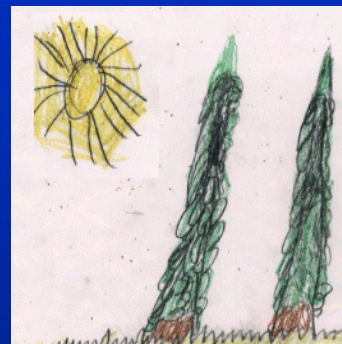
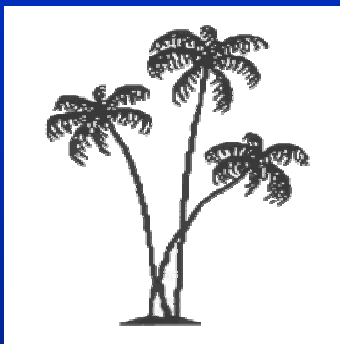
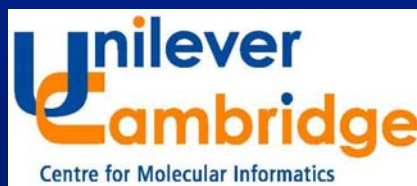


Molecular Similarity – Approaches, Advances and Illusions

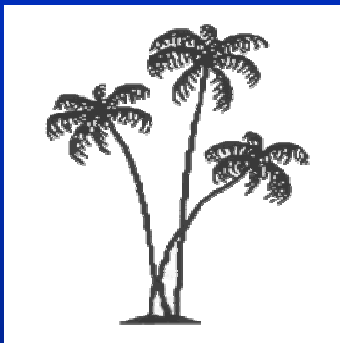


Andreas Bender, ab454@cam.ac.uk

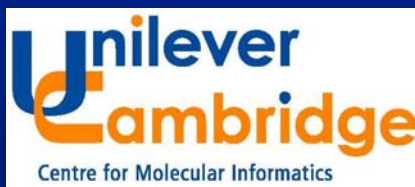


Unilever Centre for Molecular Informatics,
University of Cambridge, UK

Molecular Similarity – (Some) Approaches, (Small) Advances and (Great) Illusions



Andreas Bender, ab454@cam.ac.uk



Unilever Centre for Molecular Informatics,
University of Cambridge, UK

The Menu

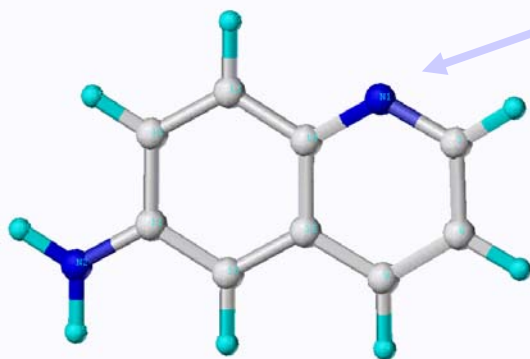
- Introduction to similarity searching
- Some of our approaches to molecular similarity
- Some thoughts on the databases and performance measures we use
- Information content of current descriptors, the “bias” of chemical libraries and their suitability for the estimation of descriptor performance

Similarity Searching

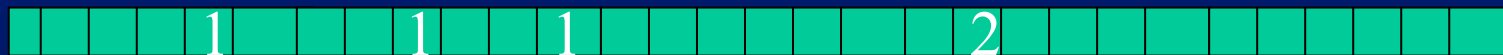
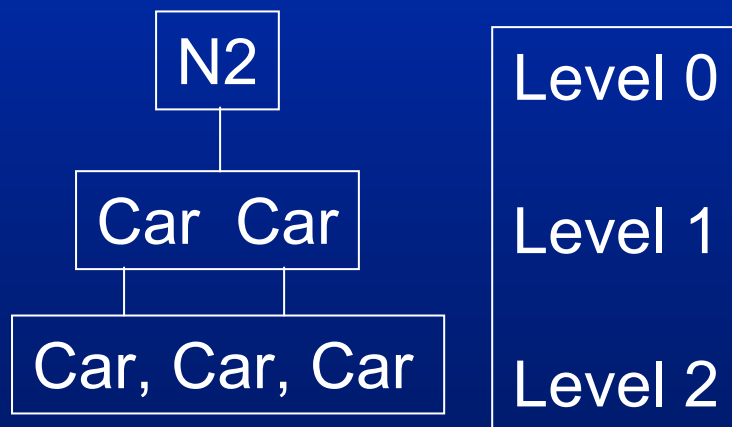
- Complementary approach to substructural searching
- In substructure searching exact retrieval of a subgraph of a molecule is performed
- In similarity searching, an abstract molecular representation in *descriptor space* is calculated which is compared to abstract representations of other molecules
- For reviews see e.g.:
 - Bender, A. *et al.*, *Ann. Rep. Comp. Chem.* 2006 (*in statu nascendi*); focus on methods validation
 - Bender, A. and Glen, R.C., *Org. Biomol. Chem.*, 2004, 2, 3204 – 3218.
(freely available from www.cheminformatics.org)

Earlier: MOLPRINT 2D (~Augmented Atoms)

- E.g. 6-Aminoquinoline



1. Assign Sybyl mol2 atom types
 2. Find connections
Find connections to connections
- Create a tree down to n levels
'Bin' the atom types for each level
->Creates a 'fingerprint' for this atom



These features are created for every (heavy) atom in the molecule
(Bender, A., *et al.*, *JCICS* 2004, 44, 170-178; *JCICS* 2004, 44, 1710-1718)

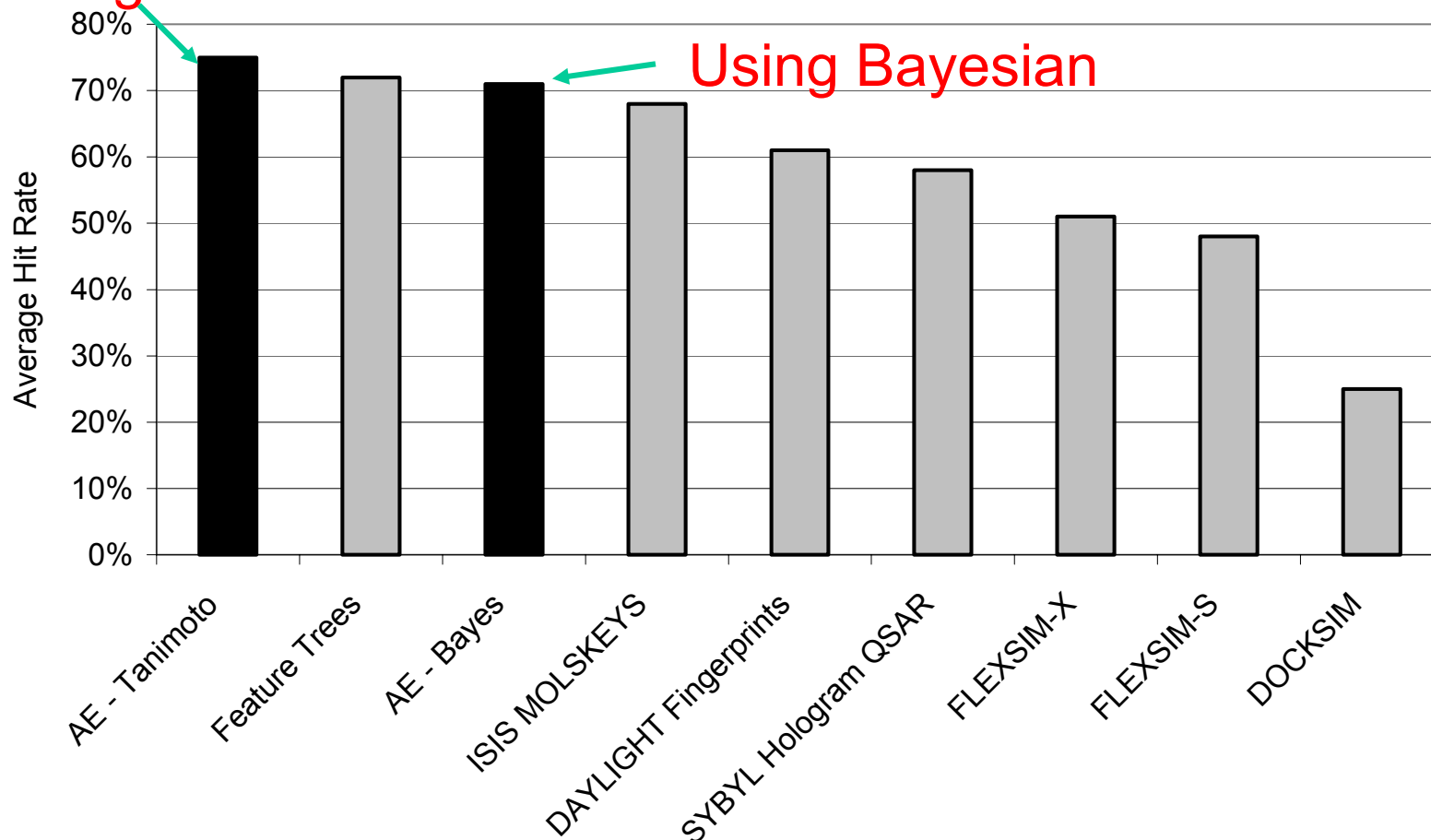
Application: lead discovery

- Database: MDL Drug Data Report (MDDR)
- 957 ligands selected from MDDR
 - 49 5HT3 Receptor antagonists,
 - 40 Angiotensin Converting Enzyme inhib. (ACE),
 - 111 HMG-Co-Reductase inhibitors (HMG),
 - 134 PAF antagonists and
 - 49 Thromboxane A2 antagonists (TXA2)
 - 574 “inactives”
- [Briem and Lessel, *Perspect Drug Discov Des* 2000, 20, 245-264.]
- Calculated Hit rate among ten nearest neighbours for each molecule

Comparison

Using Tanimoto Coefficient

Using Bayesian



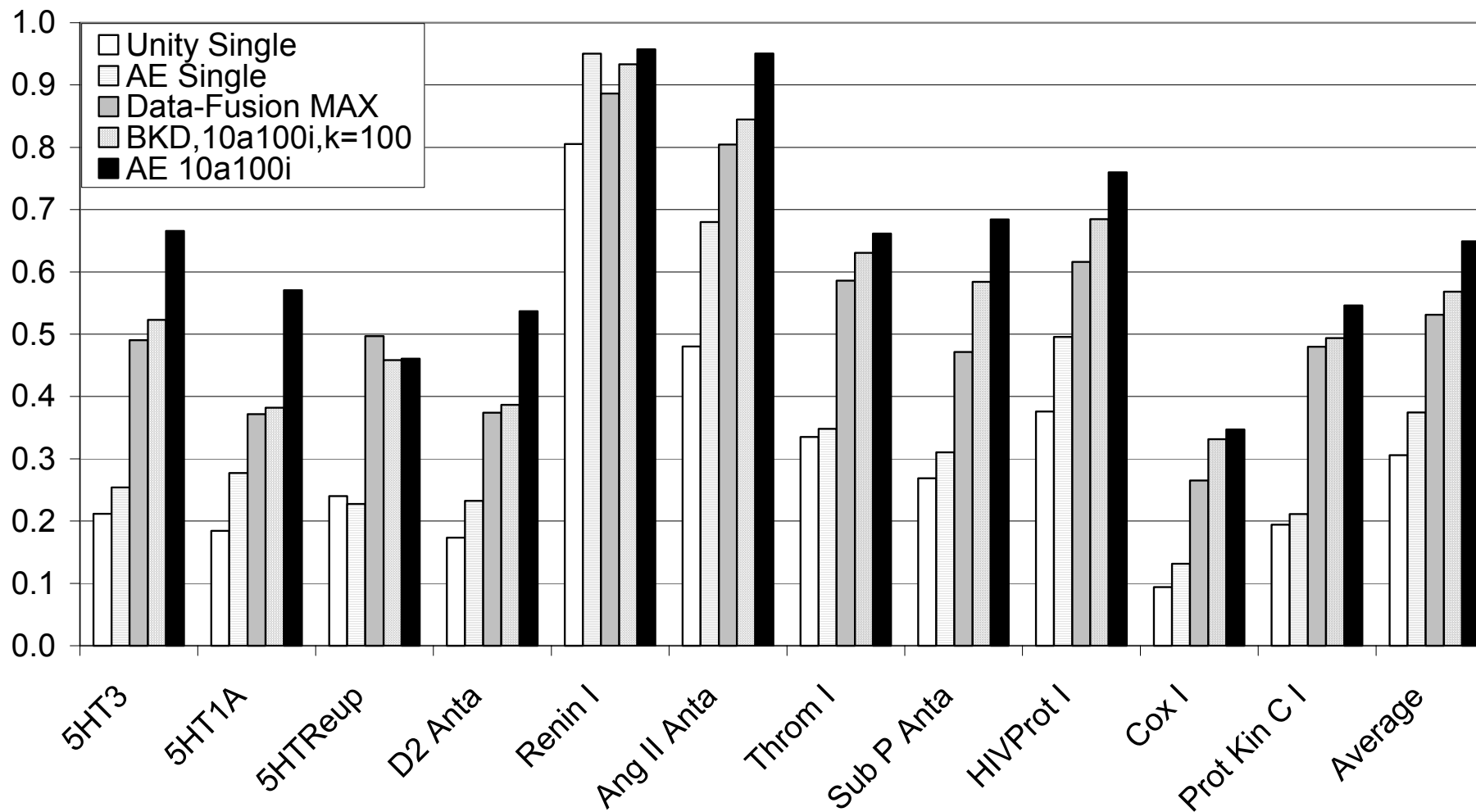
Briem and Lessel, Perspectives in Drug Discovery and Design
2000, 20, 245-264.

Comparison using Large Data Set *

- 102,000 structures from the MDDR
- 11 Sets of Active Compounds, ranging in size from 349 to 1246 entries – large and diverse data set
- Performance Measure: Fraction of Active Structures retrieved in Top 5% of sorted library
- Compared to Unity Fingerprints in Combination with Data Fusion (MAX) and Binary Kernel Discrimination
- In case of Binary Kernel Discrimination and the Bayes Classifier 10 actives and 100 inactives used for training

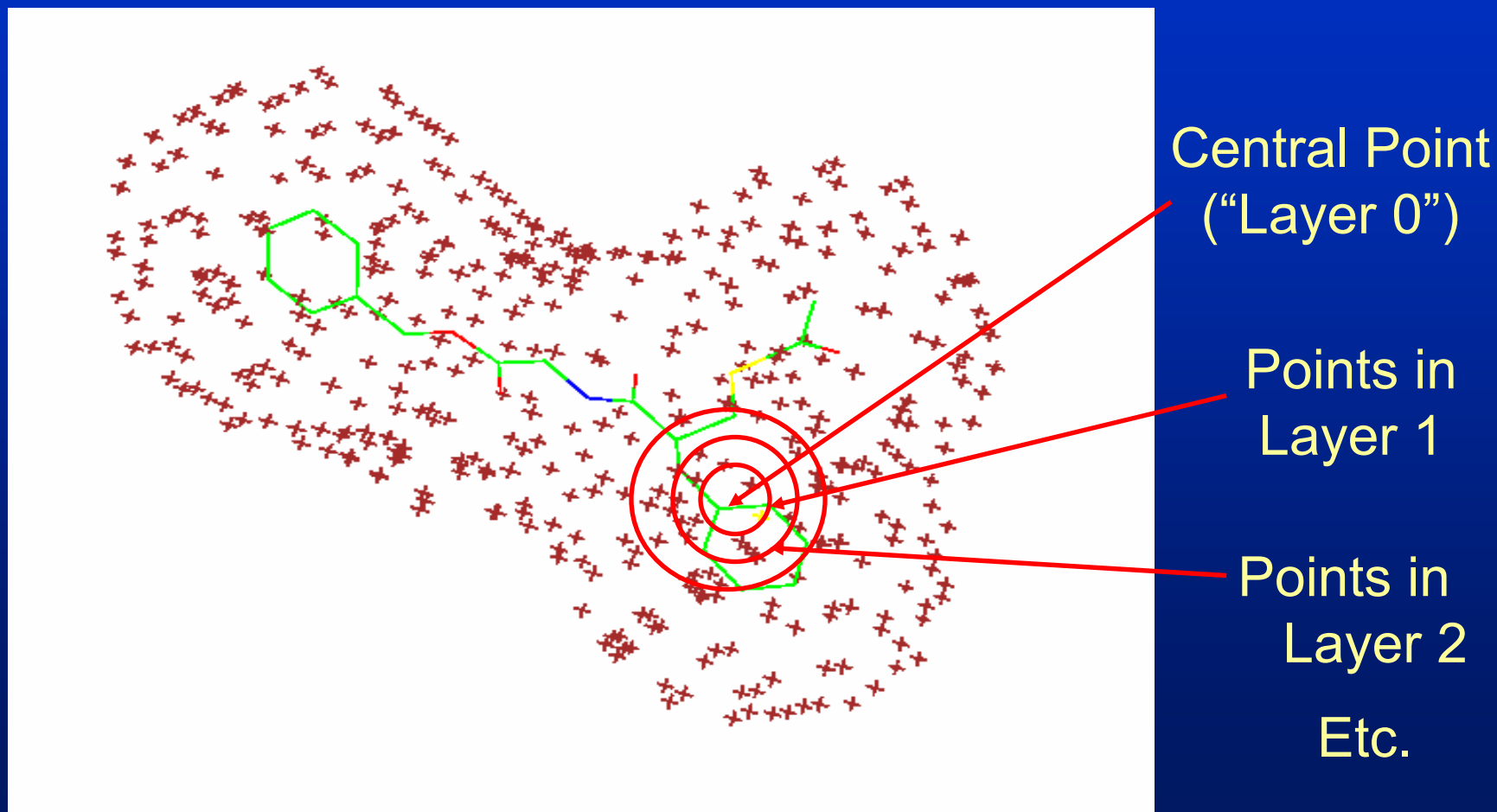
* Hert et al., J. Chem. Inf. Comput. Sci. 2004, 44, 1177 – 1185.

Comparison of Methods – MOLPRINT 2D

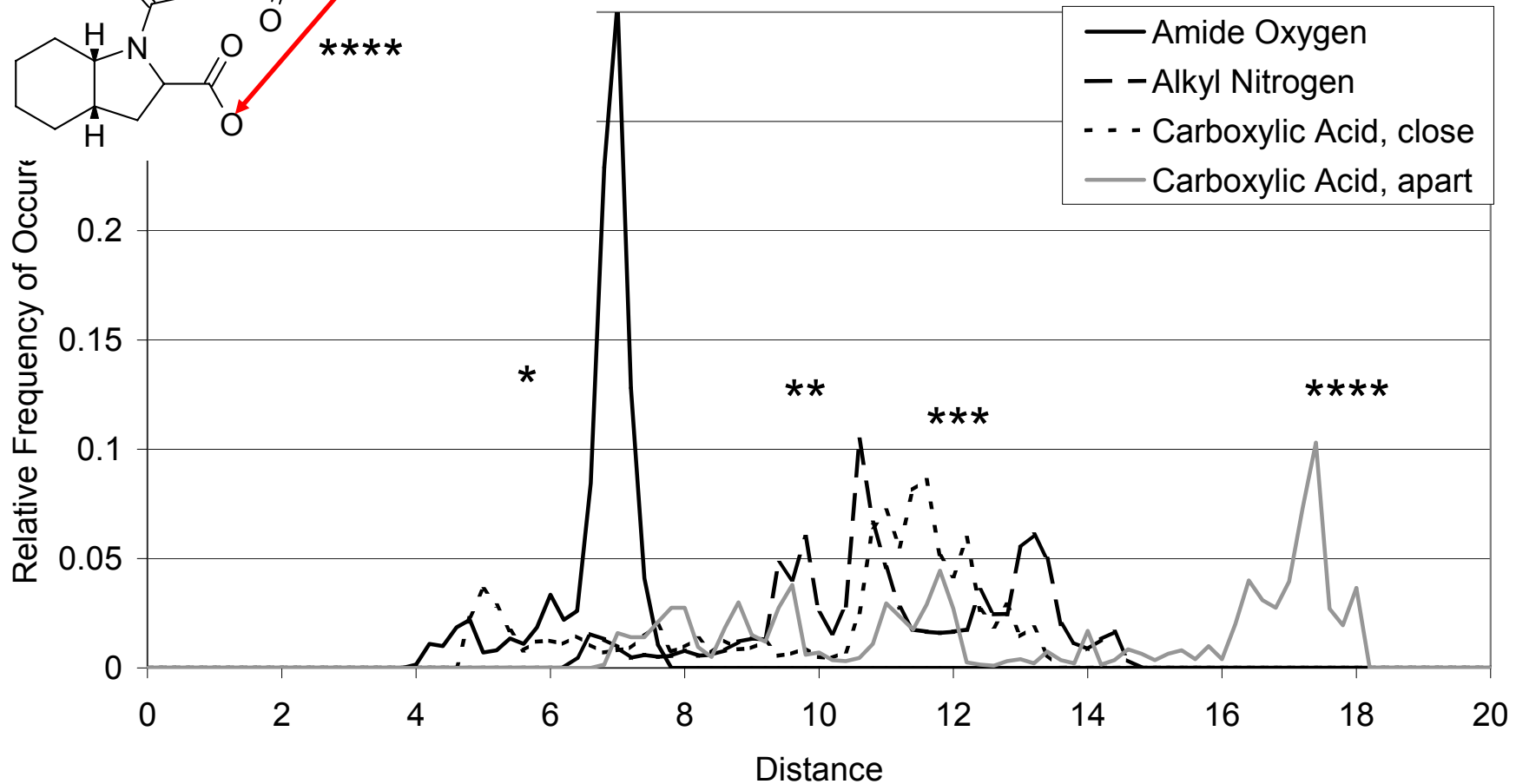
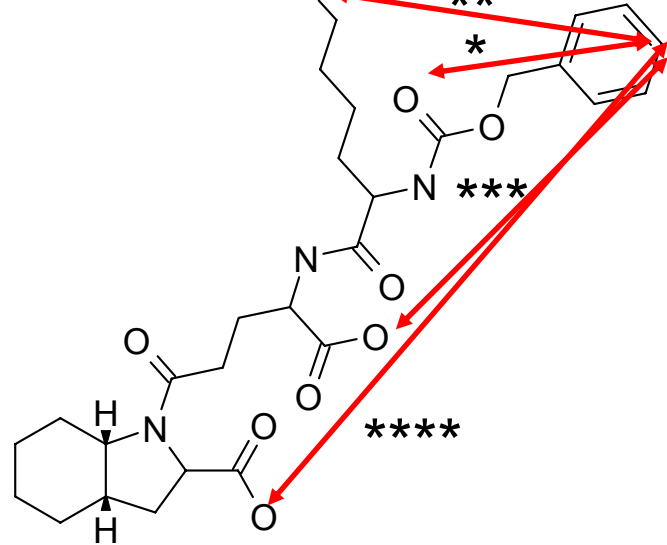


Bender, A., et al., *J. Chem. Inf. Comput. Sci.*, 2004, 44, 1708 – 1718. (Unity results from Hert, J., et al., *J. Chem. Inf. Comput. Sci.*, 2004, 44, 1177 – 1185.

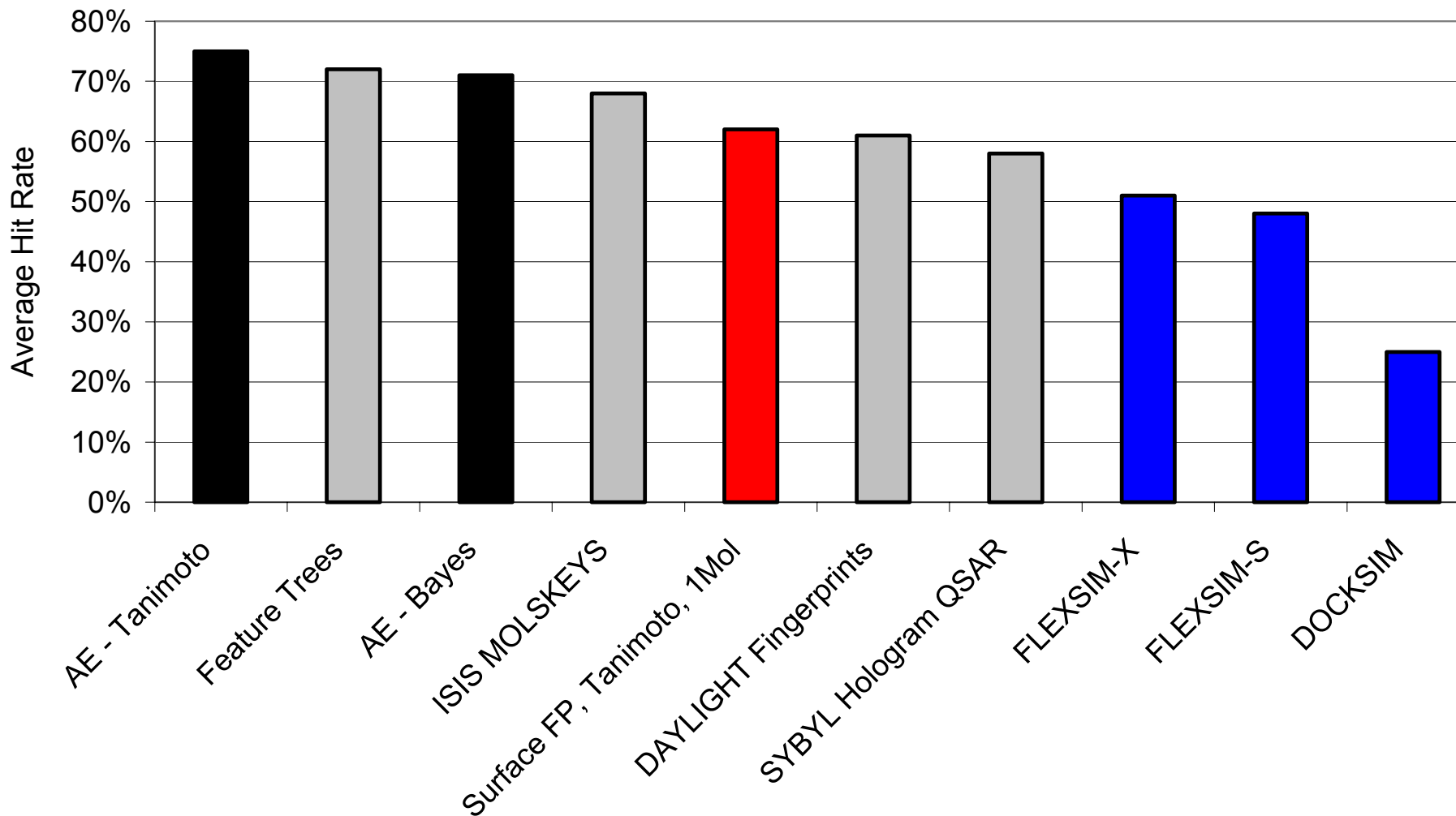
3D: Environment around a surface point: solvent accessible surface using *local* surface properties



The Conformational Problem



Overall Performance Comparable to 2D methods

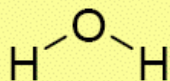


Disadvantages

- Multiple probes had to be employed to cover putative interactions sufficiently
- Force fields neglect polarization / back-polarization effects
- Force fields (usually) employ point charges, thus they don't capture directionality of some interactions such as hydrogen bonds
- -> Use more sophisticated QM method!

COSMO: Calculation of screening charges in ideal conductor

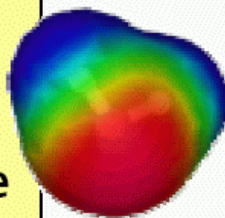
Start: Molecular Structure



Quantumchemical **COSMO** calculation



Ideally screened molecule:
Energy and screening
charge distribution on
molecular **COSMO** -surface



Database
of
COSMO -files

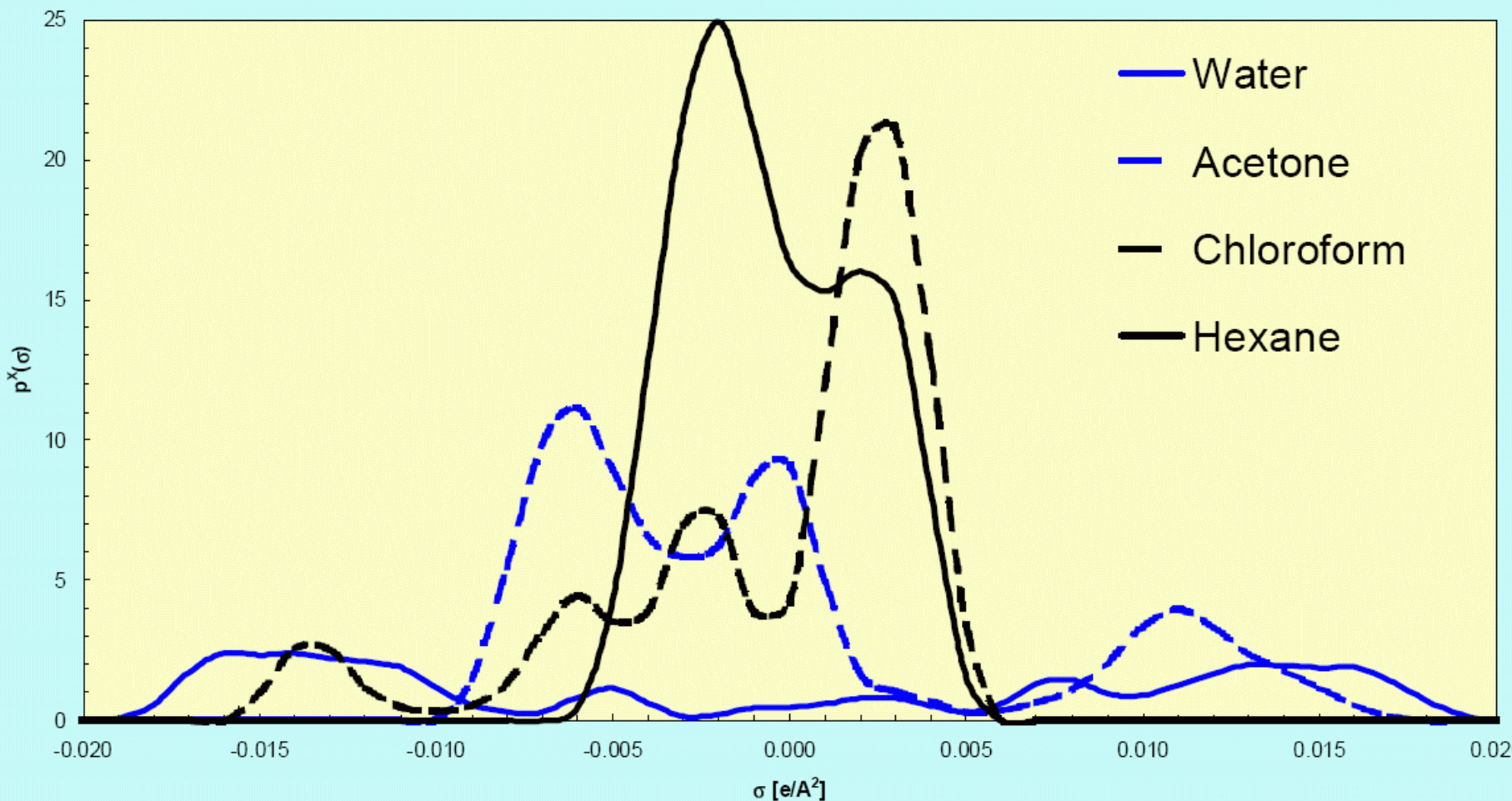
COSMO

Why COSMO-RS Properties?

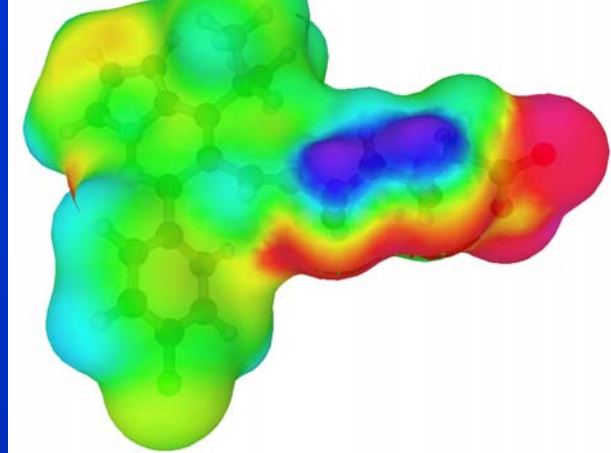
- Interactions derived from first principles *on single scale*
- Surface Properties/Directionality kept which are important for (putative) interactions
- Employs solvent model, polarization / back-polarization
- Classification in agreement with chemical intuition (e.g. =O of ester, but not –O- is H-bond acceptor)
- Gives directionality of H-acceptor lobes (unlike most force fields; exceptions are e.g. the XED force field by Andy Vinter / Cresset)
- Inaccessible atoms not used (no accessible surface)
- Secondary effects captured which are not accounted for by atom-typing

COSMO σ -Profile

σ -profiles of compounds

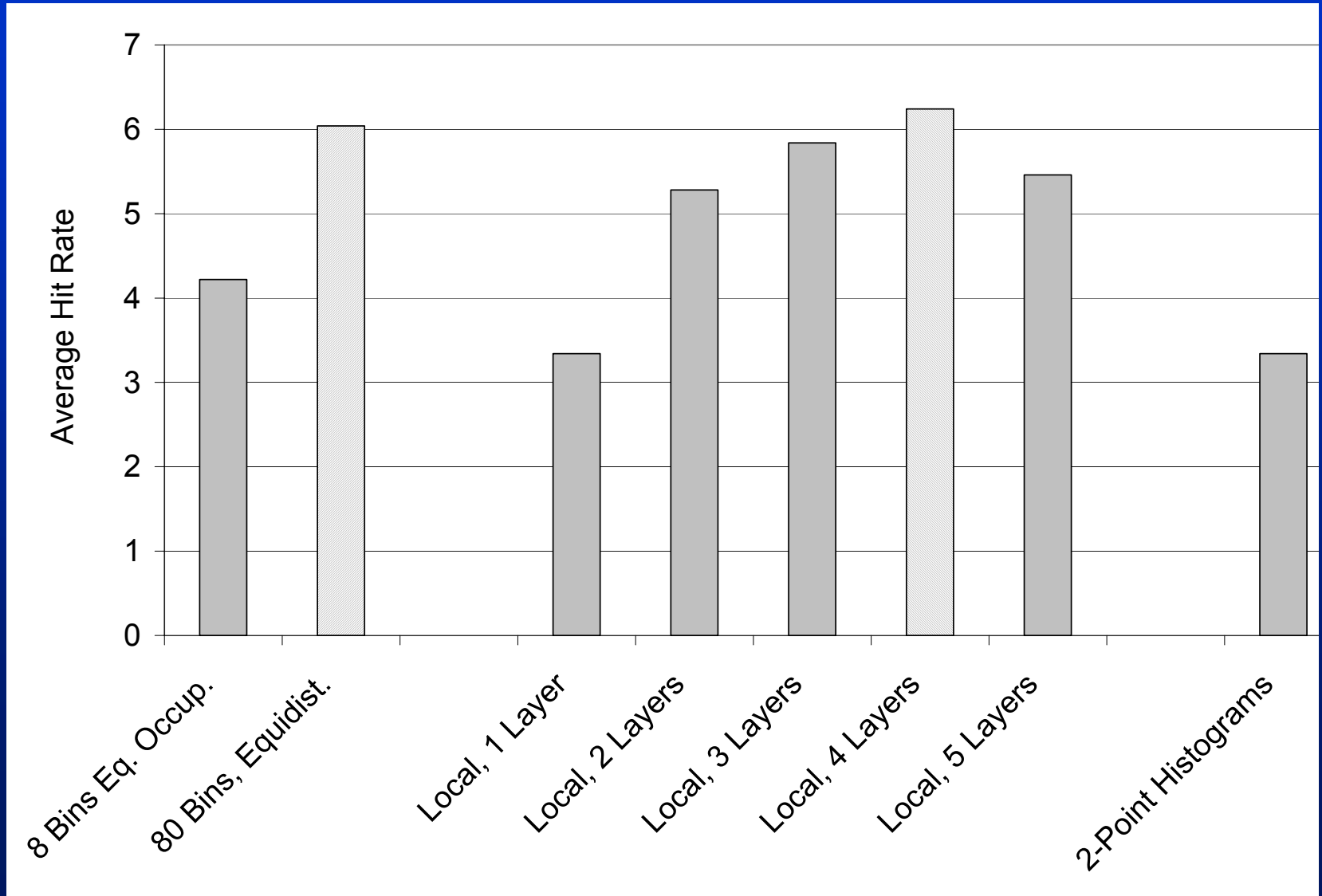


A HMG-CoA Reductase Inhibitor



- Statin binding to HMG-CoA reductase involves charge interactions of a carboxylic acid group and hydrogen bond donor/acceptor functions to the pyruvate binding site
- In addition large lipophilic groups of the ligand is required which binds to a floppy lipophilic pocket of the target protein.
- Features can be well distinguished from σ screening charges
- Carboxylate is shown to the right (purple), hydrogen bond acceptor functions beneath side chain (red)
- Hydrogen bond donor functions point towards viewer (blue) while the lipophilic bulk of the structure is given in green

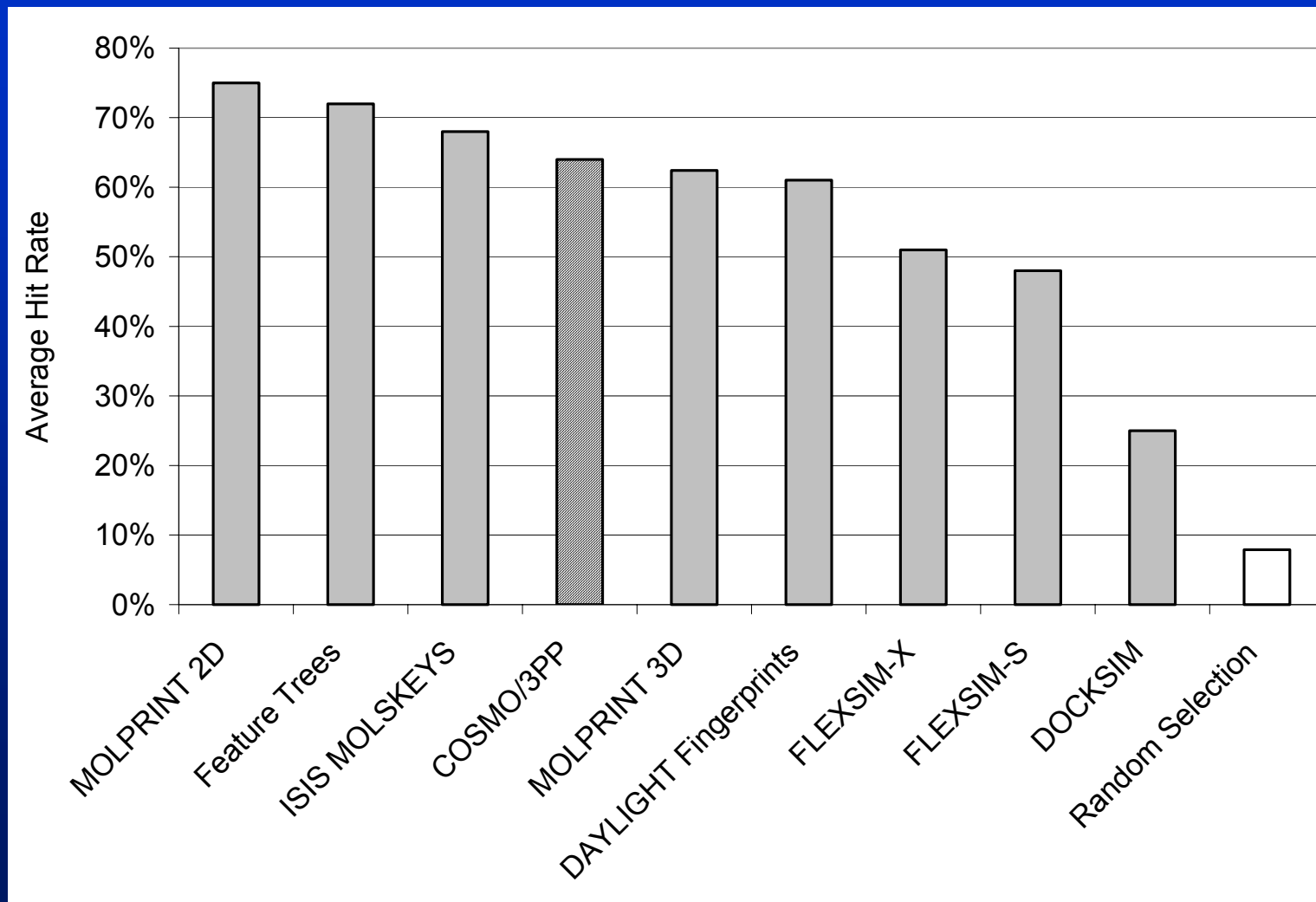
Encodings Investigated



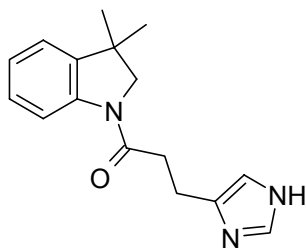
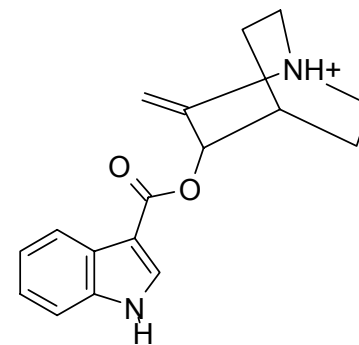
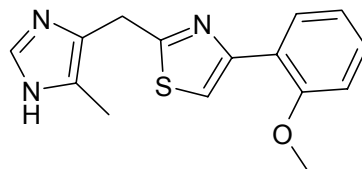
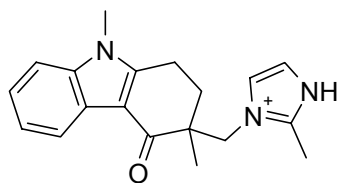
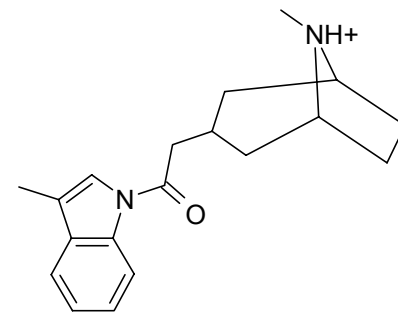
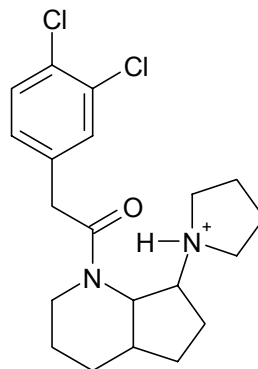
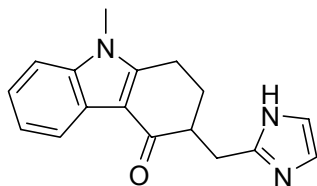
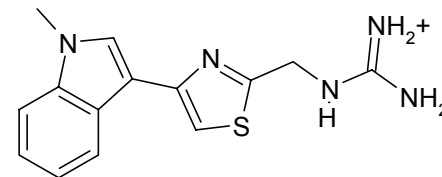
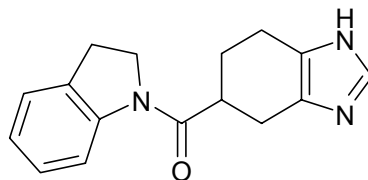
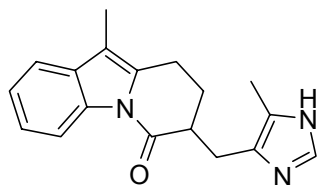
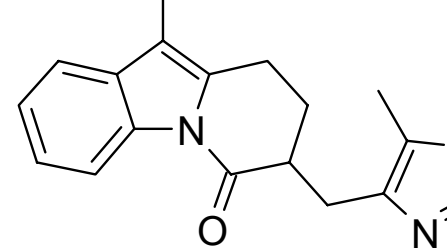
3-Point Pharmacophores: “Back to the roots”

- COSMO screening charge densities σ encoded as atom-based three-point pharmacophores (3PP)
- Average σ -values calculated for each heavy atom
- Average σ charges $> 0.014 \text{ e}/\text{\AA}^2$ classified as bearing strongly negative partial charge (type N); $0.014 \text{ e}/\text{\AA}^2 > \sigma > 0.009 \text{ e}/\text{\AA}^2$ as hydrogen-bond donors (D)
- Negative σ charges associated with atoms showing strongly positive partial charge (P) at $\sigma < -0.014 \text{ e}/\text{\AA}^2$; hydrogen-bond acceptors (A) at $-0.014 \text{ e}/\text{\AA}^2 > \sigma > -0.009 \text{ e}/\text{\AA}^2$
- Intermediate screening charge densities are lipophilic atoms (L).
- Eight bins ($>2, 3.5, 5, 6.5, 8, 9.5, 11, 13$ and 15 \AA).
- Triangles rotated to a unique orientation, counts kept
- Comparison via Tanimoto-like similarity coefficient dividing number of matching features by total number of features present (takes partially account of size)

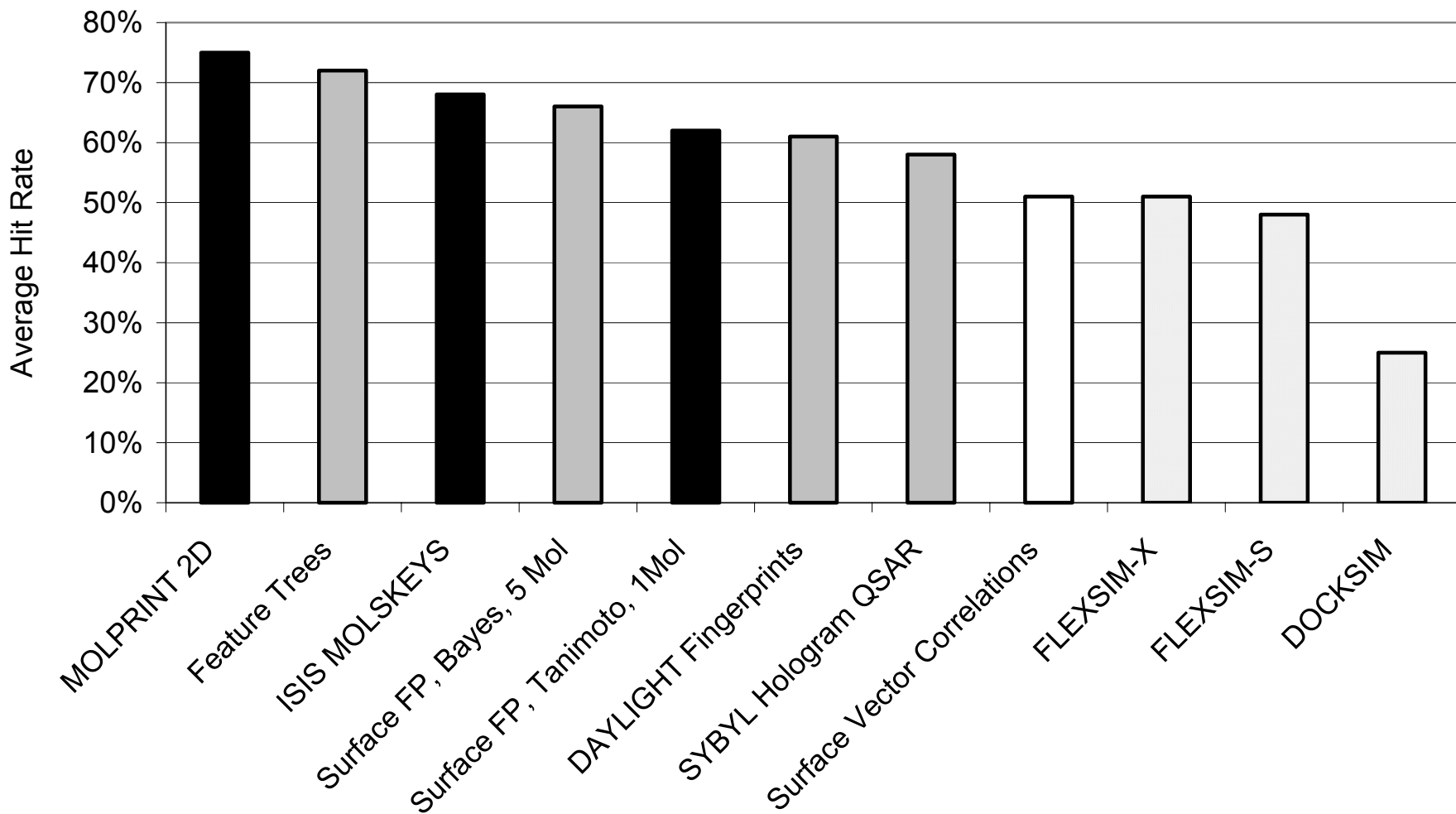
Comparison to other methods



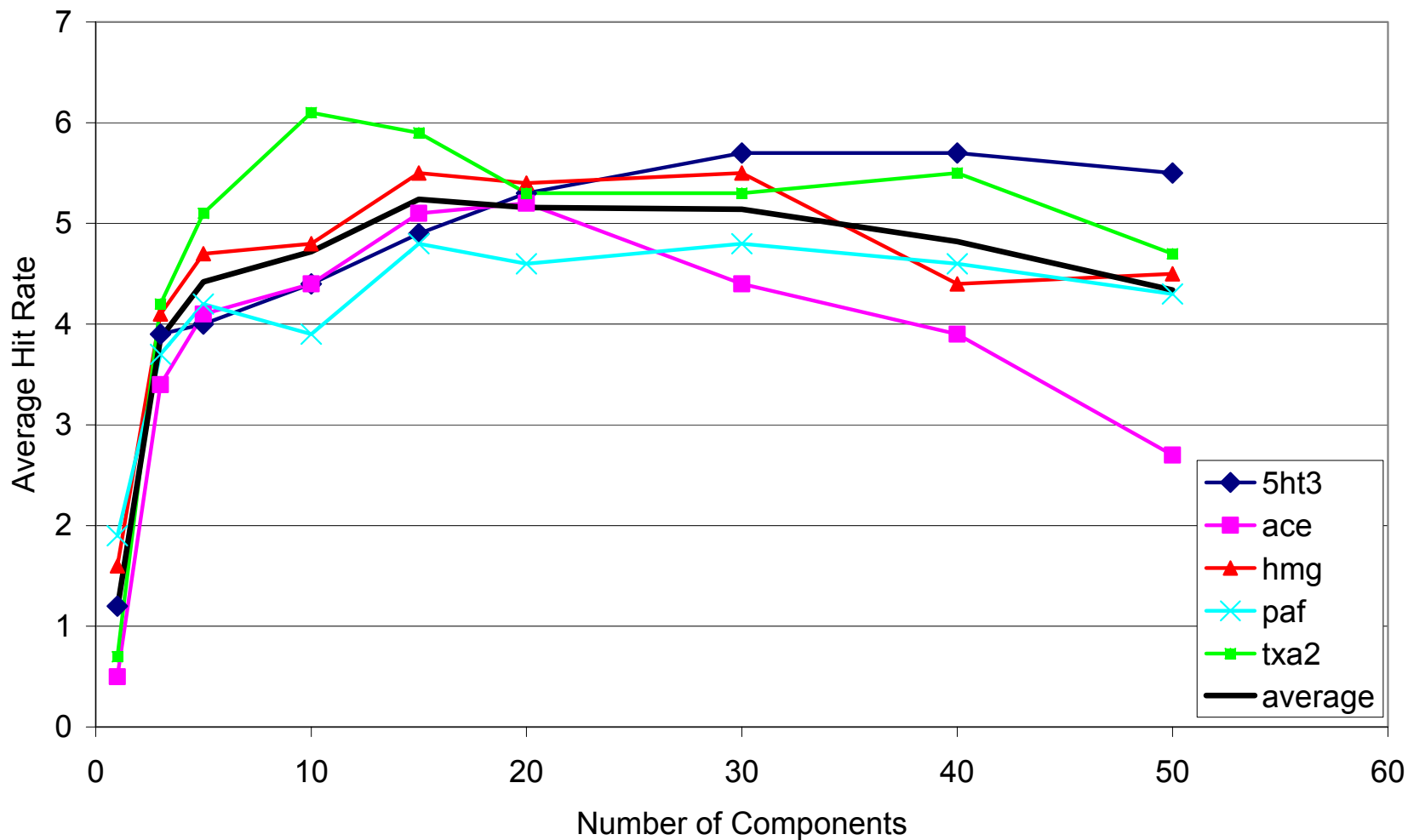
Scaffolds Found



The Difficulty of Getting High



PCA of 3-Point Pharmacophores



Some Thoughts on Performance Assessment of Virtual Screening Methods

- Many of the published databases in comparative studies were taken from current drug databases
- Examples: Briem/Lessel dataset; Hert/Willett dataset
- Two major disadvantages:
 - Large number of analogue compounds
 - No *inactive* information is contained in the database
- MDDR is synthetic dataset that was partly generated using similarity, analogue considerations – so you partly only exploit this composition
- Effects contribute to unrealistic performance measures

MDDR-Derived Datasets for Performance Assessment

- (Very) incomplete data matrix, often only single activities are reported for compounds
- Thus, activities may well be present, but just not yet be detected in assays (or *in vivo* as side effects etc.)
- Unknown positives lead to false-positives in rankings and blur performance measures
- Complete data matrices eliminate this problem, e.g. the Cerep (Bioprint) database, Boehringer Kinase dataset
- Then e.g. validation according to the 'Neighbourhood Principle' can be performed

Banal features

- Idea: Use some simple, non-structural ligand features for VS to give estimate of “added value” of “real” descriptors
- Docking known to prefer larger ligands – here ligand-based VS
- How good performs MW? Number of atoms? Count Vectors of Element Atom Types?
- Compare to circular fingerprints (MOLPRINT 2D) – gave “best” retrieval rates on large retrospective dataset, comparable to Scitegic ECFPs (Hert, J., *et al.*, *Org. Biomol. Chem.* 2004, 2, 3256.)
- In current issue of JCIM

Properties, Distance Measure

- Simple properties employed as descriptors: #atoms, MW, “Atom count vectors”
- “Atom count vectors” were calculated using the total number of atoms, the number of heavy atoms and the numbers of Boron, Bromine, Carbon, Chlorine, Fluorine, Iodine, Nitrogen, Oxygen, Phosphorus and Sulphur atoms.
- No structural information at all was contained in this 12-integer “fingerprint” representation
- Euclidean distance employed as similarity/distance measure

Previous Work

- Livingstone¹: “Overall molecular parameters which are able to discriminate between compounds showing different physicochemical or biological behavior. E.g., blood-brain barrier penetration is closely related to logP, and electron density on a nitrogen atom in the HOMO of a set of aniline mustards and tumor inhibition can be related in a simple linear fashion. “
- Pan²: “Heavier molecules are favored by docking algorithms due to the simple fact that on average more atom-atom interactions are present which contribute to the predicted binding energy. As a remedy normalization of the binding energy with respect to the number of heavy atoms per molecule was suggested.”

¹ Livingstone, D. J. The characterization of chemical structures using molecular properties. A survey. *J. Chem. Inf. Comput. Sci.* 2000, 40, 195-209.

² Pan, Y. P., *et al.*, Consideration of molecular weight during compound selection in virtual target-based database screening. *J. Chem. Inf. Comput. Sci.* 2003, 43, 267-272.

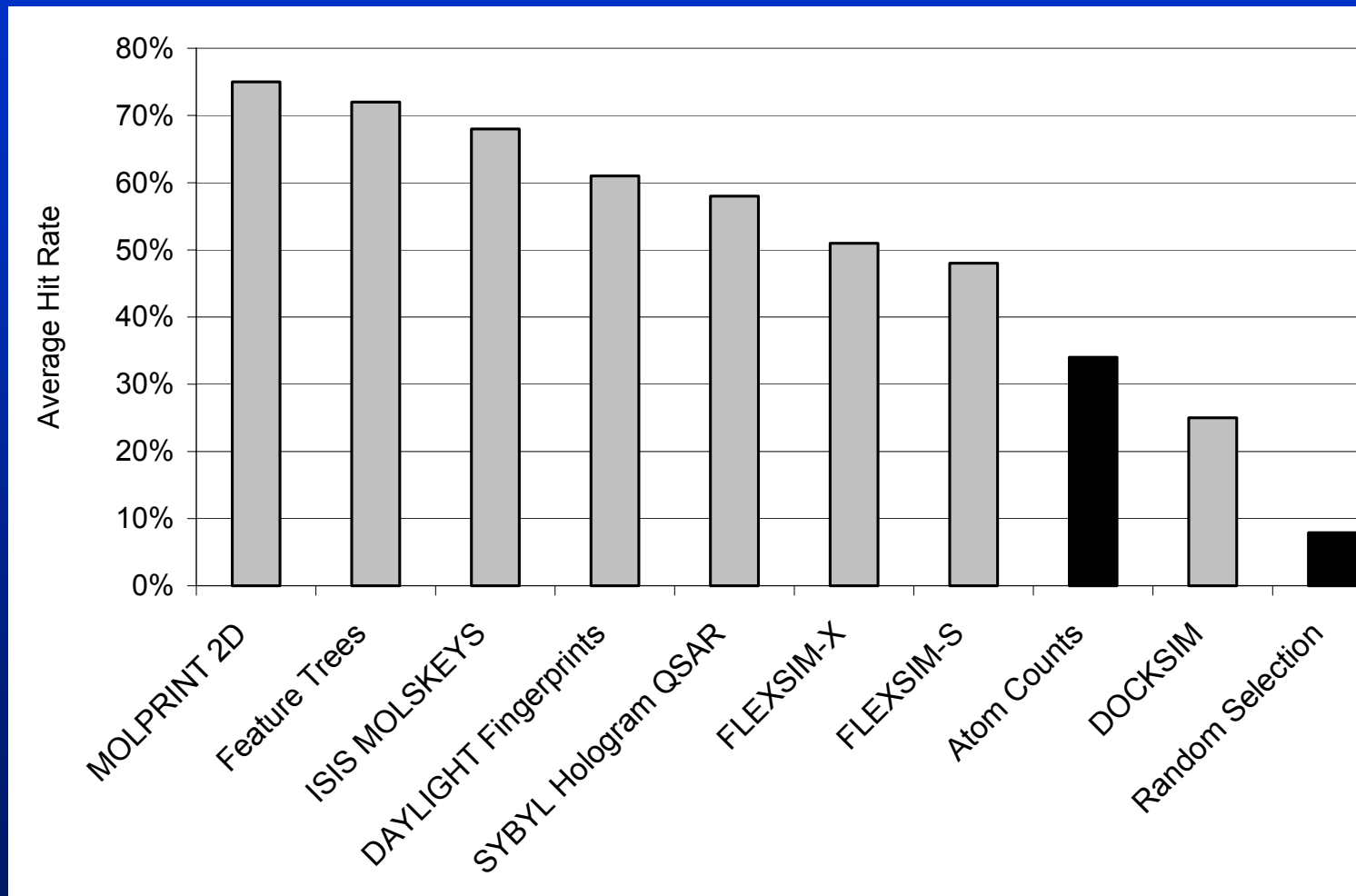
Previous Work (2)

- Gillet³: “Bioactivity profiles (BPs) include the number of H-bond donors and acceptors, MW, a kappa shape index and the numbers of rotatable bonds and aromatic rings. BPs found application in distinguishing molecules from the World Drug Index and those from the SPRESI database (which were assumed to be inactive); using single features such as the number of H-bond donors alone enrichments of up to 4.6 were found in identifying WDI molecules in a merged dataset. “
- Verdonk⁴: “Considering heavy atom counts alone on two hypothetical libraries of active compounds, which are either on average much heavier or much lighter than the whole library, was shown to give considerable enrichments. “

³ Gillet, V. J.; Willett, P.; Bradshaw, J. Identification of biological activity profiles using substructural analysis and genetic algorithms. *J. Chem. Inf. Comput. Sci.* 1998, 38, 165-179.

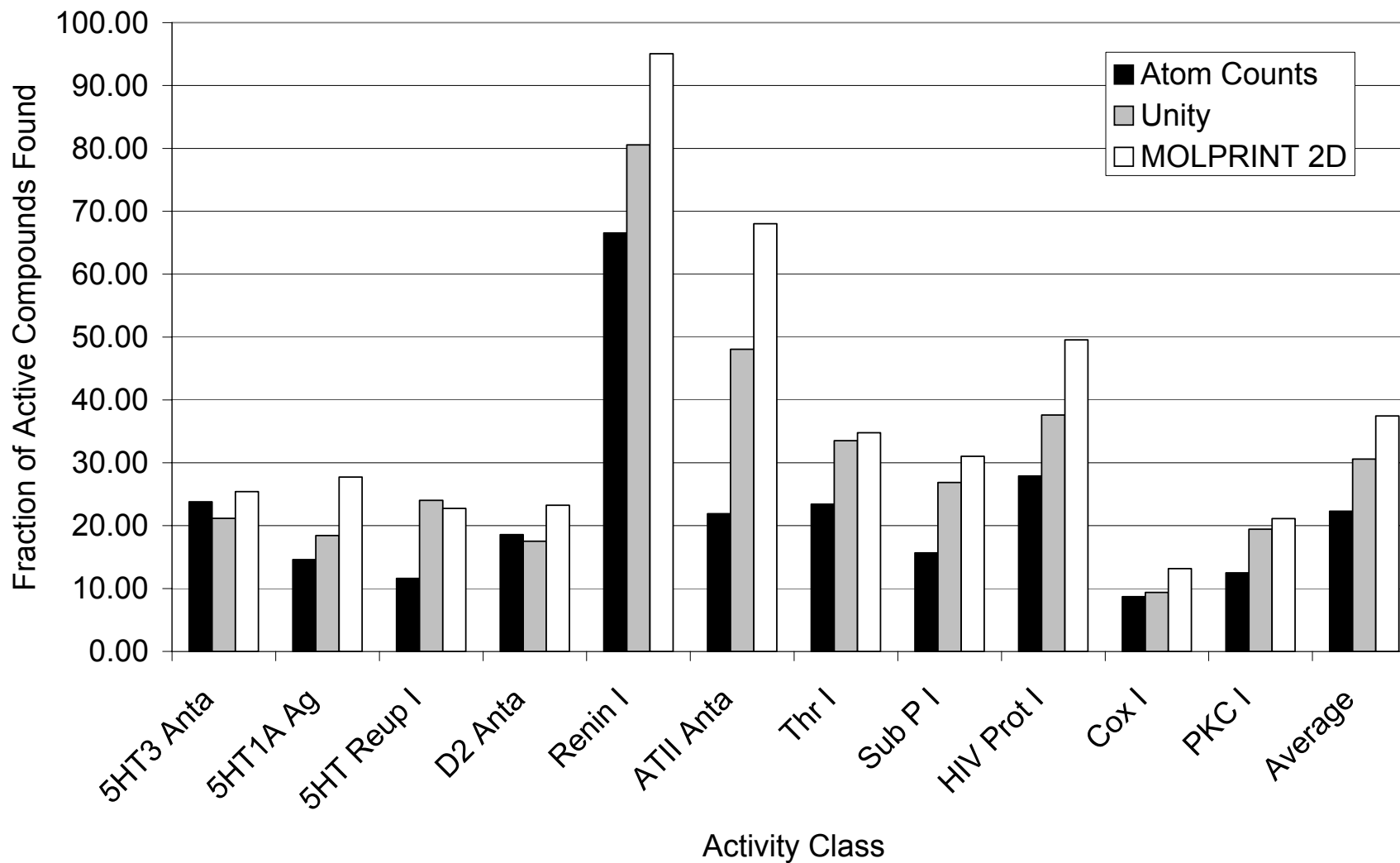
⁴ Verdonk, M. L., *et al.*, Virtual screening using protein-ligand docking: Avoiding artificial enrichment. *J. Chem. Inf. Comput. Sci.* 2004, 44, 793-806.

Briem Dataset: 4-fold "Enrichment"

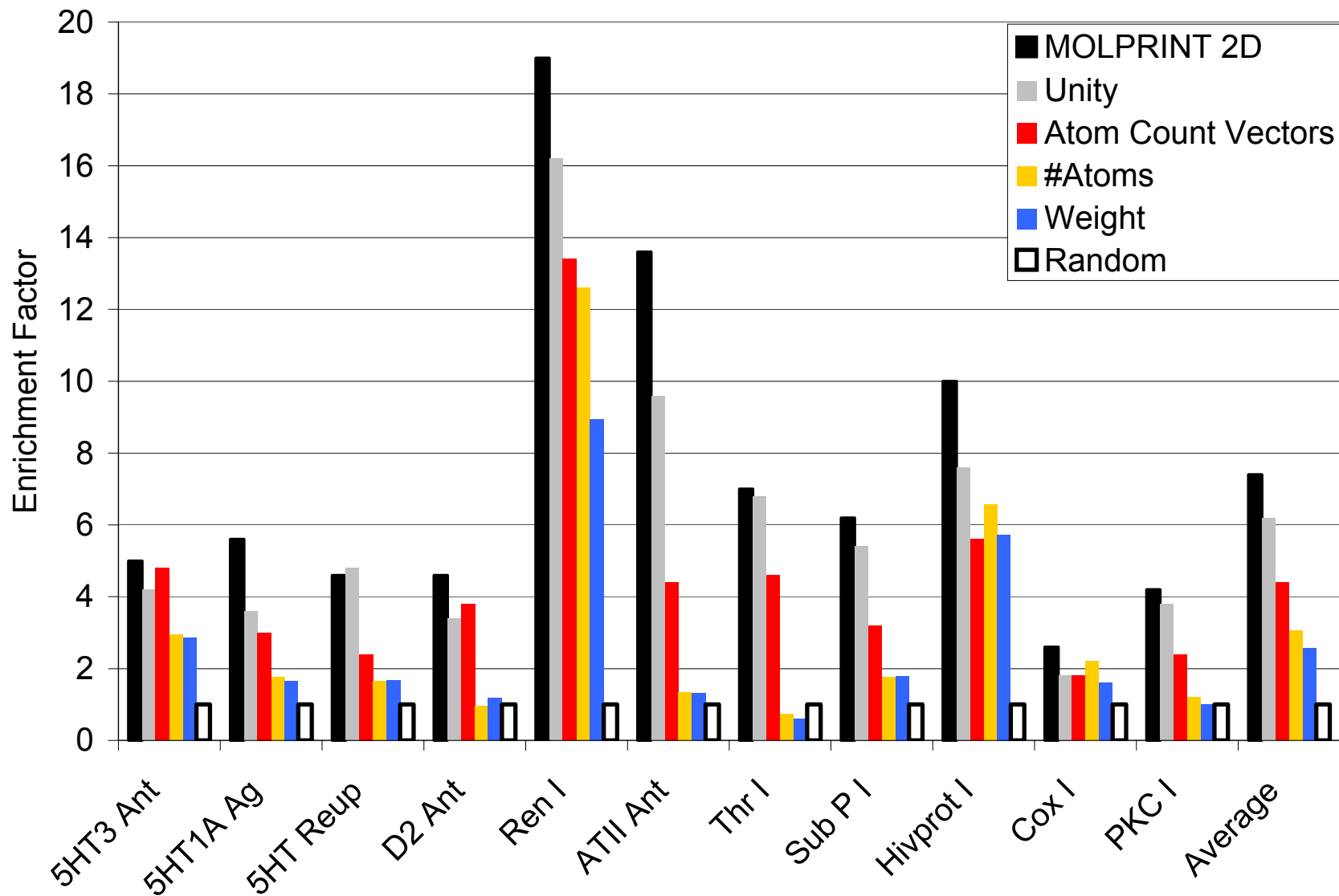


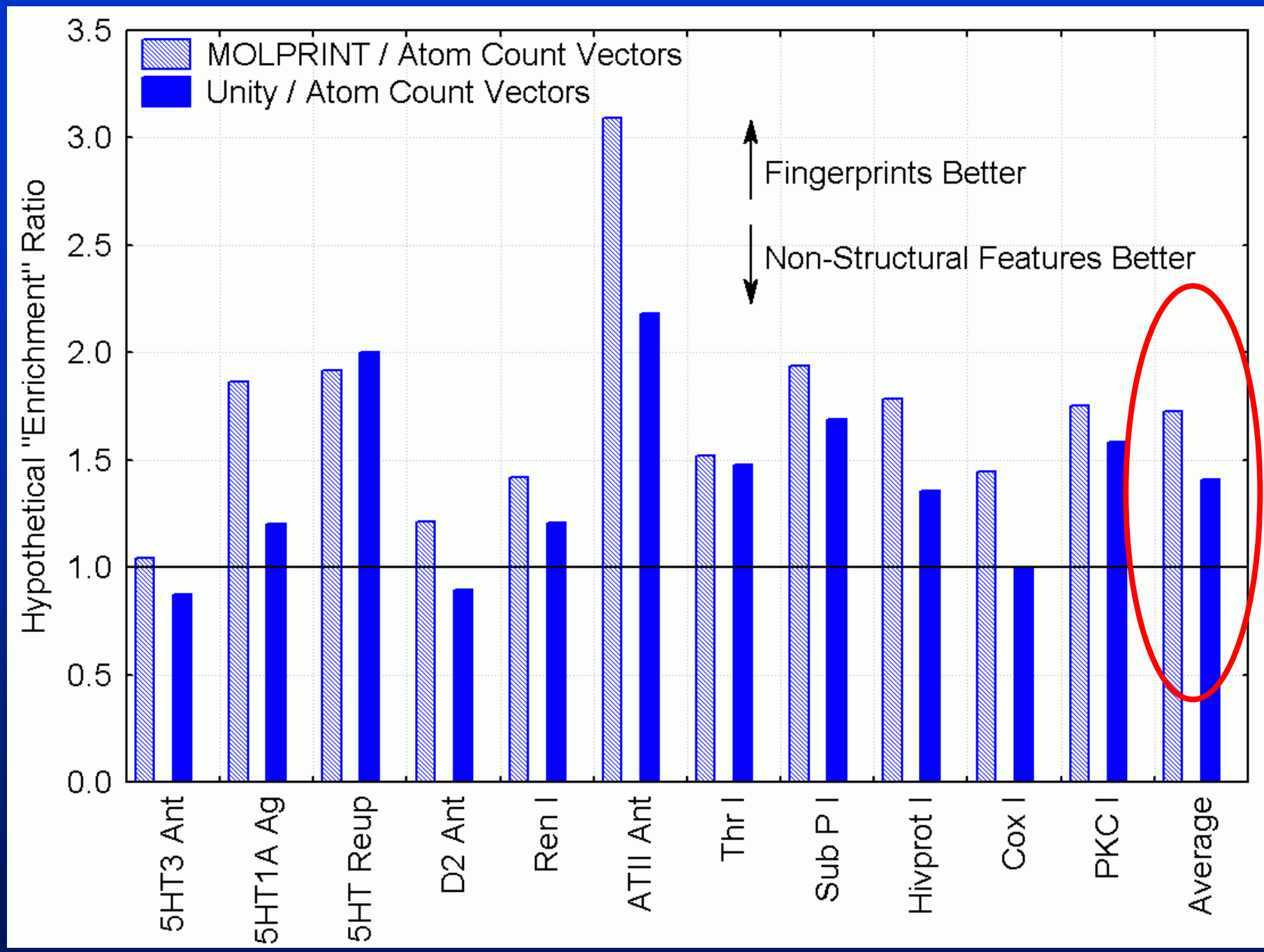
Hert Dataset

Hit Rates via Atom Counts, Unity Fingerprints and MOLPRINT 2D



Molecular Weight / #Atoms is not enough





Conclusions

- Current descriptors don't capture as much information as one would like them to
and/or
- MDDR-based (retrospective) virtual screening libraries are no suitable performance measure
and/or
- There exists a particular relation between atom count vectors and activity (can partly be explained by the different number and type of interactions)
- Be careful when evaluating virtual screening performance and put it in relation to complexity, expense

So – How do we fare today?

- 1. Similarity of molecules is both context (e.g. receptor) and location-dependent
- 2. Current descriptors treat molecules as static entities – but even by definition receptor binding involves dynamical motions of the protein
- 3. No agreement exists which kind of interaction of the ligand with the receptor actually causes (for example agonistic or antagonistic) action: Is it occupancy? Is it on-off rates? Or some completely different property?
- 4. Similarity (in the context of bioactivity) is a clearly non-linear problem, as illustrated by the recent success of for example k-NN QSAR. Descriptors don't capture this.

So – How do we fare today?

- 5. Multiple binding modes and even multiple binding sites trash the concept of ‘similar molecules = similar effect’
- 6. Protein-ligand binding is the result of the difference of two large numbers and thus a delicate equilibrium position – simple treatment such as ‘there is a hydrogen-bond donor in molecule A and one in molecule B’ so they are similar neglects subtle differences in solvation and desolvation, rendering one interaction favourable while the other is not
- 7. Is binding really an equilibrium process? Entropy consumption on even nanometre scales is known.

Summary

- 2D Method: Finds lots of active molecules – but they are similar to what is known already
- 3D Method: Find less active compounds – but enables discovery of *new chemotypes*
- Similarity searching using screening charges derived from first principles shows good performance and possesses a sound theoretical basis
- Employ full-matrix data for performance assessments with appropriate performance measures (don't worry, I will do that anyway)
- Current descriptors do not contain as much information as one (at least I) suspected

Acknowledgements

- Robert C Glen (Unilever Centre, Cambridge, UK)
- Hamse Y Mussa (Unilever Centre, Cambridge, UK)
- Andreas Klamt, Karin Wichmann (COSMOlogic, Germany)
- Michael Thormann (Morphochem, Germany)

Software

- GRID, CACTVS, gOpenMol; many, many others

Funding

- Bill Gates, Unilever, Tripos