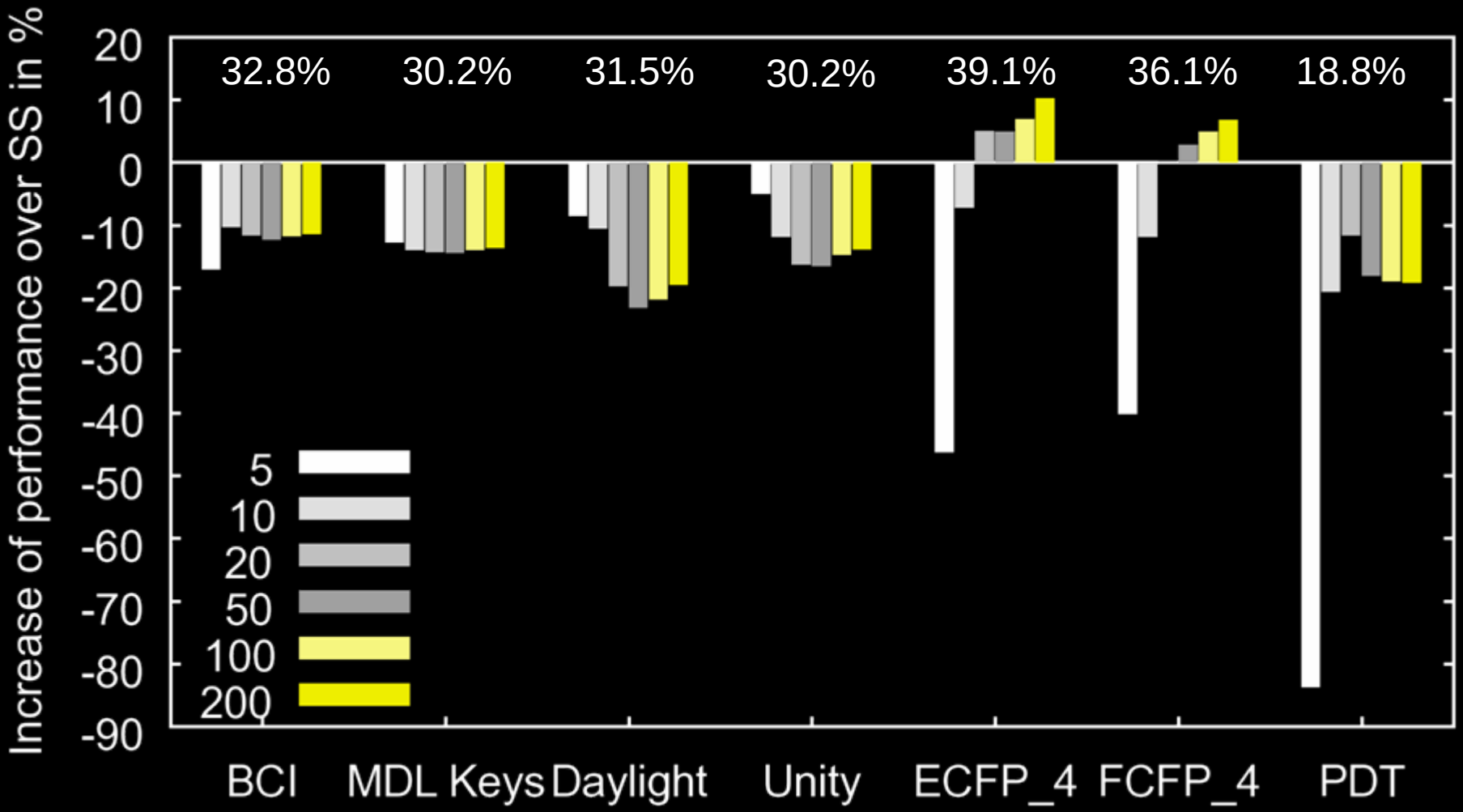
Descriptors

- Structural Keys:
  - 1052-bit BCI
  - 166-bit MDL
- Hashed Fingerprint
  - 2048-bit Daylight
  - 1024-bit Unity
- Circular Substructures
  - Scitegic Extended Connectivity ECFP_4
  - Scitegic Functional Connectivity FCFP_4
- 3D Pharmacophores
  - Pharmacophore distance triplets (PDT)

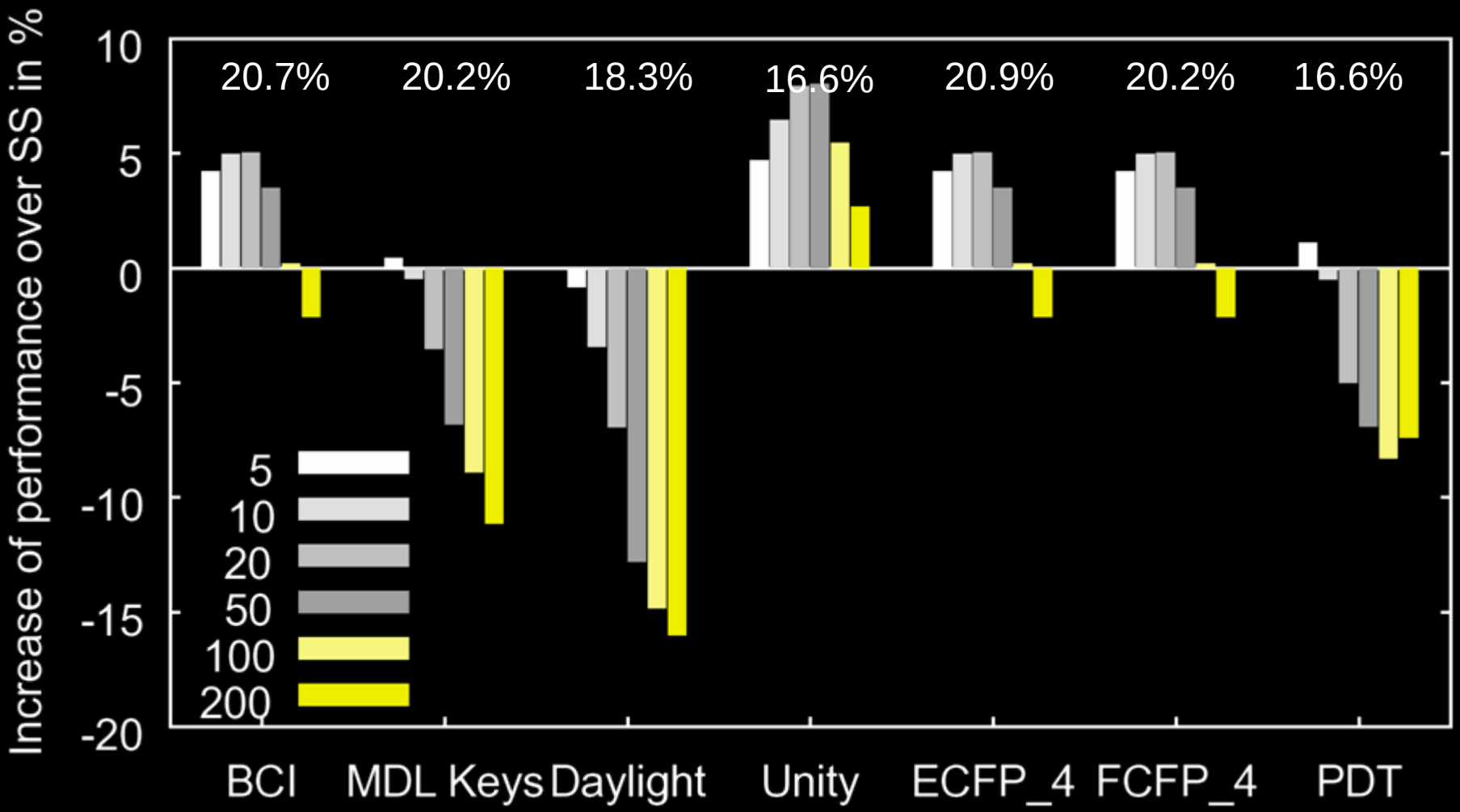
Results – MDDR Original – TSS



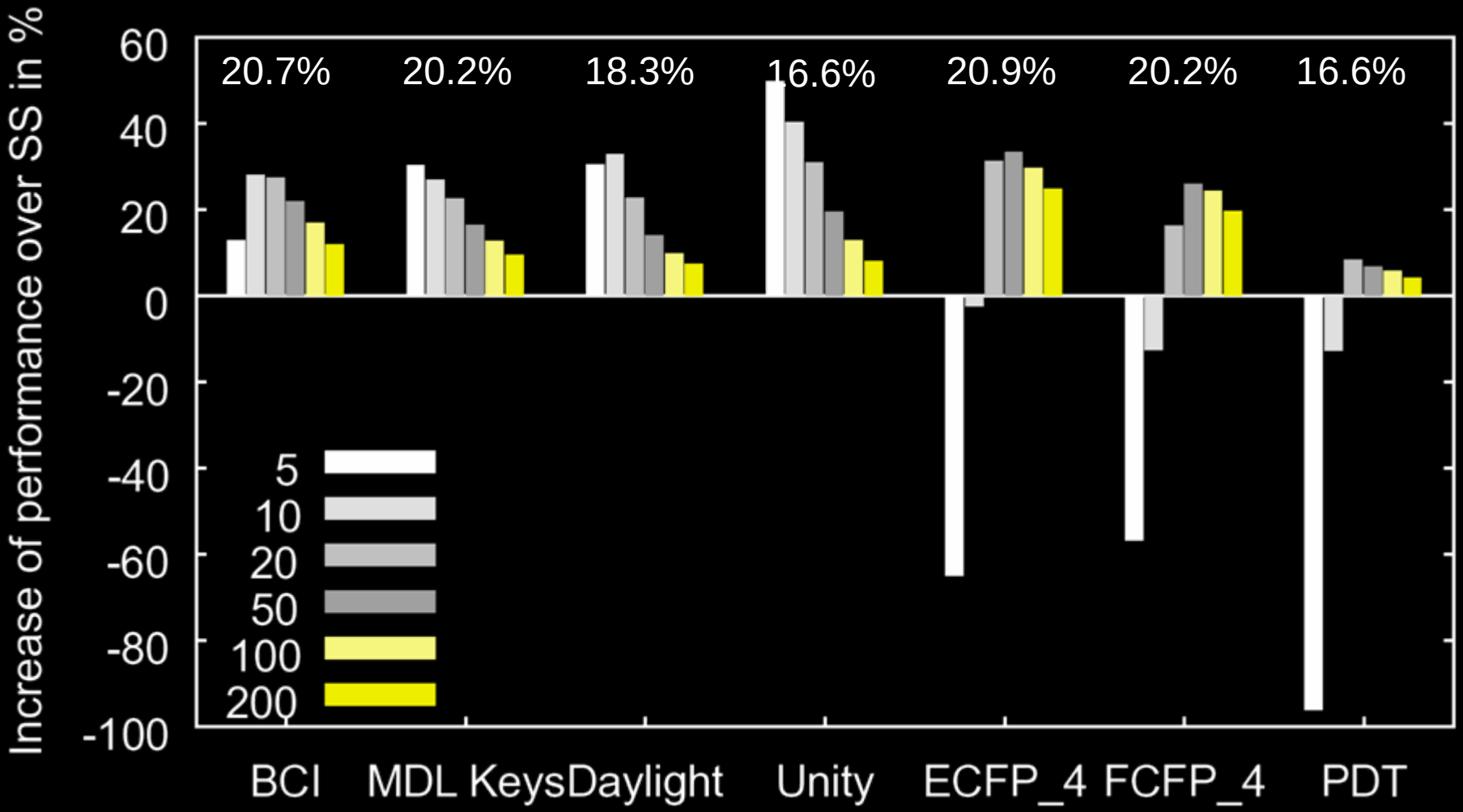
Results – MDDR Orig – TSS-SSA



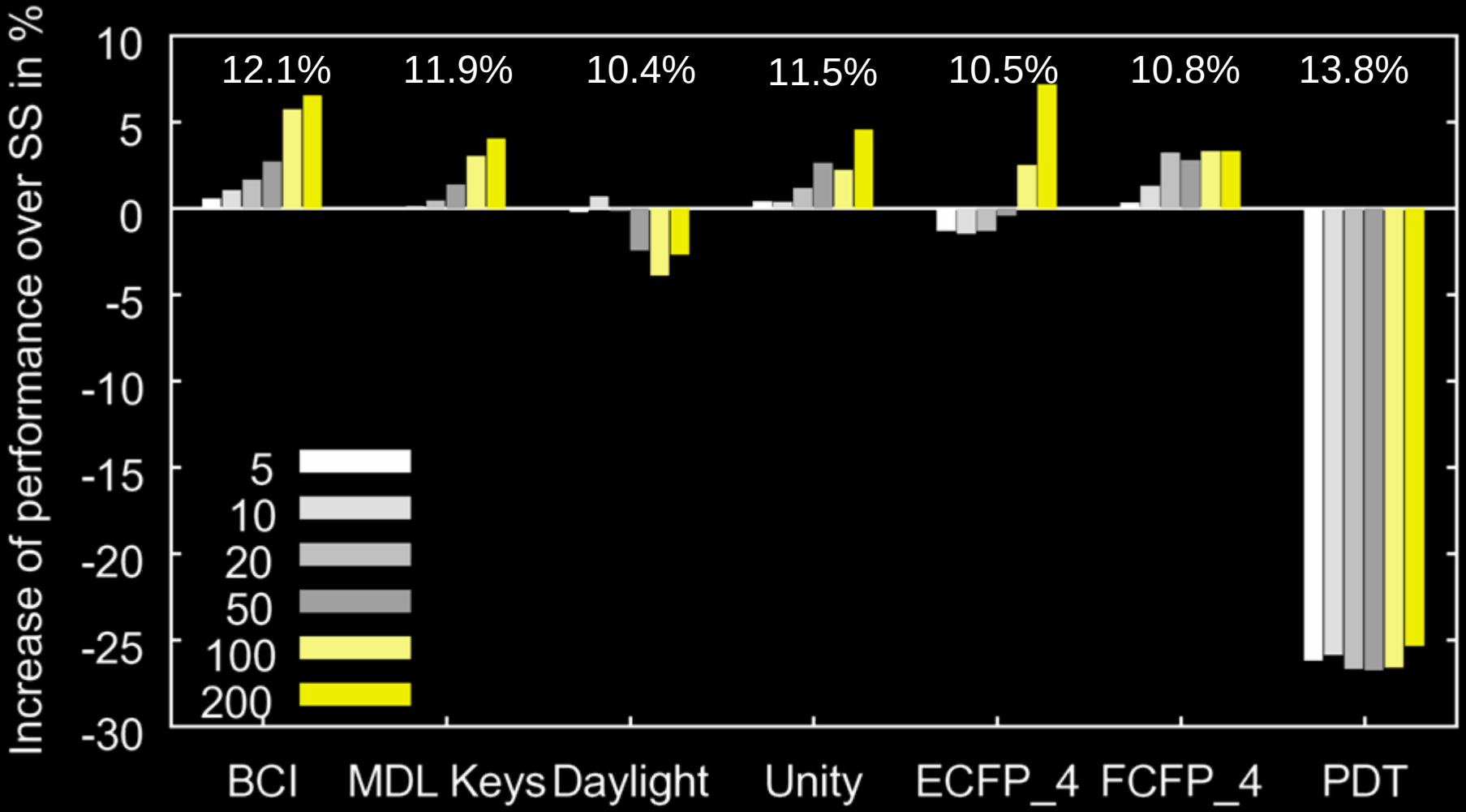
Results – MDDR Diverse – TSS



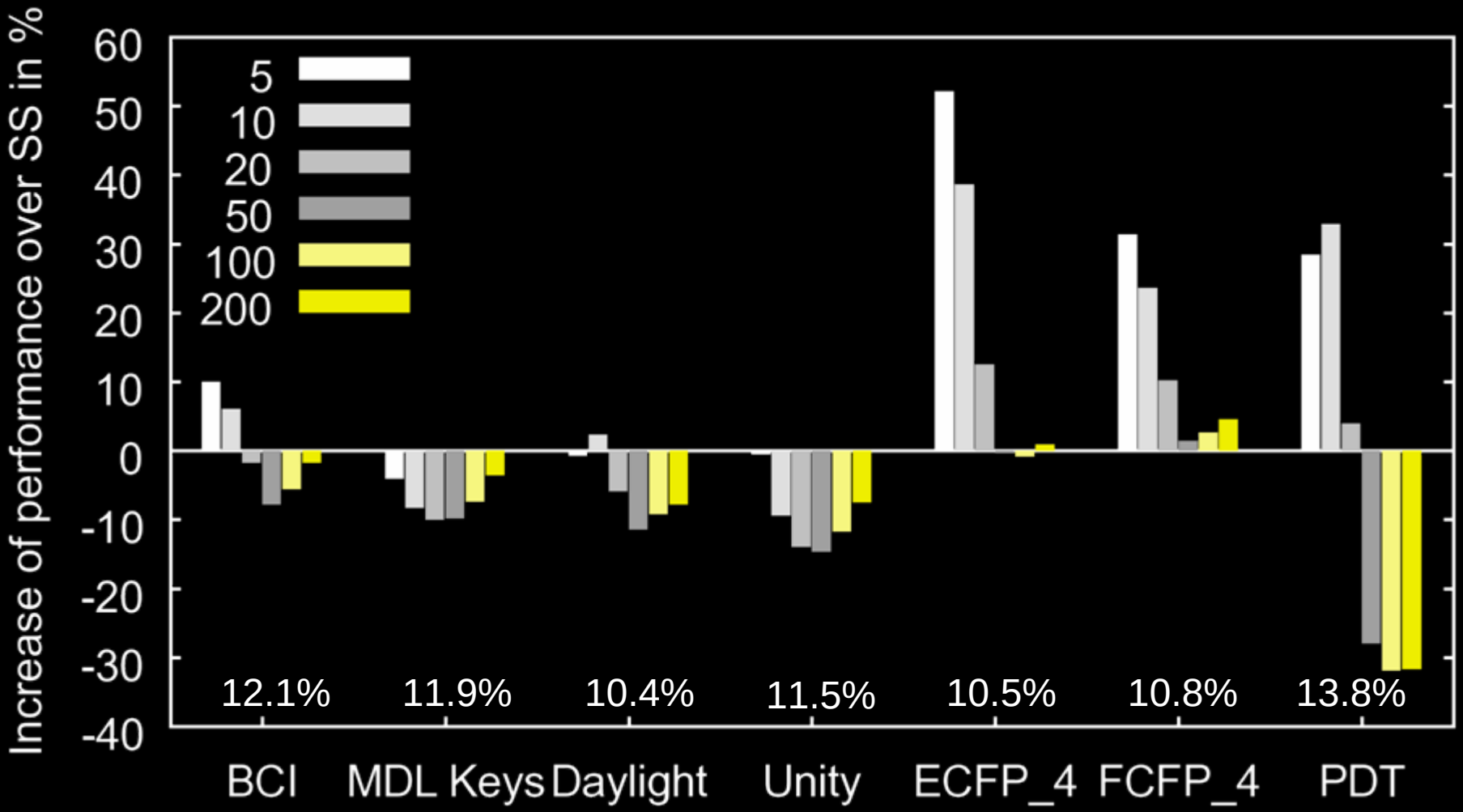
Results – MDDR Div. – TSS-SSA



Results – NCI AIDS – TSS

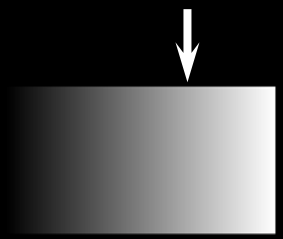


Results – NCI AIDS – TSS-SSA



Conclusions

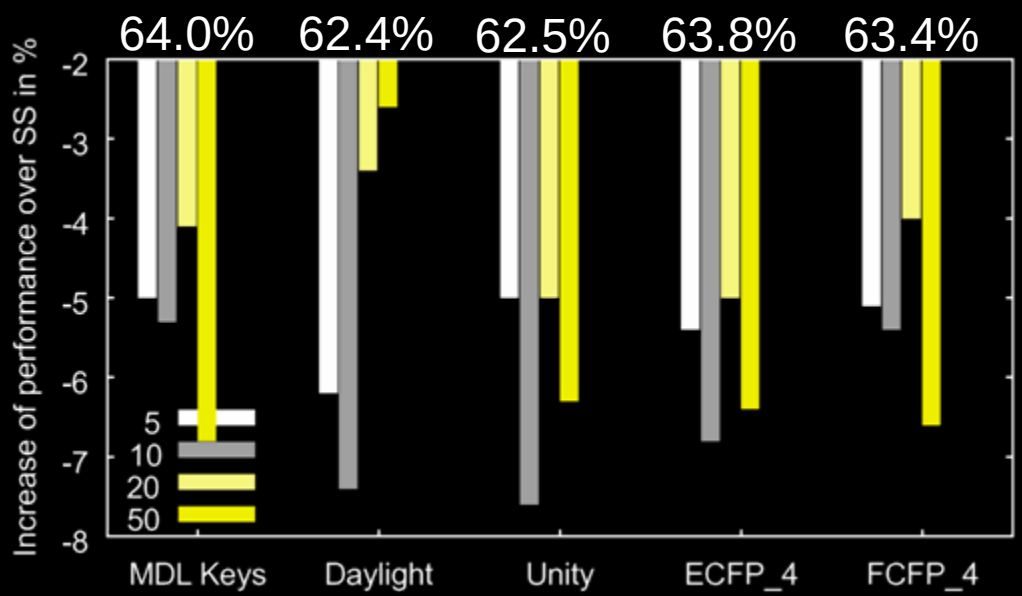
- TSS is NOT general (dataset, descriptor etc. independent).
- TSS > SS (In general)
- TSS with GF especially interesting with homogeneous datasets
- TSS with ML (SSA) especially interresting with heterogeneous datasets



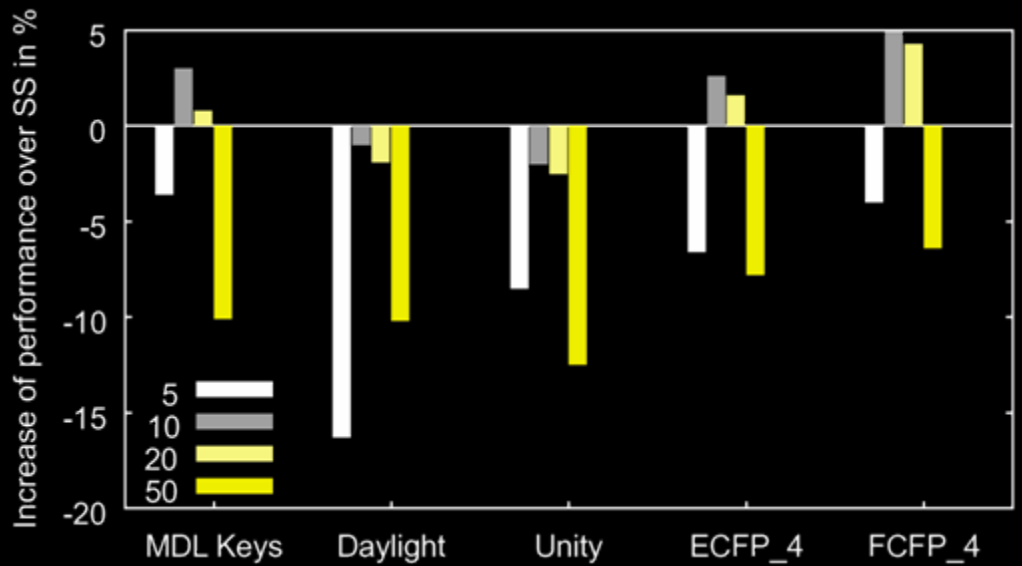
…

Results – Dock29

TSS

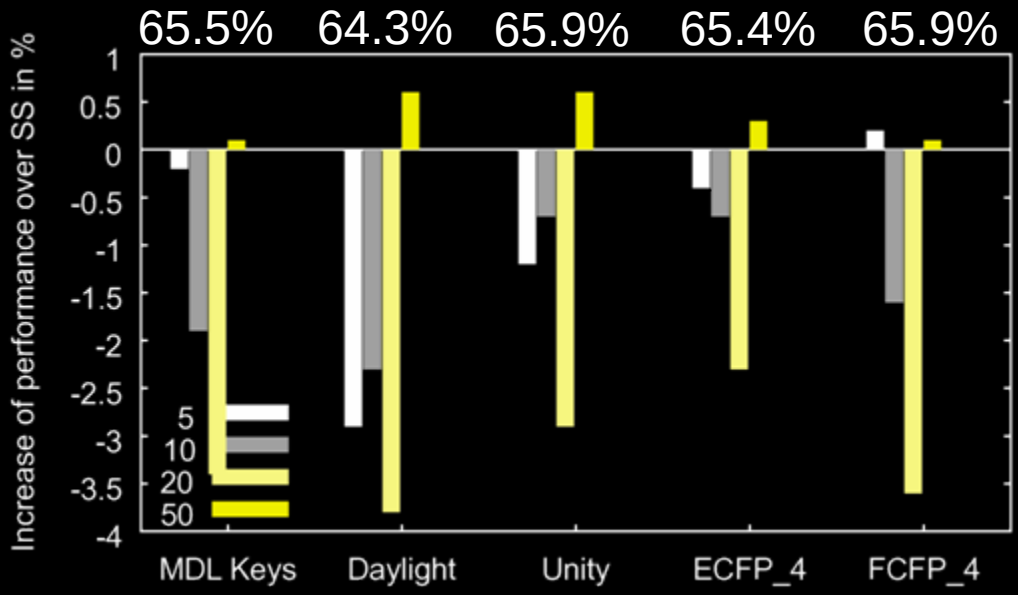


TSS-SSA

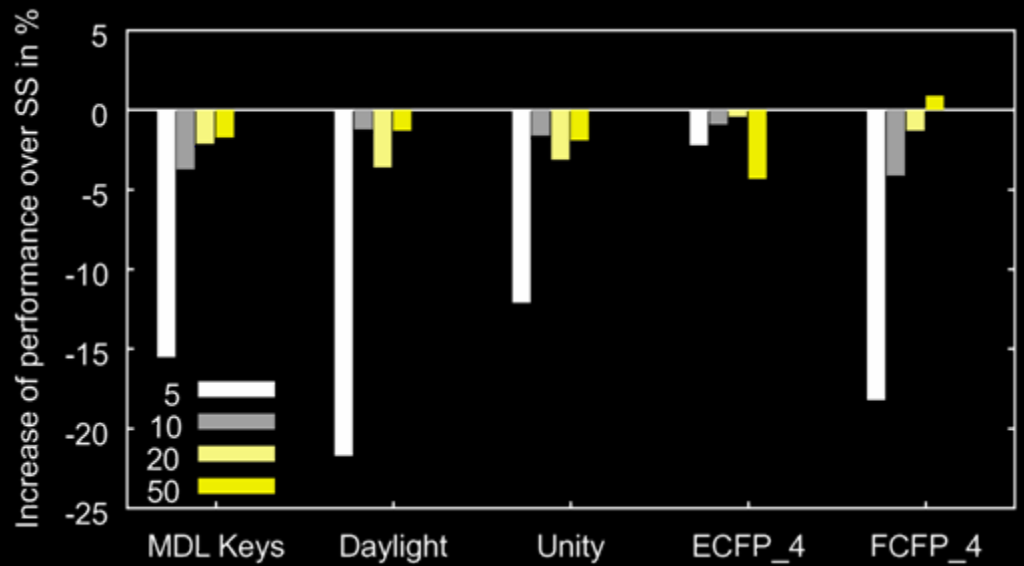


Results – Similarity22

TSS

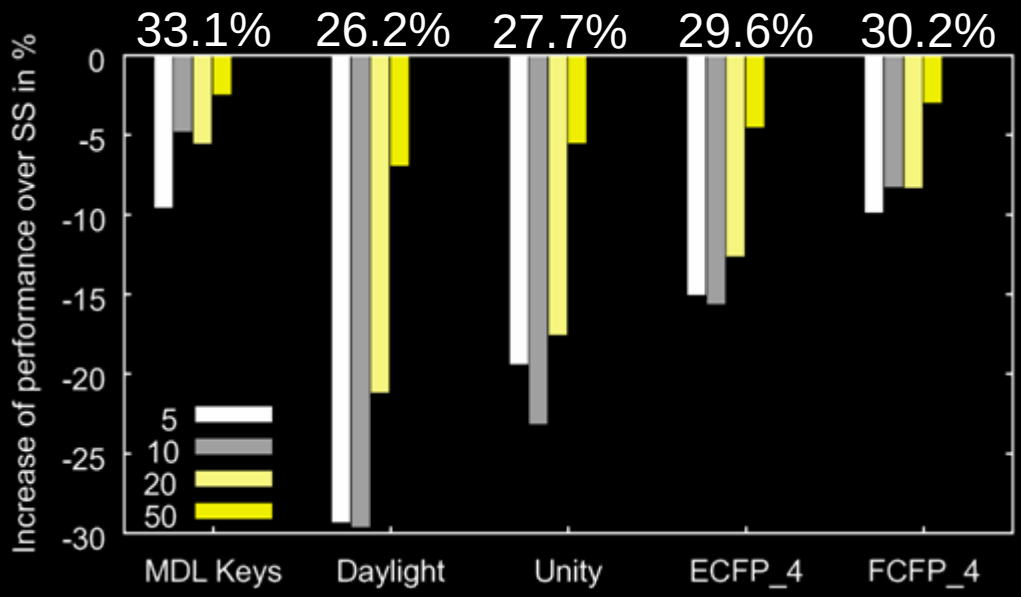


TSS-SSA

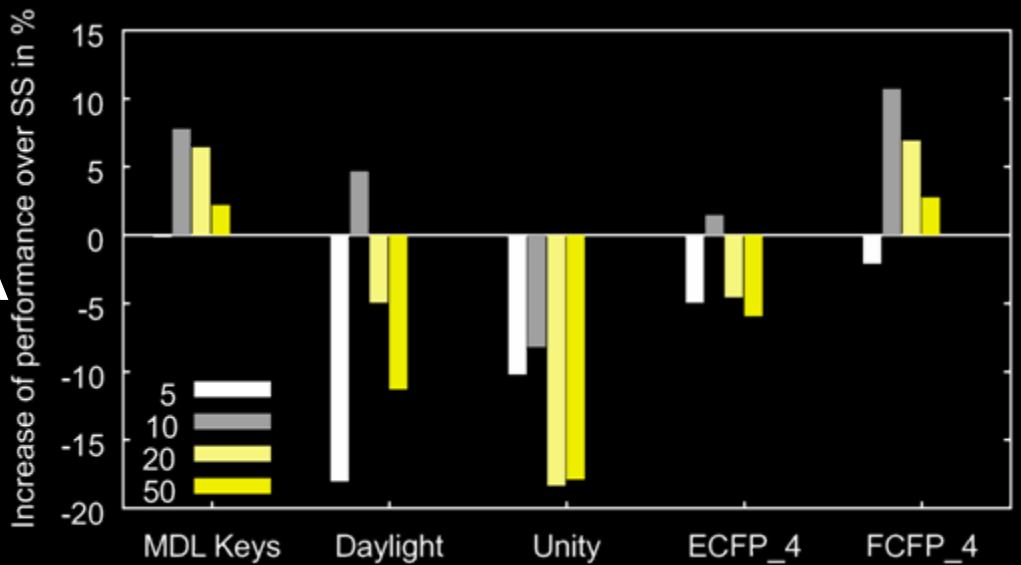


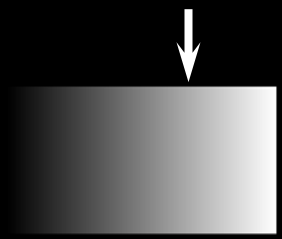
Results – Similarity4

TSS



TSS-SSA





Conclusions

- TSS is NOT general (dataset, descriptor etc. independent).
- TSS is still worth using (I would)
- TSS and TSS-SSA are alternative methods to SS.

Acknowledgements

- Peter Willett (University of Sheffield)
- David J. Wilton (University of Sheffield)
- Brian Shoichet (UCSF)