Molecular Informatics Tools for Data Analysis and Discovery

1. **Some examples of coping with Molecular informatics data**
   - **legacy data (accuracy)**

2. **Database searching using a similarity approach**
   - **fingerprints in 2D and 3D**

ChemSource 2005

Robert Glen

UNIVERSITY OF CAMBRIDGE

Unilever Cambridge
Centre for Molecular Informatics

The scale of the data in Molecular Informatics…

Biological data

- **3,488,108,873 nucleotide bases**
- over 800 organisms
- 25442 protein x-ray cryst[illegible]tures
- 12 million citations in [illegible]edline
- 40 main stream databa[illegible] fr[illegible]m EBI
- Ensemble : 24 million gene p[illegible]edictions

Chemical data

- **60,475,000 chemical substances**
- 3,700,000 chemical reactions
- 613,000 available reagents
- [illegible]n has 600,000 reaction abstract[illegible]
- [illegible]00 organic x-ray structures
- Combichem libraries of **Billions** of compounds

Patents

- **European patents – 150,000,000 pages**
- 150,000 applications/year
- 400,000 chemical patent hyperstructures
- Over 100 countries in patent cooperation treaty (PCT)

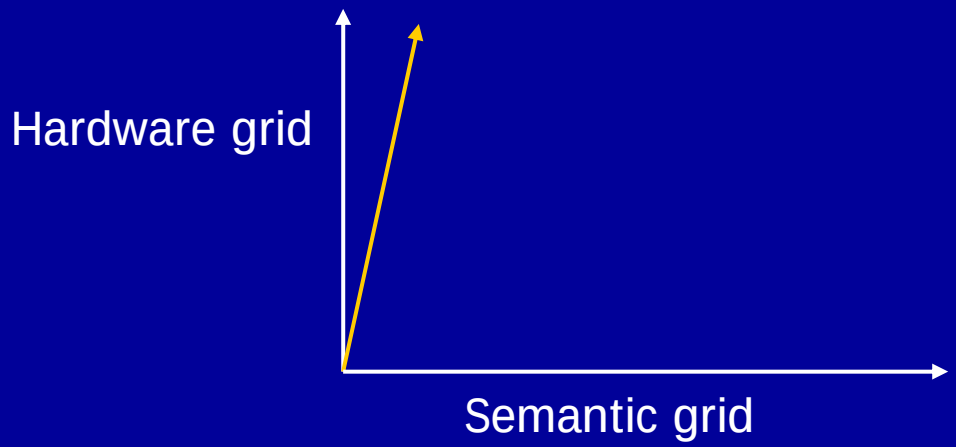
The BIG 'connection' => innovation

This implies a clear need to make maximum use of this 'sea' of data...

Robert C. Glen and S. Aldridge. *Developing tools and standards in molecular informatics*. Chem. Comm. 2002. pp2745-2747

Enabling technologies...

- The grid
  - Information and processing as accessible as electricity – plug and play
- The semantic grid
  - Language and knowledge to access the grid



Using information from Legacy data

- 100 years of chemistry – in books
- 20 years of data destruction – PDF
- Computational studies seldom available for analysis and aggregation

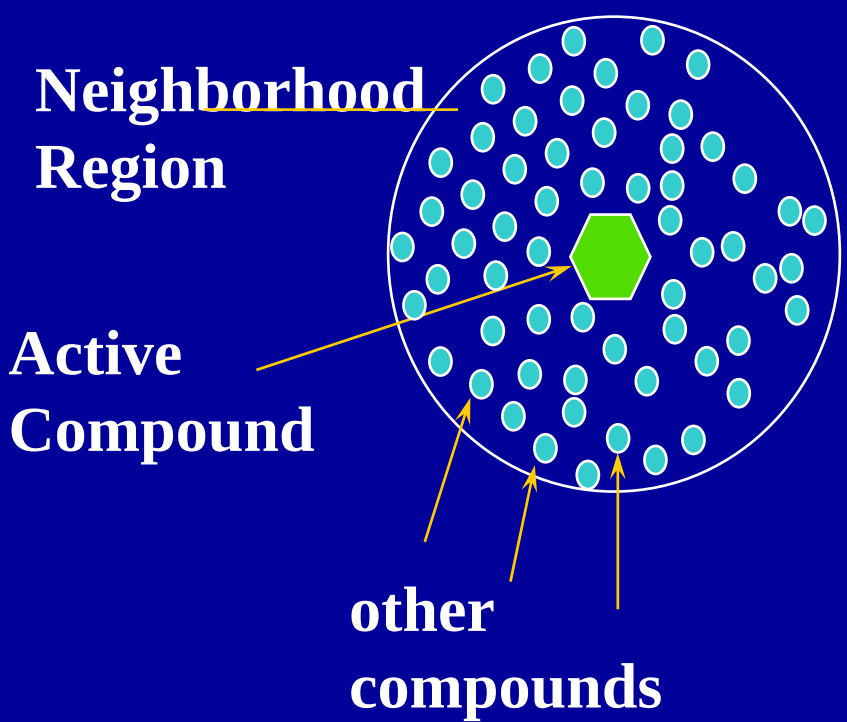
- Need data to be abstracted to a computer readable form
- Need data to be standardised and checked
- Need data to be available to ‘robots’ for processing, checking, analysis – and to ‘talk’ to other robots

RSC/UCC markup project

- Uses CML, natural language processing, knowledge of chemistry
- An authoring tool
- A data checker
- Data abstraction from chemistry papers -> computer readable database

Experimental data checker: better information for organic chemists S. E. Adams, J. M. Goodman, R. J. Kidd, A. D. McNaught, P. Murray-Rust, F. R. Norton, J. A. Townsend and C. A. Waudby *Org. Biomol. Chem.* 2004, **2**, 3067-3070. Highlighted in *Chemical Science* 2004, **1**, C33.
Chemical documents: machine understanding and automated information extraction J. A. Townsend, S. E. Adams, C. A. Waudby, V. K. de Souza, J. M. Goodman and P. Murray-Rust *Org. Biomol. Chem.* 2004, **2**, 3294-3200

Database searching Using a similarity approach



If the molecular descriptors are valid ...

the activity of a Compound is shared by *most* other compounds within its Neighborhood Region

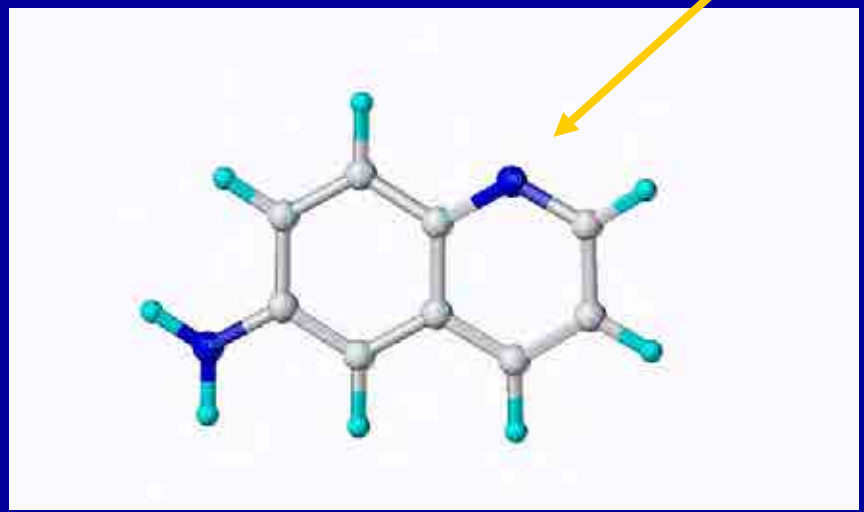
i.e. neighbors of a bioactive compound have a higher probability of behaving in a 'similar' bioactive way

Molecular similarity: a key technique in molecular informatics. Organic and Biomolecular Chemistry perspective article. R. C. Glen and A. Bender, *Org. Biomol. Chem.* 2004, **2**, 3204 - 3218.

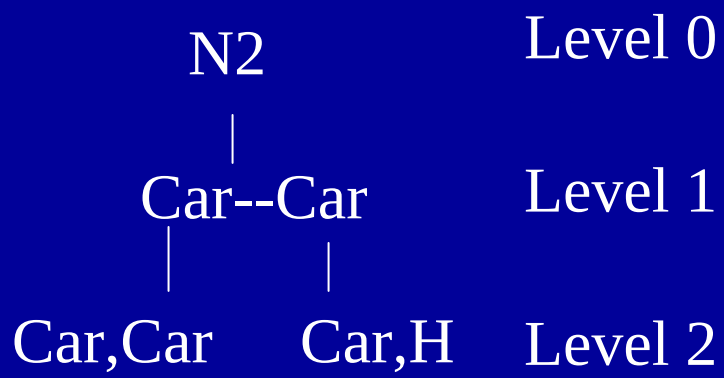
The descriptors: Similar to the environment around an ionizable center (atom environments) used previously (Xing, Clark and Glen)

- E.g. 6-aminoquinoline

Measured 5.7
predicted 5.4



Start with interesting atom
find connections
find connections to connections
create a tree down to 5 levels
'bin' the atom types for each level
create a 'fingerprint' for this atom



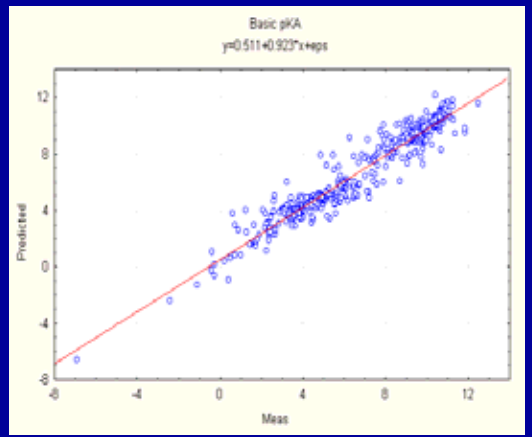
String contains a bin for each required atom type at each level,
the number of atom types is accumulated to form the string - 56 bins

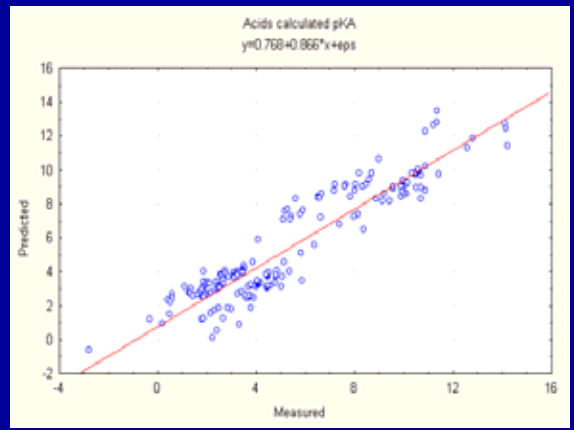
Method

- Tabulate many reliable pKa measurements
- Describe the environment around ionizable centers
- Use partial least squares to create a predictive model
- Test model with cross validation

Using the data

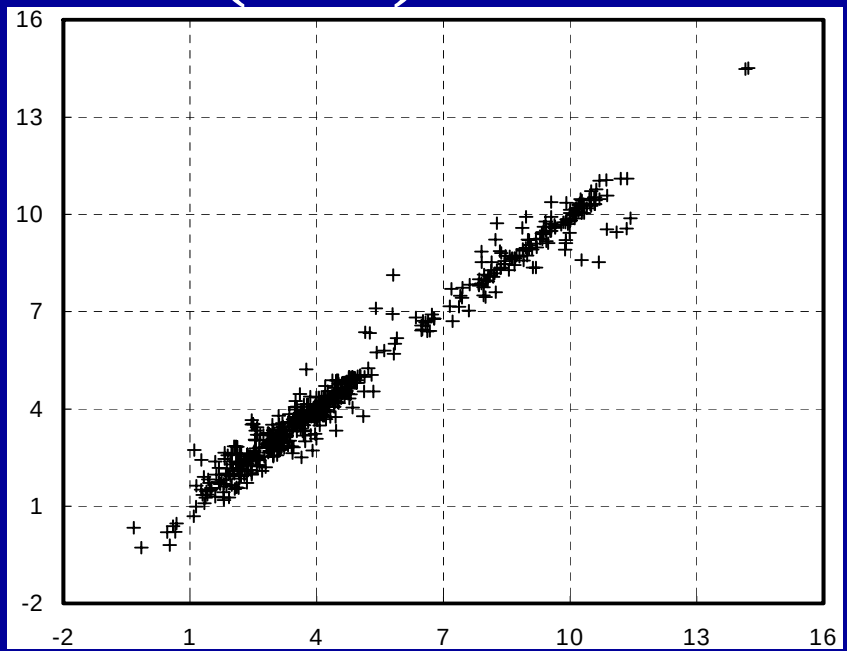
- 56 bins used to cover all the possibilities
- Used pls (partial least squares) to create a model
- $pK_a = pK_c^0 + \Sigma a_i x_i + \Sigma g_j y_j + \Sigma q_k z_k$ ...
- Used cross validation to validate the model
- *Novel methods for the prediction of pKa, logP and logD,* Xing L. and Glen R.C.. J. Chem. Inf. Comput. Sci.; **2002**; 42(4); 796-805





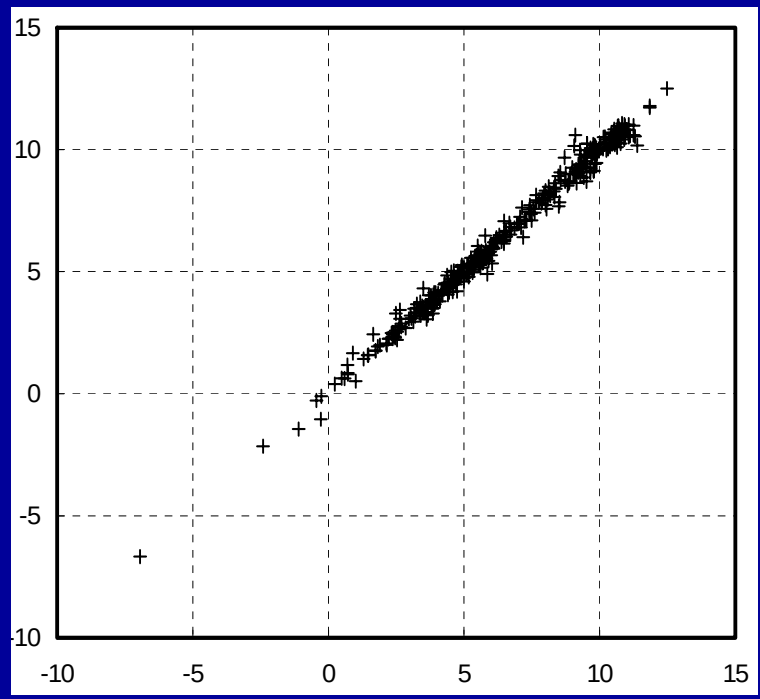
- Recently refined model to improve accuracy

pKa of acids (625)



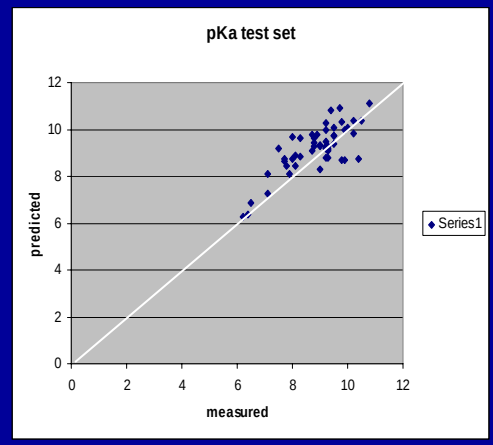
R2=0.98
Std.Err.=0.405
N=625
Q^2=0.92

pKa of bases (412)



R^2=0.99
Std.Err.=0.302
N=412
Q^2=0.95

Conclusions



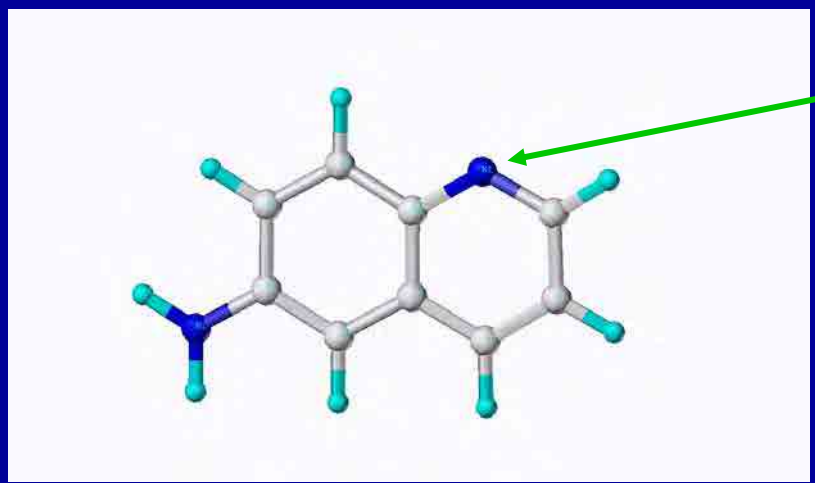
- Surprisingly good results - fast
- Predictive for most pK's
- Useful in biological setting in estimating Pharmacokinetics, active species, metabolism etc.
- Predicts for all types - sometimes get odd results though, if outside parameter set or the 'atom types' are miss-set
- Applying method to other problems e.g. similarity

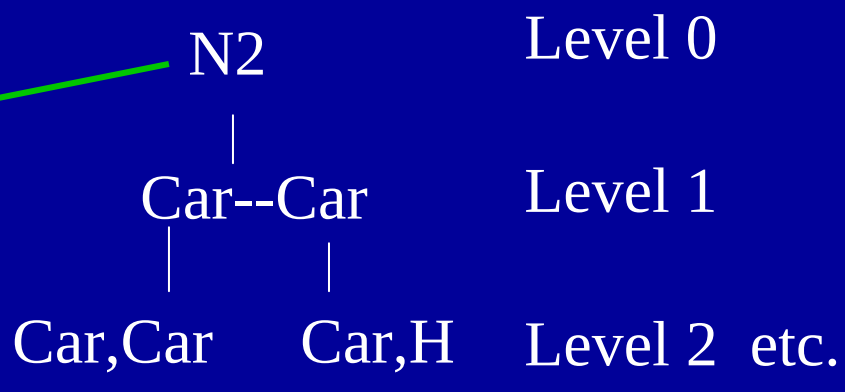
Novel methods for the prediction of pKa, logP and logD, Xing L. and Glen R.C.. J. Chem. Inf. Comput. Sci.; **2002**; 42(4); 796-805

Predicting pKa by Molecular Tree Structured Fingerprints and PLS. Xing L.,Glen R. C. and Clark, R. D. J. Chem. Inf. Comput. Sci. **2003**, 43(3), 870

1. Atom centred fingerprints

- We created a descriptor suitable as a similarity index by looking at all atoms in turn in a molecule and for each atom, generating a depth-3 atom environment. No hashing was involved. These are then binned into an integer string - a ‘fingerprint’ for each atom centre





2. Information-Gain Based Feature Selection

- We wish to select the important features.
- To do this we calculate the entropy of the data as a whole and for each class.
- This is used to select those features with the highest discrimination, e.g. active or inactive or toxic and non-toxic molecules

$$S = -\sum p \log_2 p$$

$$I = S - \sum_v \frac{|S_v|}{|S|} S_v$$

3. Classification

- The next step is to identify which molecules belong to which class.
- To do this we use a Naïve Bayesian Classifer using the features (atom environments) we have identified as being important.

3. Naïve Bayesian Classifier

- Include all selected features f_i in calculation of

$$\frac{P(CL_1 \mid F)}{P(CL_2 \mid F)} = \frac{P(CL_1)}{P(CL_2)} \prod_i \frac{P(f_i \mid CL_1)}{P(f_i \mid CL_2)}$$

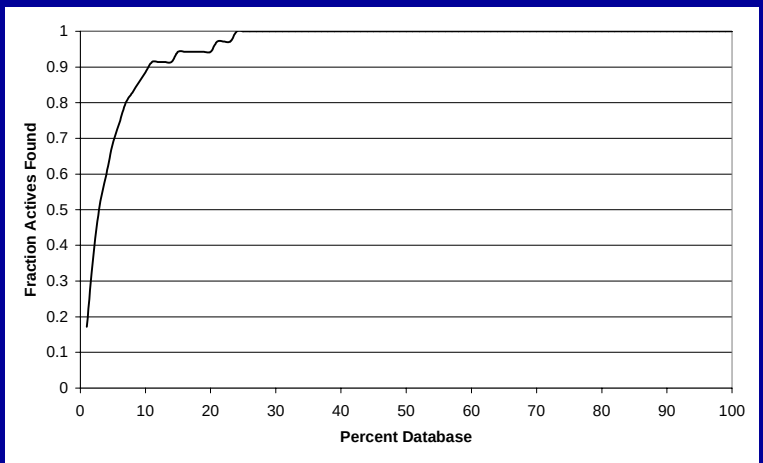
- Ratio > 1: Class membership 1
- Ratio < 1: Class membership 2
- F: feature vector
- f_i: feature elements

MDDR – lead discovery

- MDDR test run: 957 ligands from MDDR
  - 49 5HT3 Receptor antagonists, 40 Angiotensin Converting Enzyme inhibitors (ACE), 111 HMG-Co-Reductase inhibitors (HMG), 134 PAF antagonists and 49 Thromboxane A2 antagonists (TXA2)
- A) Hit rate among ten nearest neighbours for each molecule
- B) 20-fold Cross Validation, 5 Molecules for query generation

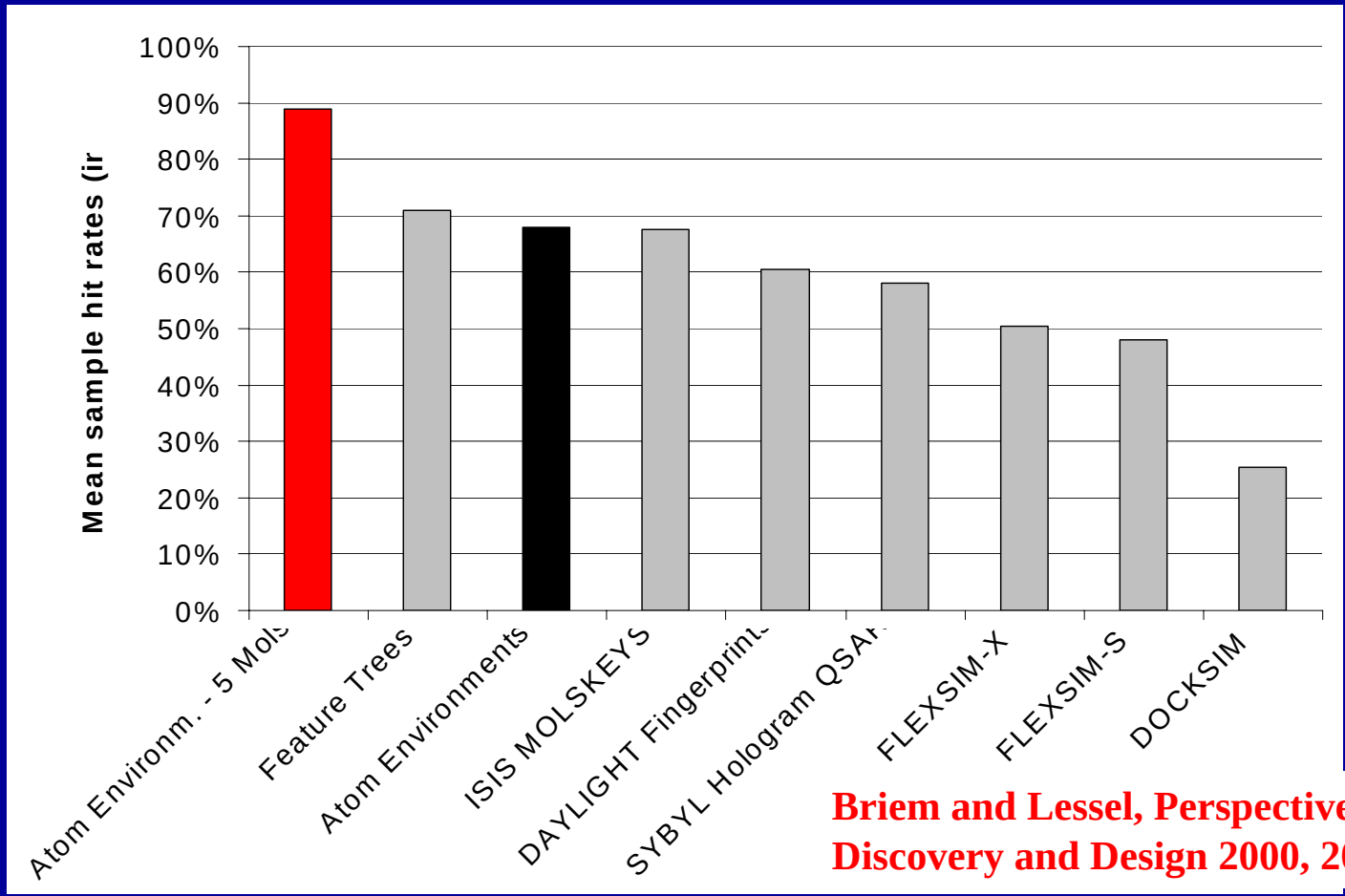
MDDR database searches

Performance of the Atom Environment Approach, Selecting 20 Features						
Group of Active Compounds	5HT3	ACE	HMG	PAF	TXA2	Overall
Expected Hit Rate for Random Selection	0.50	0.41	1.15	1.39	0.50	0.79
Hit Rate for this Method	5.82	5.85	8.33	7.29	6.47	6.75
Enrichment Factor	11.6	14.3	7.24	5.24	12.9	8.54



e.g. ACE: We found about 80% of the active molecules among the first 10% of the library

Combining data and search performance



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

Molecular Similarity Searching using Atom Environments, Information-Based Feature Selection and a Naïve Bayesian Classifier
Andreas Bender, Hamse Y. Mussa and Robert C. Glen, University of Cambridge
Stephan Reiling, Aventis Pharmaceuticals

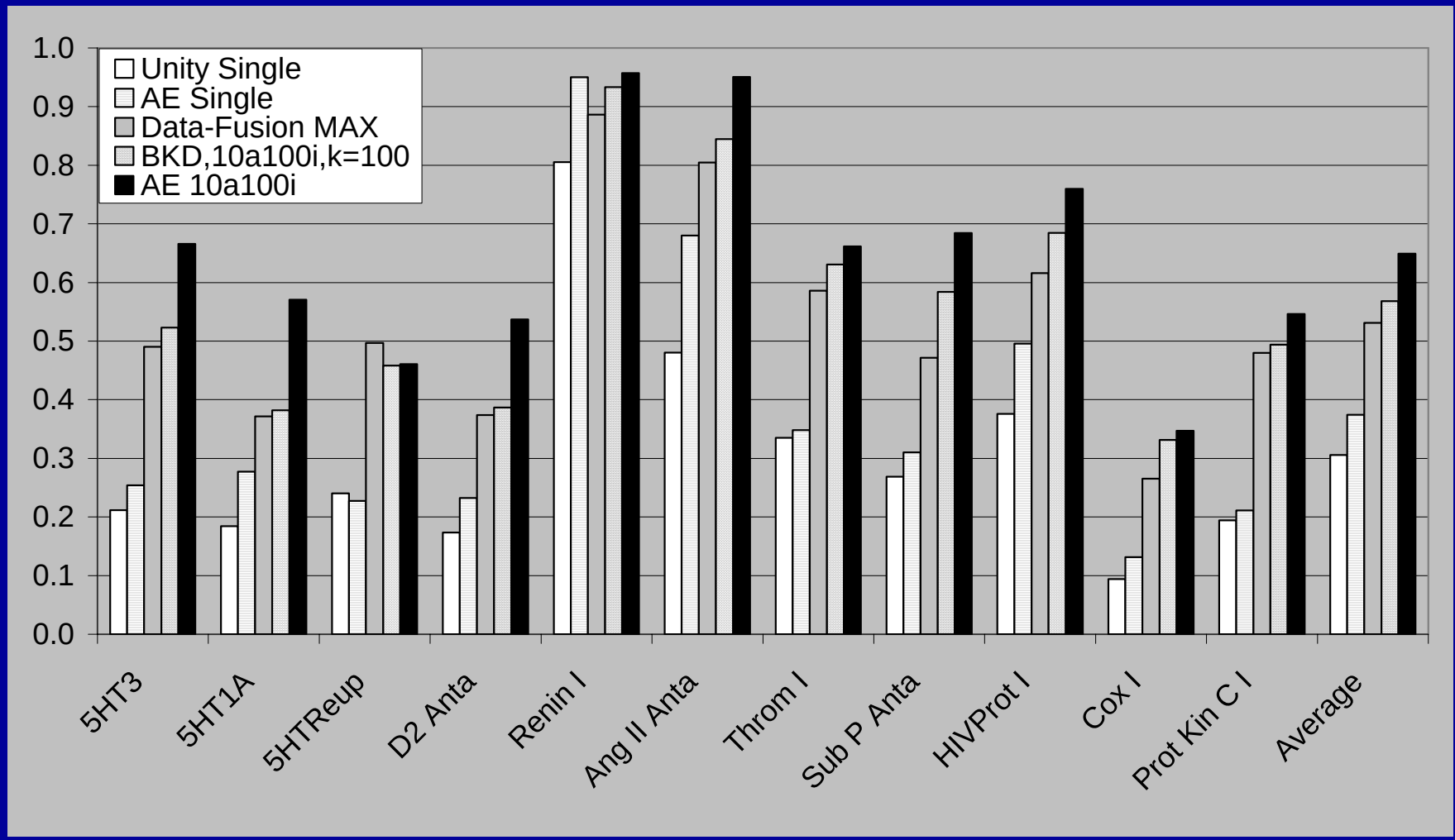
J. Chem. Inf. Comp. Sci. , 2004; 44(1); 170-178

Comparison using Larger 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
- Atom Environments were 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 J, Willett P, Wilton DJ: Comparison of fingerprint-based methods for virtual screening using multiple bioactive reference structures. J Chem Inf Comput Sci 2004, 44:1177-1185.

Comparison of Methods

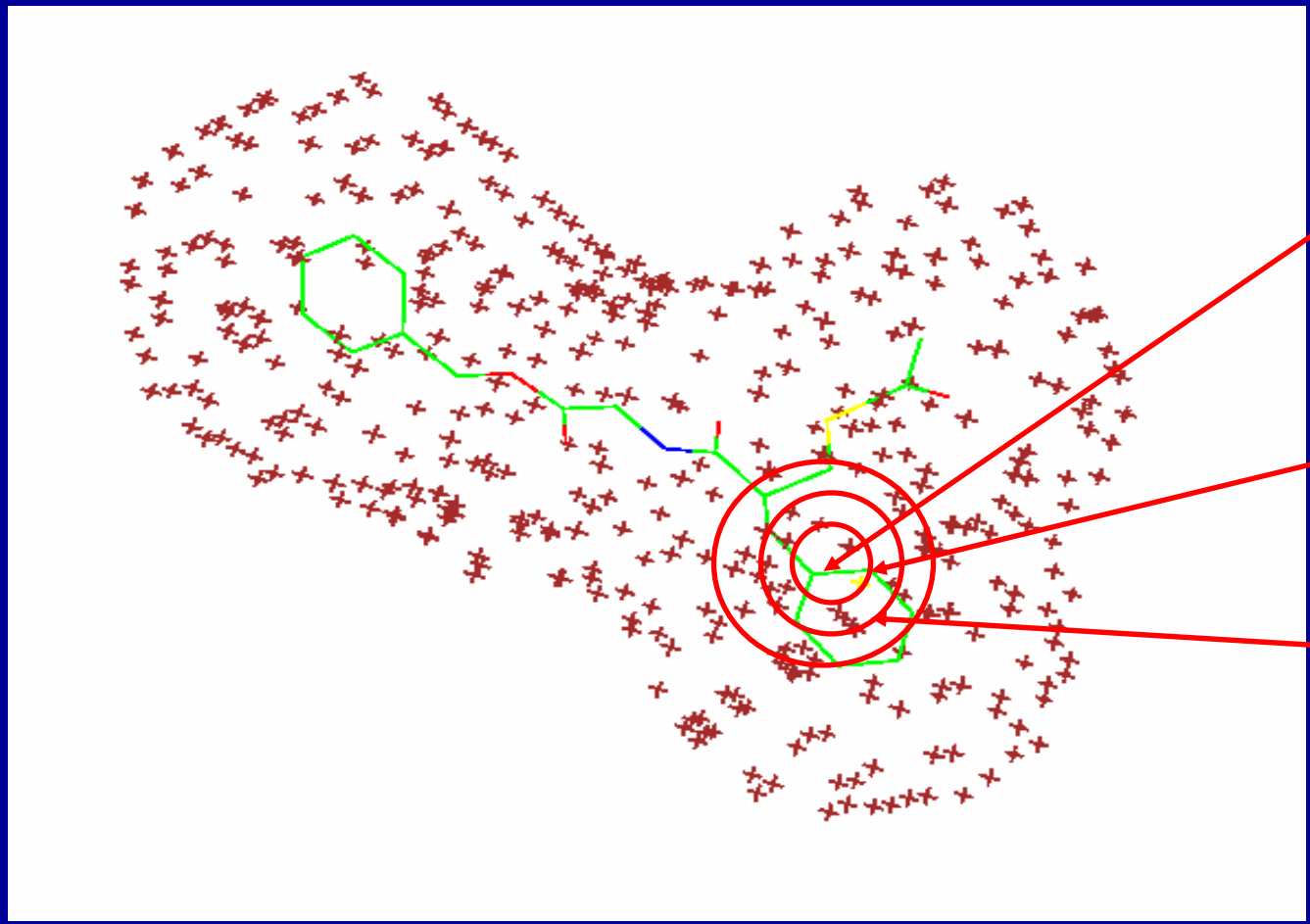


Similarity Searching of Chemical Databases Using Atom Environment Descriptors (MOLPRINT 2D): Evaluation of Performance. Bender, A.; Mussa, H. Y.; Glen, R. C.; Reiling, S.J. Chem. Inf. Comput. Sci.,2004; 44(5); 1708-1718.

Transformation to 3D

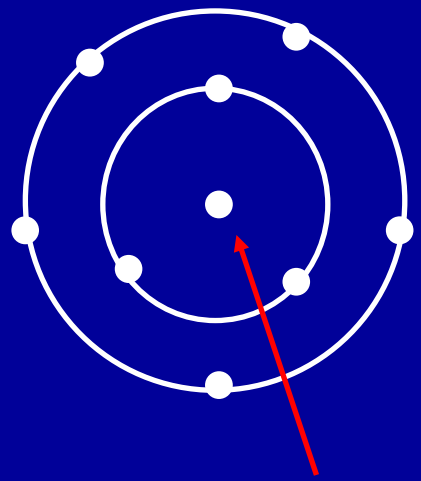
- Two parts: Interaction fingerprint and shape description; here results using only interaction fingerprints are shown, shape description under development
- Information was merged from multiple molecules by using information-gain feature selection and the Naïve Bayesian Classifier

3D: Environment around a surface point: solvent accessible surface



Algorithm

Interaction Energies at Surface Points, one Probe at a time



-0.35 EU

Binning Scheme

-1.0
-0.45
-0.4
-0.3
-0.1
0.0
0.2

-0.35

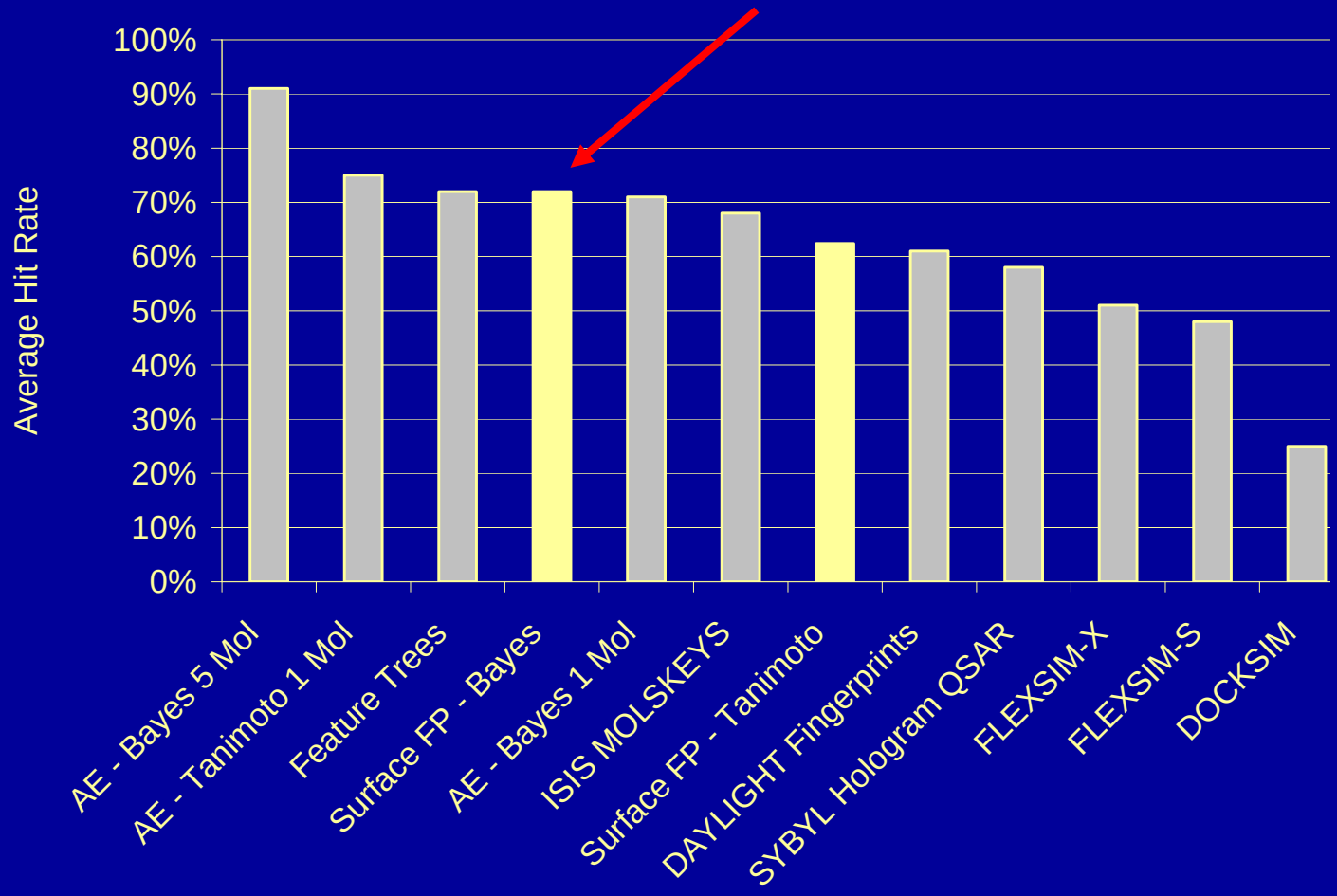
Surface Point Environment

00010000 – 01100010 - 011101100

Algorithm Flow

Step	Program used	Parameters
Generation of 3D coordinates	Concord	
Calculation of Surface Points	msms	Sphere radius, probe size, triangulation density
Calculation of Interaction Energies	GRID	Probe (and various others)
Transformation of interaction energies into descriptors	Perl script	Binning, number of bins, threshold levels

Surface Environments – comparison with 2D and other methods



Conformational Variance

- MDDR Dataset (5HT3, ACE, HMG, PAF, TXA2)
- 10 Randomly selected compounds each
- 10 Conformations generated by GA search with large window (10° for rigid 5HT3, 100° for ACE, HMG, PAF, TXA2), giving diverse conformations
- One force field optimized conformation (Concord-generated) used to find other conformations of the same molecule in whole database of 937 structures, using Tanimoto Coefficient

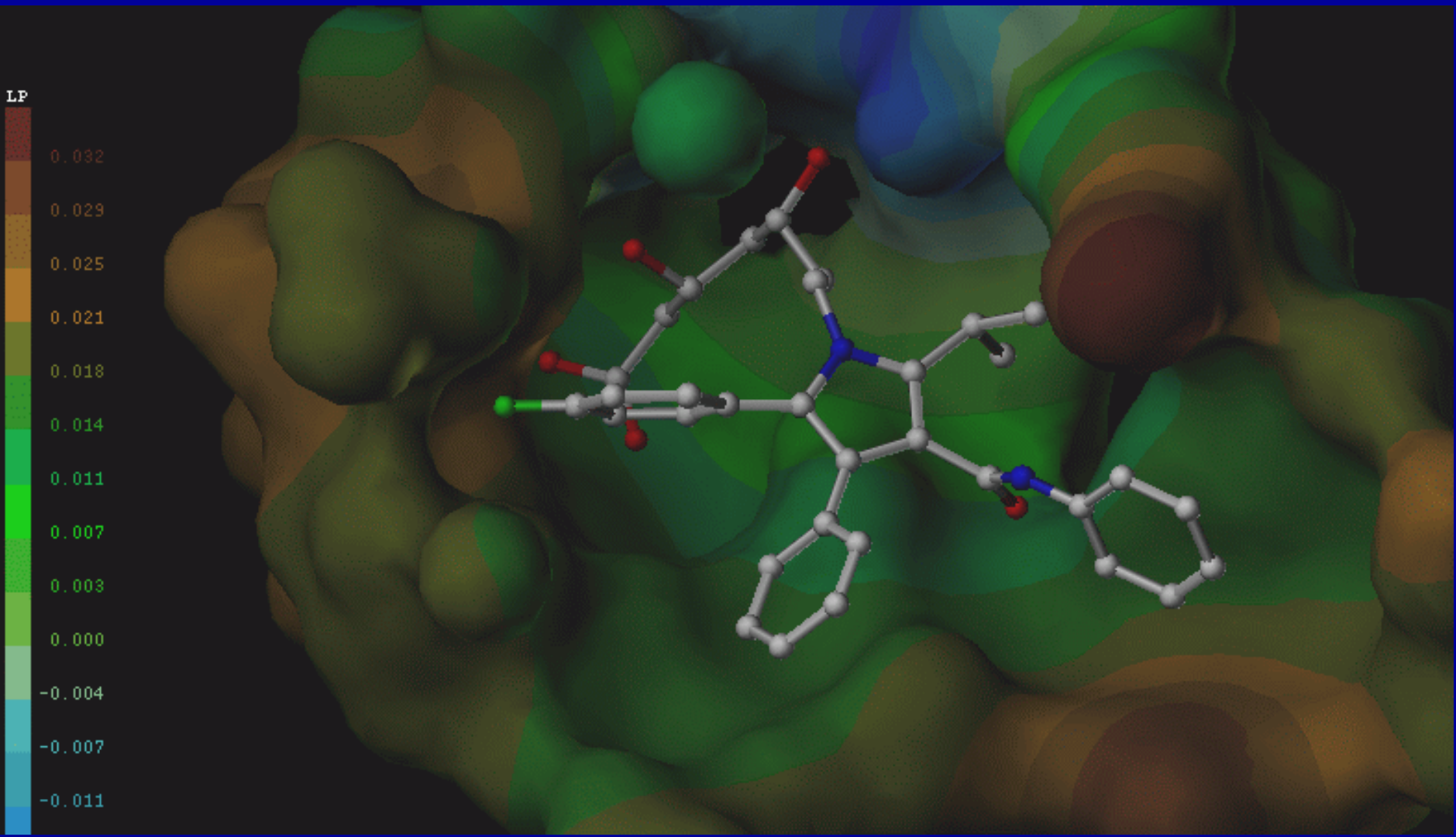
Overall findings

- 64% of conformations found at the top 10 positions -> 2/3 of compounds identified as being most similar (among list of > 900 structures and 40-134 structures of same active dataset)
- >90% of conformations found in Top 5% of sorted database
- Conclusion: If molecules with the right features are present in the database, they will not be missed (in most cases) because they are represented by a particular conformation

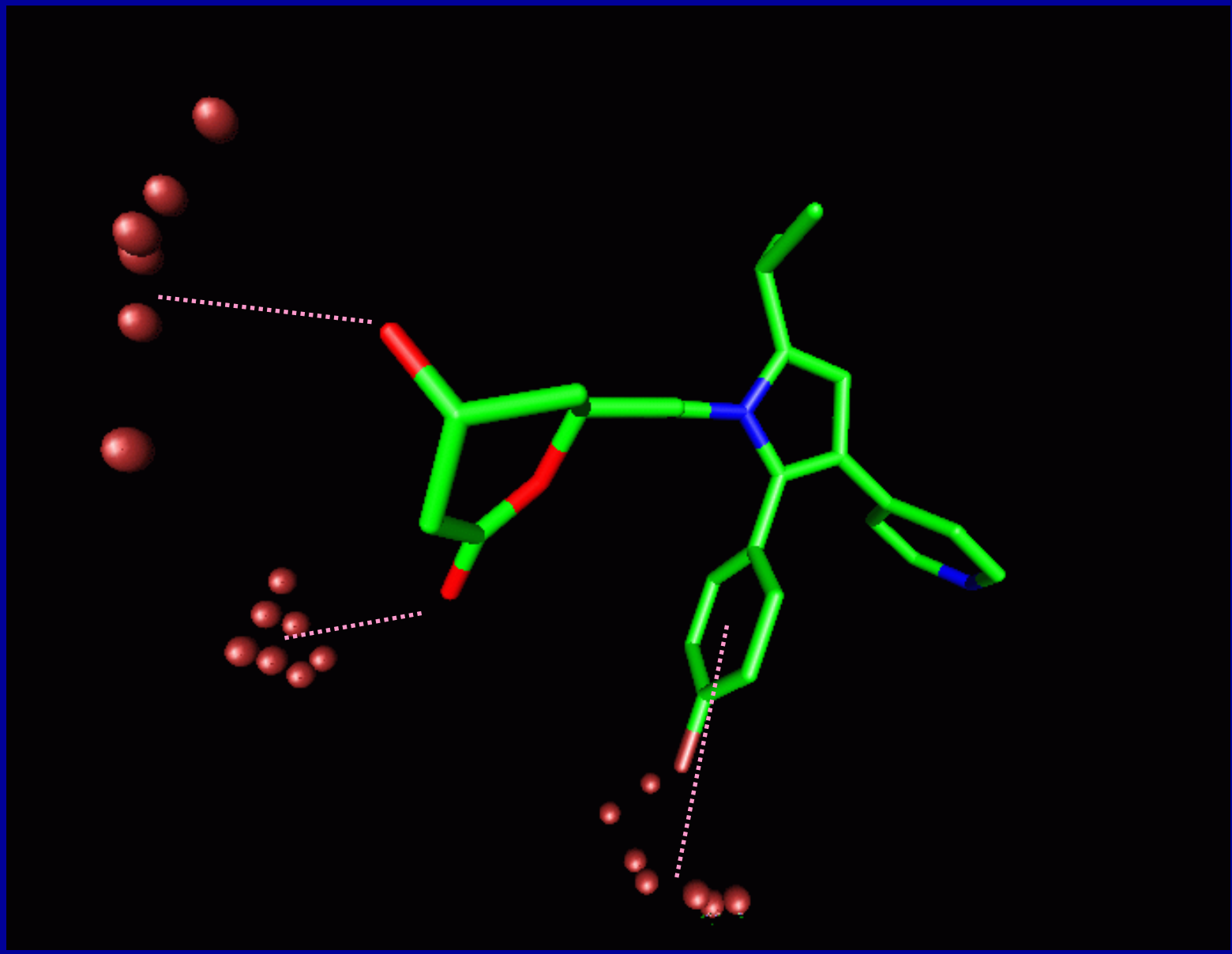
Which features are selected for classification?

- Even if your classifier works, do the selected features make *sense*?
- Set of active vs. inactive molecules
- Information Gain calculated for each feature, those which are much more frequent among actives are “suspicious” and might constitute the pharmacophore
- Look at features from ACE, HMG and TXA2

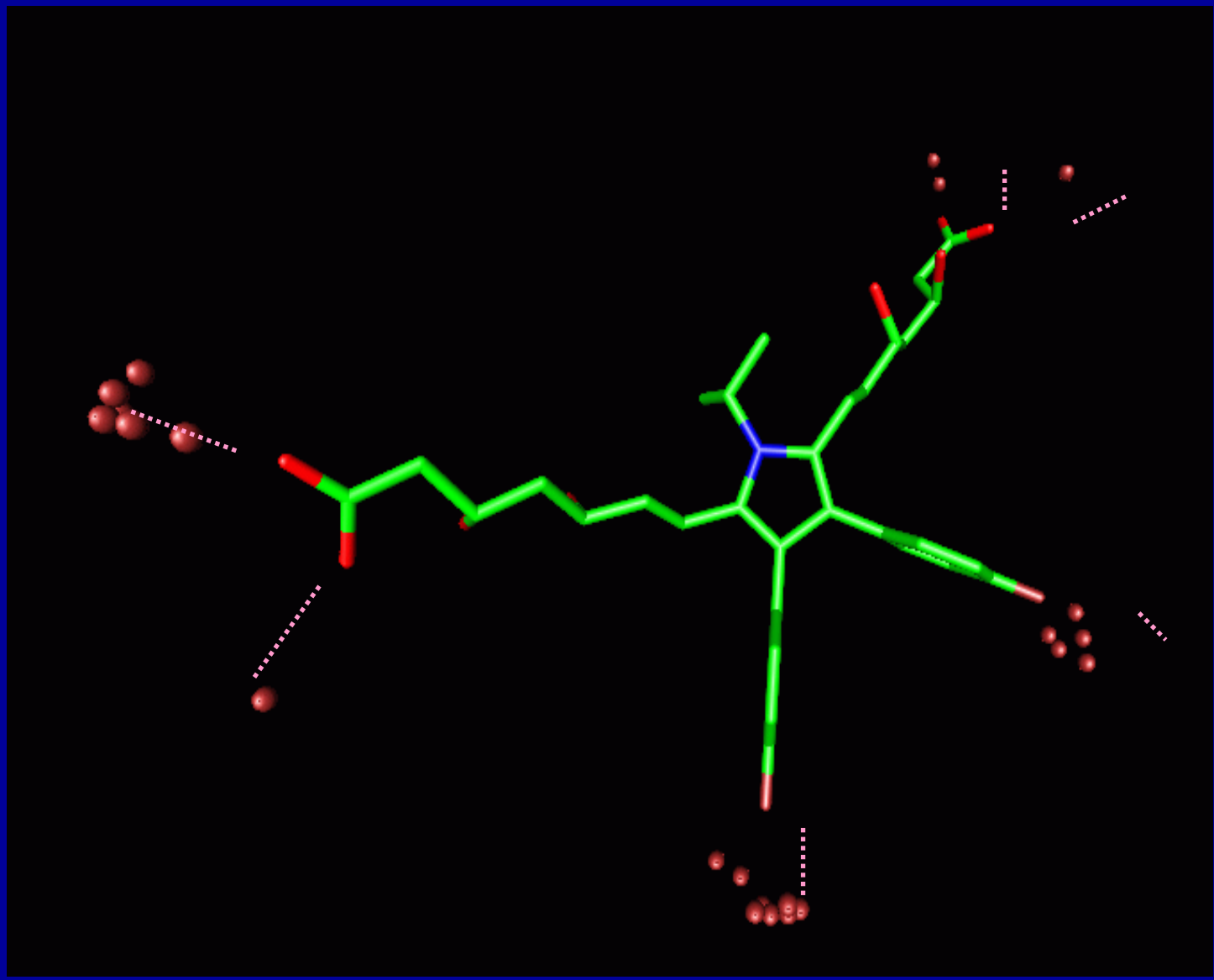
Selected Features - HMG



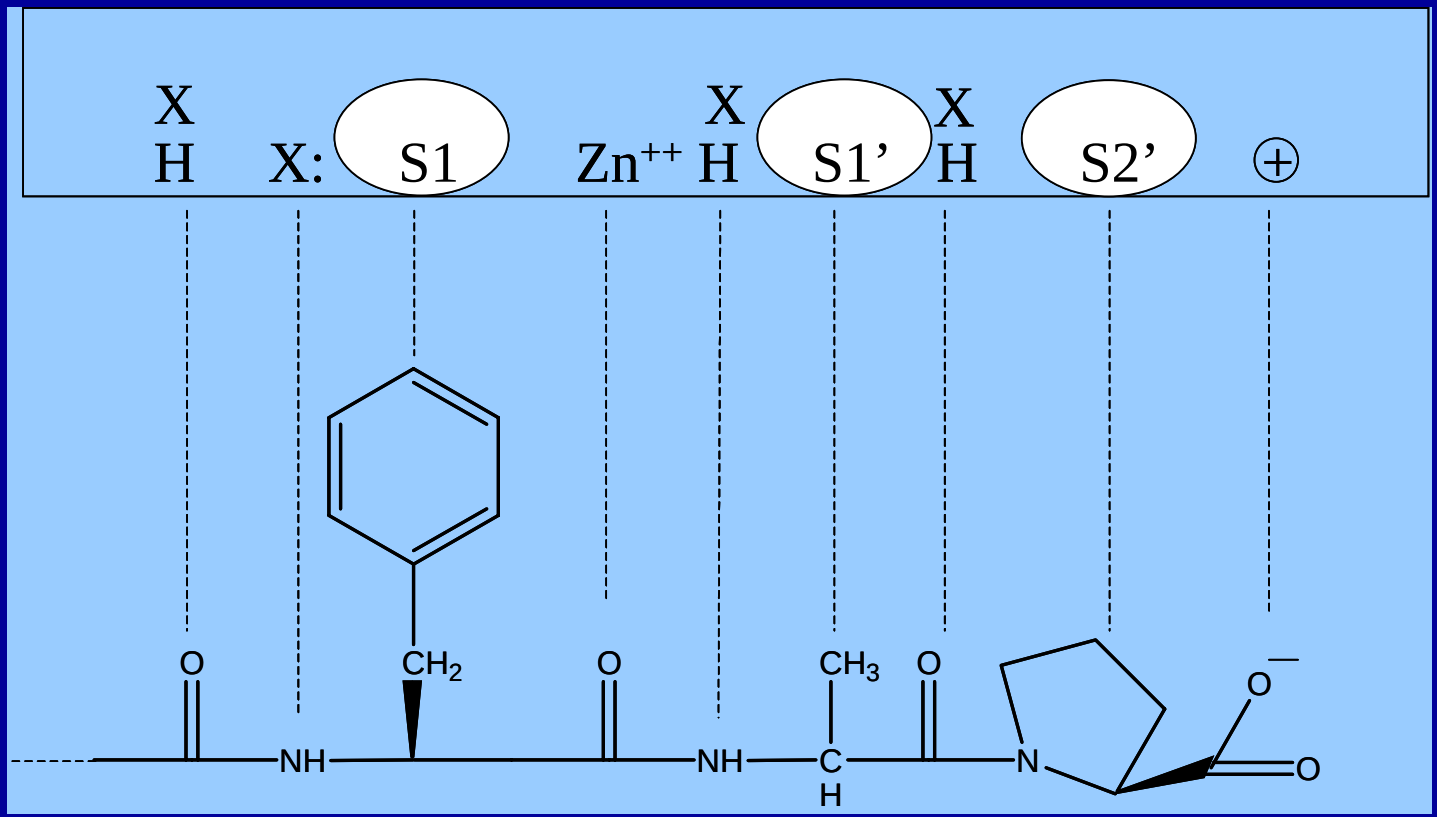
HMG-15



HMG-19

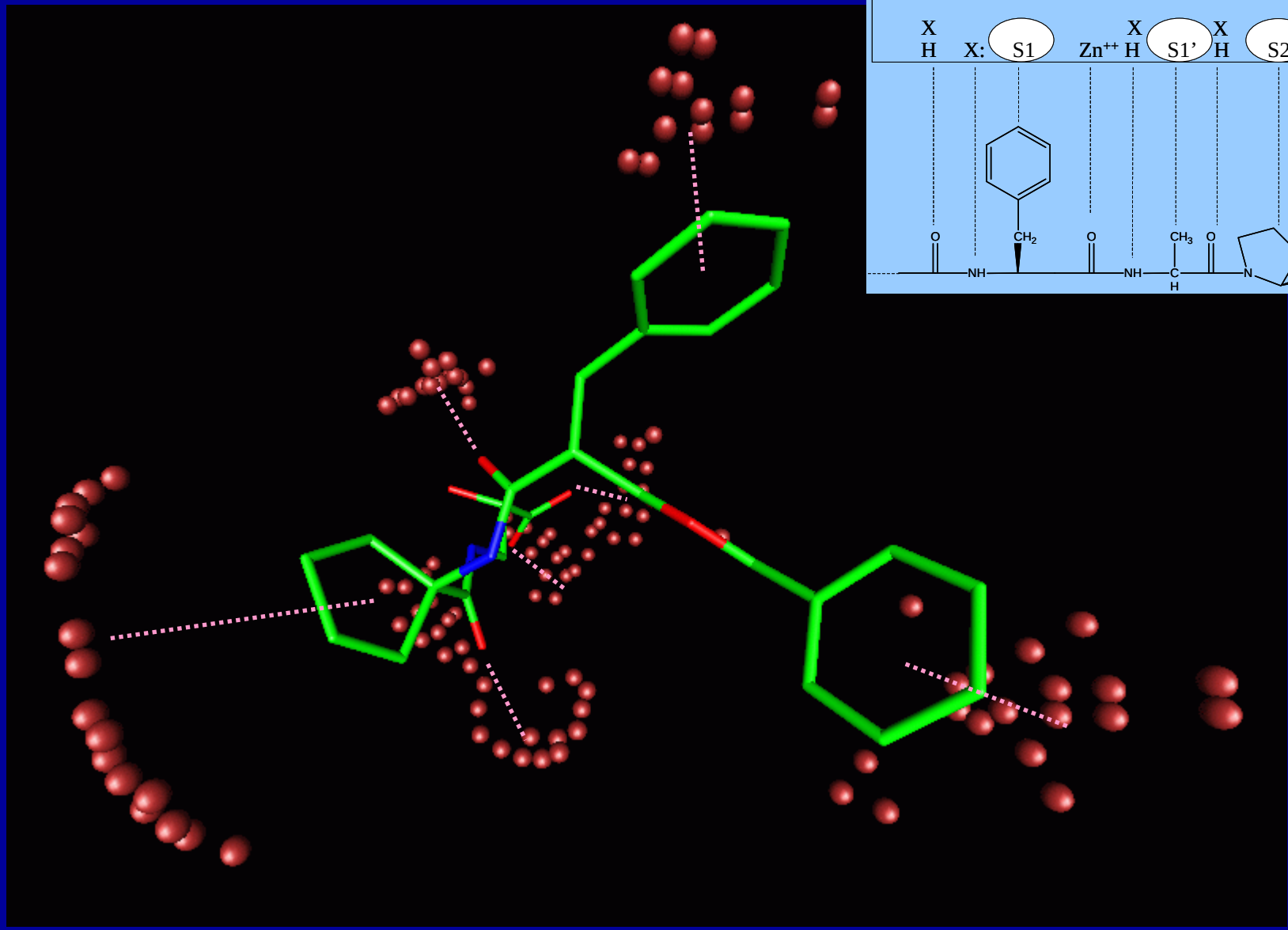


ACE – Binding Site

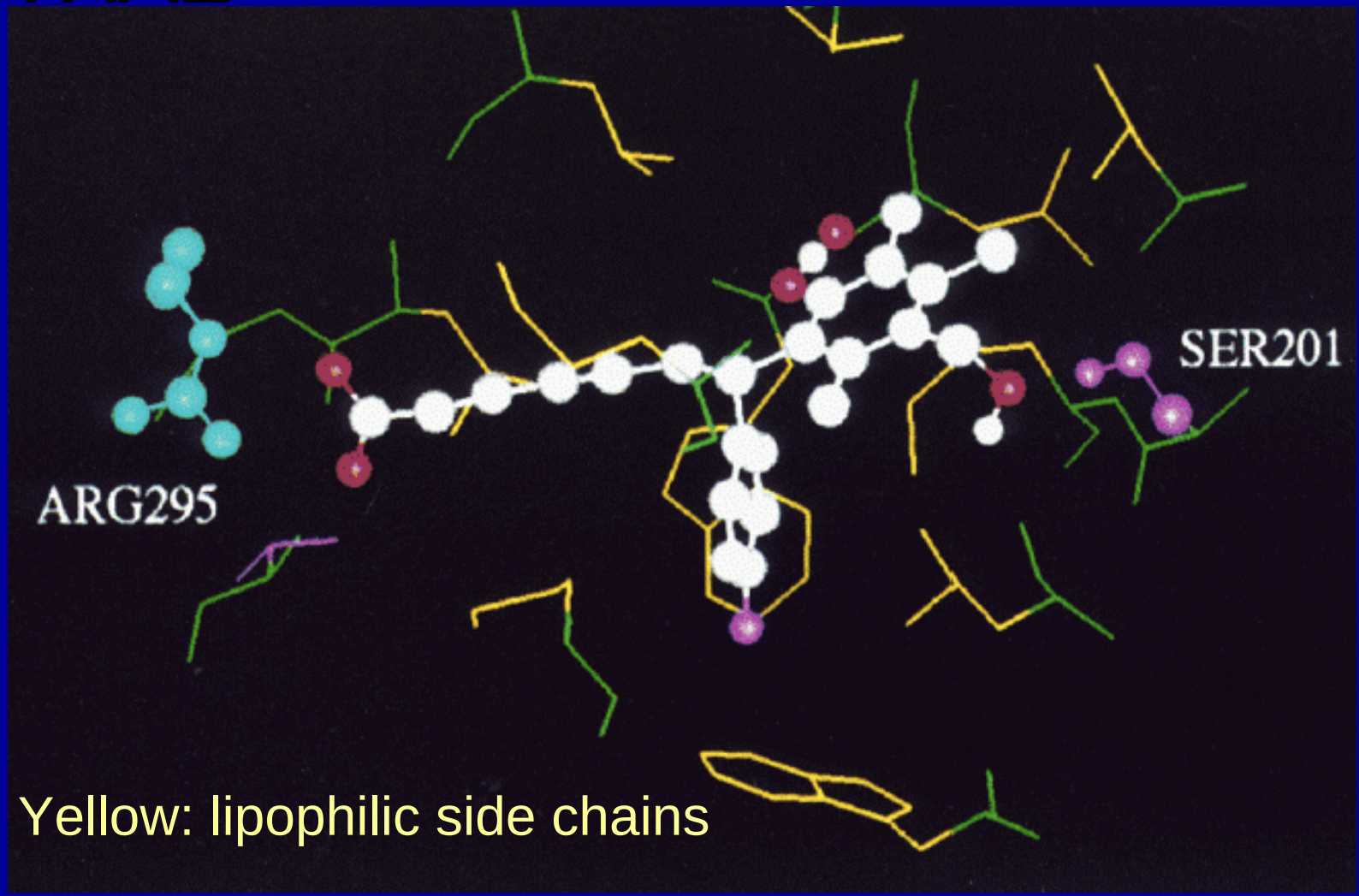


Snake venom peptide analog with putative binding motif to angiotensin used in early compound design (Cushman et al., Biochemistry (1977), 16, 5484-5491.) – recent crystal structure available

Selected Features – ACE-31

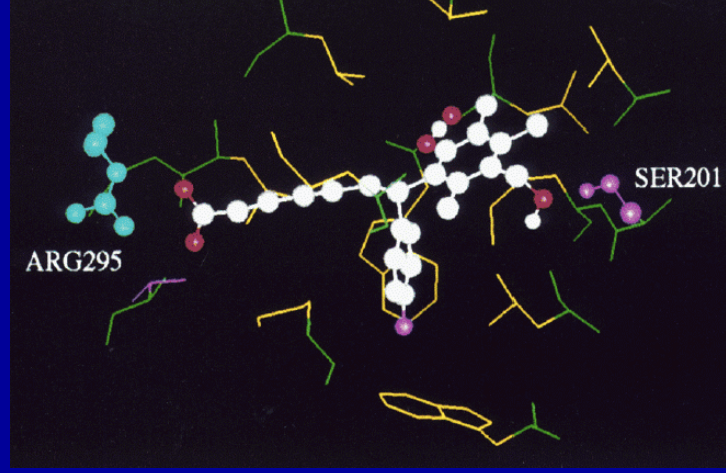


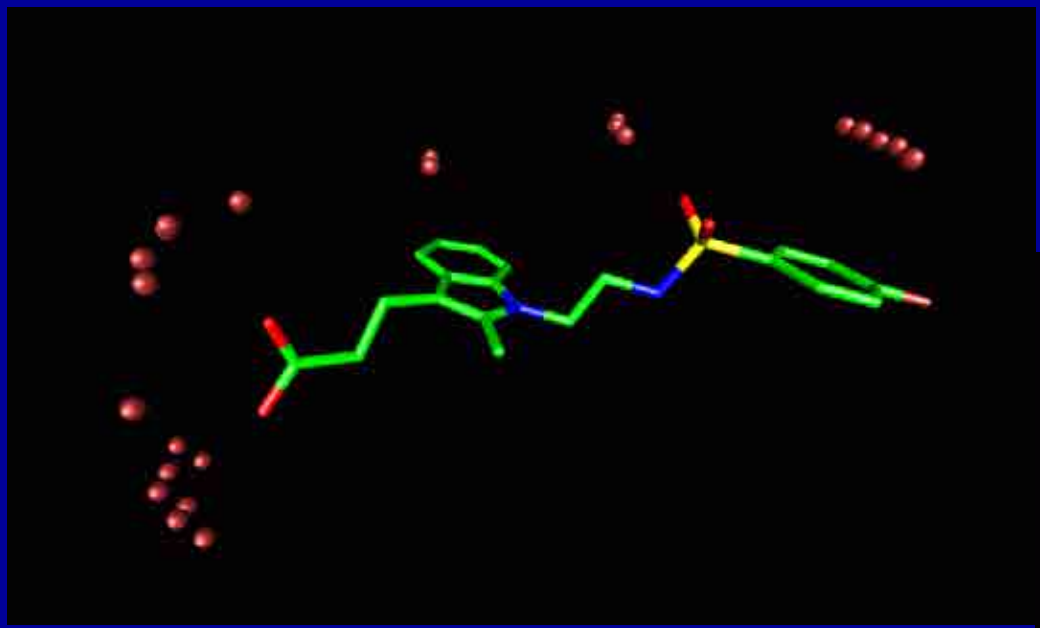
TXA2

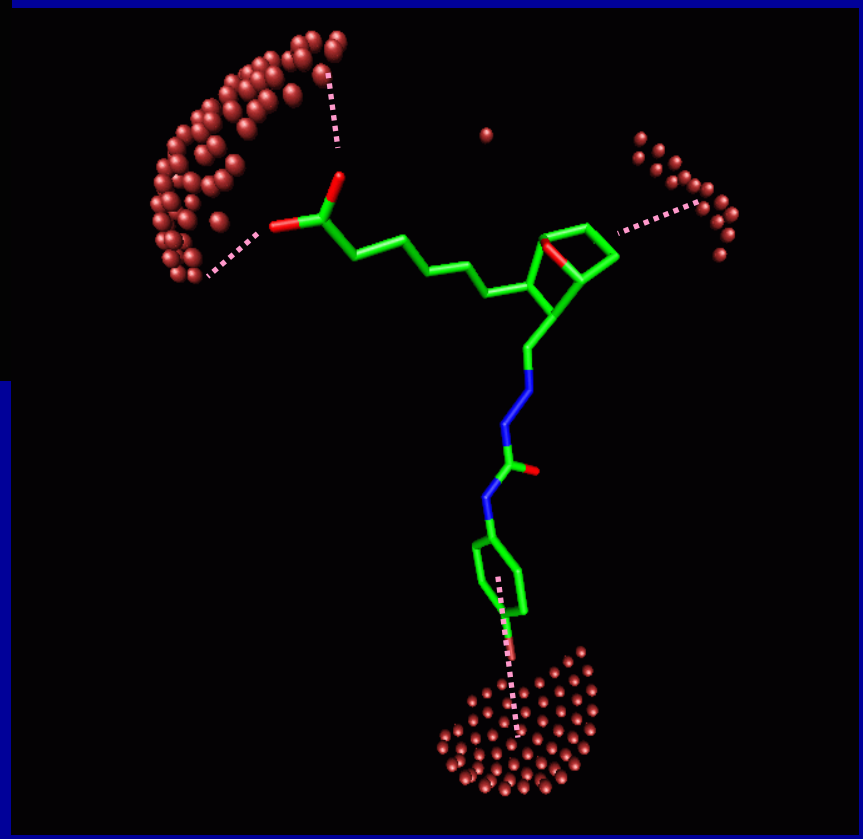


- Yamamoto et al., J. Med. Chem. 1993 (36) 820

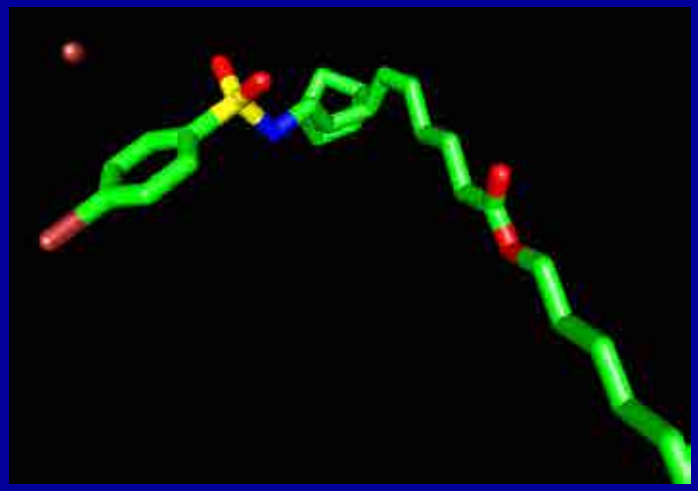
TXA2- 7, and 44

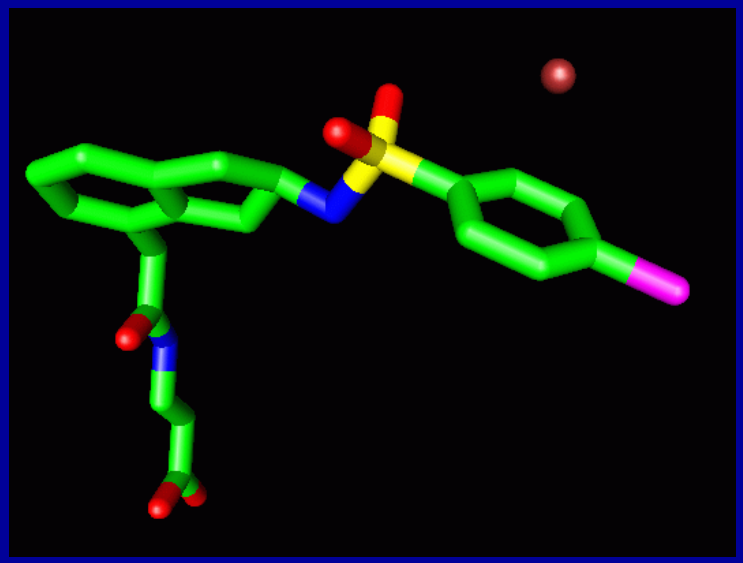


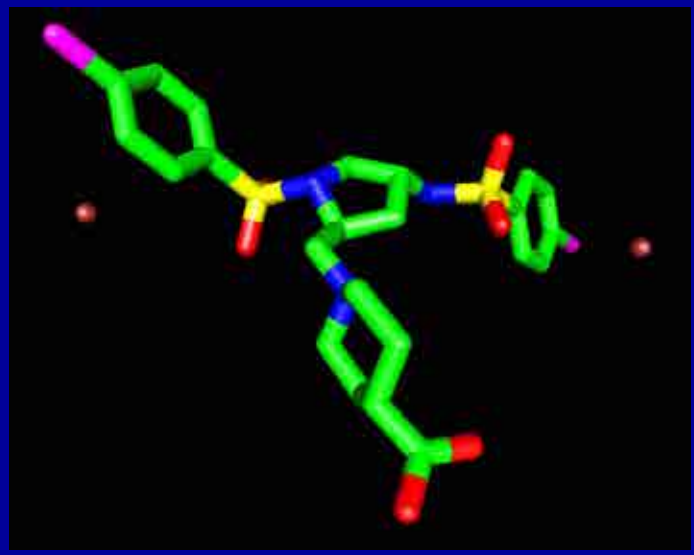


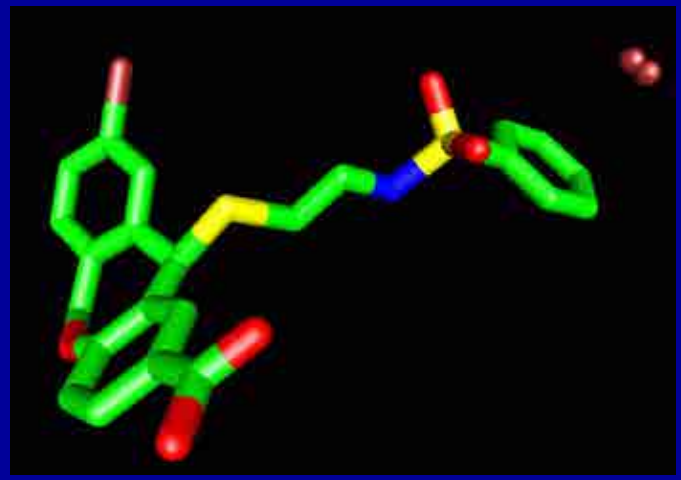


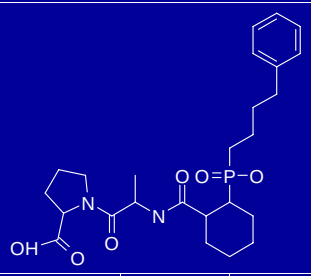
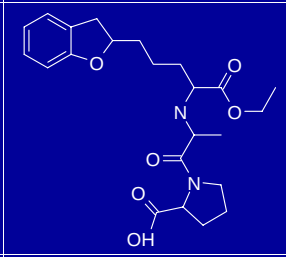
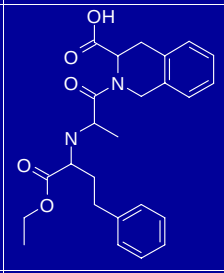
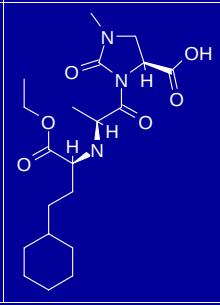
“Feature Hopping”

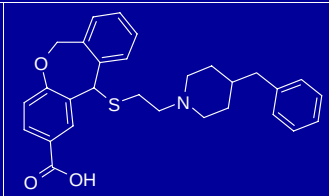
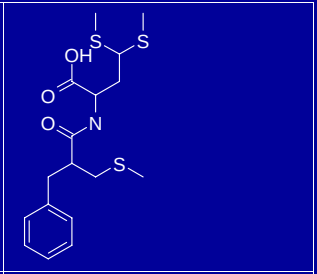
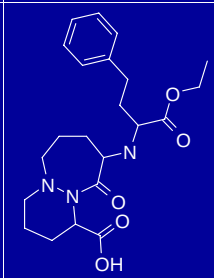
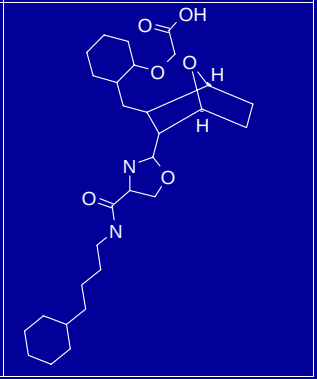








Query			
			
Ranking position	Structure	Ranking position	Structure
1		2	
3		4	
5		6	

7		8	
9		10	

Query (ACE inhibitor) used to screen the database and the highest ranked structures found (out of which all except no. 6,7 and 10 are classified as being ACE inhibitors in the MDDR database). Five of the active structures found (no. 3, 4, 5, 8 and 9) were not found by any of the other seven methods employed.

Molecular Surface Point Environments for Virtual Screening and the Elucidation of Binding Patterns (MOLPRINT 3D). Bender, A.; Mussa, H. Y.; Gill, G. S.; Glen, R. C. J. Med. Chem. 2004, 47(26), 6569-6583.

HTS Data Mining and Docking Competition 2005 at McMaster University (Ontario)

A competition to take 50,000 dihydrofolate reductase inhibitors of known activity (Training Set) and to (blindly) predict the activity of 50,000 new compounds (Test Set) in a high throughput screen.

32 groups took part. We obtained the ‘best’ results.

MOLPRINT 2D, was employed for virtual screening of *E. coli* dihydrofolate reductase (DHFR) inhibitors.
Using an original training set of 49,995 compounds, enrichment factors (between one and three) could be achieved on a test library, comprising 50,000 structures

We think that these results are poor. Reasons are described below.

Results

Number of Active Compounds Identified in Each Group's Ranked List

Group	# Submitted[a]	Consensus Residual Activity[b]		Average Residual Activity[c]		Comment[e]
		Active[b]	Well-Behaved[d]	Active[c]	Well-Behaved[d]	
1	50000	**4**	**1**	**6**	**2**	---
2	495	0	0	0	0	---
3	22	0	0	0	0	---
4	50	0	0	**1**	0	---
5	2000	0	0	**3**	0	---
6	127	0	0	0	0	---
7	50000	**2**	0	**7**	**2**	---
8	150	0	0	0	0	---
9	20	0	0	0	0	---
10	200	0	0	0	0	---
11	30	0	0	0	0	---
12	77	0	0	0	0	---
13	59	0	0	0	0	---
14	294	0	0	0	0	---
15	46901	**1**	**1**	**4**	**1**	---
16	344	0	0	0	0	---
17	10	0	0	0	0	---
18	21	0	0	0	0	---
19	105	0	0	0	0	---
20	50000	**1**	**1**	**1**	**1**	---
21	59	0	0	0	0	---
22	44	0	0	0	0	**YES**
23	6	**1**	0	**1**	0	**YES**
24	40	0	0	0	0	---
25	28	0	0	0	0	---
26	21	0	0	0	0	---
27	121	0	0	0	0	---
28	601	**1**	0	**2**	0	---
29	46720	**2**	**2**	**13**	**5**	**YES**
30	439	0	0	**1**	**1**	---
31	26	0	0	0	0	---
32	1000	0	0	0	0	---

MolPrint2D

a Total number of ranked compounds submitted by each group
NOTE: If group submitted a larger list, only the top 2500 ranked compounds were used

b Cutoff set at 75% residual activity for both replicates

c Cutoff set at 75% residual activity for the average of the replicates

d Subset of active compounds for which a dose response curve could be obtained

e General comment made by group indicating knowledge either that the test set and the training set were from different areas of chemical space or that the test set would perform worse in this assay

Data Set :

High-throughput screening of 49,995 compounds was performed by Zolli-Juran *et al.*, identifying 32 hits (defined by less than 75% residual activity in both of two screening runs) comprising several novel scaffolds.

Objective:

The extraction of the structural 'knowledge' from the compounds and their activities from the first screening ('training set') and to make predictions about the inhibitory activities of a second set of 50,000 compounds that was to be screened subsequently (42 'hits' subsequently found in the 'test set).

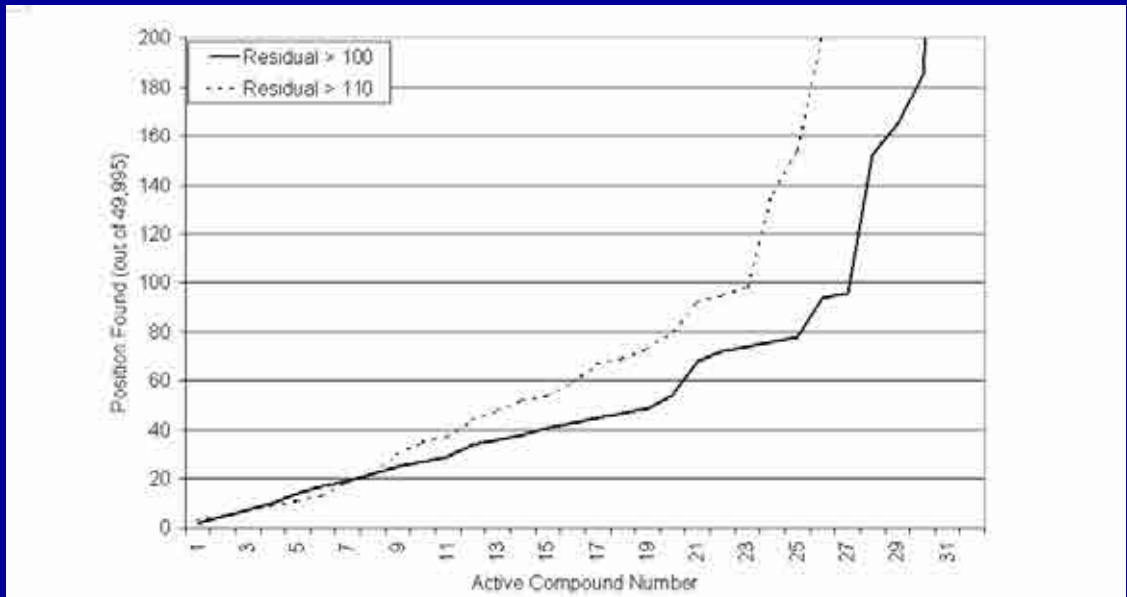
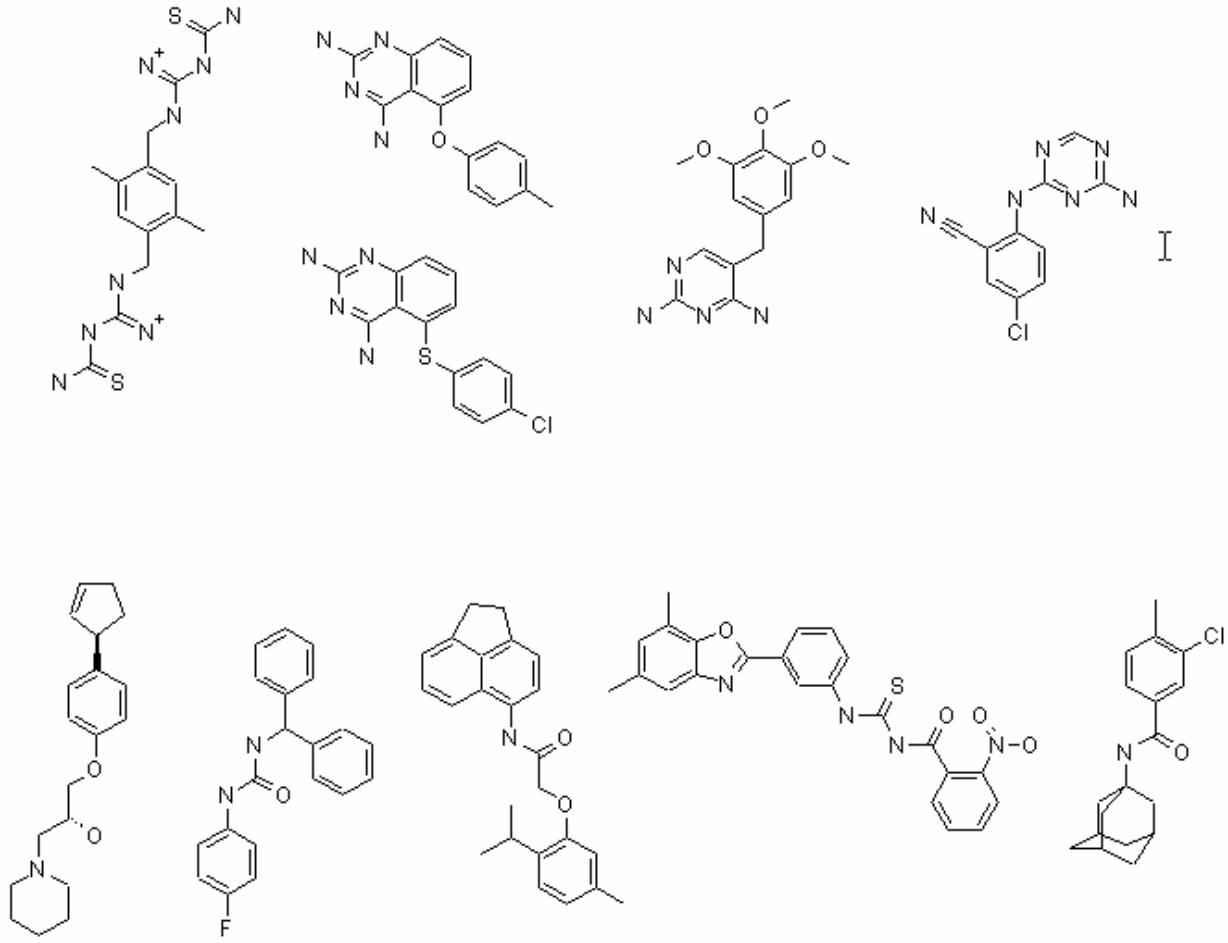


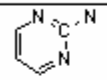
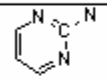
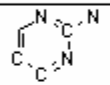
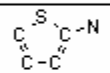
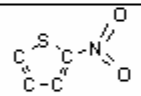
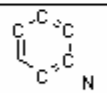
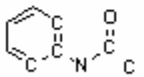
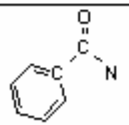
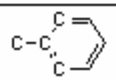
Figure 2. The first 30 active compounds (out of 47; 64%) are found in the top 200 out of 49,995 positions (0.4%) of the ranked training set

Our results show ca. 3 fold enrichment in the first 200 compounds ranked. However, this reduced to just over one in the complete set – why ?



Comparison of the most potent inhibitors of the training set (upper half) and the most potent inhibitors of the test set (lower half). Structural differences can already be identified on this small subset of compounds, for example the high number of pyridazine rings and guadinium groups in the training set.

Features showing highest information-gain in discriminating active structures of the training set from those of the test set. Features characteristic for the most active compounds of each set are also more frequent in the whole set; this ratio is even more apparent among the active structures of each set.

Characteristic Features of Training-Set Active Structures				
	Number in Training Set	Number in Test Set	Among Actives in Training Set	Among Actives in Test Set
	416	159	10	0
	295	72	10	0
	40	72	8	0
	106	12	10	0
	136	0	9	0
Characteristic Features of Test-Set Active Structures				
	9077	18645	14	133
	6580	19186	4	126
	2449	10685	0	76
	9191	16043	12	115
	15202	25851	22	160

The 'Test Set' and the 'Training set' contains chemically different structures.
Therefore, the method does not always recognise new features in the new set as contributors to activity.

We repeated the analysis by randomizing the data and predicting using cross validation.

Training and test set were pooled in a second step and randomly split into training and test of equal size again, thus evening out the different chemical characteristics of both libraries.

In a ten-fold cross validation study on the new training and test sets, typically 10-fold enrichment could be found in the first 96 positions, 4-fold enrichment in the first 384 positions and 3-fold enrichment in the first 1536 positions, corresponding to 6, 10 and 28 hits (out of a total of 307), respectively.

Training and test set were pooled in a second step and randomly split into training and test of equal size again, thus evening out the different chemical characteristics of both libraries.

	Hit Rates			Enrichment Factors		
First … positions	96	384	1536	96	384	1536
Actives < 80% activity; Inactives > 100% activity, 200 Features	2	4	10	3.4	1.7	1.1
Ten-fold Random Validation Actives < 85% activity, Inactives > 100% activity, 200 Features	6.0 (0.7)	10.2 (2.4)	28.0 (3.0)	10.2 (1.2)	4.2 (1.0)	3.0 (0.3)

'Blind study'

after randomization note big increase in success

In a ten-fold cross validation study on the new training and test sets, typically 10-fold enrichment could be found in the first 96 positions, 4-fold enrichment in the first 384 positions and 3-fold enrichment in the first 1536 positions, corresponding to 6, 10 and 28 hits (out of a total of 307), respectively.

Conclusions :

On the one hand the work presented here shows that exact-fragment-matching similarity searching methods are not capable of finding completely novel hit structures. Still, they are able to combine knowledge from multiple active structures to give novel combinations of features, as shown previously. On the other hand this work emphasizes the need for an even distribution of "chemistry" between the training and the test set. 'Lead hopping', moving from one chemical space to another thus requires analysis based on chemical descriptors (not the structural diagram), which is generally a much more compute intensive calculation.

Summary

- 2D Method: Performs about as well as other 2D methods for single molecule searches, outperforms them by a large margin when combining information from multiple molecules
- 3D Method: TR invariant, conformationally tolerant; combines high enrichment factors with scaffold hopping – discovery of new chemotypes
- Features shown to correlate with binding patterns
- Performance (at least in part) due to Bayesian Classifier, which is able to take multiple structures as well as active *and* inactive information into account
- Chemically similar training and test sets required for 2D method

Acknowledgements

- Peter Murray-Rust, Jonathan Goodman, Hamse Mussa, Andreas Bender, Joe Towsend, Yong Zhang, Simon Tyrrell
- Unilever, the Royal Society of Chemistry, the Newton Trust, the Department of Trade and Industry, the EPSRC, the BBSRC.