Computer Aided Drug Design

——Pharmacophore

Qin Xu
http://cbb.sjtu.edu.cn/~qinxu/CADD.htm

Pharmacophore

- IUPAC (International Union of Pure and Applied Chemistry)
  - An ensemble of steric and electronic features that is necessary to ensure the optimal supramolecular interactions with a specific biological target and to trigger (or block) its biological response

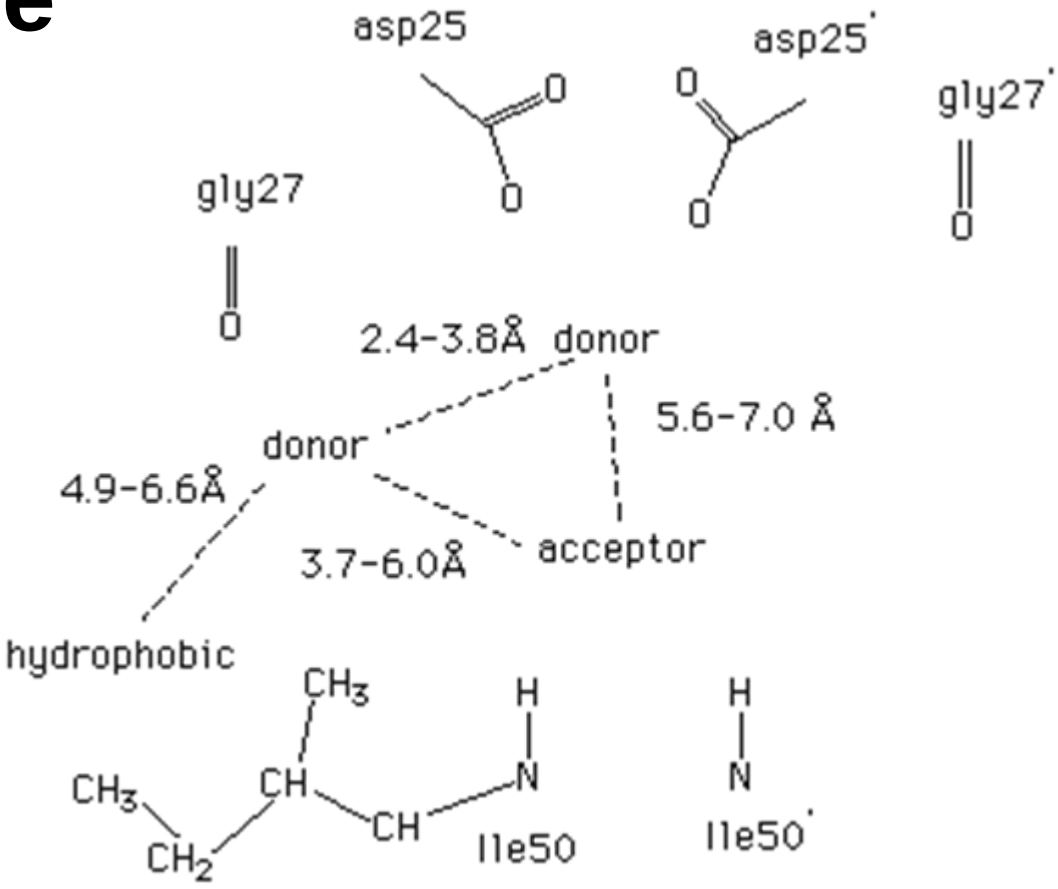
Pharmacophore

- Molecular features
  - For molecular recognition between
  - a ligand
  - a biological macromolecule
- Structural analysis
  - Superimposed active compounds
  - Binding site of the receptor

Pharmacophore

- Applications
  - Vitural screening
  - 3D-QSAR
  - De novo drug design
- Softwares
  - Sybyl
  - Discovery studio
  - MOE

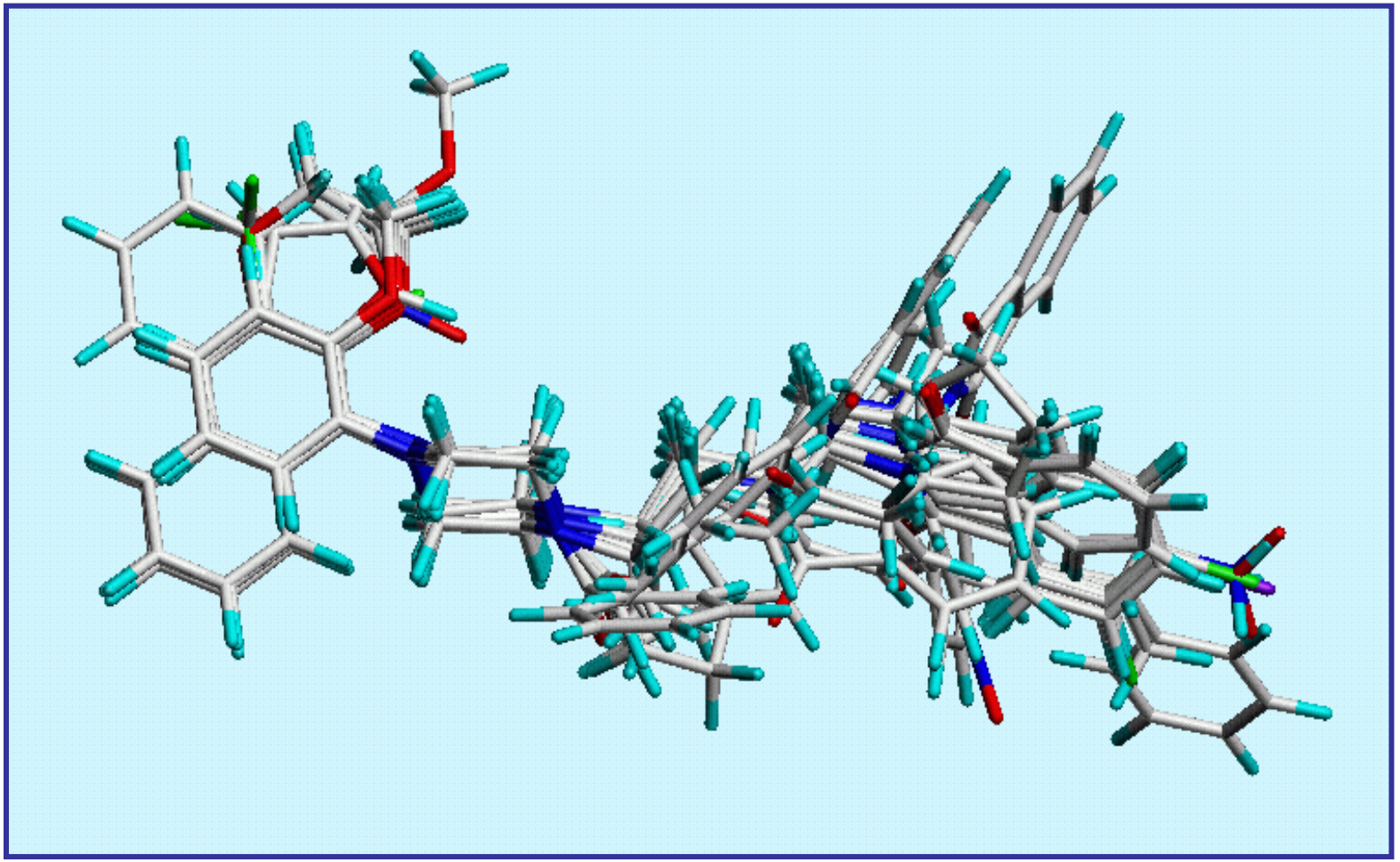
Pharmacophore in 2D



HIV-1 protease pharmacophore and active site

Geometric arrangement of functional groups of the ligand that are required for “activity”

Pharmacophore in 3D



A pharmacophore is a spatial arrangement of atoms or functional groups believed to be responsible for biological activity

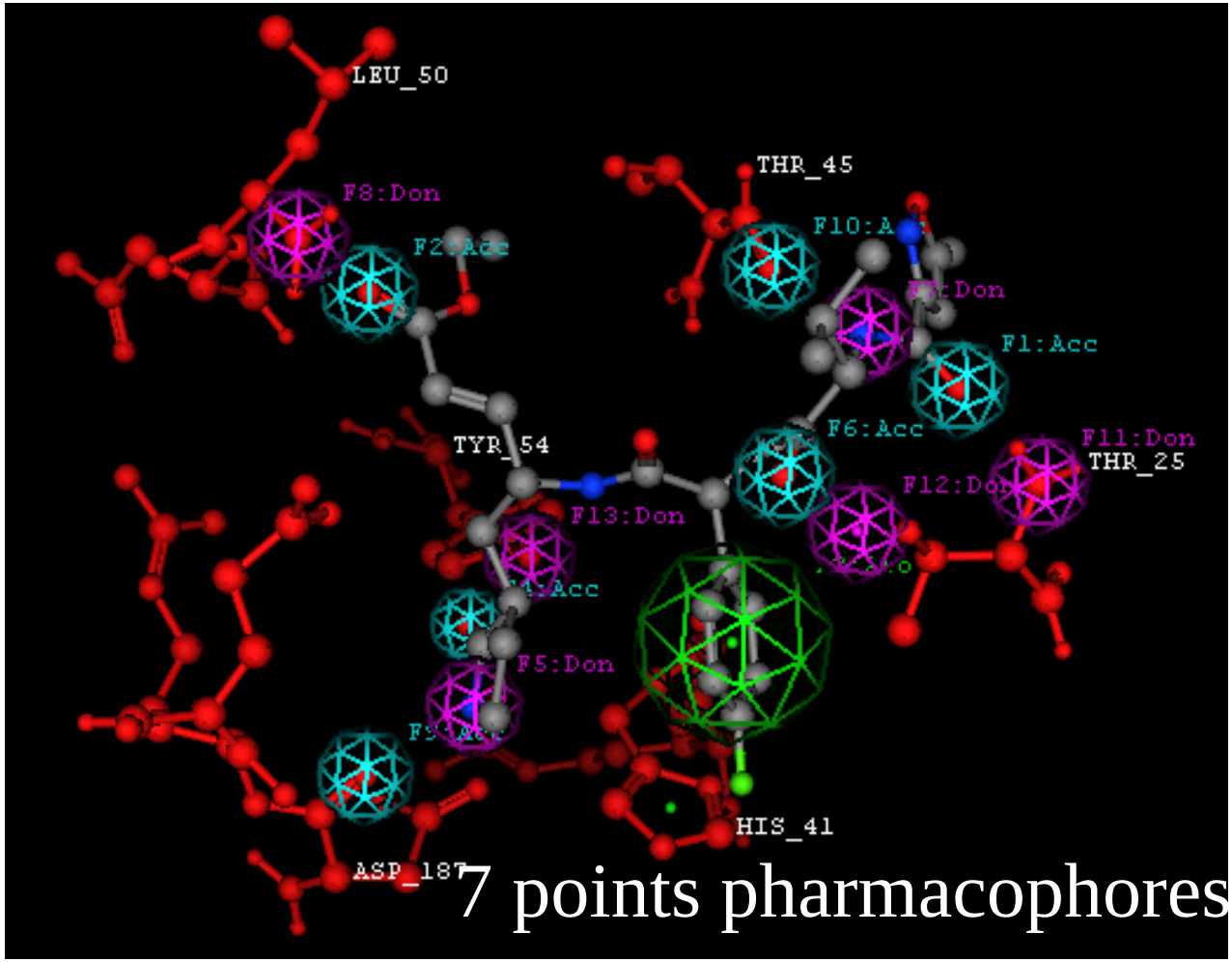
A *pharmacophore scheme*

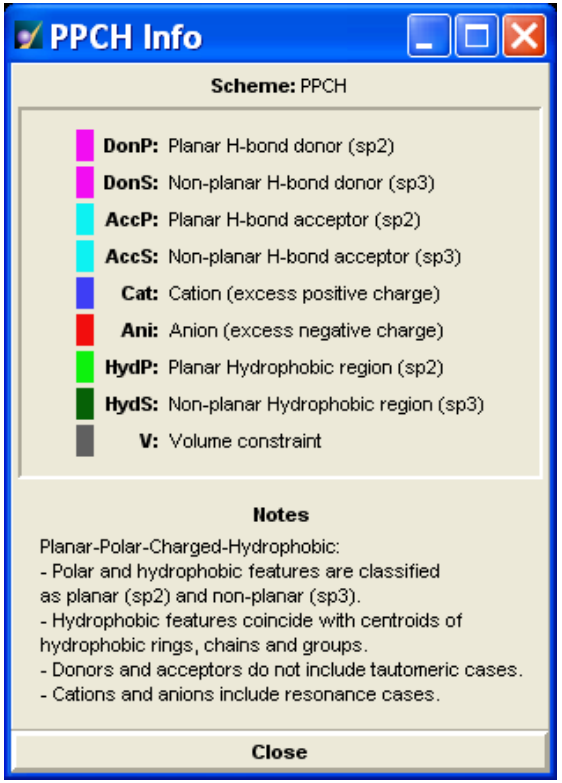
A collection of functions that define the meaning, appearance and methods of calculation of ligand annotation points and their attached labels. The scheme defines how each ligand in the searched database is annotated. A typical scheme is *PCH* (Polarity-Charge-Hydrophobicity).

Label	**Definition**
Don	Hydrogen bond donors, including tautomeric donors.
Acc	Hydrogen bond acceptors, including tautomeric acceptors.
Cat	Cations, including resonance cations.
Ani	Anions, including resonance anions.
Hyd	Hydrophobic areas.
Aro	Aromatic centers.

Pharmacophore of KZ7088

The ligand **KZ7088 in the active site of SARS-CoV M^{pro}**



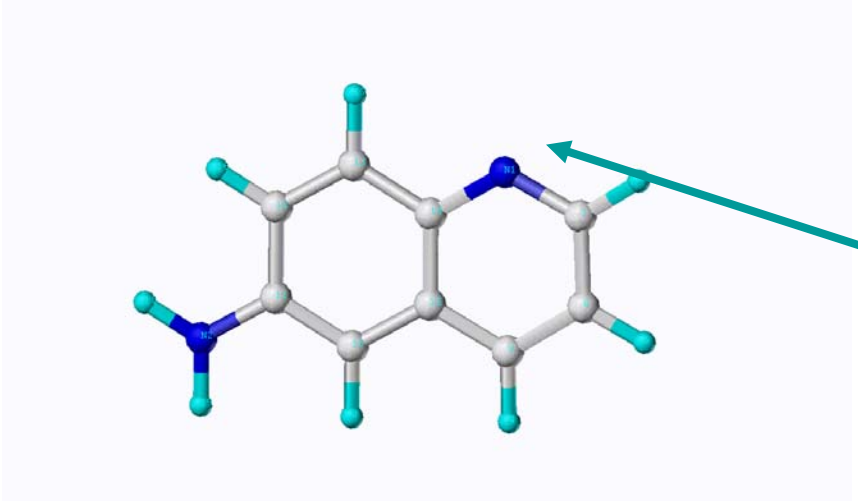


Outline

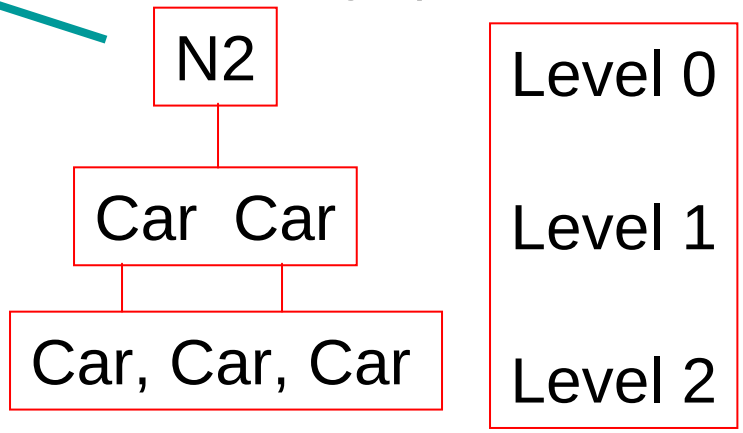
- Objective: More efficient searching of chemical databases
- New methods developed to detect molecules with similar biology: One is based on connectivity (2D), the other on surface points (3D)
- Details of the algorithms presented here, starting with the 2D type
- Results: Lead Discovery – finding new drugs, finding new chemotypes
- Feature: Discovering Binding Patterns

2D: Environment around an atom

- 6-aminoquinoline



Assign Sybyl mol2 atom types
find connections
find connections to connections
create a tree down to n levels
'bin' the atom types for each level
create a 'fingerprint' for this atom



These features are created for every (heavy) atom in the molecule

Information-Gain 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 and inactive molecules.

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

Information gain (to be maximized)

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

Entropy of the whole set

Entropy of subsets

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.

*Naïve Bayesian Classifier
("classification by presumptive evidence")

- 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

Application: lead discovery

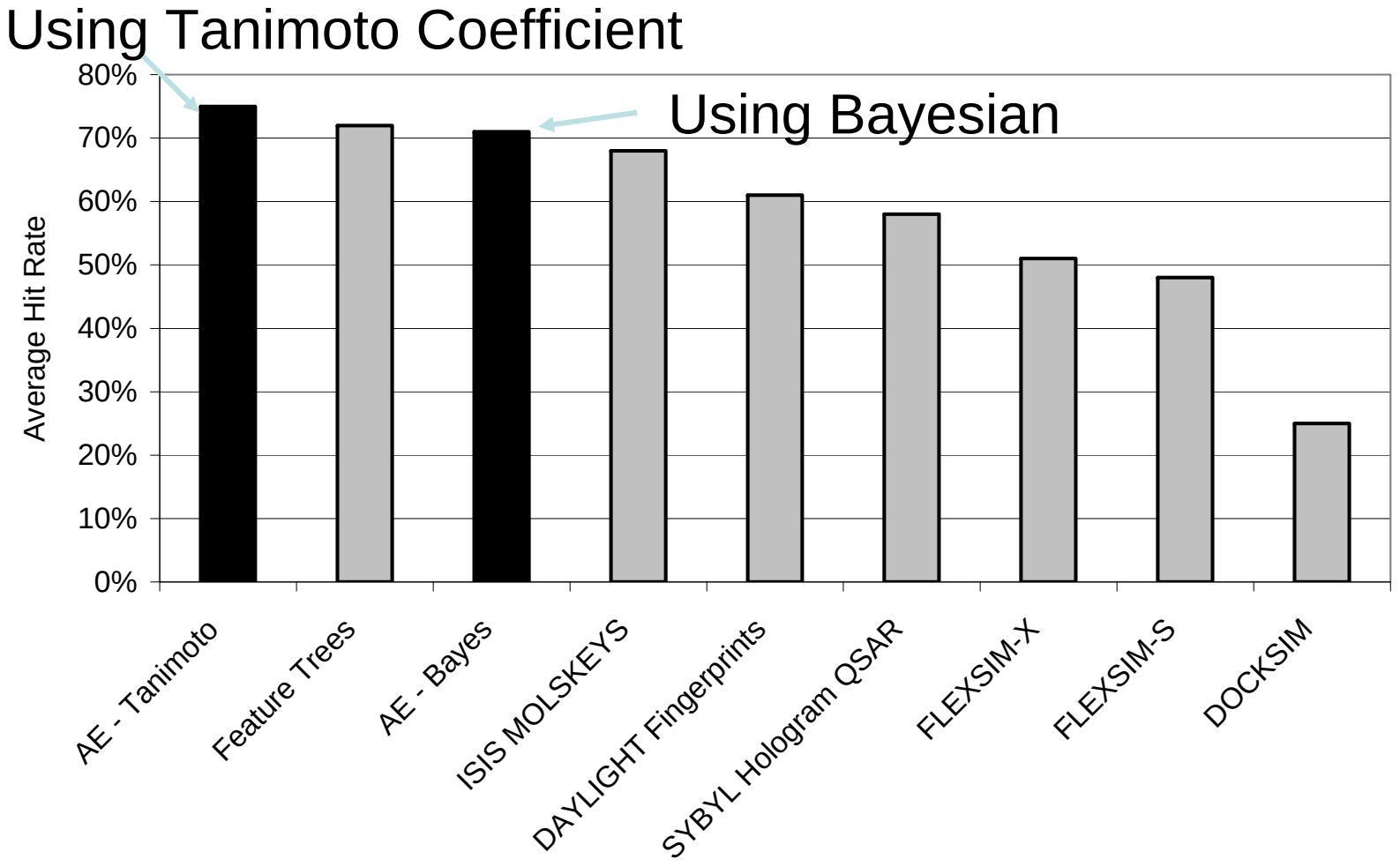
- Database: MDL Drug Data Report (MDDR)
- 957 ligands selected from MDDR

  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

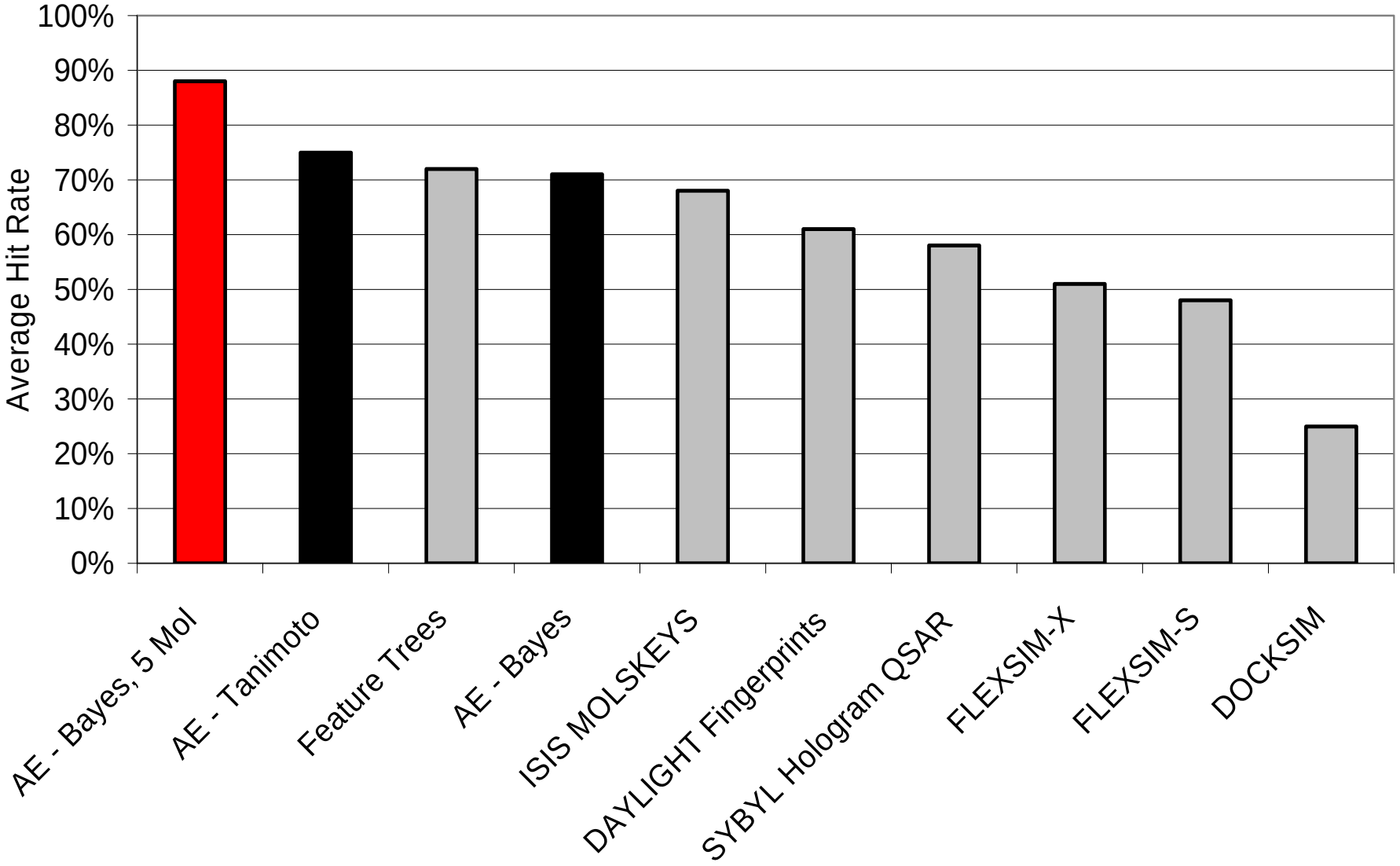


Briem and Lessel, Perspectives in Drug Discovery and Design 2000, 20, 245-264.

Combining Information in Molecules

- In this method, we can extend the approach by extracting from a set of molecules those features having the best information gain
- This can describe patterns in molecules much better than individual cases

Combining Information of 5 "Actives"

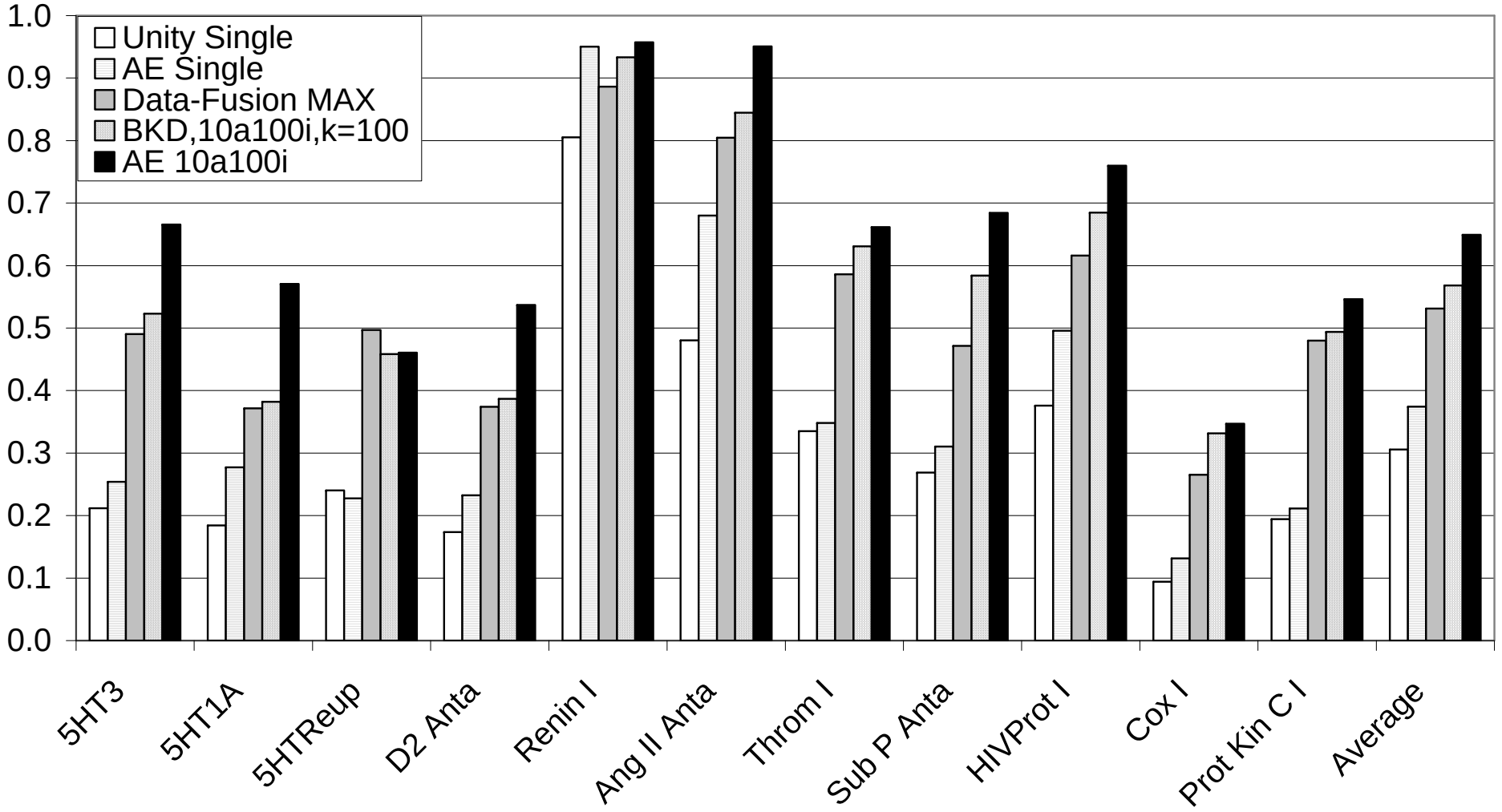


Comparison using Large Data Set *

- 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
- Atom Environments were compared to Unity Fingerprints in Combination with Data Fusion (MAX) and Binary Kernel Discrimination (BKD)
- 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 (ASAP Article)

Comparison of Methods



Conclusions 2D Method

- Atom Environments suitable descriptor, perform well with Tanimoto
- Atom Environments / Bayesian Classifier outperform Unity Fingerprints in combination with Data Fusion and Binary Kernel Discrimination on a Large Dataset -> information fusion prior to screening superior
- Average Hit Rate ~ 10% higher (65% vs. 57%) than the second best method
- Results on diverse targets may imply that method is generally applicable at high performance levels

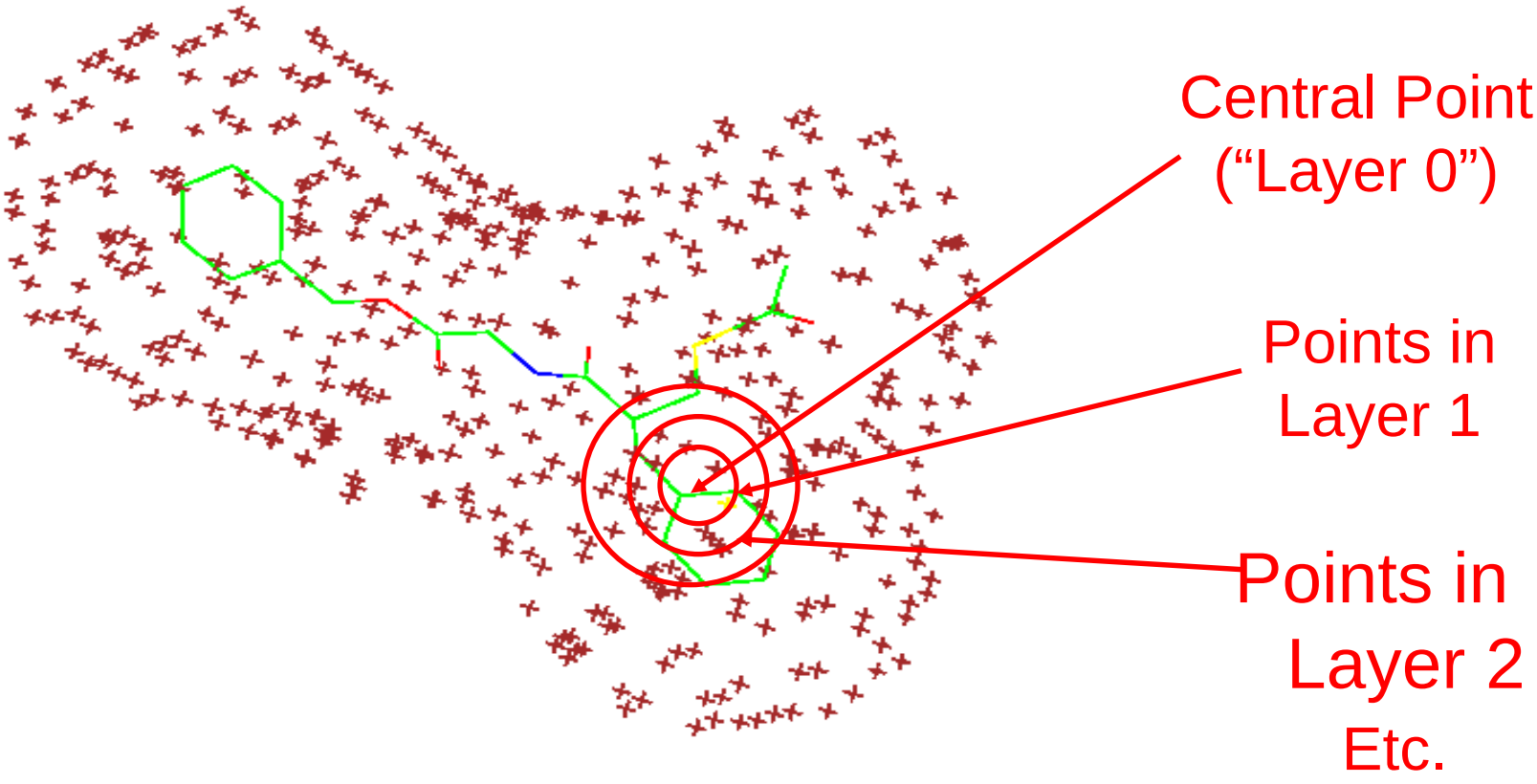
Transformation to 3D

- Idea: To develop an analogous translationally and rotationally invariant (TRI) descriptor based on surface points
- Advantage: Switching from element atom types to interaction energies gives more general model -> scaffold hopping?
- In Addition: *Local* Description hopefully less conformationally dependent
- Approach to Fingerprint Surfaces; Tanimoto and other methods become applicable (until now mainly used for 2D fingerprints)

Transformation to 3D

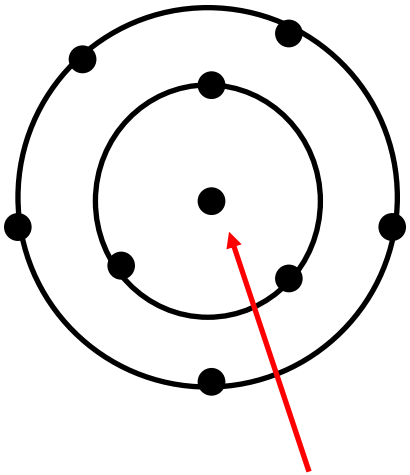
- Two parts: Interaction fingerprint and shape description; here results using only interaction fingerprints are shown, shape description under development
- Information was merged from multiple molecules by using information-gain feature selection and the Naïve Bayesian Classifier

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

Relation to other algorithms

- Surface Autocorrelation: Averaging of interaction energies – Here a favourable and unfavourable interaction in a given layer will both remain in the fingerprint
- GRIND: continuous variables from GRID; entire field of interaction energies; simplified; only maximum product enters descriptor
- MaP: categorical variables, counts are kept – size description
- (In addition the feature selection and scoring are handled differently)

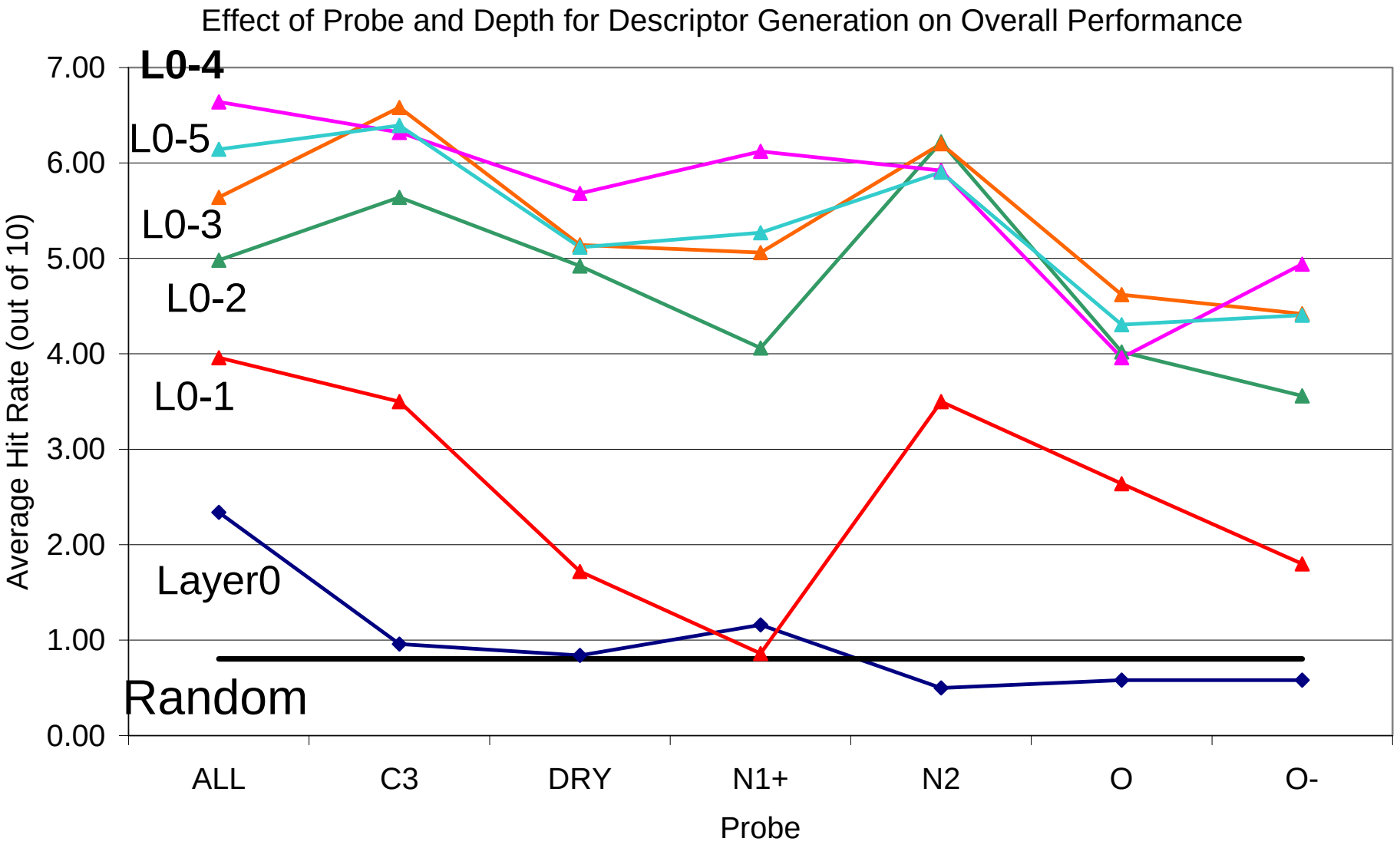
Algorithm Flow

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

Standard Parameters

- MSMS: Probe radius 1.5 Å, Density 0.5-2.0 Points/ $Å^2$, double Van-der-Waals radii for atoms, giving effectively solvent accessible surface
- GRID: DRY, C3, N1+, N2, O, O- probes, otherwise standard parameters
- Binning: Using variable number of layers, 8 bits, cutoffs were set that equal frequencies are observed

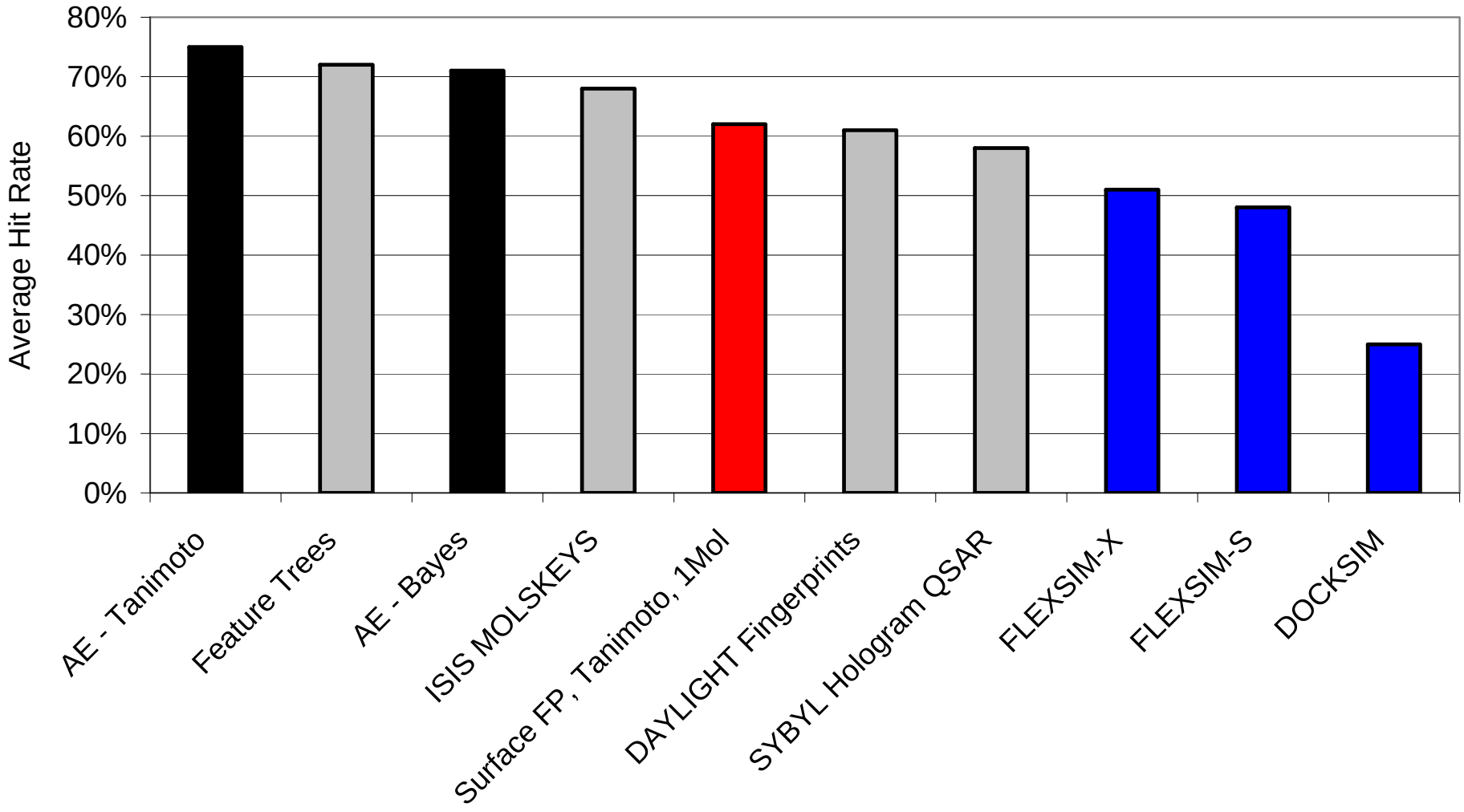
Parameterisation – Effect of Probe Type and Number of Layers (Briem Dataset, 5 Actives)



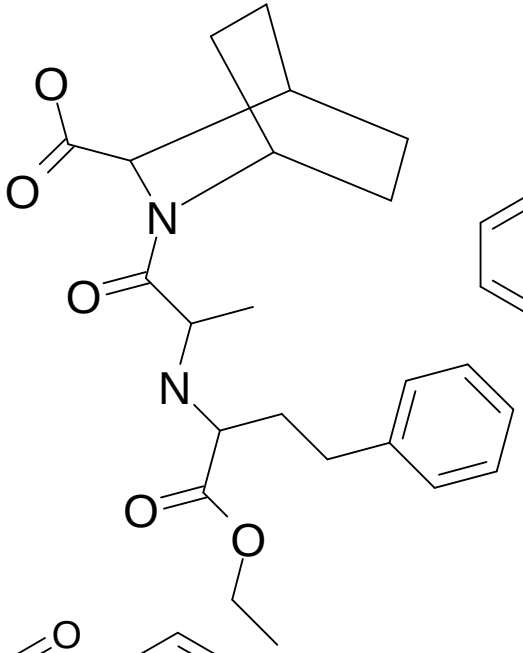
Surface Fingerprints & Tanimoto

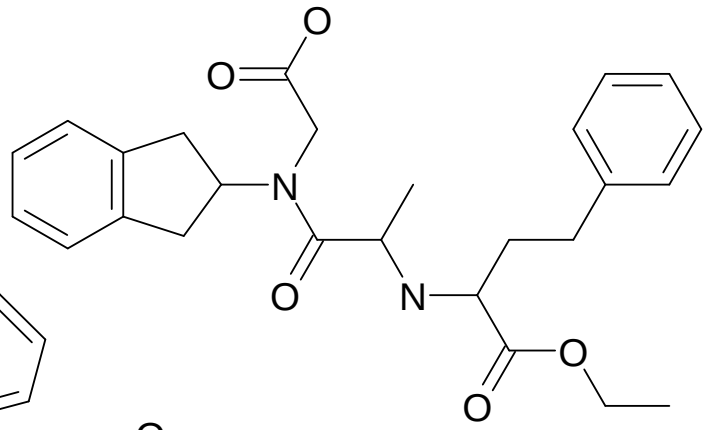
- Tanimoto coefficient used for 2D fingerprints in combination with a variety of descriptors, here applied to surfaces
- Random Selection of single active compounds from MDDR dataset
- Calculation of average hit rates of Top 10 list for whole dataset (5HT3, ACE, HMG, PAF, TXA2)
- Question: Is “scaffold hopping” observed?
- Examples: ACE, TXA2

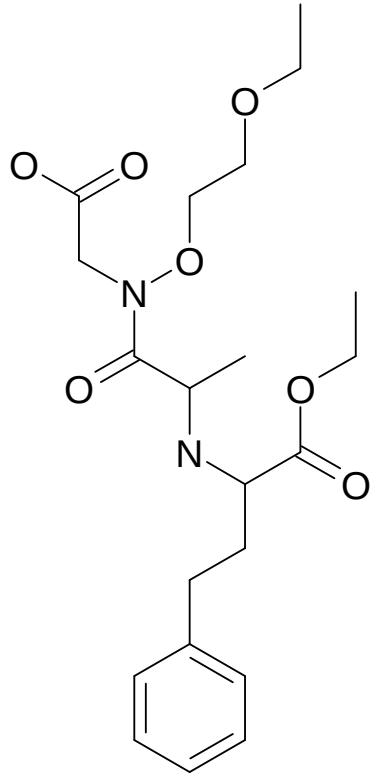
Overall Performance Comparable to 2D methods

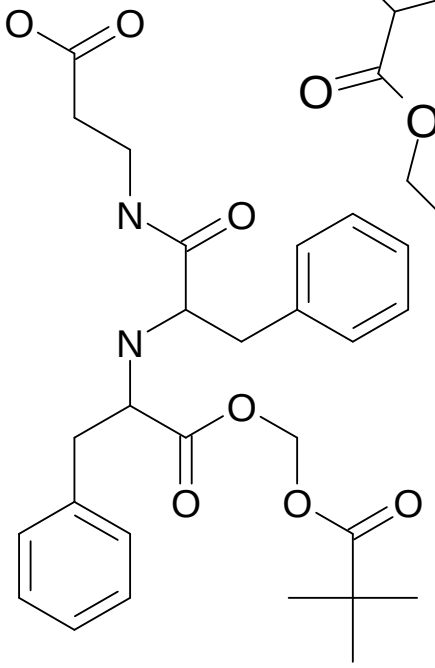


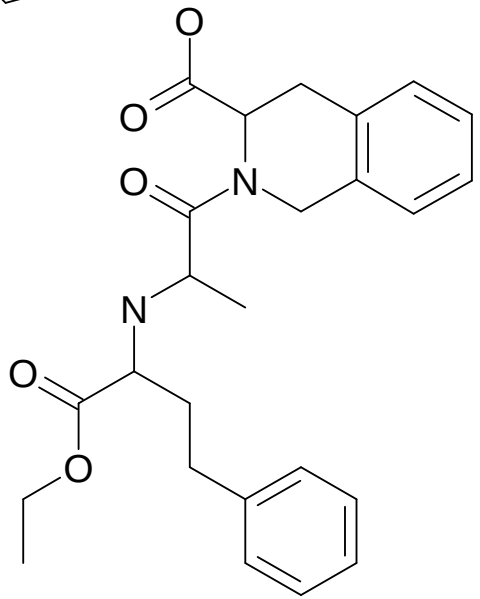
Example: ACE, Query, Actives Found in Top 10, sorted

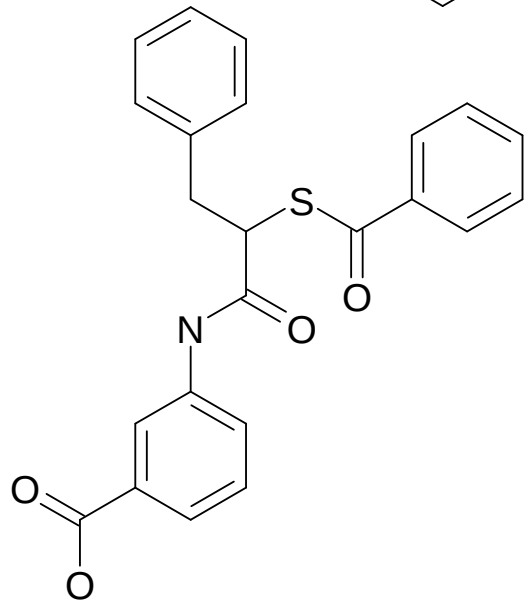




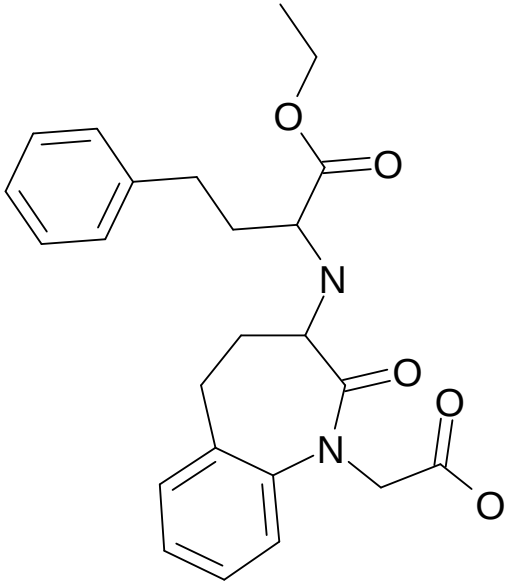


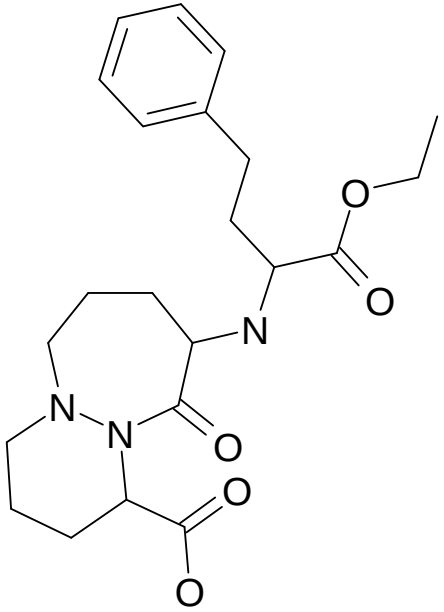


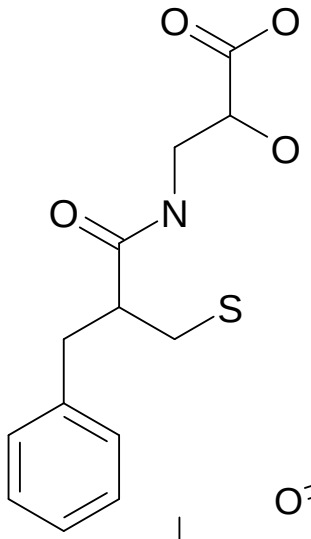


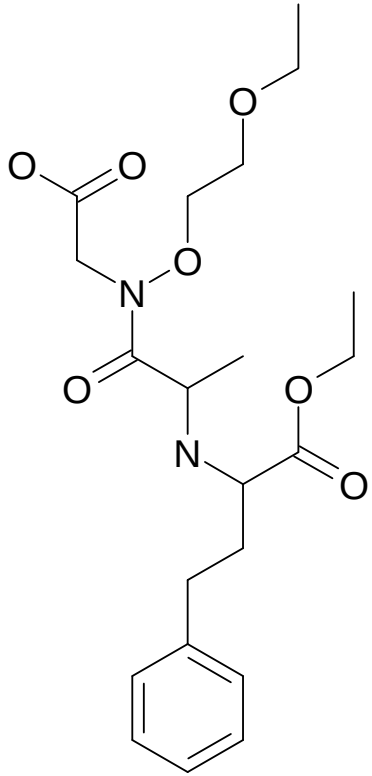


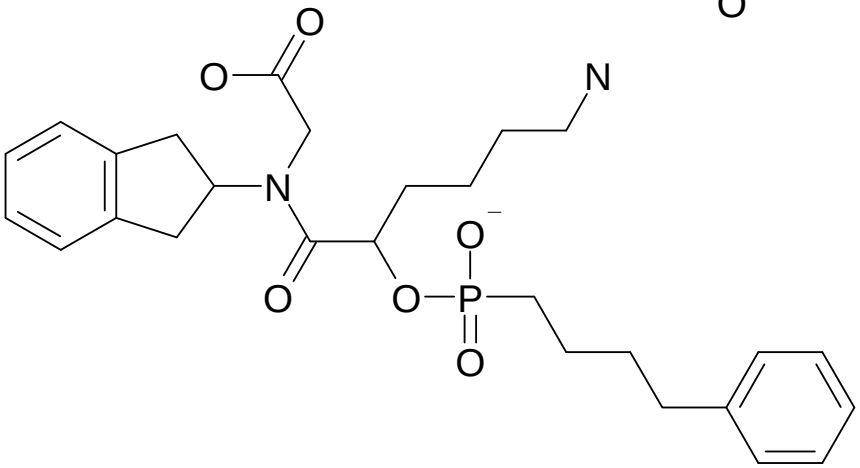
Example: ACE, Query, Actives Found in Top 10, sorted

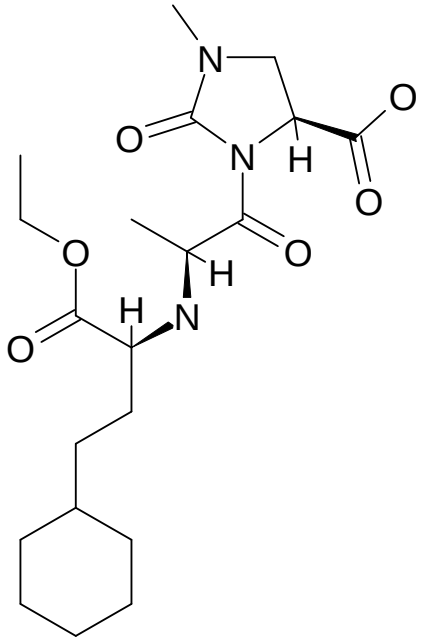




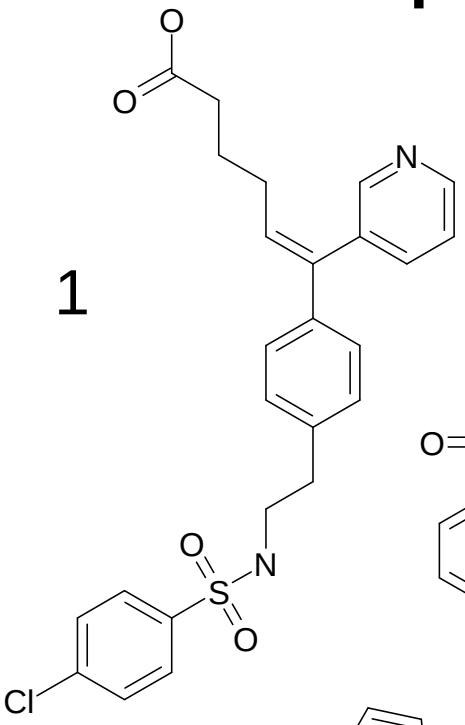


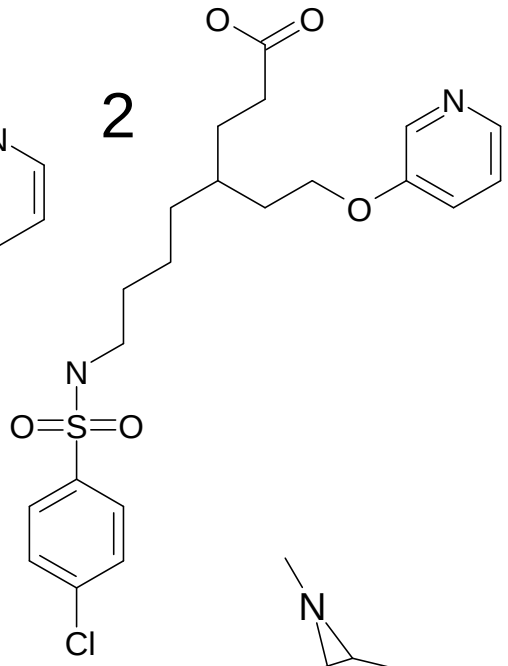


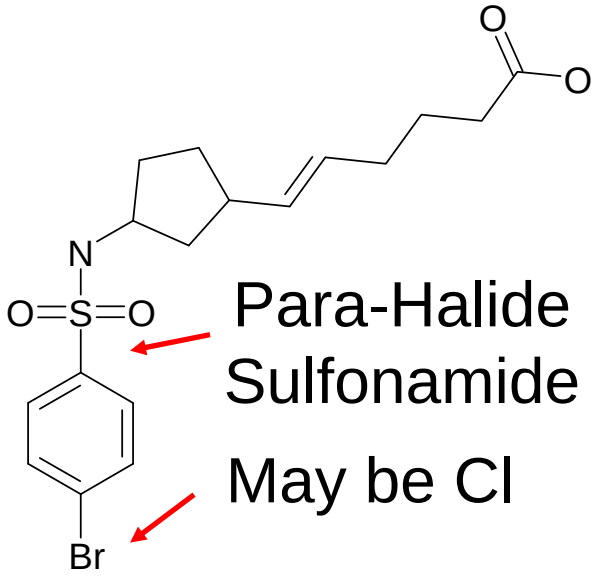




TXA2, 10 Hits among Top 10 (Sorted)

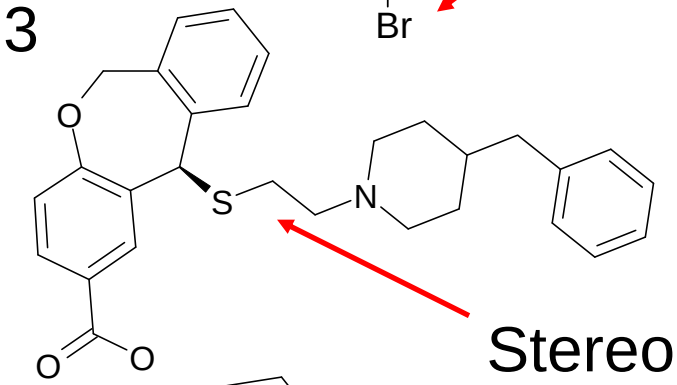




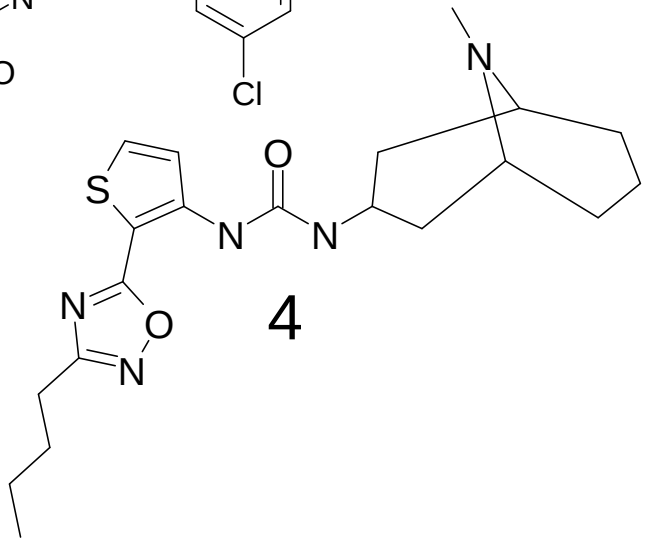


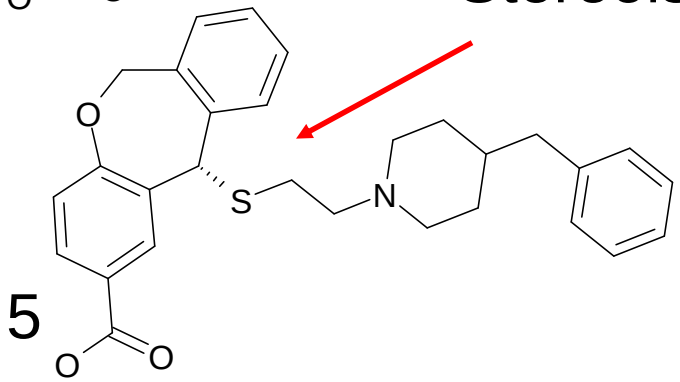
Para-Halide Sulfonamide

May be Cl

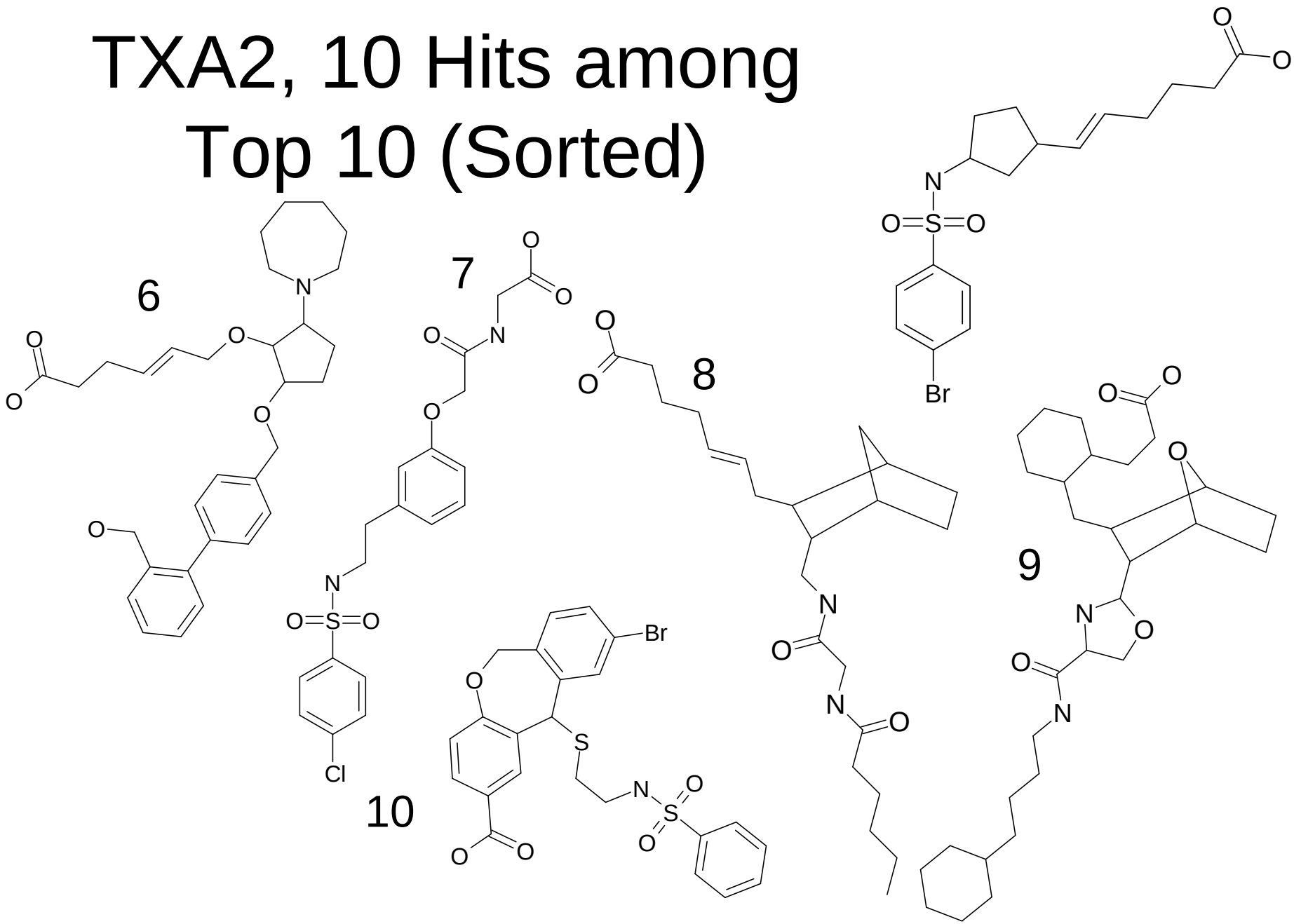


Stereoisomers

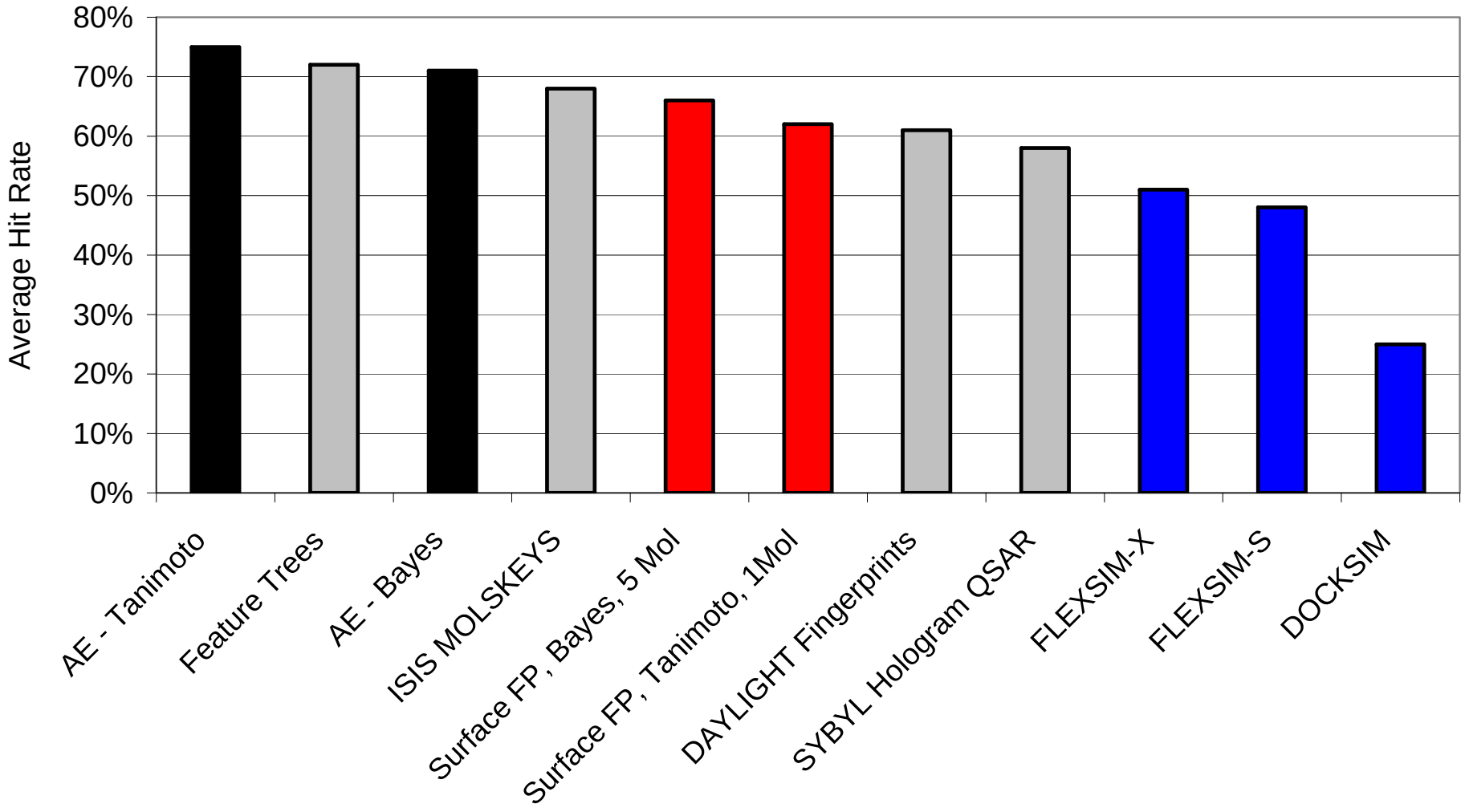




TXA2, 10 Hits among Top 10 (Sorted)



Surface Environments – Merging Information



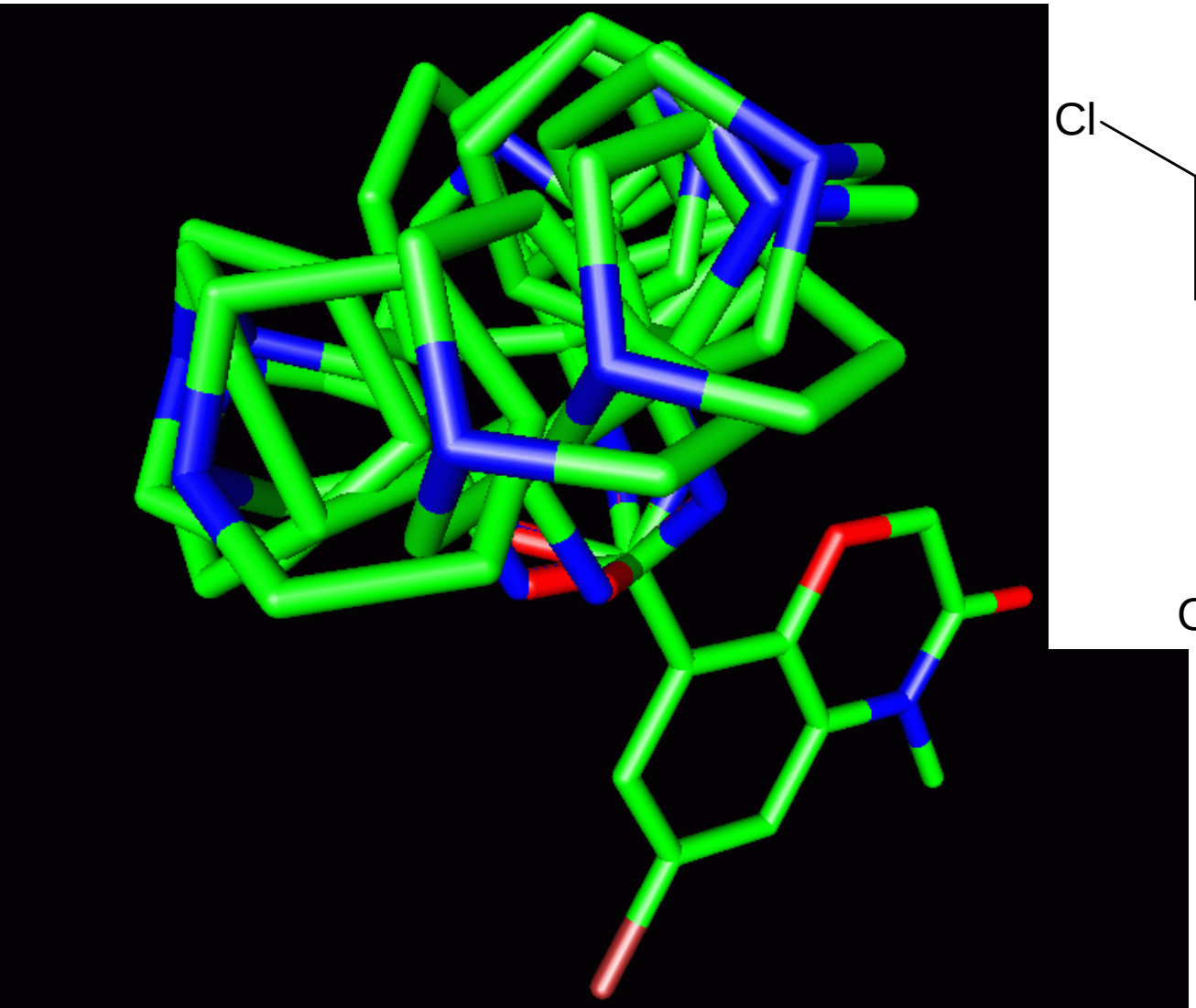
Conformational Variance

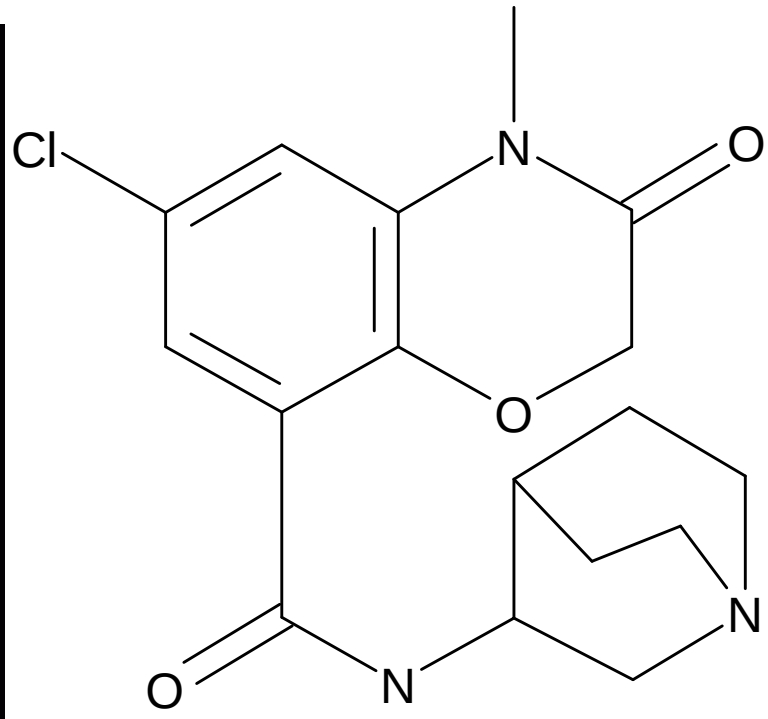
- MDDR Dataset (5HT3, ACE, HMG, PAF, TXA2)
- 10 Randomly selected compounds each
- 10 Conformations generated by GA search with large window (10° for rigid 5HT3, 100° for ACE, HMG, PAF, TXA2), giving diverse conformations
- One force field optimized conformation (Concord-generated) used to find other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

Overall findings

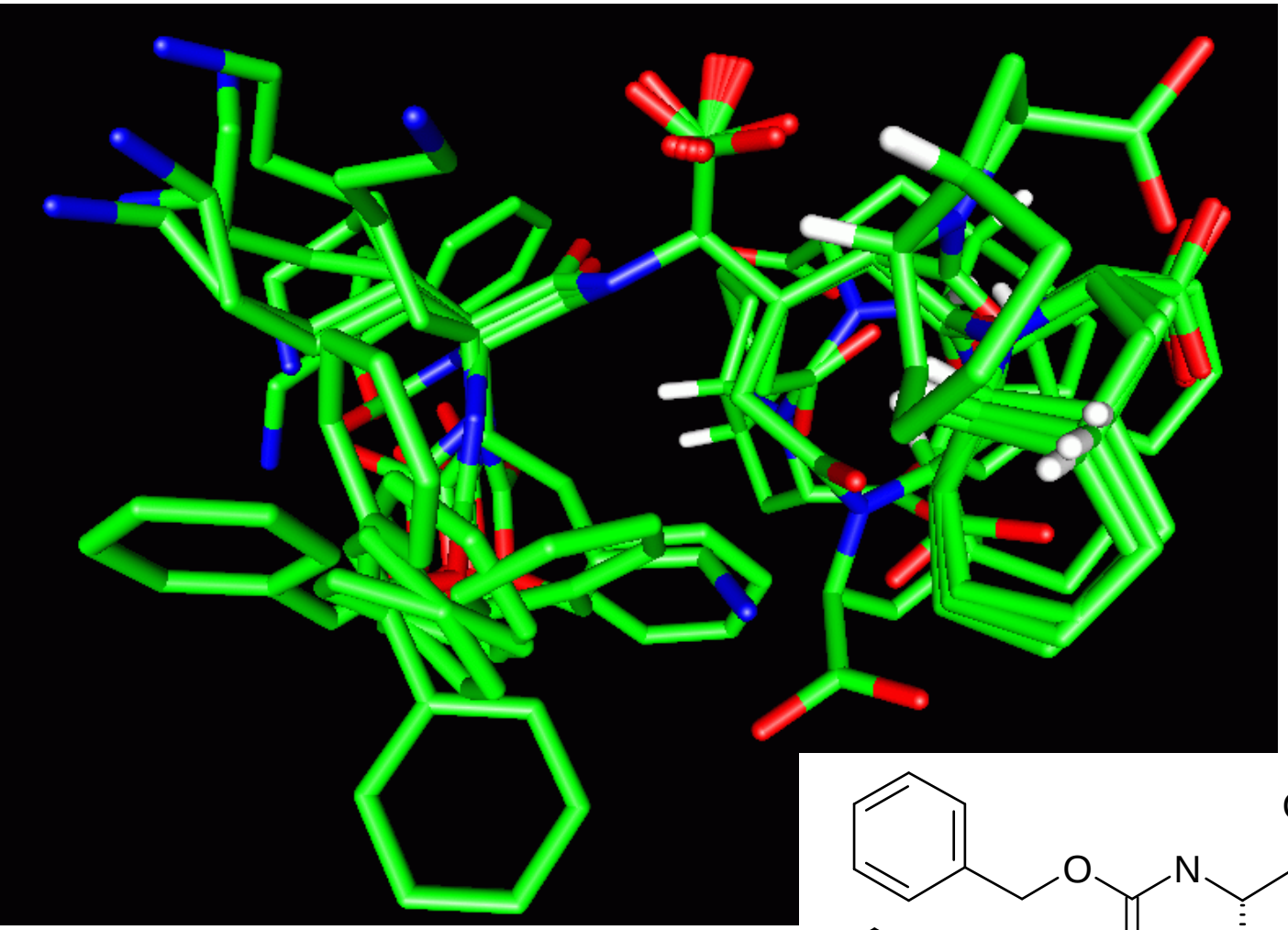
- 64% of conformations found at the top 10 positions -> 2/3 of compounds identified as being most similar (among list of > 900 structures and 40-134 structures of same active dataset)
- >90% of conformations found in Top 5% of sorted database
- Conclusion: If molecules with the right features are present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

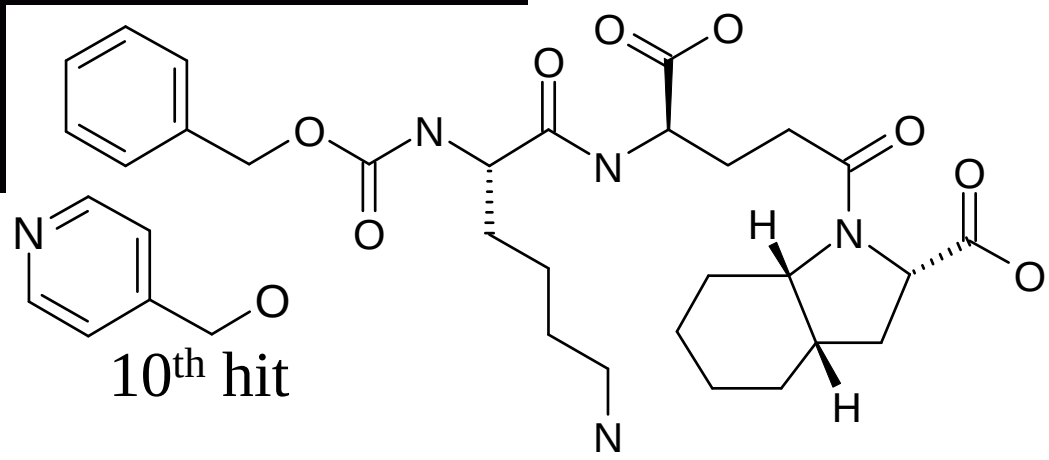
Example: 5HT3-0 (Rigid) – all 10 Conformations identified as identical





ACE-7 – 9 Conf. identified as identical



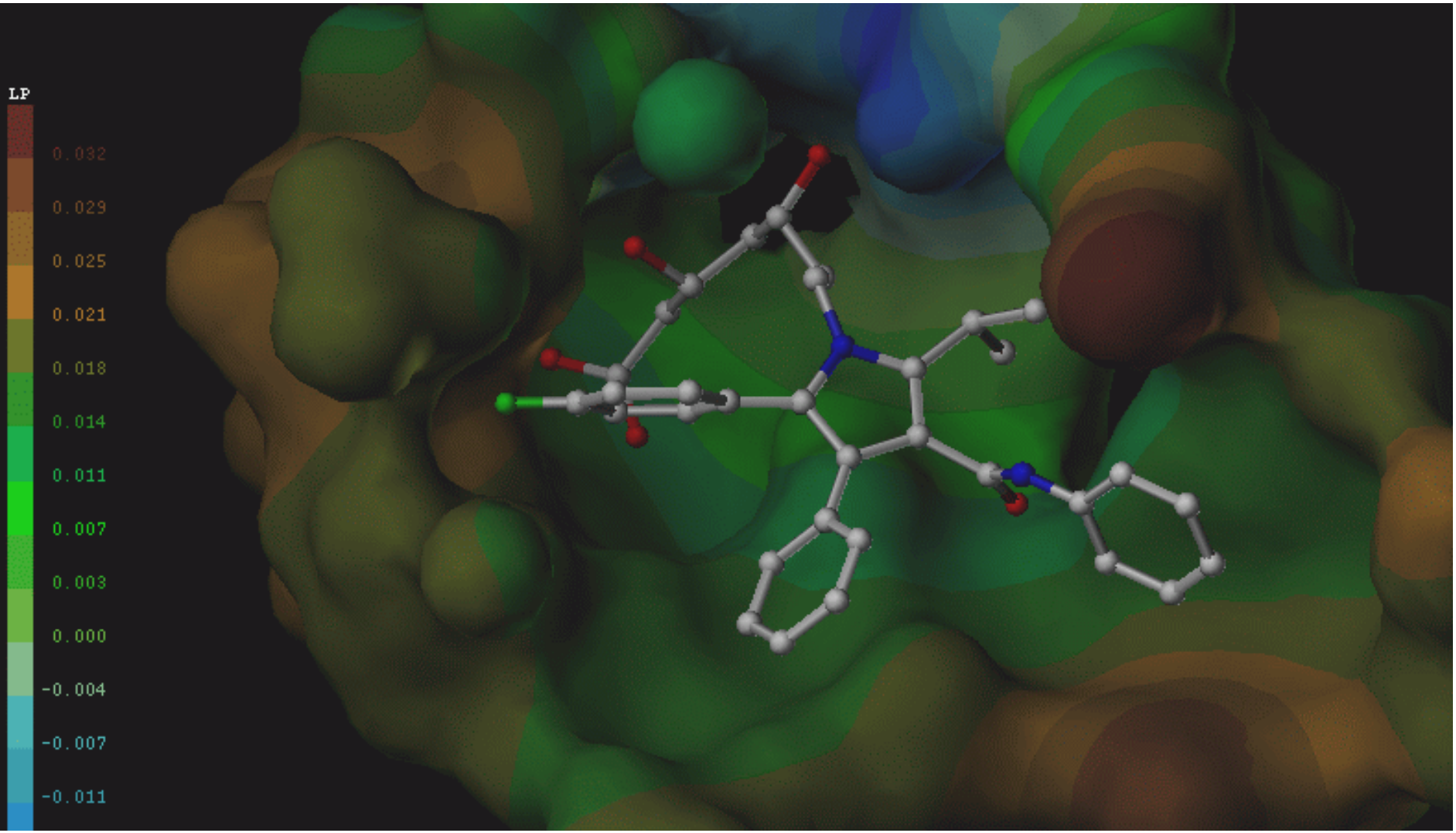


10th hit

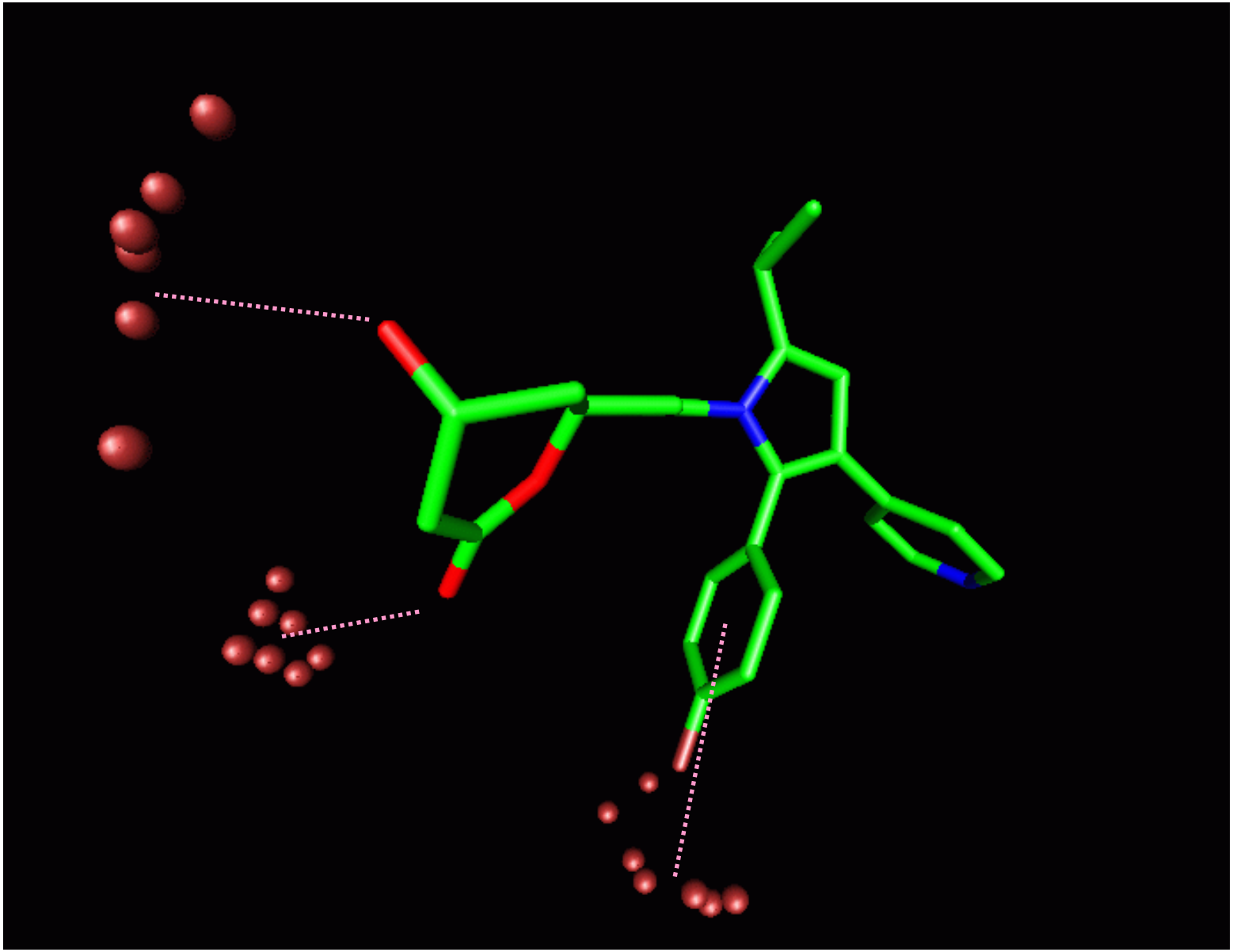
Which features are selected for classification?

- 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

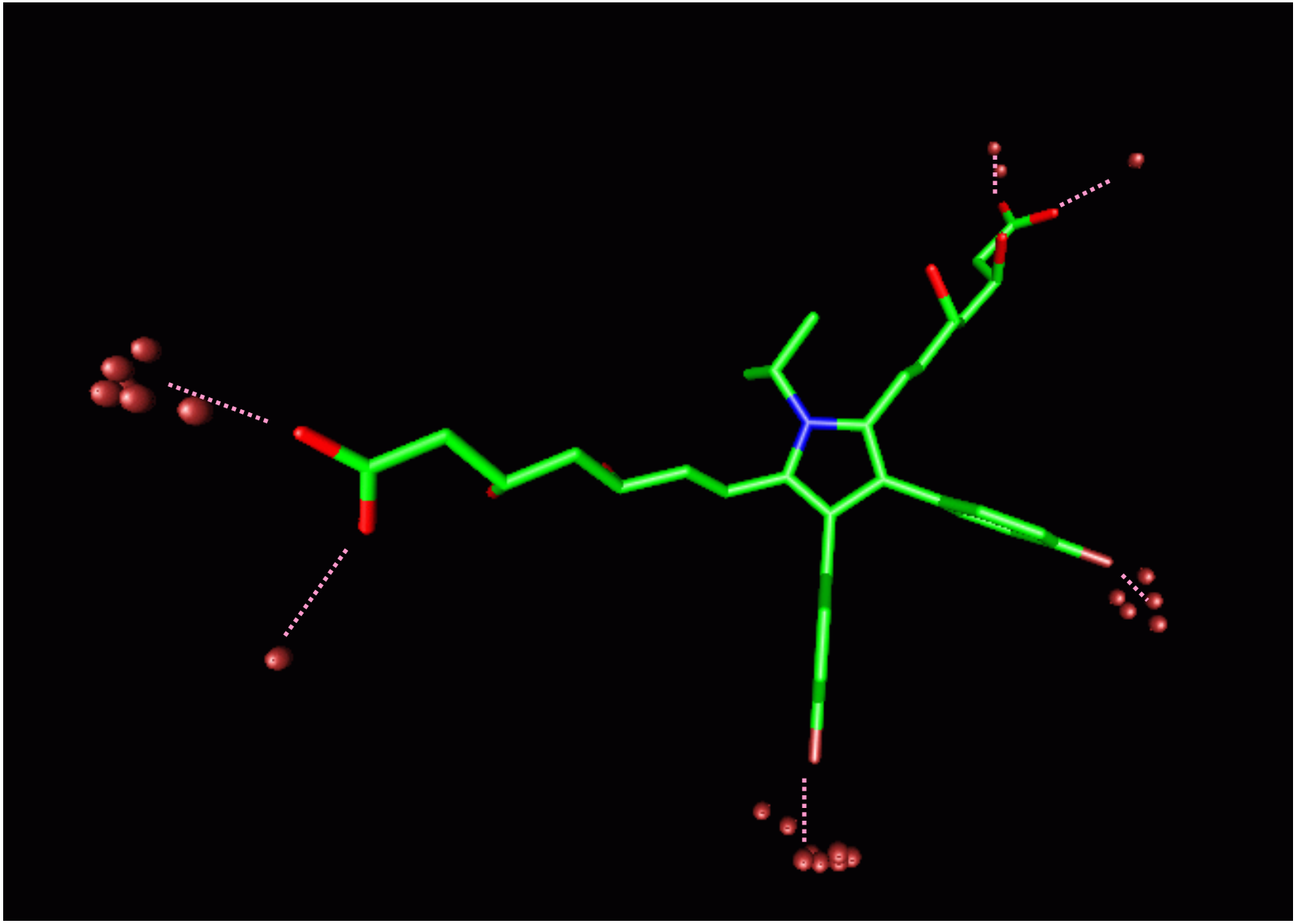
Selected Features - HMG



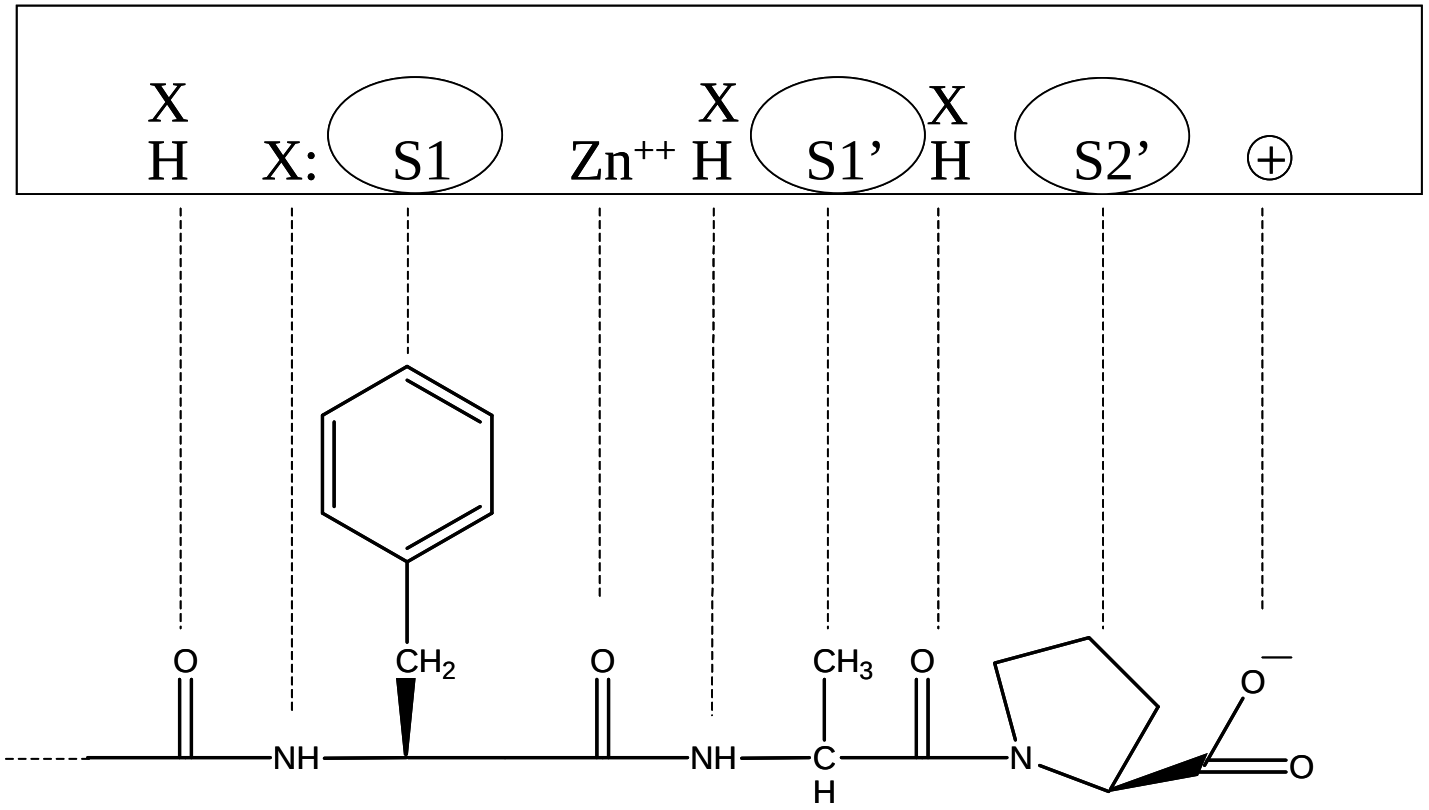
HMG-15



HMG-19

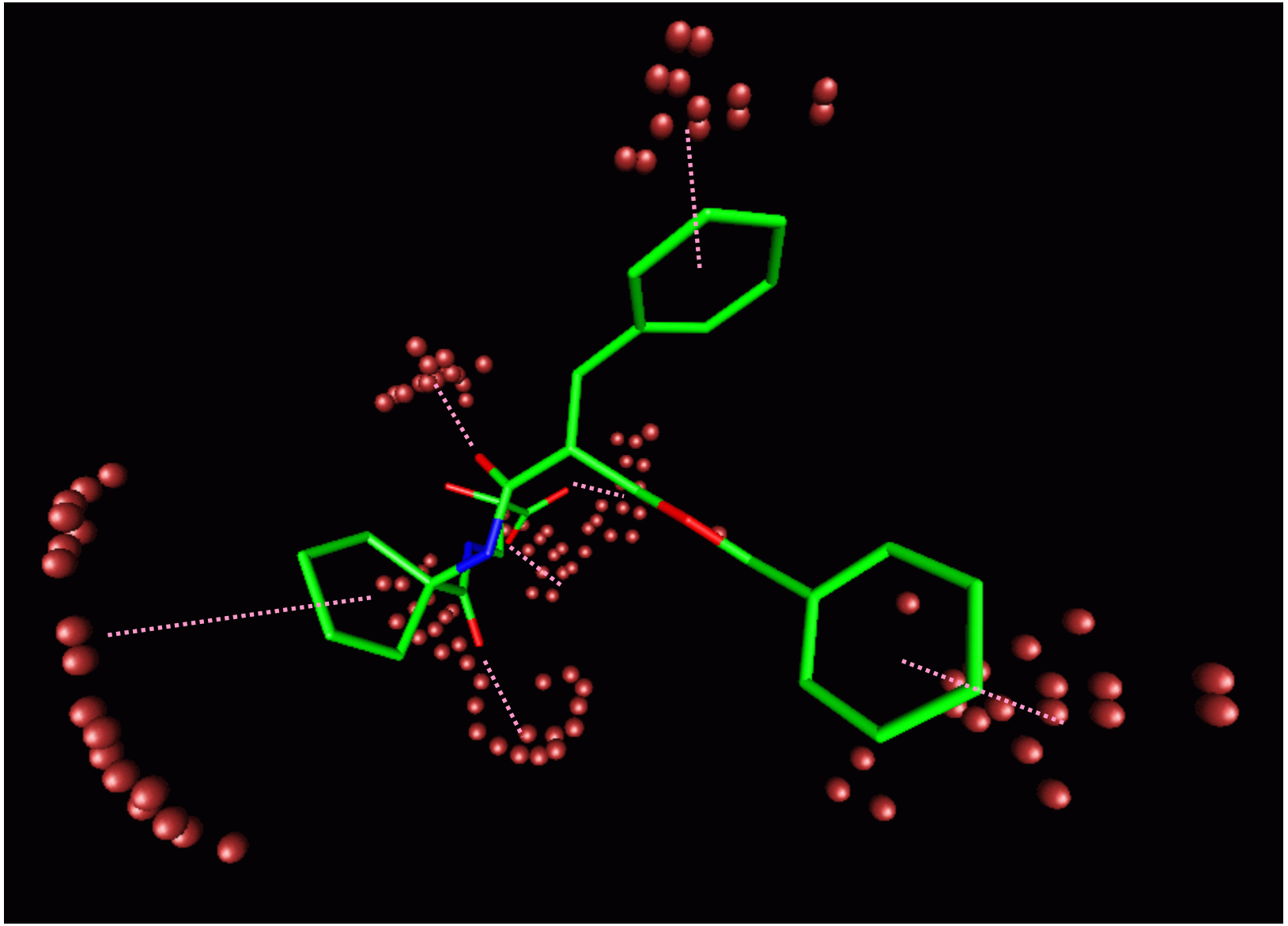


ACE – Binding Site

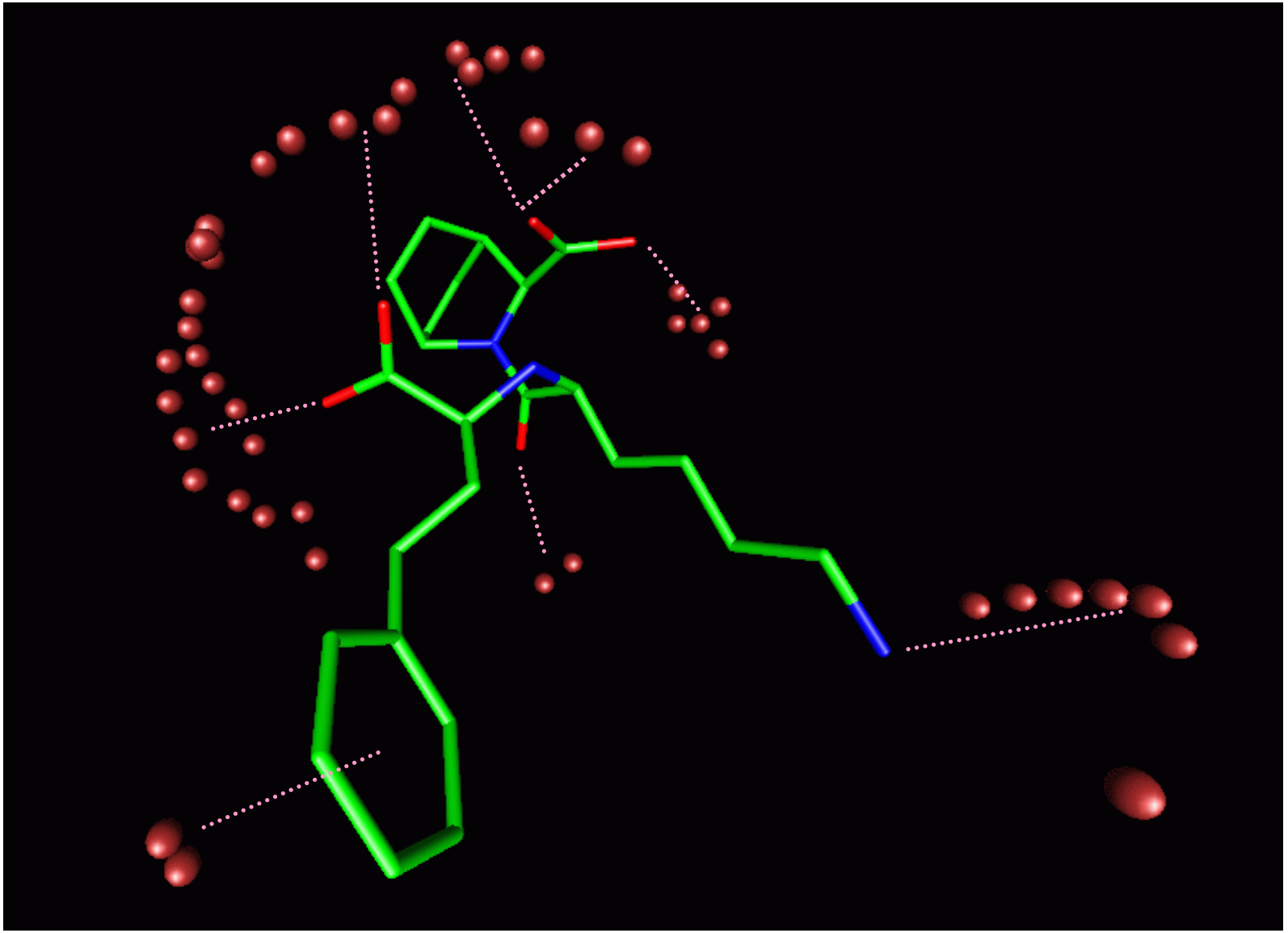


Snake venom peptide analog with putative binding motif to angiotensin used in early compound design (Cushman et al., Biochemistry (1977), 16, 5484-5491.)

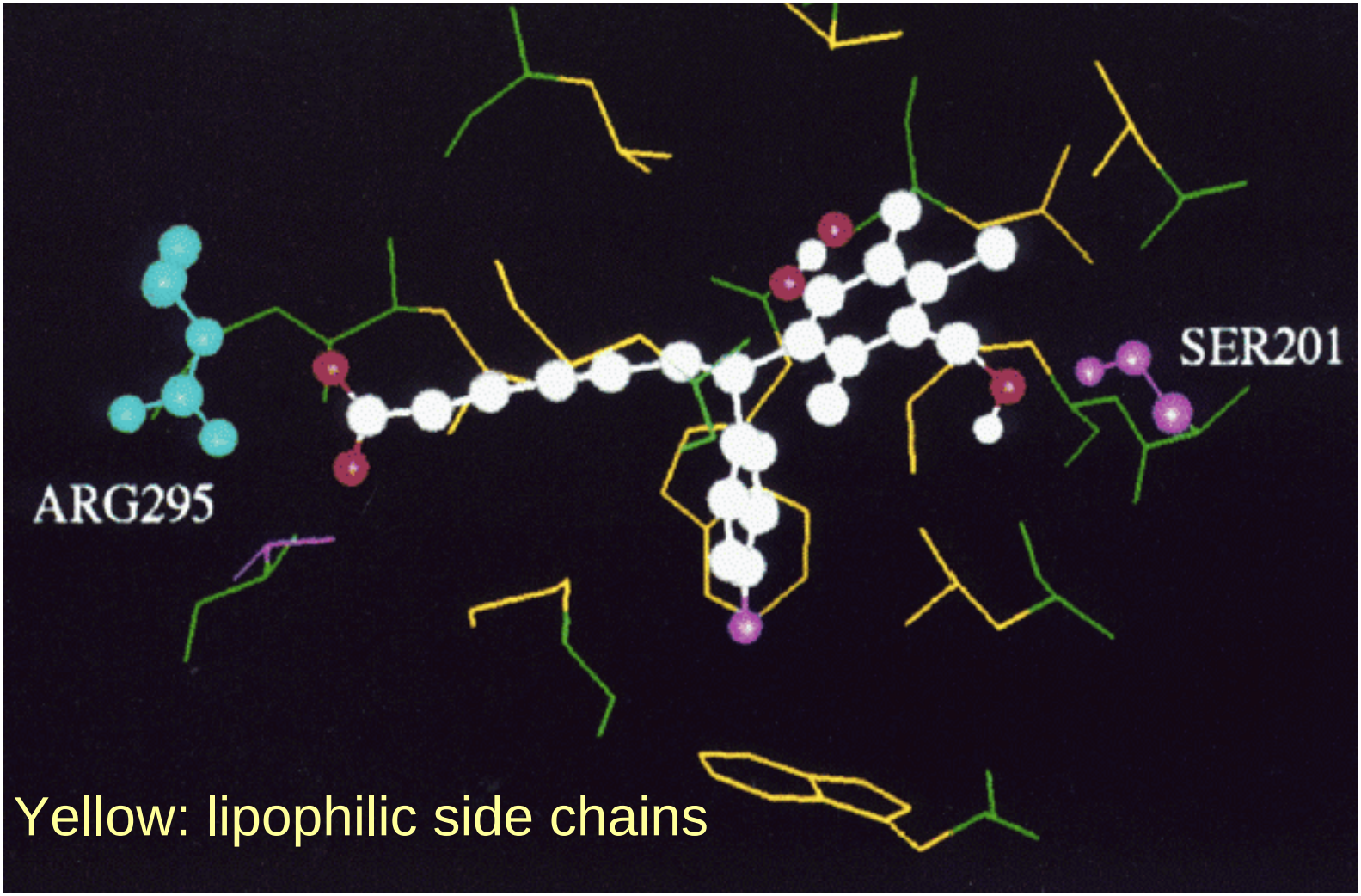
Selected Features – ACE-31



Selected Features – ACE 39

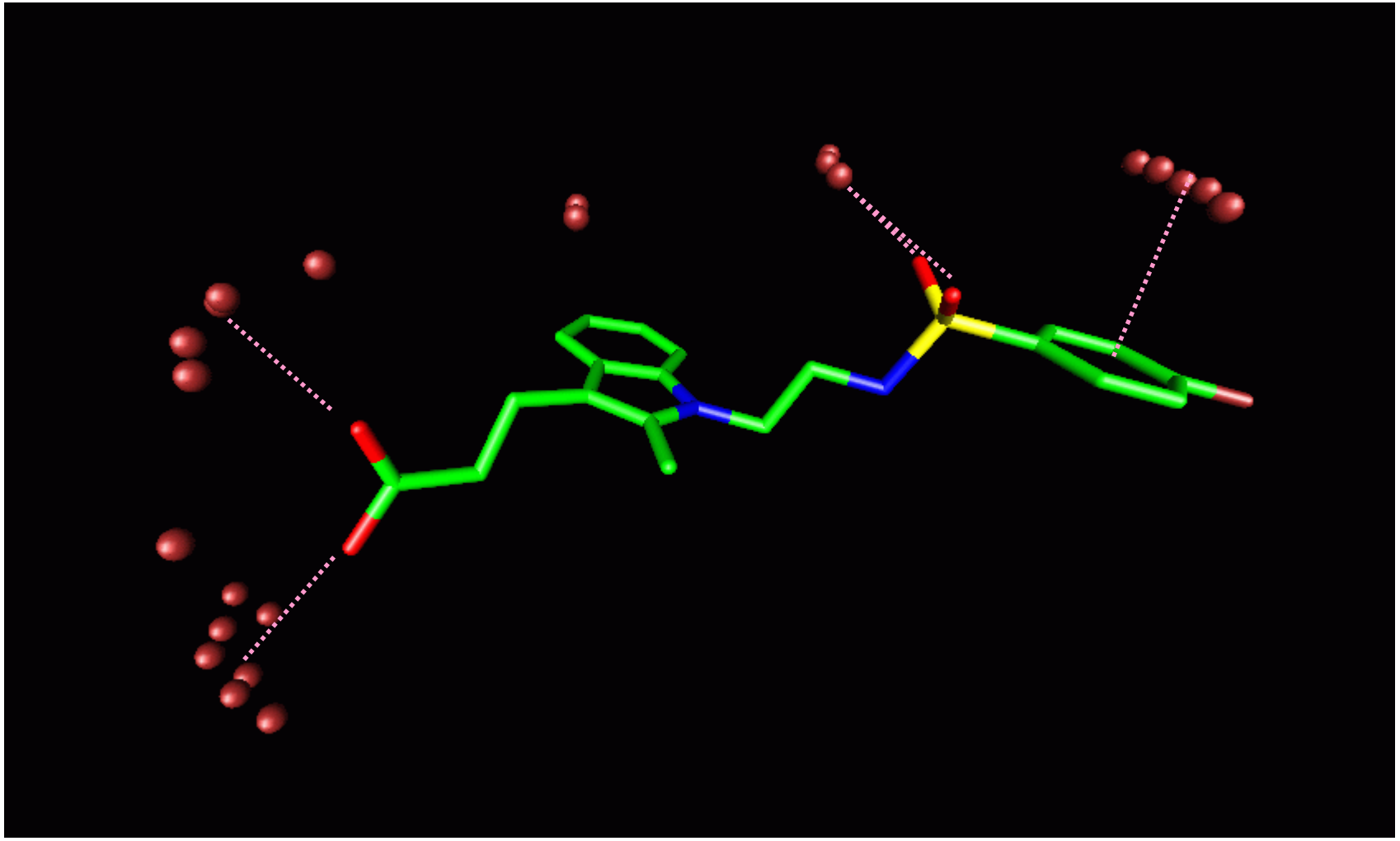


TXA2

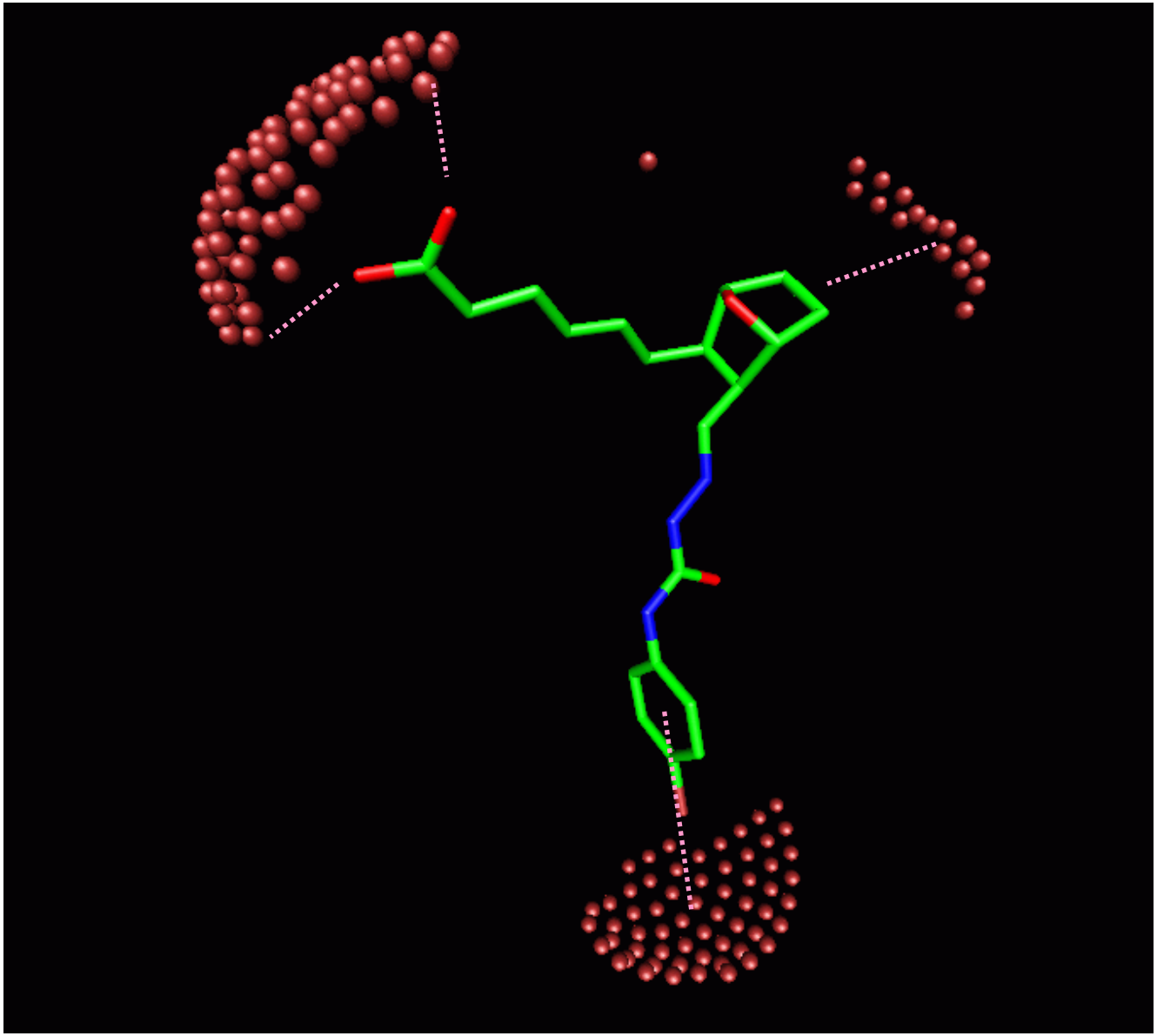


Yamamoto et al., J. Med. Chem. 1993 (36) 820

TXA2-44



TXA2-7



Summary

- 2D Method: Performs about as other 2D methods for single molecule searches, outperforms them by a large margin when combining information from multiple molecules (published in *J. Chem. Inf. Comput. Sci.* (2004) 44, 170-178)
- 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 and active *and* inactive information into account

Course Outline

- Introduction and Case Study
- Drug Targets
  - Sequence analysis
  - Protein structure prediction
  - Molecular simulation
- Molecular Docking
- Drug Design
  - QSAR
  - Pharmacophore
  - De novo Drug Design
  - Combinatorial library