Summer School on Chemoinformatics, Obernai, 2008

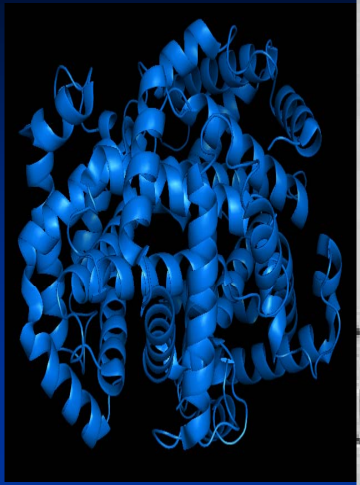
Lessons learned from modelling bioactivity - what works and what doesn't

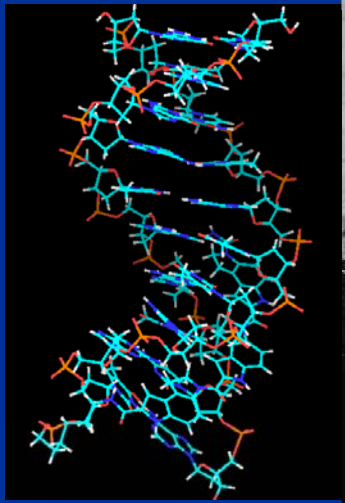
- dynamic pharmacophores

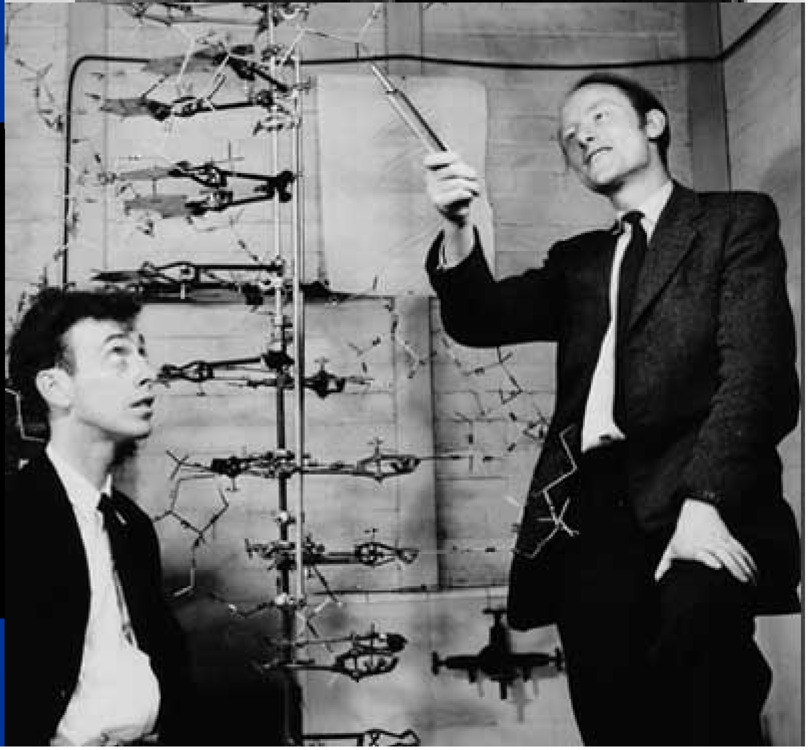
- Property vs. activity data models

They knew about ‘chemoinformatics’ before the word was coined

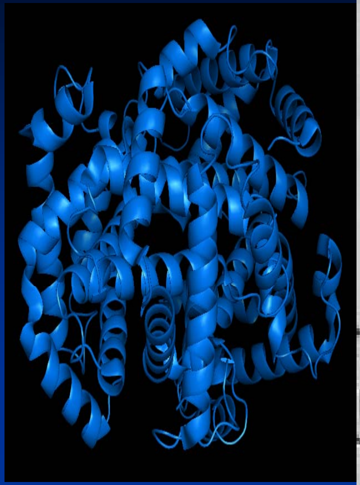
Max Perutz and John Kendrew admire the structure of haemoglobin, and Watson and Crick with DNA.



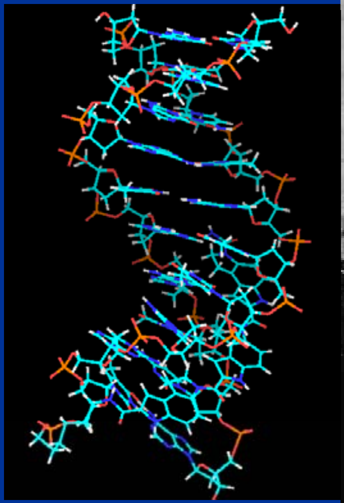


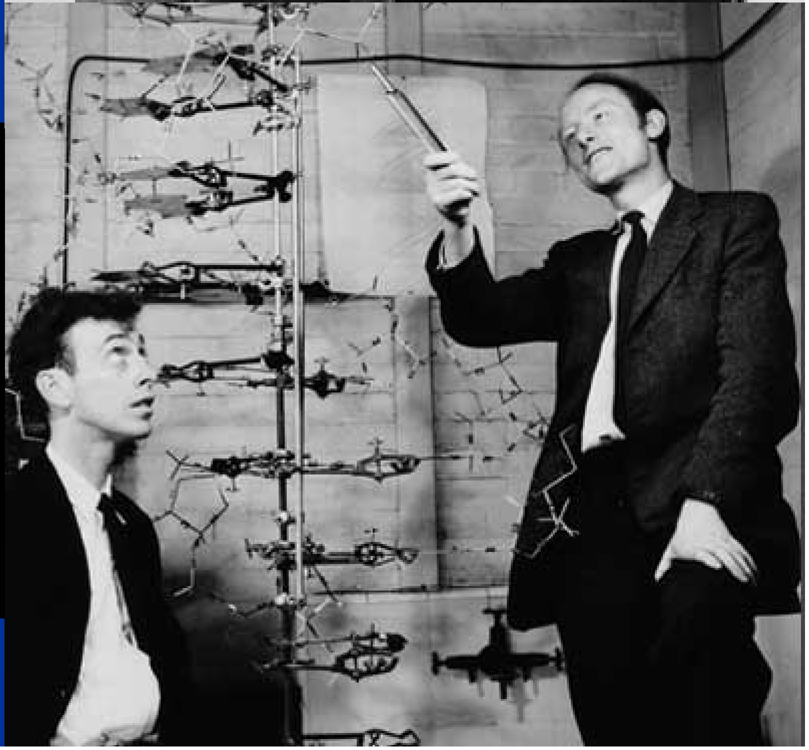


They knew about ‘chemoinformatics’ before the word was coined



Max Perutz and John Kendrew admire the structure of haemoglobin, and Watson and Crick with DNA.





The fundamental idea is that the use of models, built from experimental data and theory, can profoundly influence our philosophy of science

– this is what we spend most of our time doing with chemical data, and it’s easier now with computers – especially as data is available as never before…

Dynamic pharmacophores in the 5HT1b/d GPCR

James Bell and Robert Glen

University of CAMBRIDGE

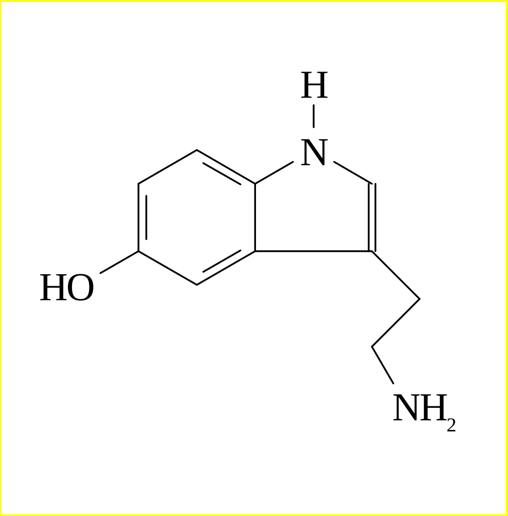
Unilever Cambridge
Centre for Molecular Informatics

Contents

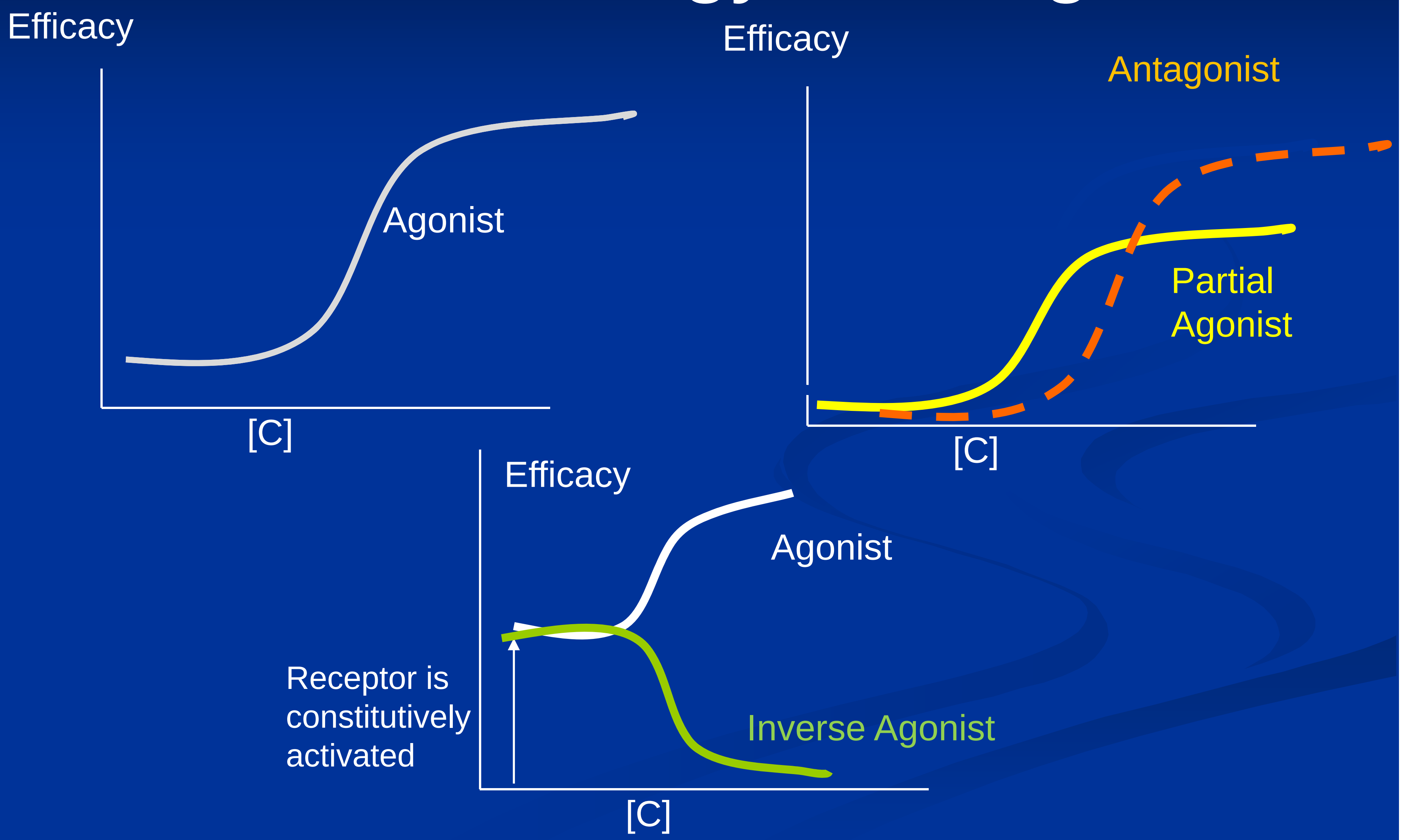
- An introduction to designing 5-HT_{1B} ligands
- 5-HT_{1B} pharmacophore development
- Homology modelling based on the beta-2 crystal structure
- Dynamic pharmacophores
- Efficacy models

5-HT Receptors

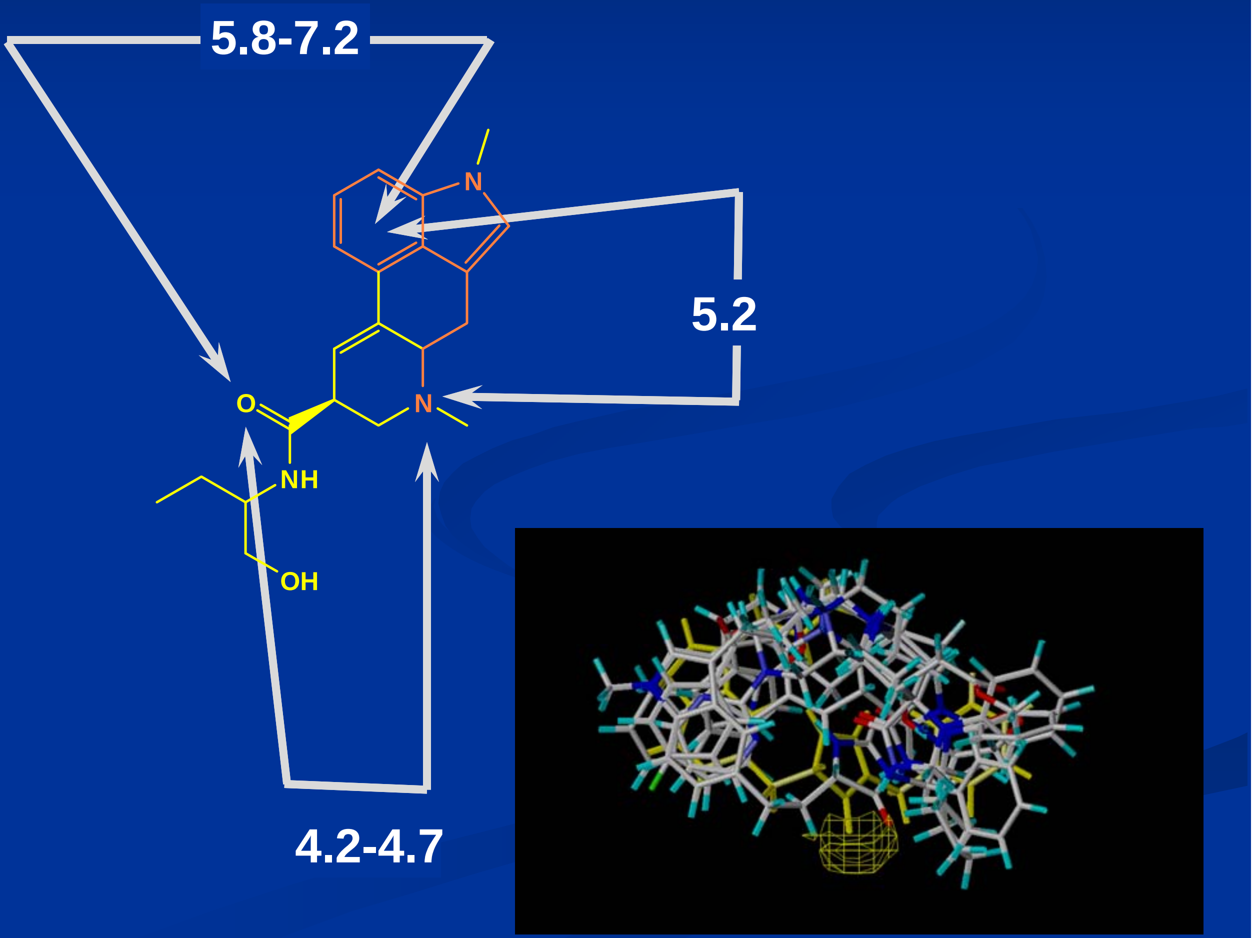
- Bind 5-hydroxytryptamine, 5-HT (Serotonin)
  - Natural hormone
  - Controls mood, muscle tone, blood pressure etc.
- 7-transmembrane helix structure, Monoamine G-coupled protein receptor (GPCR)
  - Cell signaling device
- 14+ receptor subtypes (5-HT_{1-7})
- Excellent drug targets
  - Migraine, depression, blood pressure



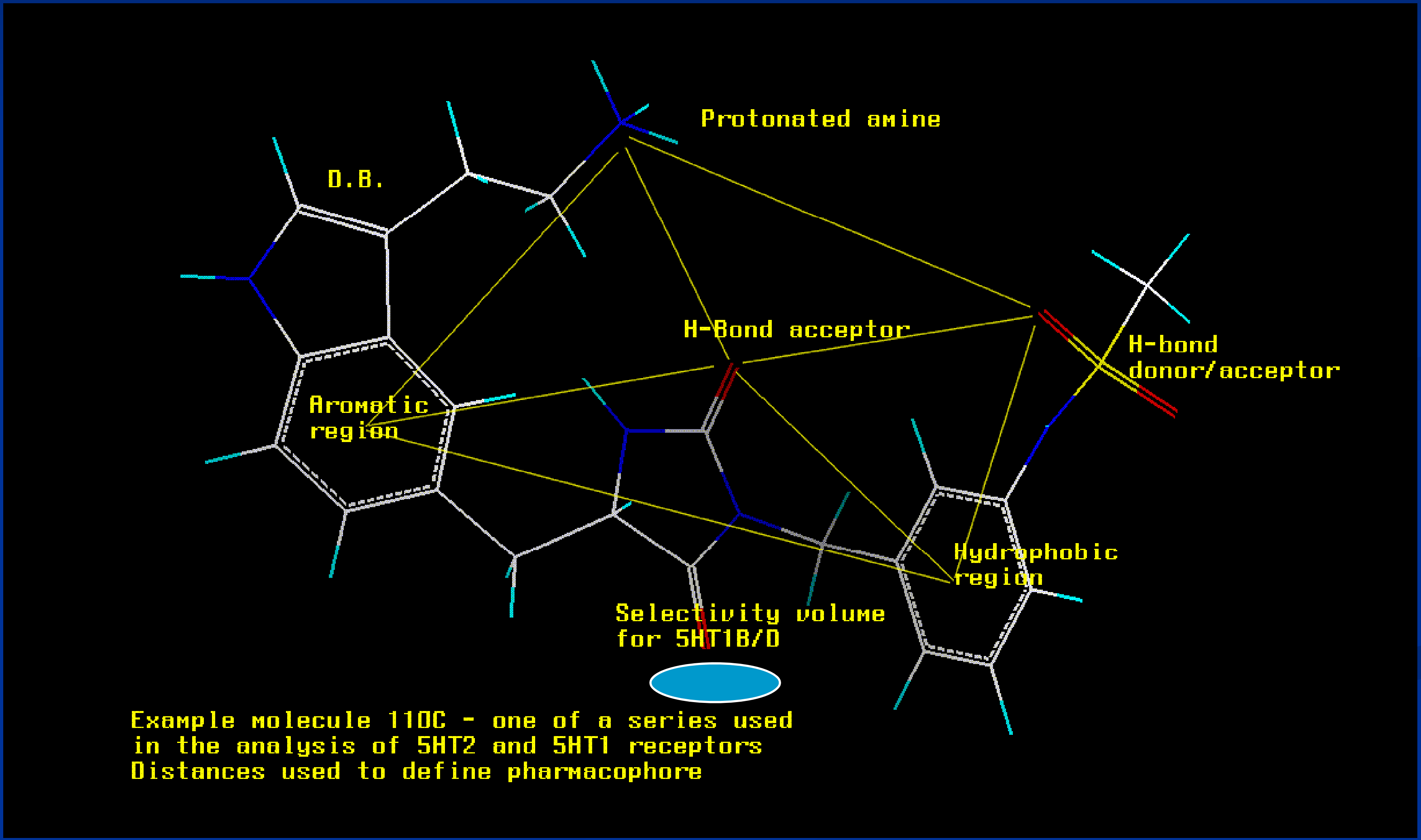
Dummy’s guide to the Pharmacology of Drugs



Methysergide – constrained analog
Useful to help determine pharmacophoric points
Approximate heteroatomic distances for low energy
conformers of methysergide (Angstroms) – plus many analogs, gives rise to a pharmacophore



End point – a topological model for affinity – distances between binding groups used to fit old and new structures



Example molecule 110C - one of a series used in the analysis of 5HT2 and 5HT1 receptors
Distances used to define pharmacophore

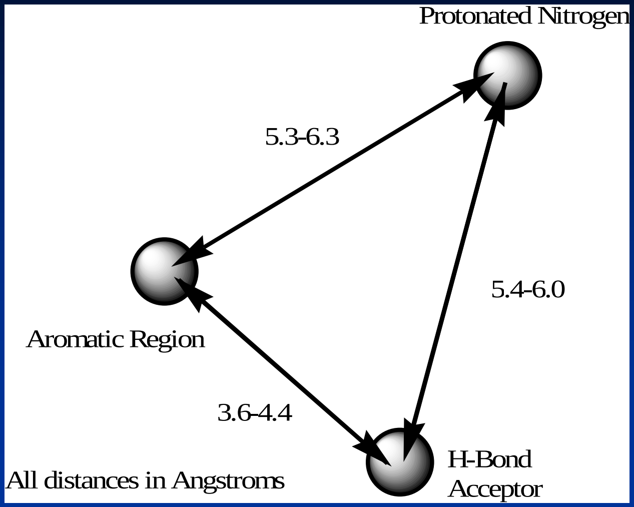
Glen, R. C. et al. *J Med Chem* **1995**, *38*, 3566-3580.

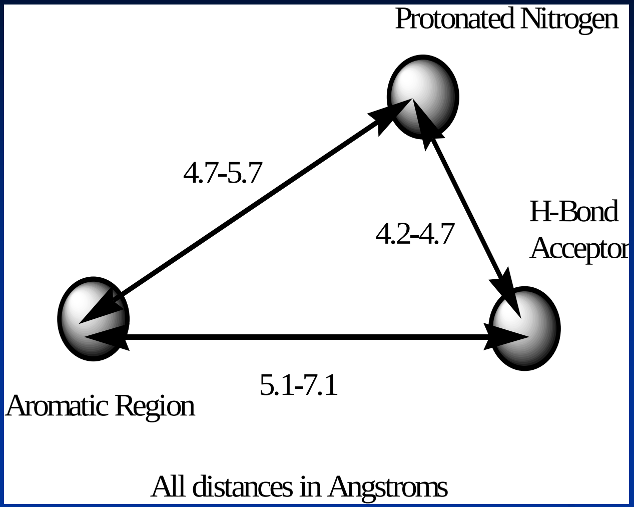
Evaluating the Pharmacophore

- Pharmacophore successfully used to develop anti-migraine drug Zolmitriptan
- Strong evidence that the amine to aromatic ring distance is accurate
- However there are several issues:
  - Most subsequent 5-HT$_{1B}$ ligands don’t fit well
  - Several binding regions were inferred from very flexible compounds
  - Overlaying the H-bond acceptor requires high-energy conformations and conflicting H-bond directions
  - It doesn’t accommodate serotonin…

Updating the Model

- Same 3-point scheme was used with distances generated by:
  - Selection of ligands with different constrained substructures
  - Systematic searches of the distances
  - Minimum necessary intersection of distances used
- Database of 800+ diverse known 5-HT_{1B} ligands
- UNITY 3D flex searches used to compare model performance





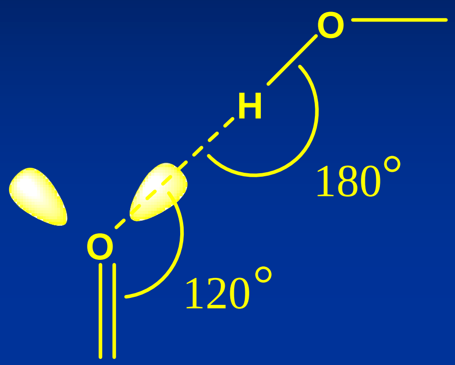
- New Model
  - 69 % recall
  - Tighter Ar-Acceptor constraint

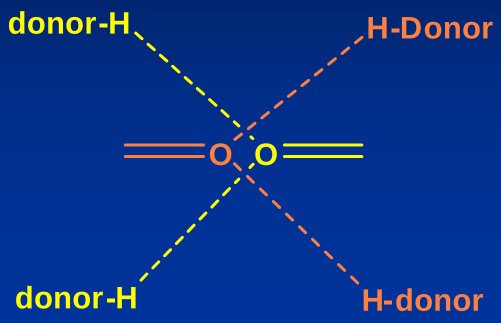
- Old Model
  - 52 % recall

What about the missing 30 %?

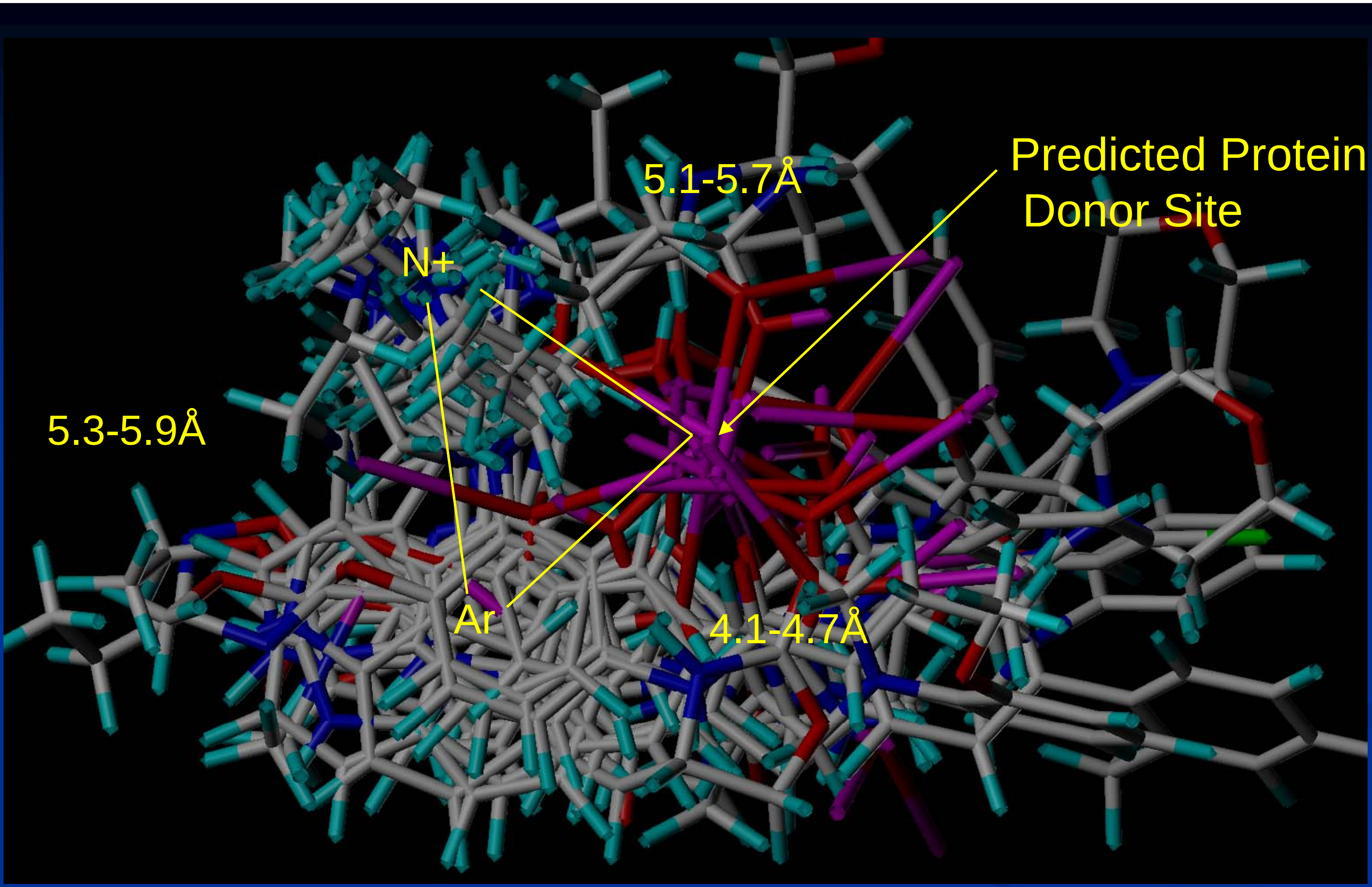
And how can such different H-bond acceptor positions be rationalised?

Virtual-Site Pharmacophores





- Hydrogen bonds are highly directional
- Less important where the acceptor is than whether it can form the H-bond
- Conversely, identically positioned acceptors may only be able to H-bond completely different donor locations
- Multiple acceptor positions and orientations satisfy the ideal H-bond requirements
- Instead of predicting the position of the acceptor, identify the donor location and require the acceptor to be able to H-bond it



- Ligand lone pairs extended to 3Å
- Molecules overlaid so the lone pairs coincided to bond a single H-bond donor

- Updated model now fits 88% of ligands
- Structurally diverse ligand classes all fit the same pharmacophore
- H-bond problems solved and serotonin fits!
- Suggests a single H-bond donor is responsible and even locates it…

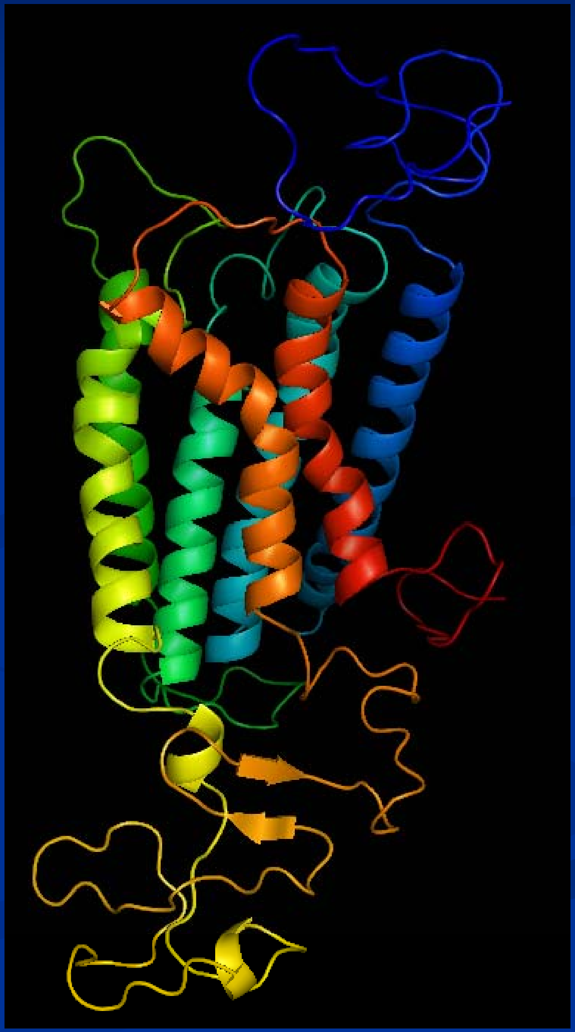
- So how can we ‘validate’ this ?

Receptor-Based Approaches

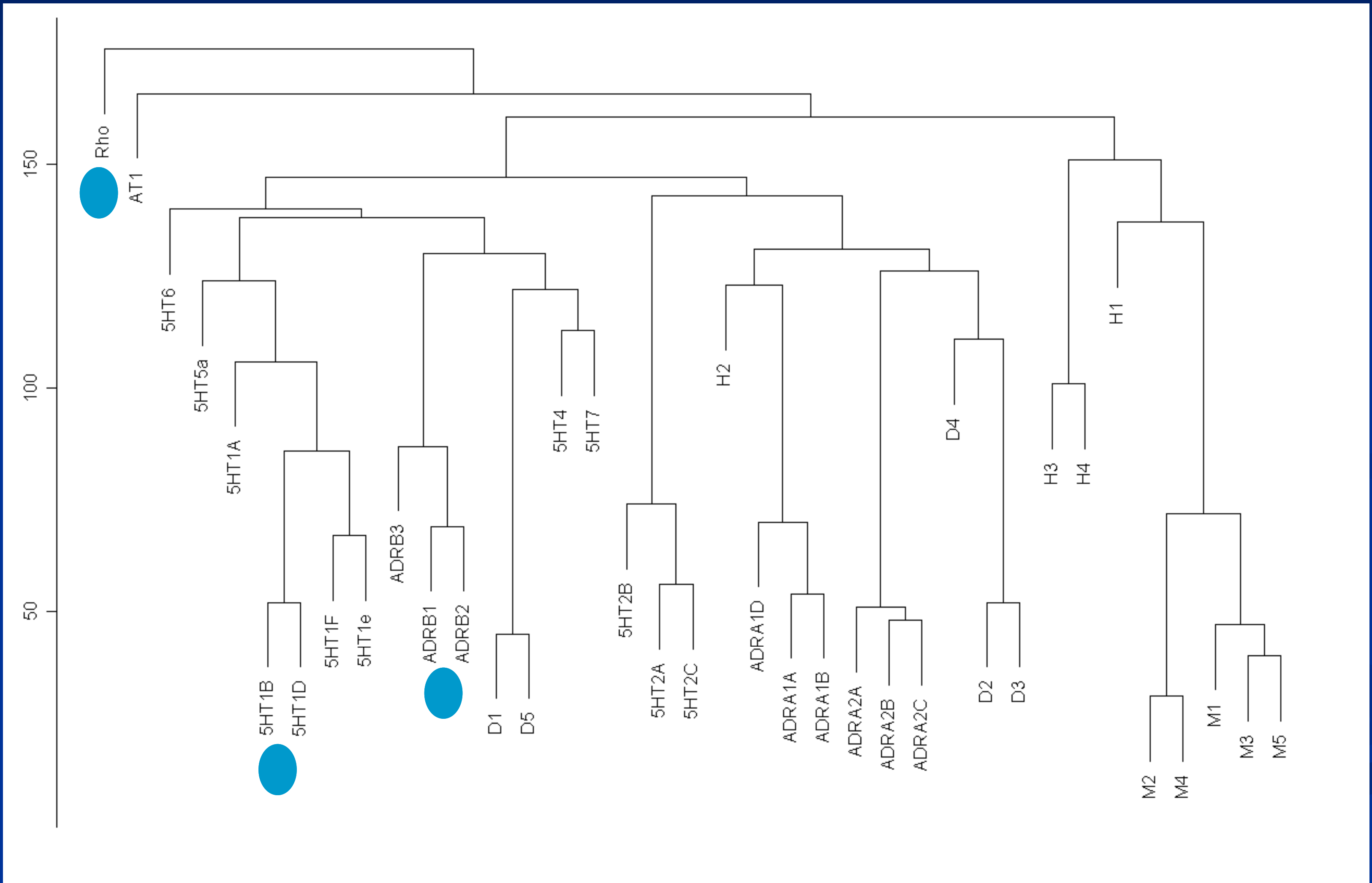
- Homology modelling of 5-HT_{1B} based upon rhodopsin
- Molecular dynamics used to refine model
  - Inserted into an explicit DPPC membrane
  - GROMACS with the 53a6 forcefield, modified lipid charges and SPC water used
  - Equilibration (30ns), then 3x100ns simulations

Rhodopsin-Based Homology Model

- Several problems arose:
  - No significant loop homology
  - Helices started unwinding
  - No binding crevice observed
  - Ligand docking failed
  - ‘Best’ binding mode for ligands had the longer ones binding sideways towards the membrane
  - Poor homology and rhodopsin’s endogenous ligand make it insufficiently close to use as an effective template



Monoamine GPCR Helix Homology

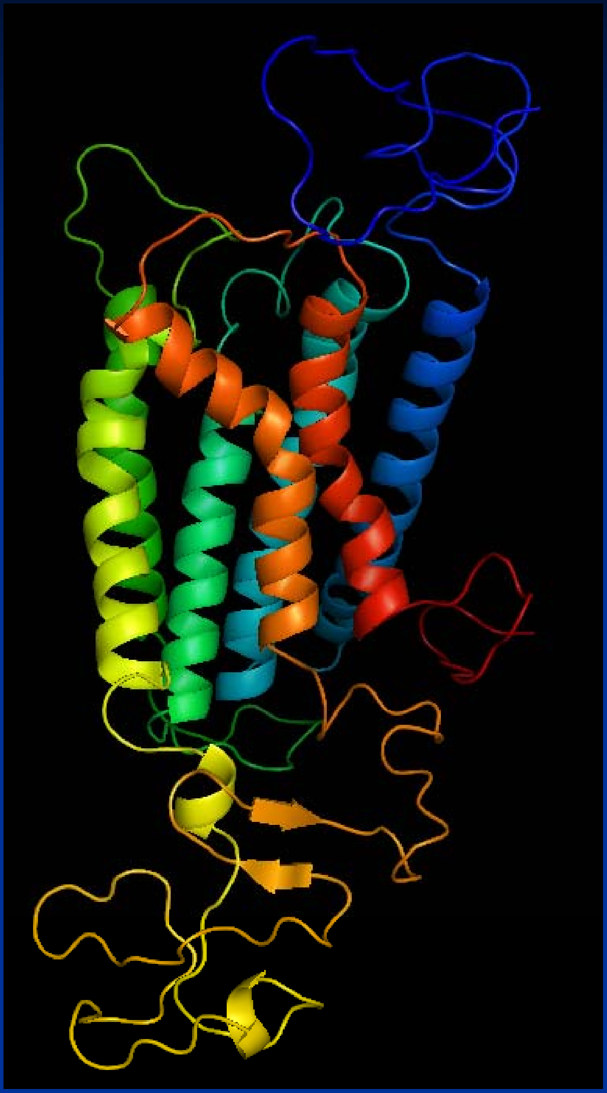


Beta-2 Model

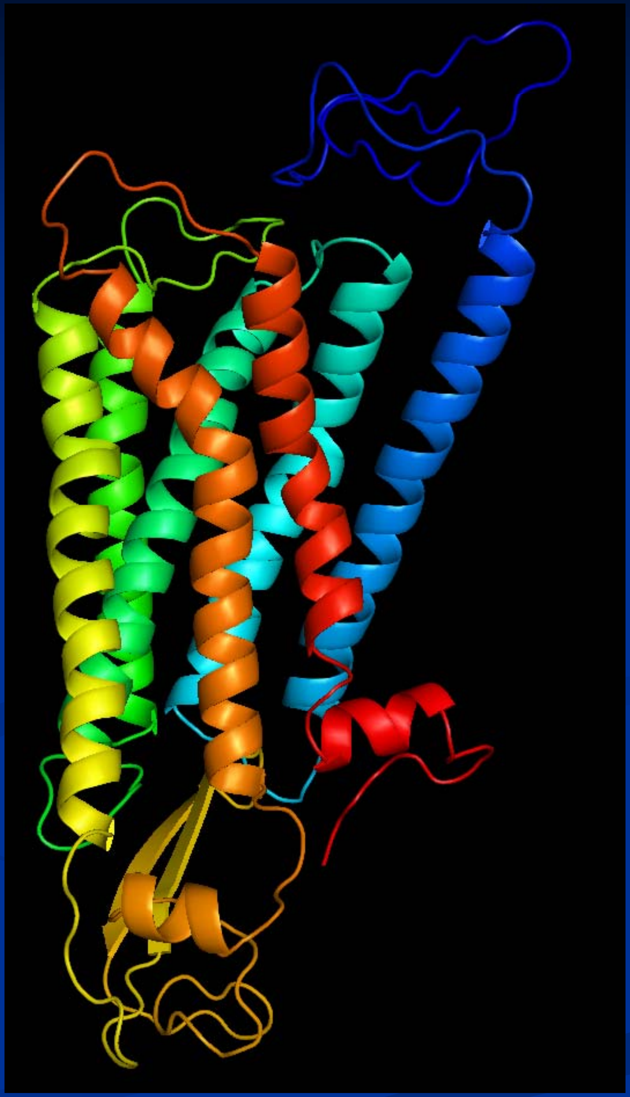
- Advantages:
  - Much higher transmembrane homology: 40% vs 22%
  - Functionally similar
  - Shares several key residues in the binding site
  - Higher loop homology makes them suitable as a template too
- Disadvantages:
  - Crystals are incomplete and have foreign entities attached
  - Lower resolution crystals
  - Only became available in November!
- 2RH1 chosen: Higher res, more complete, functional

Cherezov *et al*. **2007** Nov 23;318(5854):1258-65

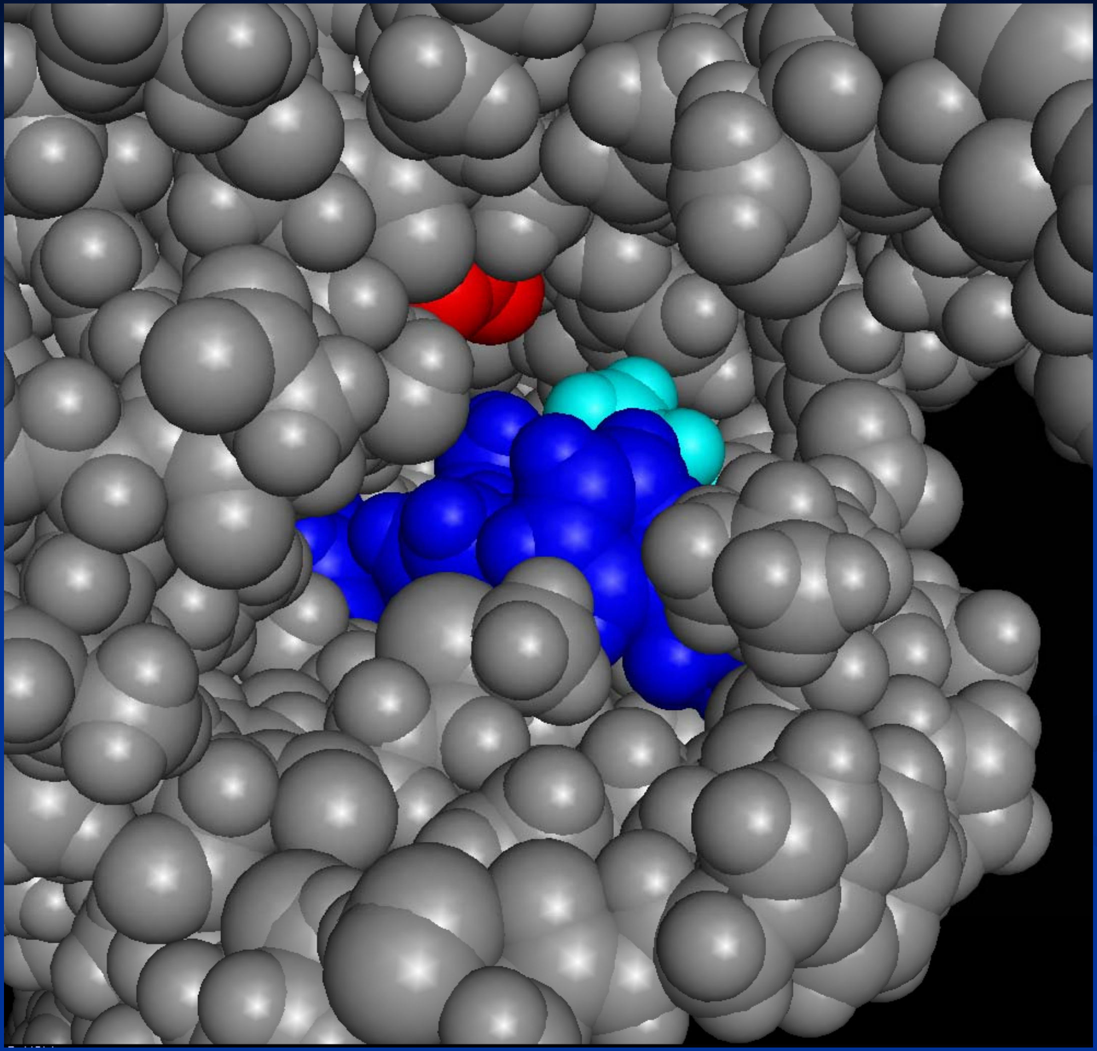
- Sequence analysis shows conserved Cxx(x)C in ECL3 – disulfide formed in model
- Two cysteines in N-terminus were also bridged
- ICL1,2 and ECL1 and the 7 helices were homology modelled
- ECL2 and 3 were built using loop searches, conserving the TM3-ECL2 disulfide link
- N- and C-termini were 'self assembled' from straight-chain by 20ns of MD
- ICL3 was built by secondary structure prediction methods and loop searches to be as compact as possible
- 20ns of Positioned-restrained helix backbone MD was used to 'settle' the model



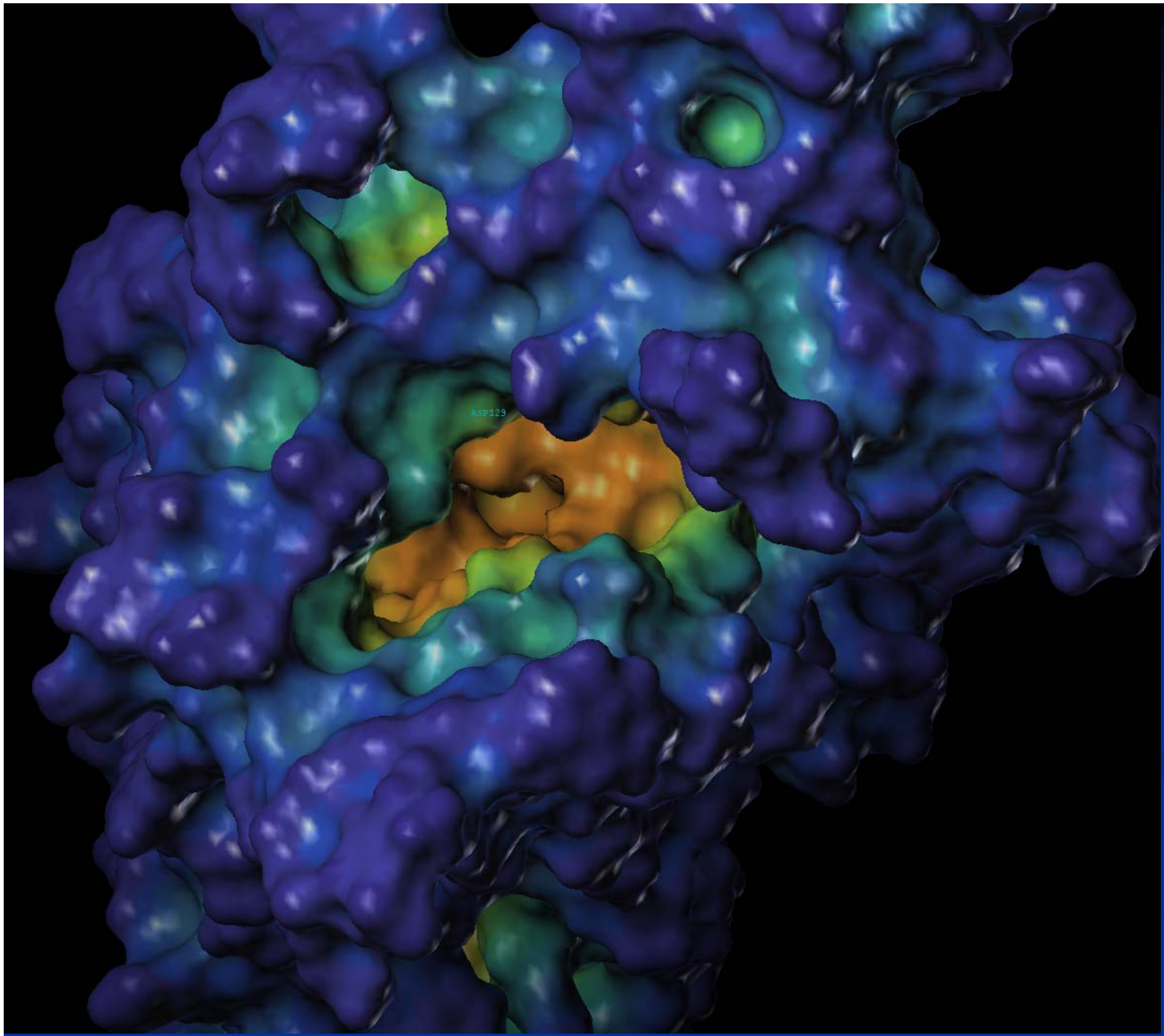
Rhodopsin Model



Beta-2 Model

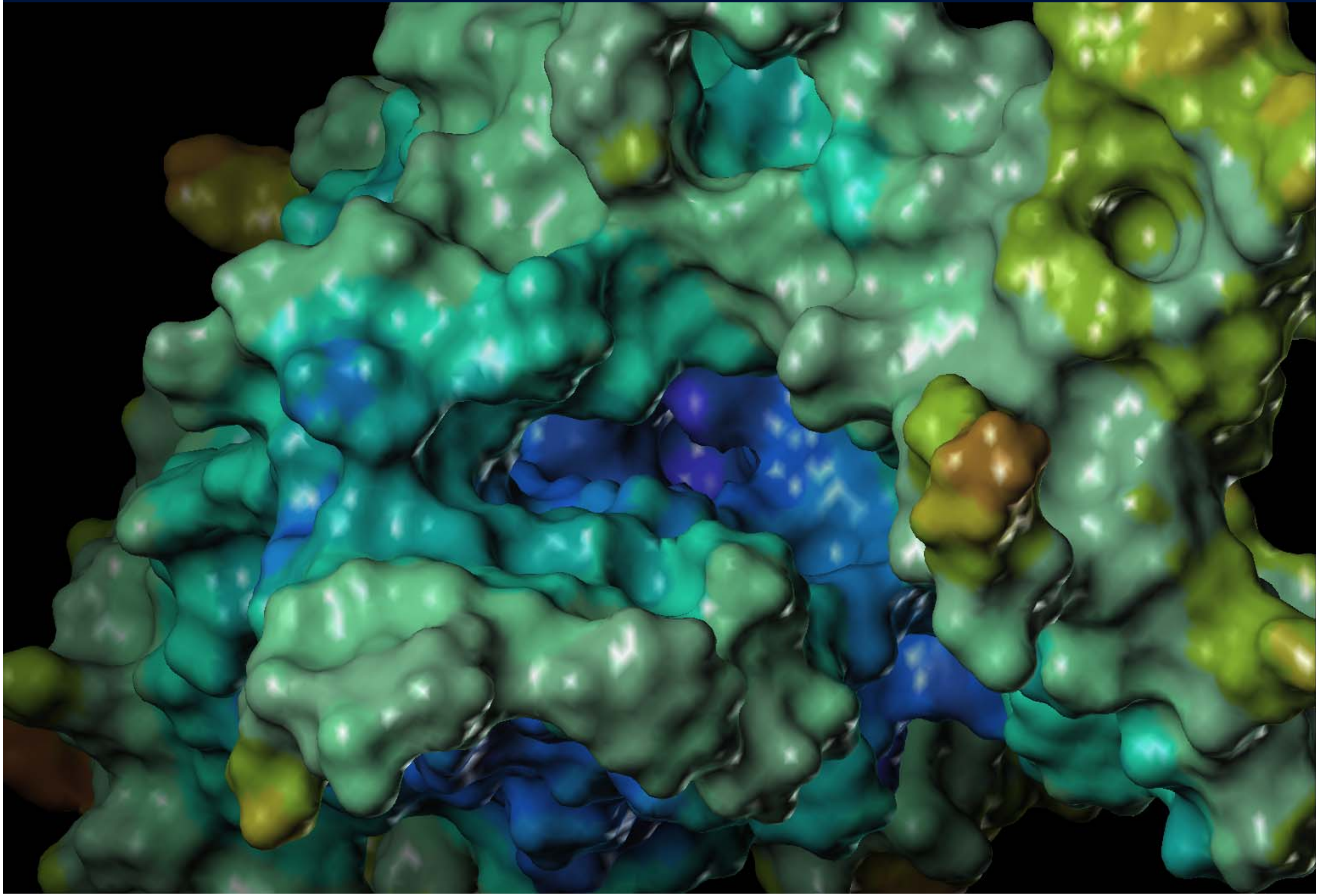


- The binding site is clearly much larger and exposed to the extracellular side of the membrane
- Asp129 (red) is accessible at the bottom of the pocket
- Aromatic residues (blue) cover the side of the cavity
- Thr355, the potential H-bond donor is in cyan

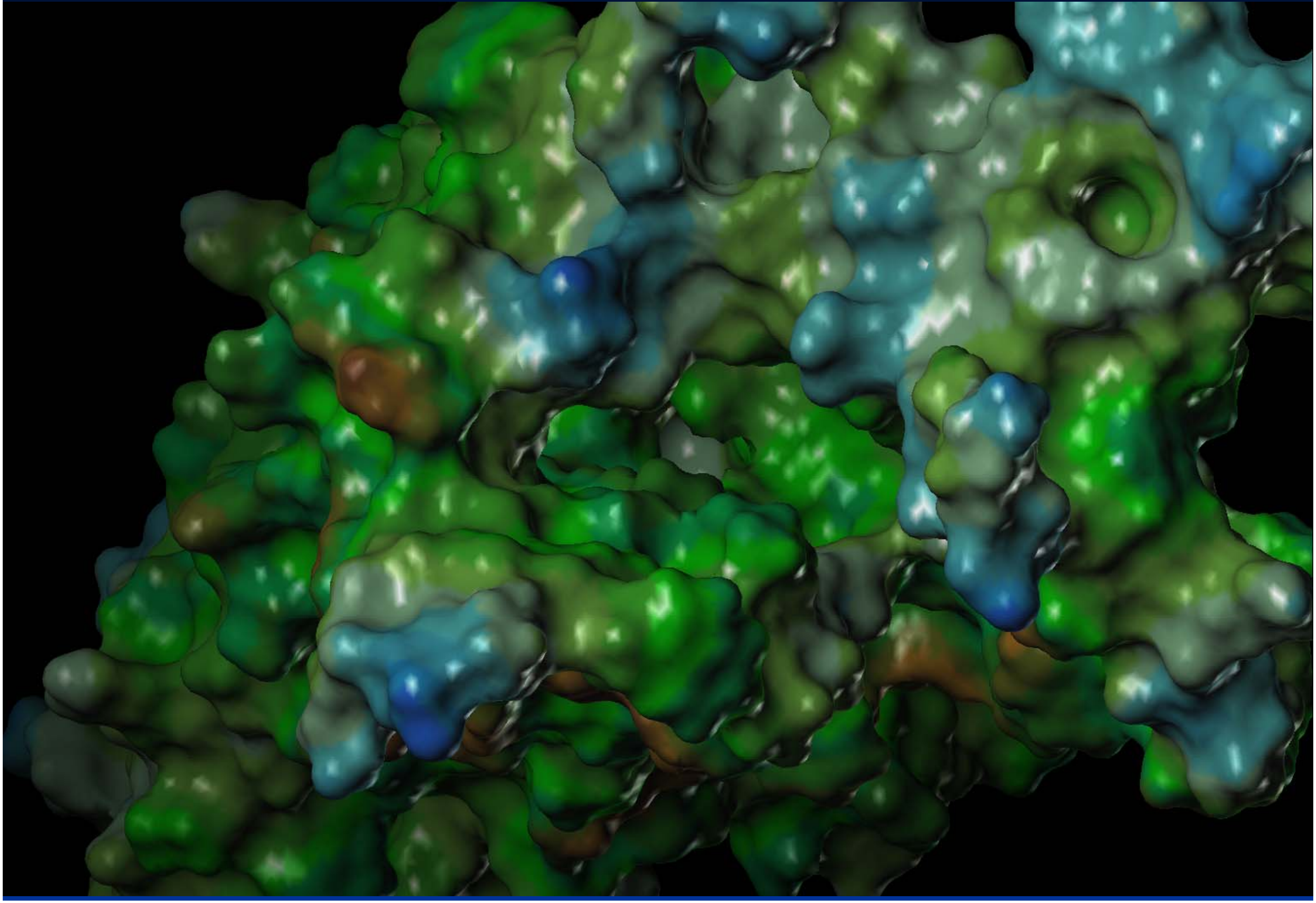


Depth Probe (shows binding crevice)

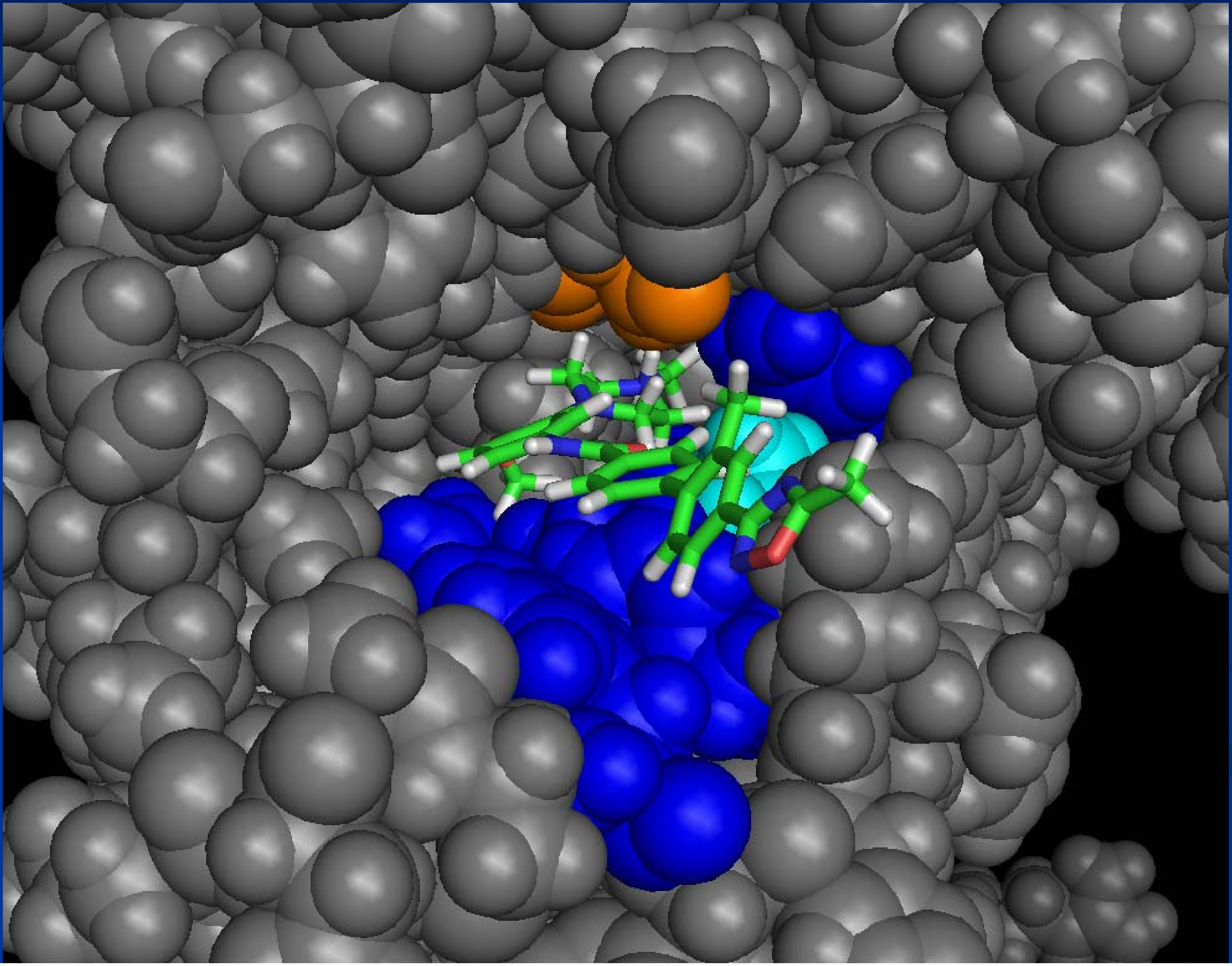
Binding Site Electrostatics: blue = neg, red = pos

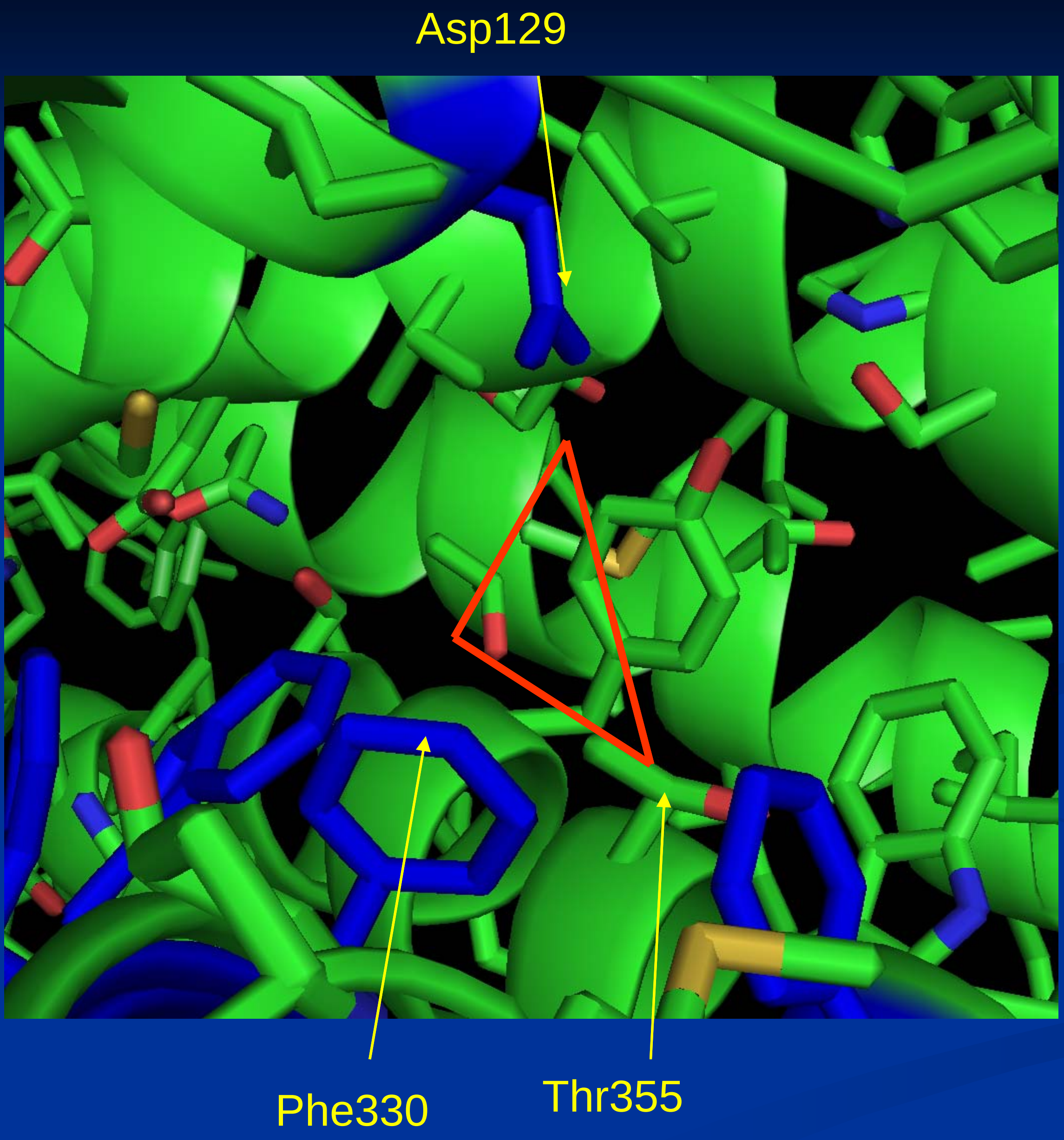


Binding Site Lipophilicity: blue = hydrophilic, brown = hydrophobic



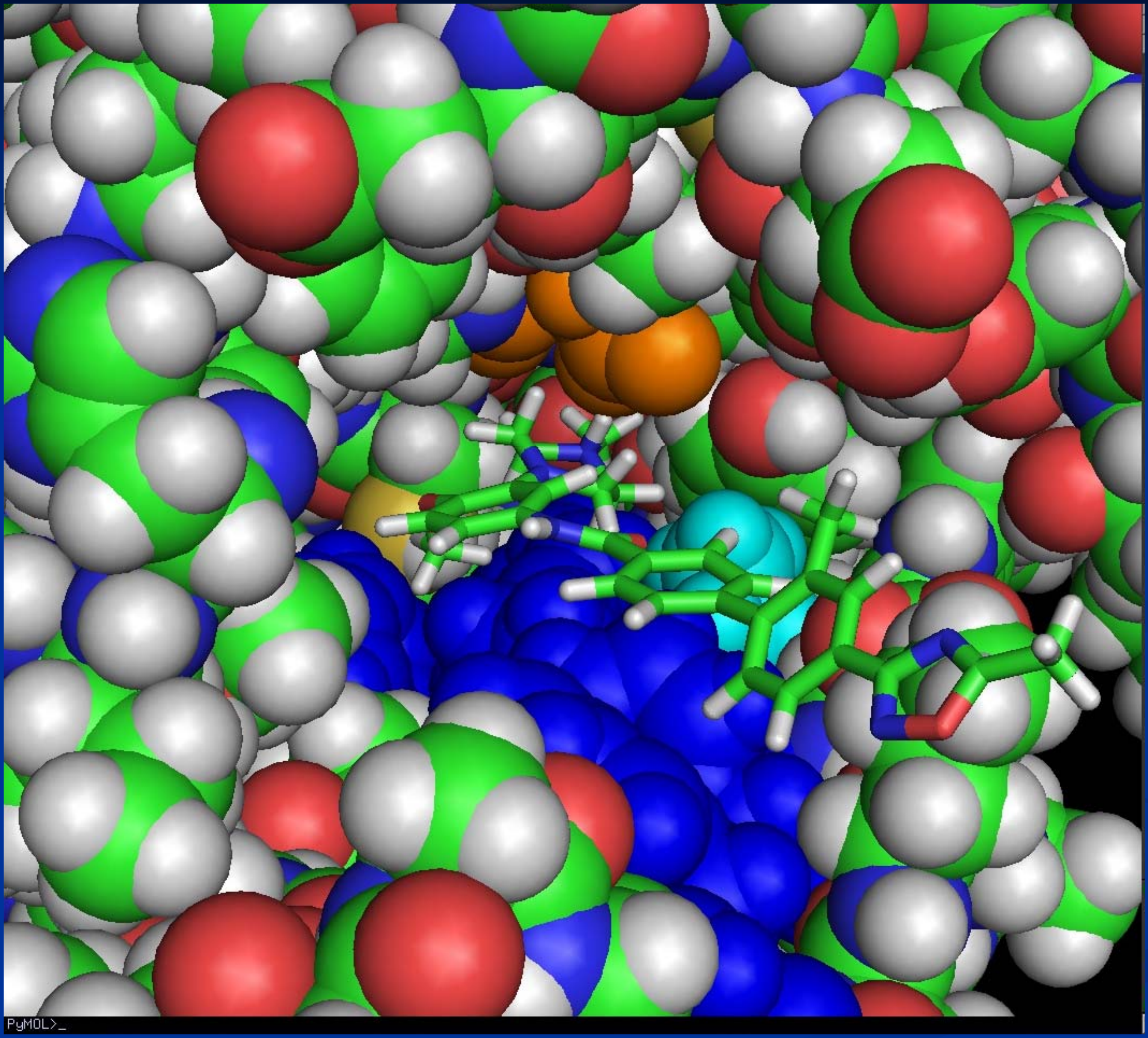
GR127935 in binding site





- Asp129 is known to bind the protonated amine
- A shelf of aromatic residues pointing in/out of the receptor can bind the aromatic groups (blue)
- Phe330 is suspected to form the main aromatic interaction
- Inserting the pharmacophore (red) suggests Thr355 is the hydrogen bond donor
- Thr355 is slightly too ‘low’, but it is within the margin of error
- T355N mutant receptors (as in rats) show very significantly different pharmacology

Pharmacophore —

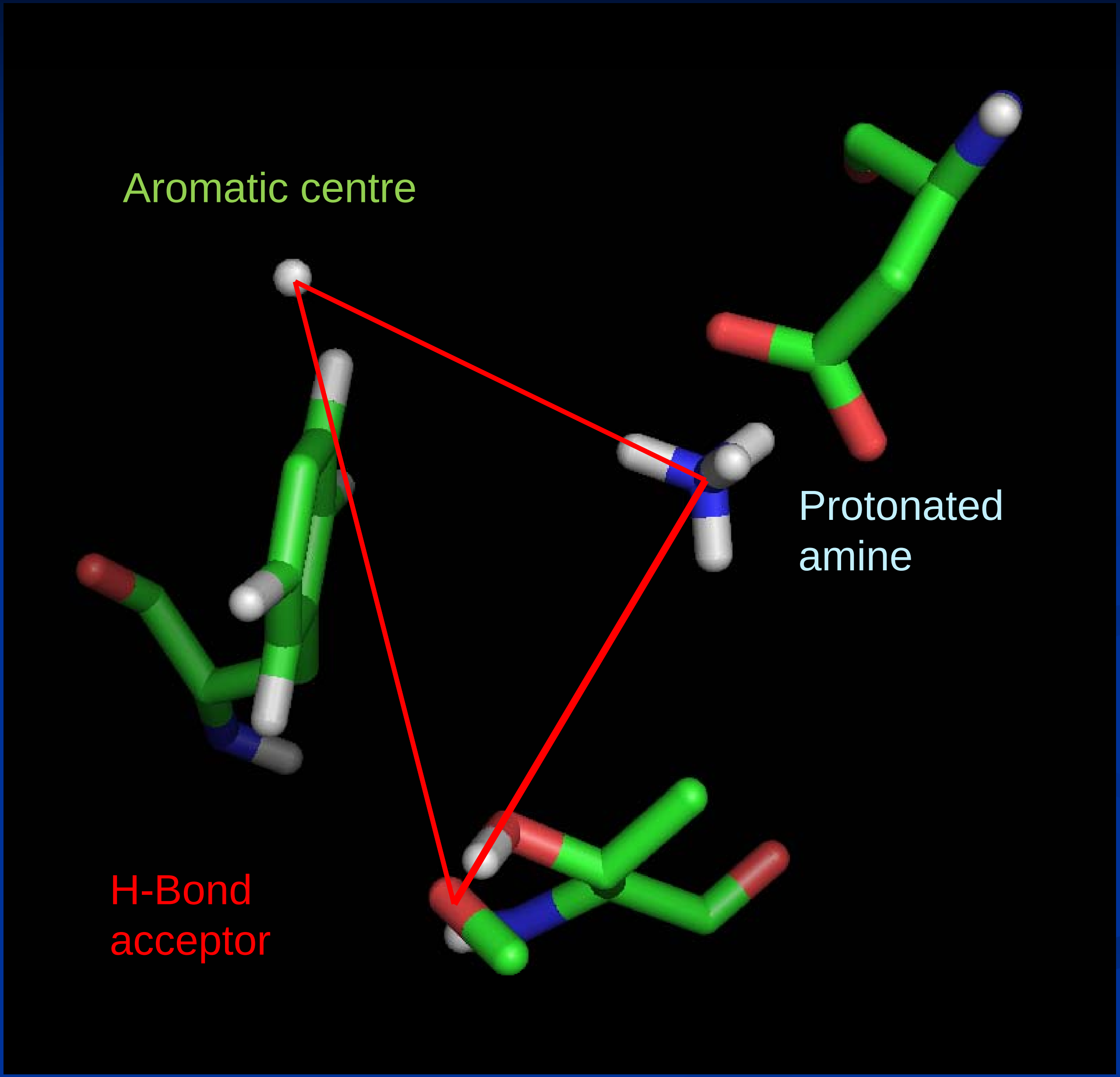


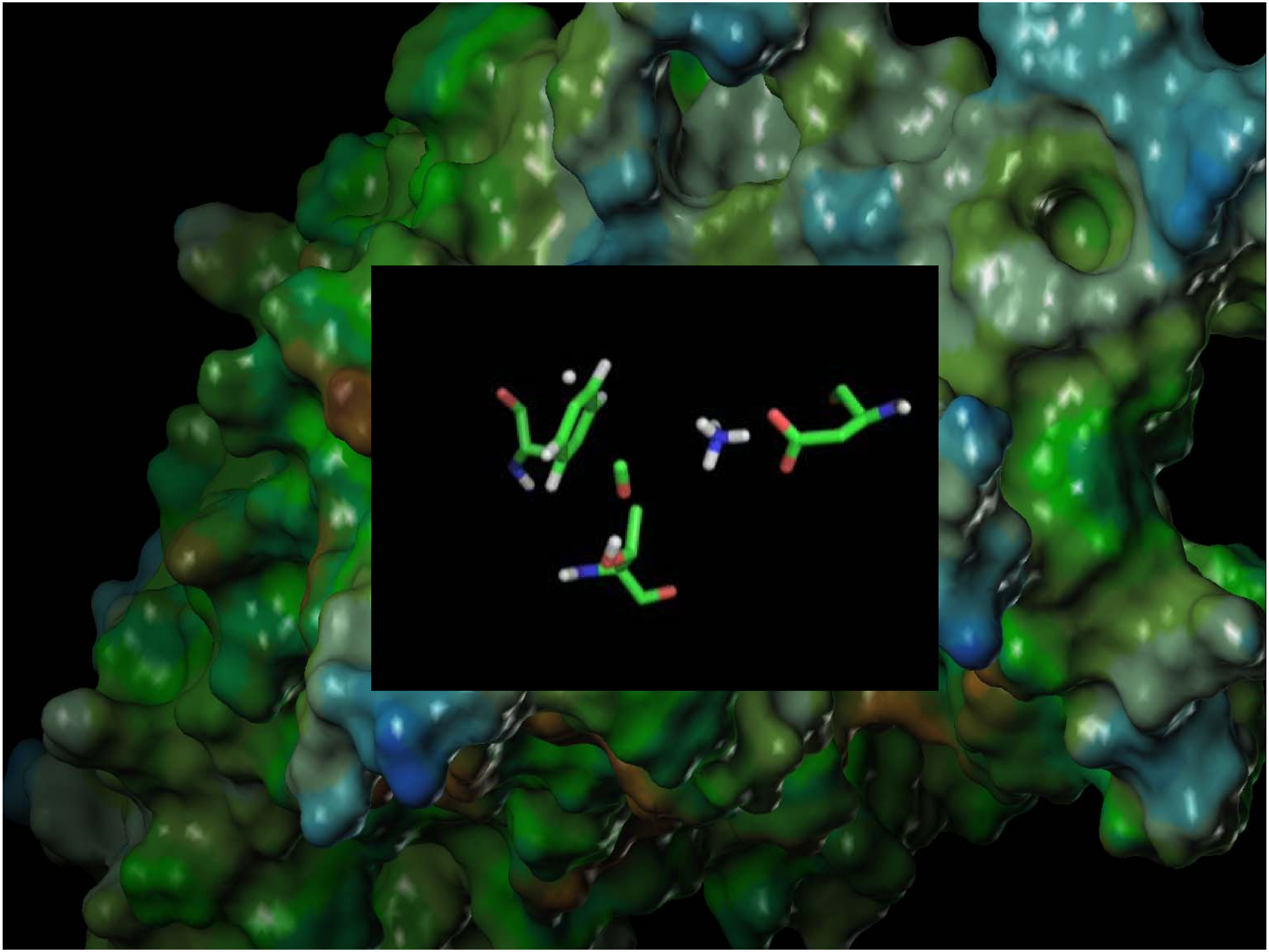
- GR127935 in the binding site
- Asp129 in orange, aromatic residues in blue, Thr355 in cyan

Dynamic Pharmacophores

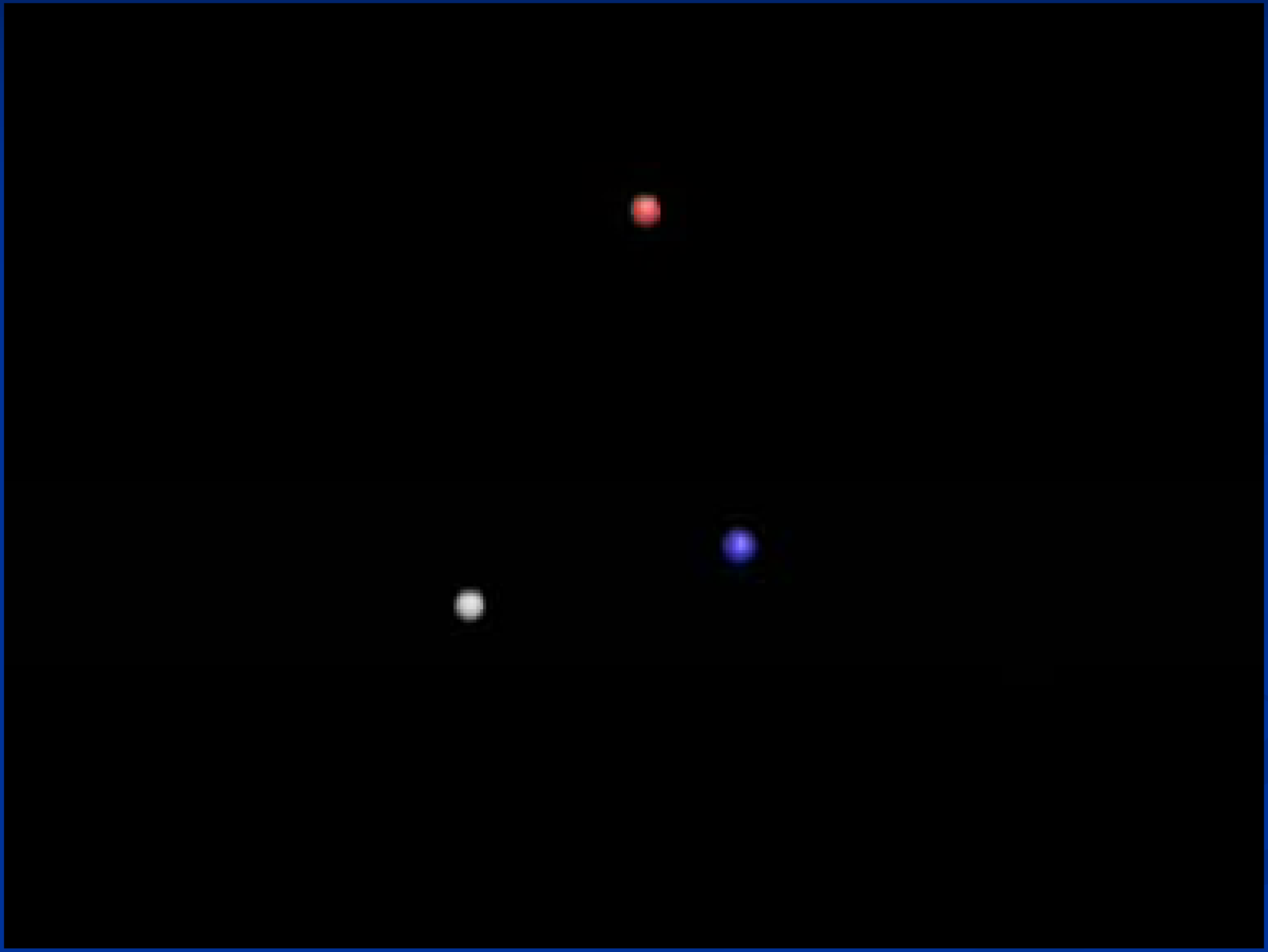
- The pharmacophore fits the homology model's residues and the shape of the binding crevice
- Asp129, Phe330 and Thr355 appear to produce the pharmacophore
- How can we use the receptor model to study the pharmacophore?
  - Proteins are dynamic, especially their sidechains
  - Lowest energy conformations are not necessarily the binding conformations for either ligands or proteins
  - Pharmacophore points are virtual sites of the protein
- Multiple protein conformations must be studied and the changes in the virtual sites followed
- Conformations where key residues are inaccessible must also be avoided

- Virtual sites were covalently attached and parameterised for the 53a6 force-field
- Bond lengths, angles and torsions were derived from ideal hydrogen bonding and pi-pi interaction parameters
- NH_4^+ was attached to Asp129
- C=O was likewise attached to Thr355
- To form the T-stacked aromatic virtual site, the C-H bond of the key hydrogen of Phe330 was lengthened
- The fragments' bulk keeps residues pointing towards the binding crevice
- Using these 'extended' amino acids, molecular dynamics was carried out, with position restraints on the protein backbone


Aromatic centre
Protonated
amine
H-Bond
acceptor



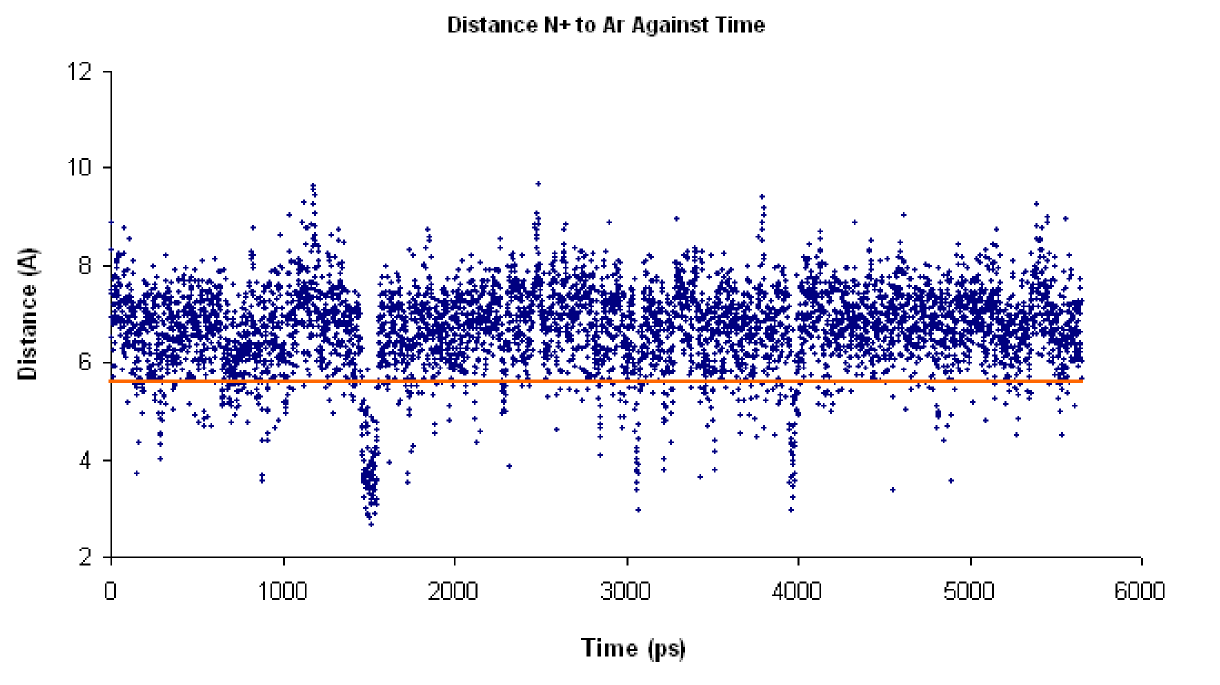
Molecular dynamics of pharmacophore fragments associated with binding



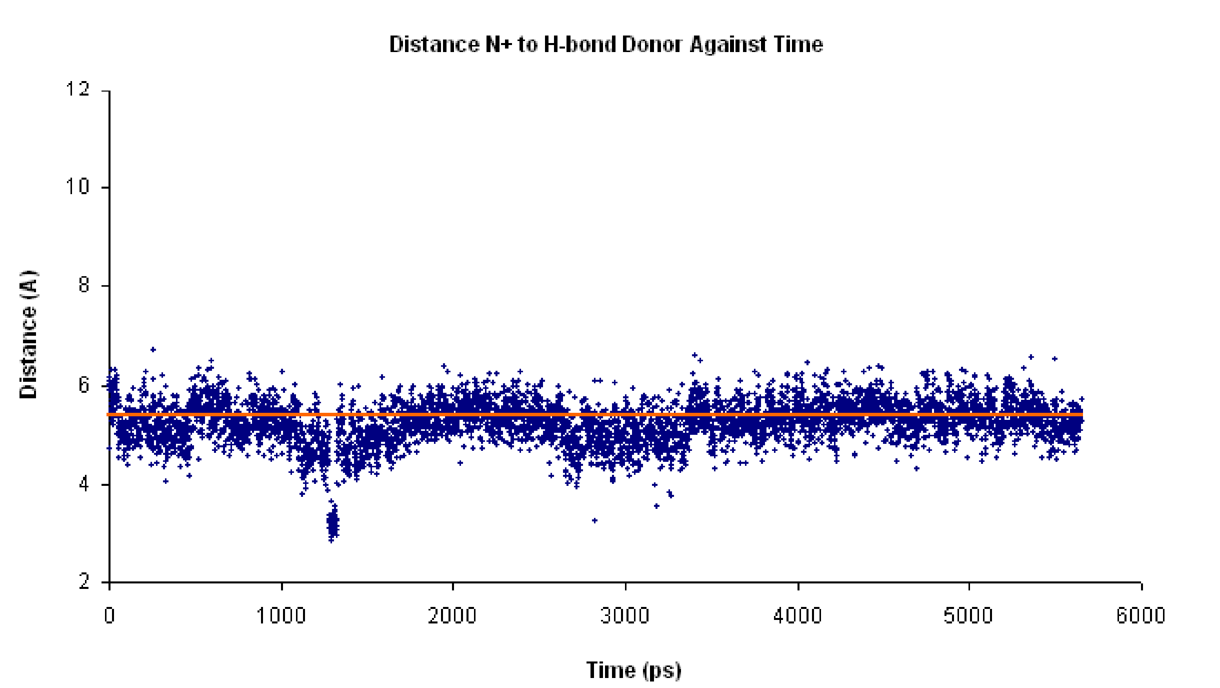
Protonated amine

H-Bond acceptor

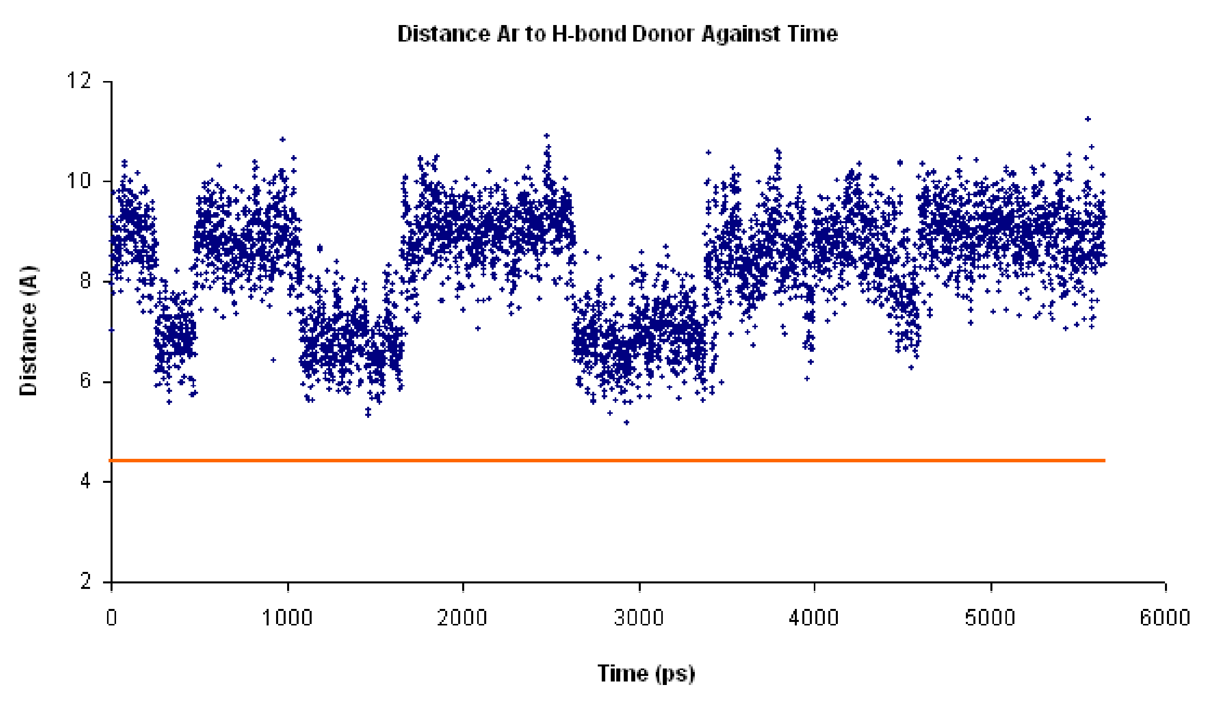
Aromatic ring centre



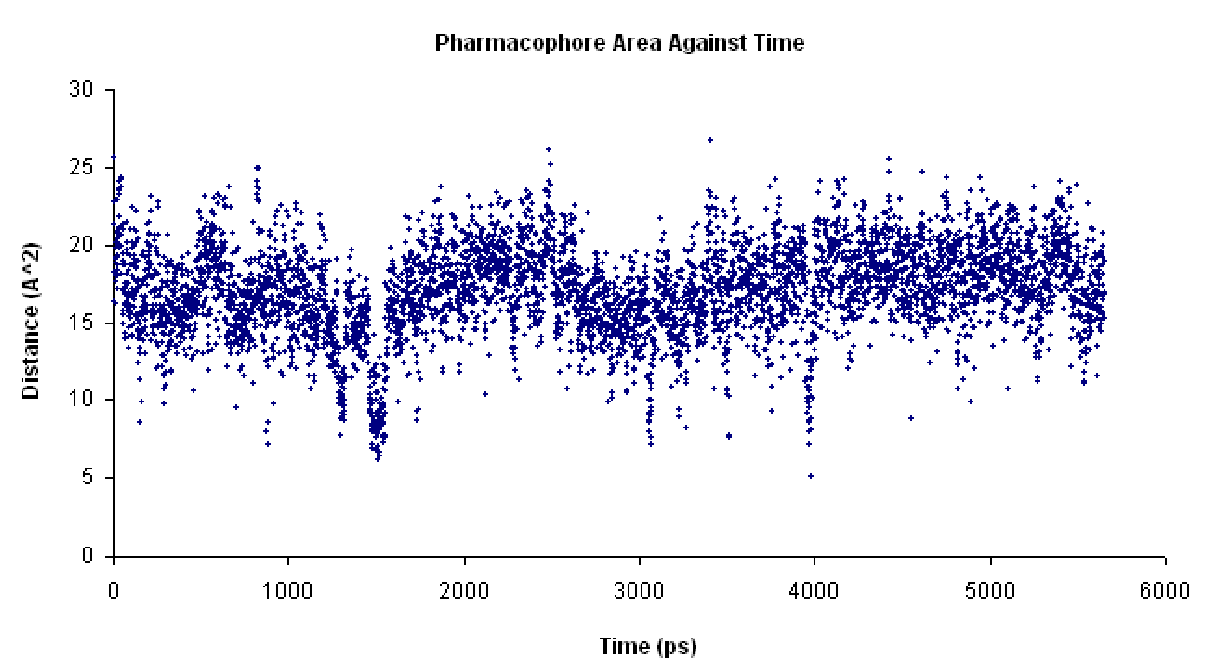
Distance N+ to Ar Against Time



Distance N+ to H-bond Donor Against Time



Distance Ar to H-bond Donor Against Time



Pharmacophore Area Against Time

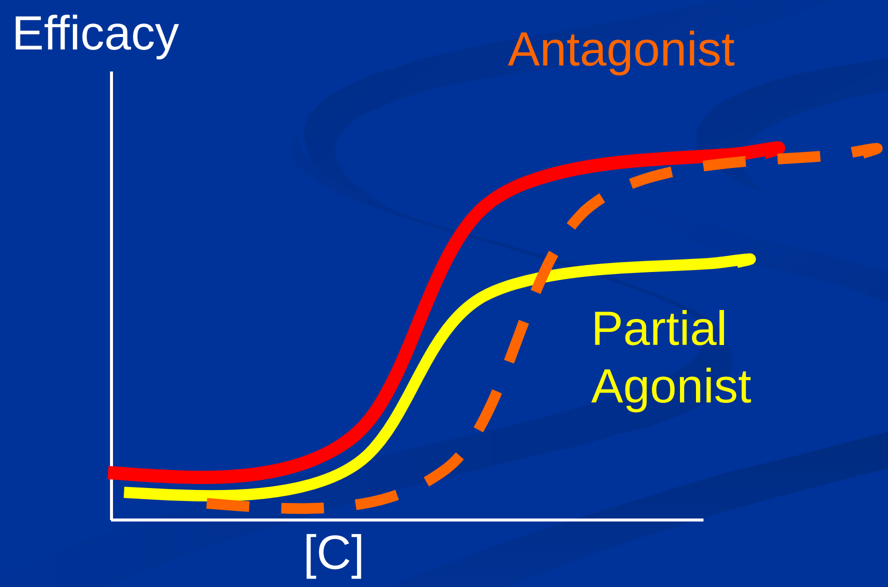
- Every effort taken to minimise variation:
    - Ideal interactions heavily favoured in parameterization
    - Protein backbone restrained
    - ‘Bulky’ virtual sites
- Yet each distance shows 1.5-2.0Å ‘background’ fluctuation
- Thr355 shows two possible conformations altering one distance by an additional 2Å
- Other significant deviations observed of up to 4Å
- Pharmacophore RMSD from starting structure averaged 0.6Å and peaked at 1.5Å

- 2 of the 3 distances fit the ligand-generated pharmacophore
- Thr355 appears too ‘low’ by about 2.5Å
  - Helix 7 may be slightly incorrect in the homology model
  - Some constrained ligands such as GR127935 find it hard to form the hydrogen bond
  - The less constrained indoles still fit the homology model pharmacophore

Efficacy Prediction

Determinents of agonism - can we design selective 5HT$_{1B}$ antagonists
Useful in angina, vasospastic disease

-the Efficacy part of the drug action

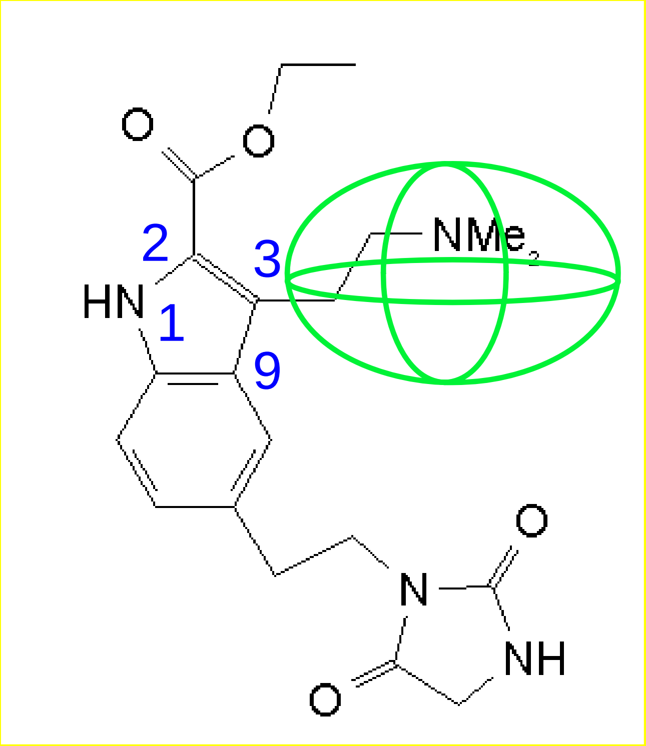


Through analysis of the efficacy of analogues and the computation of a large number of molecular properties,
The best descriptors for a fitted molecule which separated agonism and antagonism were :

➔The principal axes of the 3-subsitiuent (if on indole)

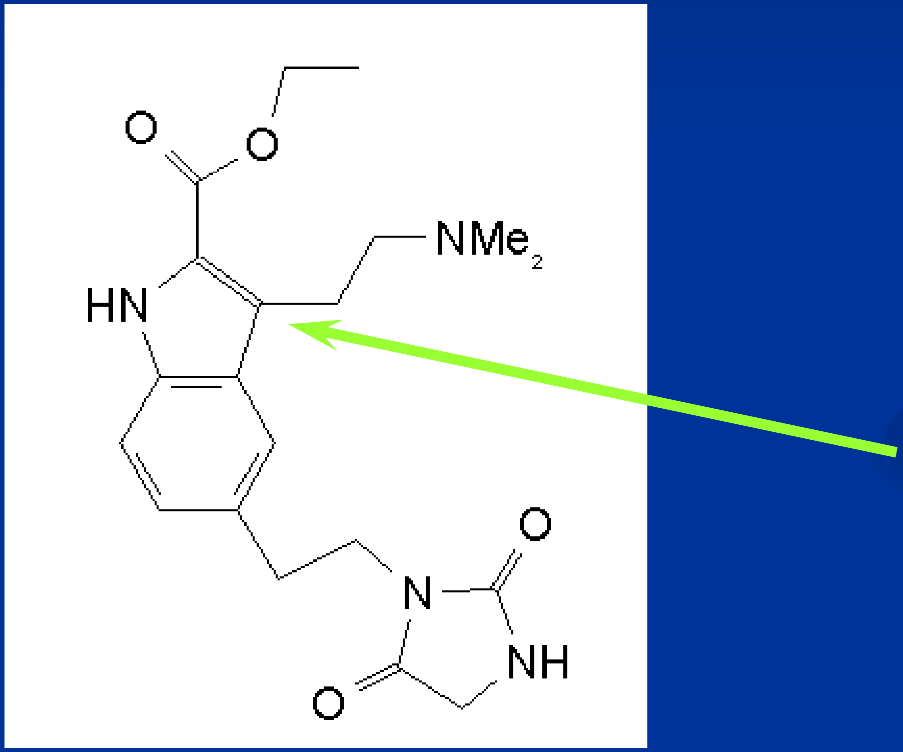
➔The electrophilic superdelocalizability of atoms 1,2,3,9 on the indole (or nearest atoms)

$$SD_e = \sum_{i=1}^{n} \frac{c_i^2}{e}$$



Jandu *et al*. *J Med Chem* **2001**, *44*, 681-693.

For antagonists,
Simple interpretation of the results :



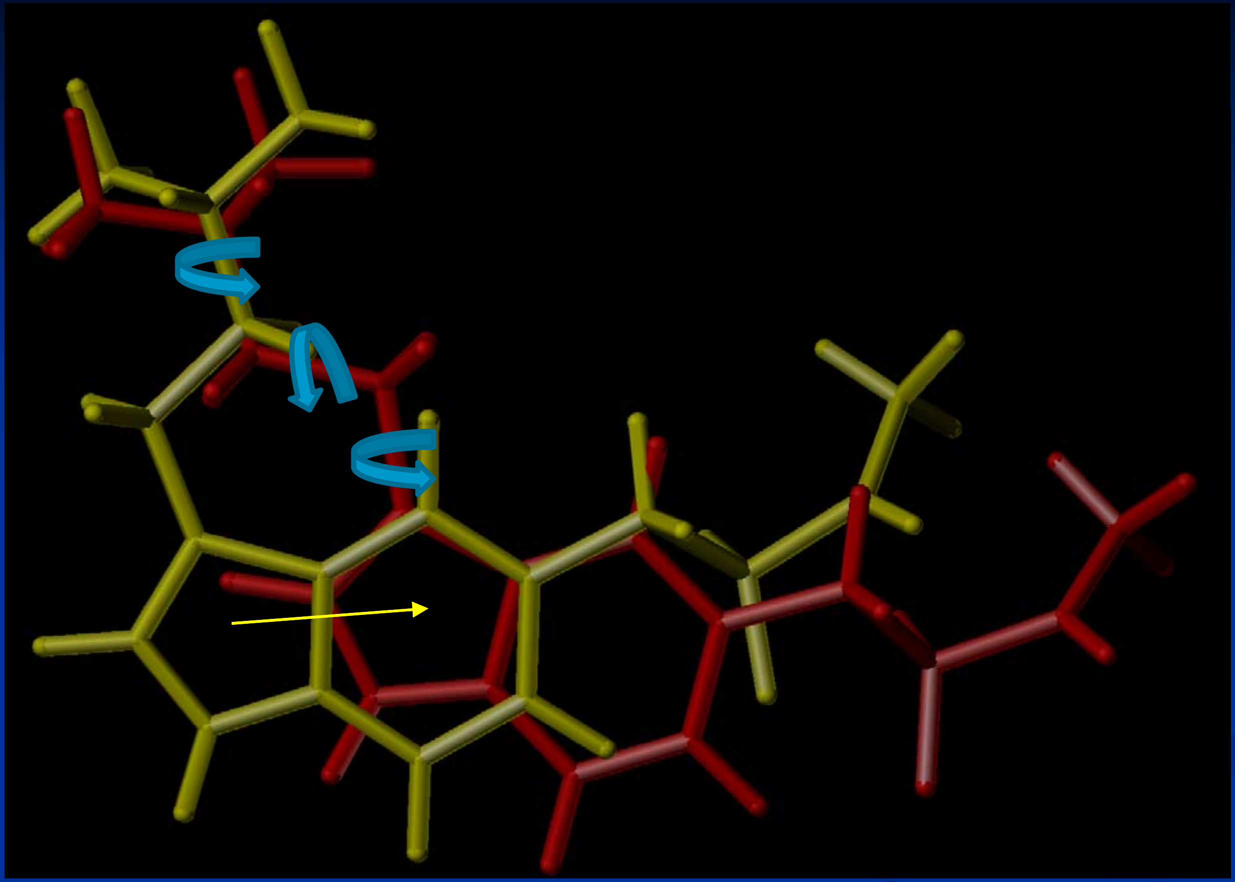
Displacement/reduction of the pi-electron density of the d.b. gives antagonism – maintains affinity, selectivity

However…

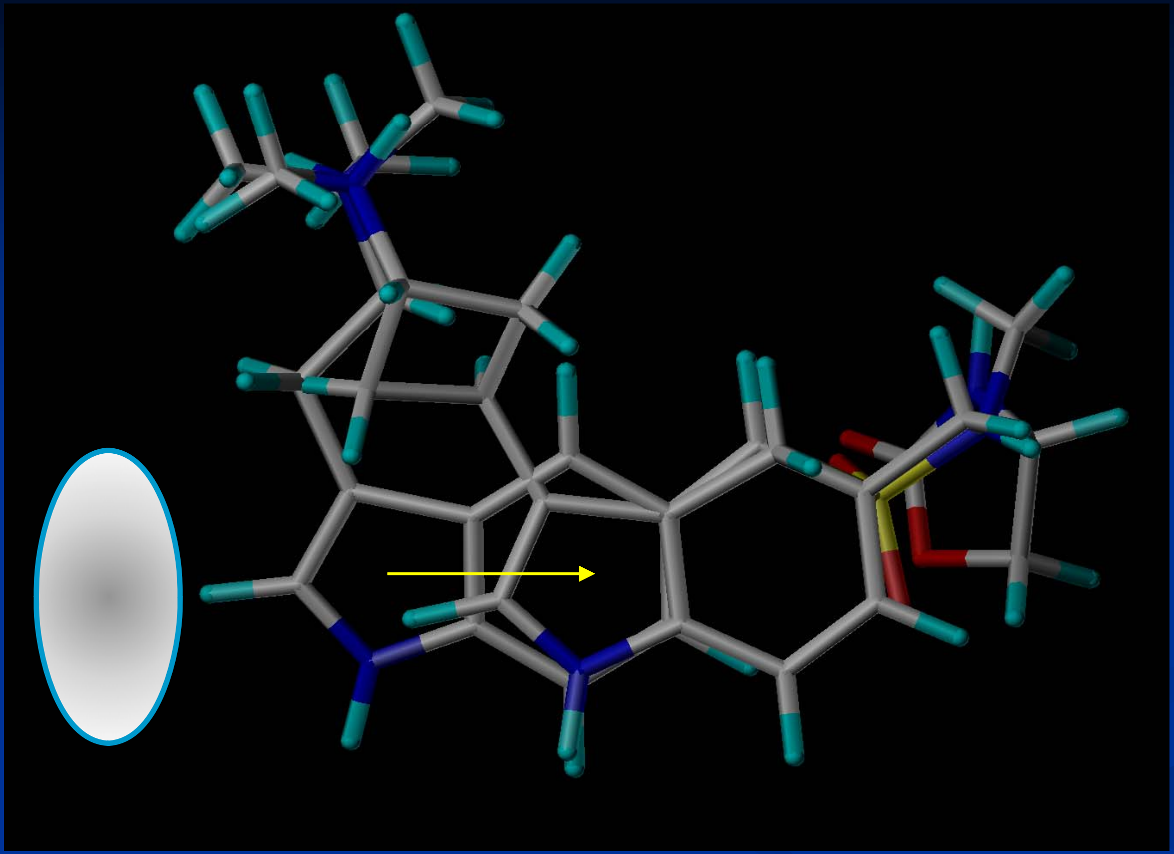
- Original data was skewed by a large number of 2-substituted esters and amides which were all antagonists…

Efficacy: Ligands and Receptors

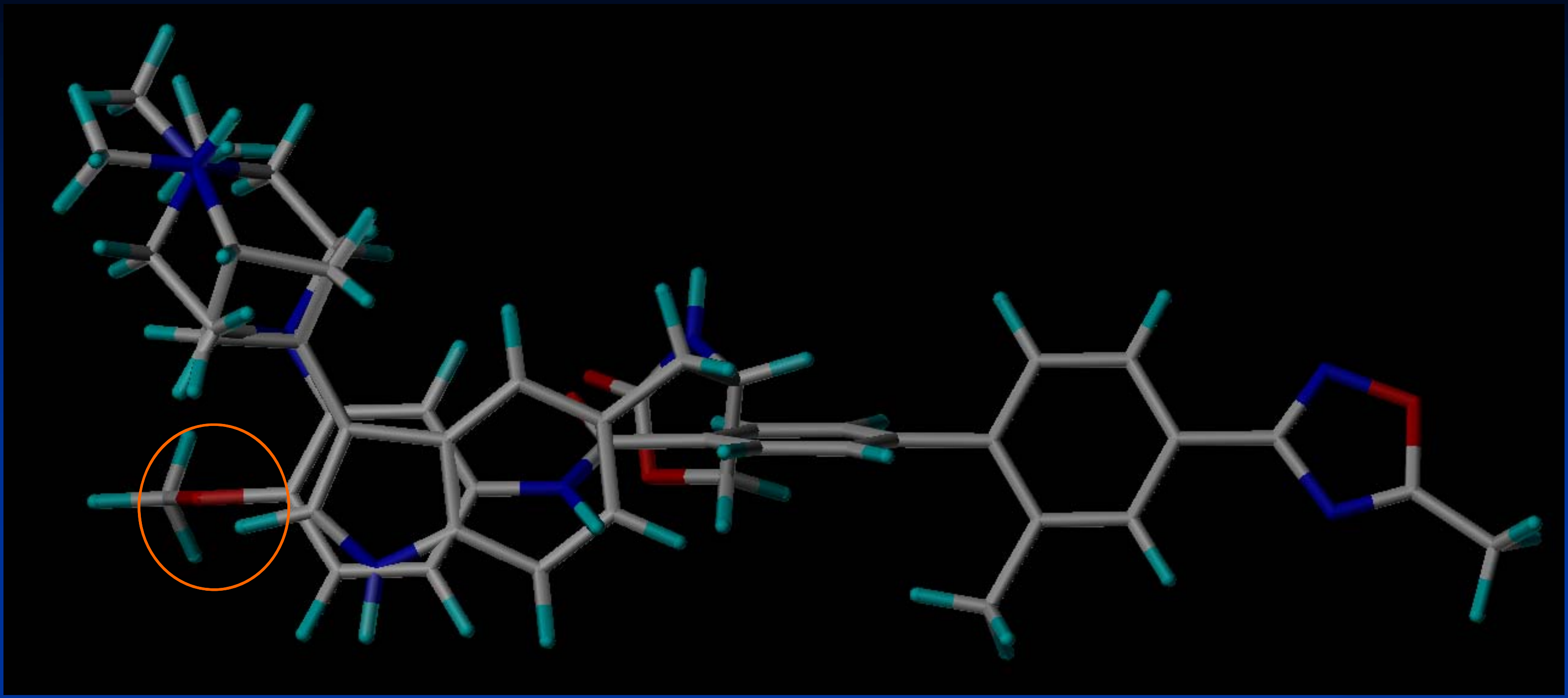
- 5-HT_{1B} ligands without the indole ring are generally antagonists or weak partial agonists
- Those with the indole ring are usually agonists
- Adding a 2-substituent to the ring mostly creates antagonists
  - The substituent may displace the ring (sterics)
  - Alternatively, removing the hydrogen may disrupt a pi-pi T-shaped interaction



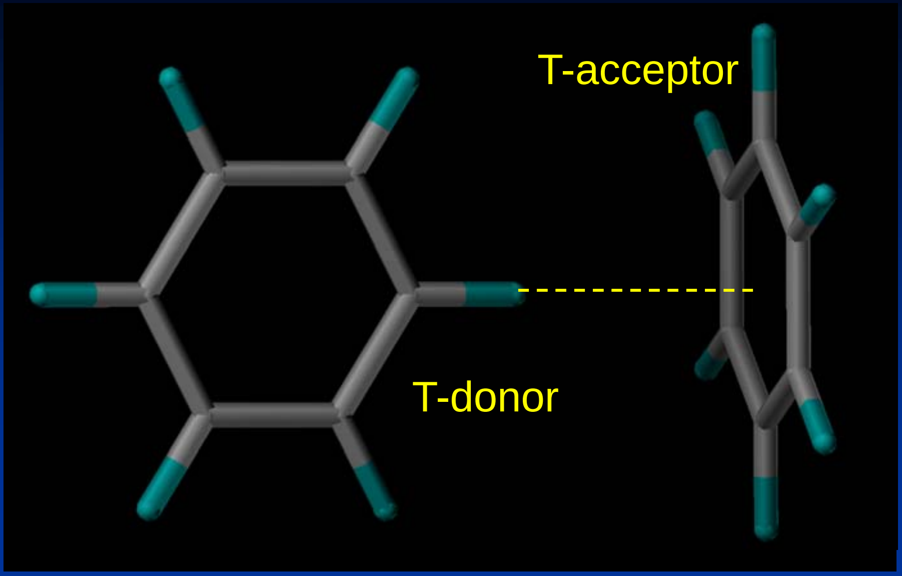
- Flexible sidechains allow indoles such as sumatriptan to adopt multiple conformations
- An extended conformation can shift the molecule as much as one ring across compared to other conformations
- Both positions satisfy the N+ to Ar distance constraint of the pharmacophore, but with different rings



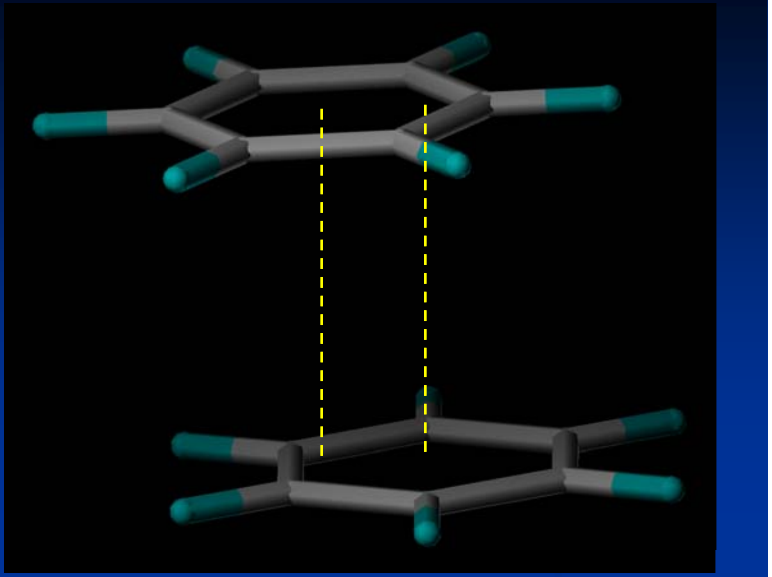
- 2-substituents appear to cause a steric clash and force the molecule into an extended conformation
- Fixing the conformation of the amine chain as a trans-cyclobutane ring mimics the extended conformation and gives antagonists



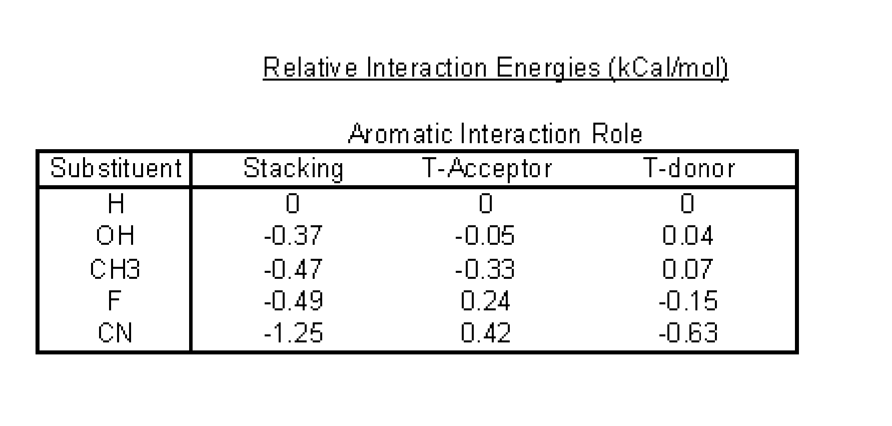
- Extended chain conformations also mimic that of piperazine
- Arylpiperazine antagonists cannot easily fit an aromatic ring in the site occupied by indole agonists
- This indicates there is a second aromatic binding site present only in the active receptor conformation



T-Shaped Interaction



Pi-stacking



Relative Interaction Energies (kCal/mol)

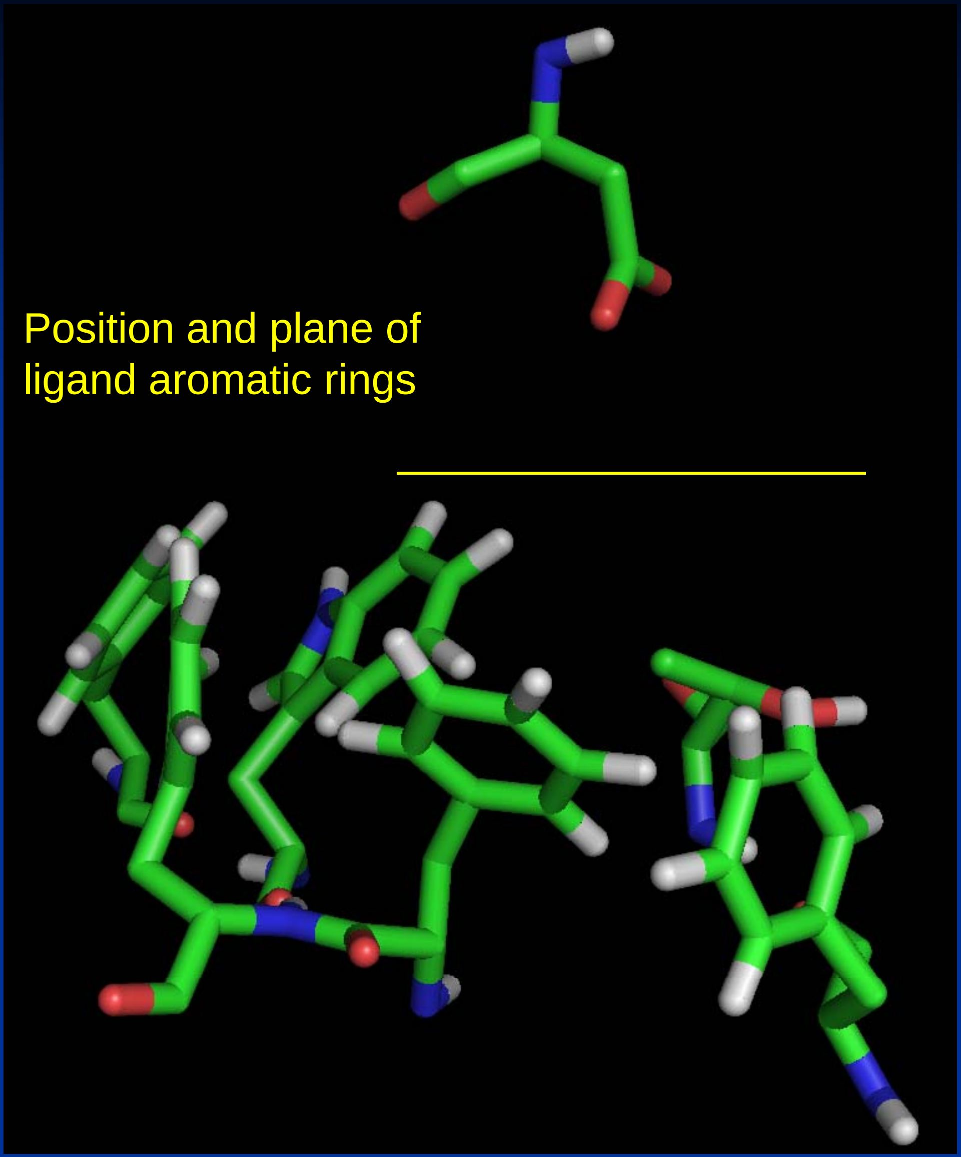
Aromatic Interaction Role

Substituent	Stacking	T-Acceptor	T-donor
H	0	0	0
OH	-0.37	-0.05	0.04
CH3	-0.47	-0.33	0.07
F	-0.49	0.24	-0.15
CN	-1.25	0.42	-0.63

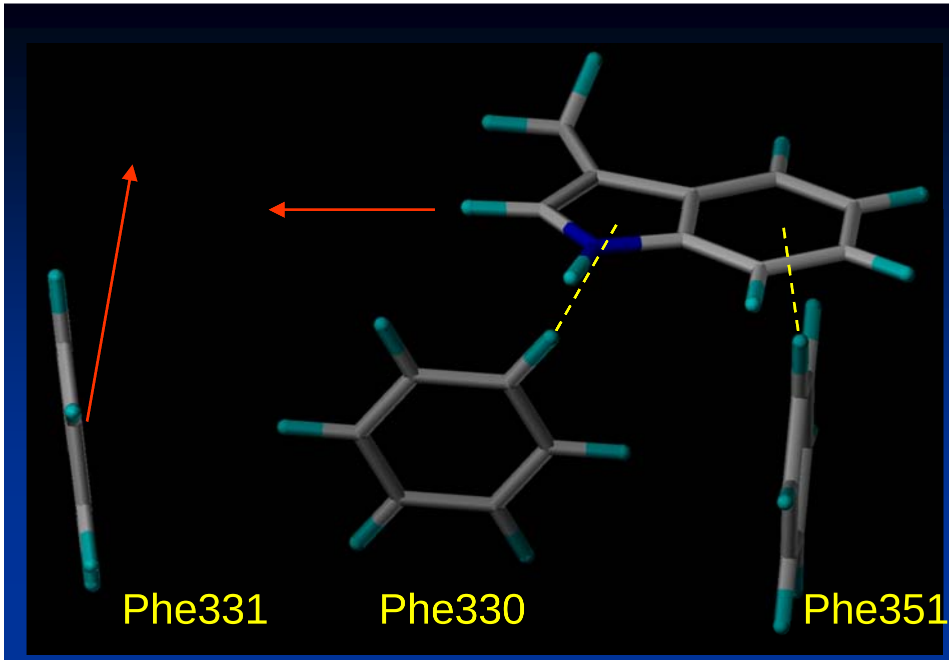
Sinnokrot et al. JACS **2004**, 126, 7690-7697

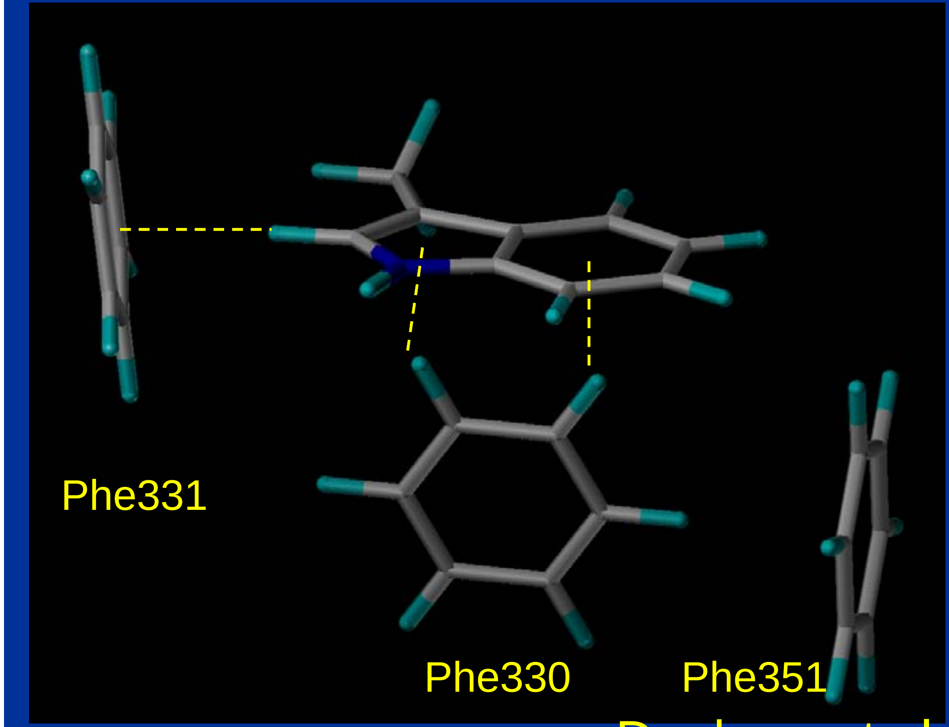
- Substituents on aromatic rings affect pi-pi interaction energies.
- Our own ligands show a 2-4x preference for methoxy over fluoro substitution

– ligands appear to be acting as T-acceptors



- Binding site has a row of aromatics all orientated with hydrogens pointed towards potential ligands
- Homology model is based upon an inactive conformation of Beta-2
- All antagonist pi-pi interactions will be T-shaped with the ligand as the ‘acceptor’





- Antagonists:
  - Shifted ‘right’
  - Indole interacts with Phe330 and Phe351
  - Room for 2-substituents
- Agonists:
  - Indole sits over Phe330
  - Phe331 relatively moves upwards to T-accept for the ligand
  - Causes/caused-by a clockwise turn and movement of TM6 seen in the activation of rhodopsin
  - Ligand stabilises protein active conformation

Dunham et al *J Biol Chem* **1999**, *274*, 1683-1690

Evidence for Pi-Pi Based Activation

- Phe330-Phe331 pair is fully conserved amongst serotonin, dopamine and adrenaline receptors
- Adjacent Pro329 is fully conserved in GPCRs and implicated in amplifying conformational change by altering the angle of the bend it induces in the alpha helix
- Substituting a cyano group for a methoxy group in an arylpiperazine has been shown to convert partial agonism into full agonism, and to increase binding affinity, consistent with strengthening the ligand as a T-donor
- Substituting N for C-H in napthylpiperazines resulted in position-dependent 20x loss of binding affinity combined with loss of partial agonism (electrostatic repulsion ?)

Sansom *et al. TiPS* **2000**,*21*, 445-451

Kling *et al*. *Bioorg Med Chem Lett* **2005**, *15*, 5567-5573

How to Cause Antagonism…

- Stabilise the extended ligand conformation
  - Longer inflexible protonated amine sidechain
  - Altering H-bond acceptor orientation and chain length
  - Suitable aromatic substituents
- Destabilise the protein active conformation
  - Steric bulk at the 2-position (or equivalent)
  - Replace aromatic hydrogens involved in pi-pi interactions
  - Suitable aromatic substituents
- Does it work ? Yes, we have used this to create a new antagonist class, being patented.

Conclusions

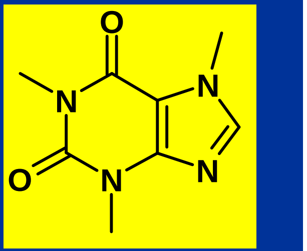
- Introducing virtual sites into pharmacophores makes them more realistic
- Dynamic pharmacophores demonstrate the huge variation of binding within the site
- The new 2RH1 beta-2 crystal structure is a superior starting point for monoamine receptor homology modelling
- Ligand and receptor-based approaches are much more powerful when combined
- Efficacy models must consider conformational changes in the protein

Moving on to Structure-Activity/Property models - some observations

- The objective here is to relate measured or computed parameters to some new property, e.g. bioactivity at a target, absorption, melting point, solubility…
- The first issue is data quality.
  - Biological data is always problematic as it is often not possible to reliably reproduce, isolate the variables, combine data. Physical data is easier to measure (in general) and there is a lot more of it.
  - Our experience with a common physical property, solubility

How reliable are solubility data ?

Caffeine solubility

	Temperature	Solubility g/l	Year
	25	2.132	1926 [1]
	25	896.2	1985 [2]
	25	21.0	2002 [3]
	25	49.79	Merck Index
	25	18.67	2005 [4]
	25	21.6	SRC PhysProp Database

[1] Oliveri-Mandala, E. (1926), *Gazzetta Chimica Italiana 56*, 896-901

[2] Ochsner, A. B., Belloto, R. J., and Sokoloski, T. D. (1985), *Journal of Pharmaceutical Sciences 74*, 132-135

[3] Al-Maaieh, A., Flanagan, D. R. (2002), *Journal of Pharmaceutical Sciences 91*, 1000-1008

[4] Rytting, Erik, Lentz, Kimberley A., Chen, Xue-Qing, Qian, Feng, Venkatesh, Srini. *AAPS Journal* (2005), 7(1), E78-E105.

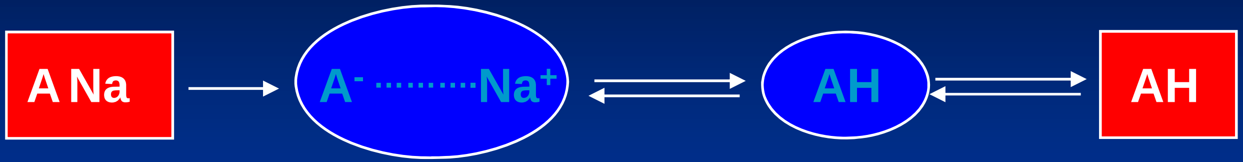
‘Solubility’ in the literature

- Katritzky observed for a diverse set of 411 compounds an average standard deviation of 0.58 log units. Jorgensen and Duffy suggested the average uncertainty of 0.6 log units. For even simple compounds such as chlorobenzenes, measured solubility values vary by ca. 1.5 log units.
    - data can have wide ranges in the literature : guanine has -3.58 and 1.86 – take your pick.
- Recent study by Dearden, re-measured 113 organic drug-like compounds,
    - 22 differed by >0.5 log unit
    - 9 differed by >1.0 log unit
    - 1 differed by >2.0 log units
- Thus, any computational method that gives estimates (usually based on SAR) better than 0.5 log units is over fitted – many are!
- Dearden J.C. Expert Opin. Drug Discov. (2006), 1(1).

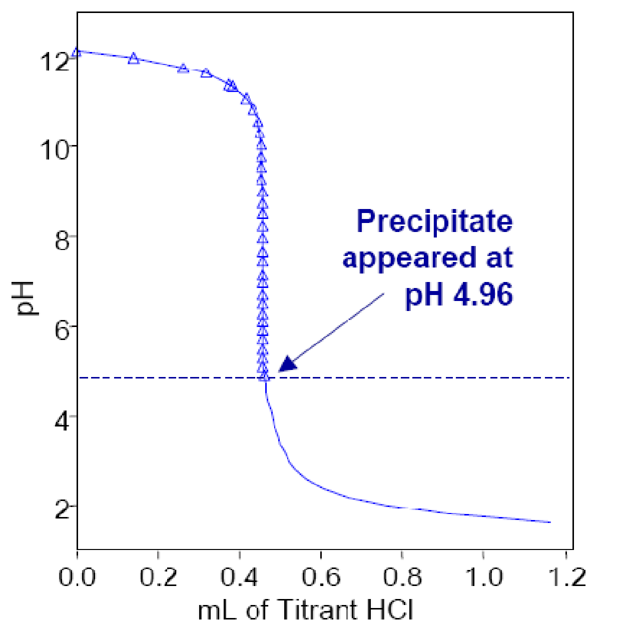
- The lit. data usually has no information on the experimental method, the material whose solubility is being studied, or the definition of the reported solubility – and commonly, many datasets are combined to build models.

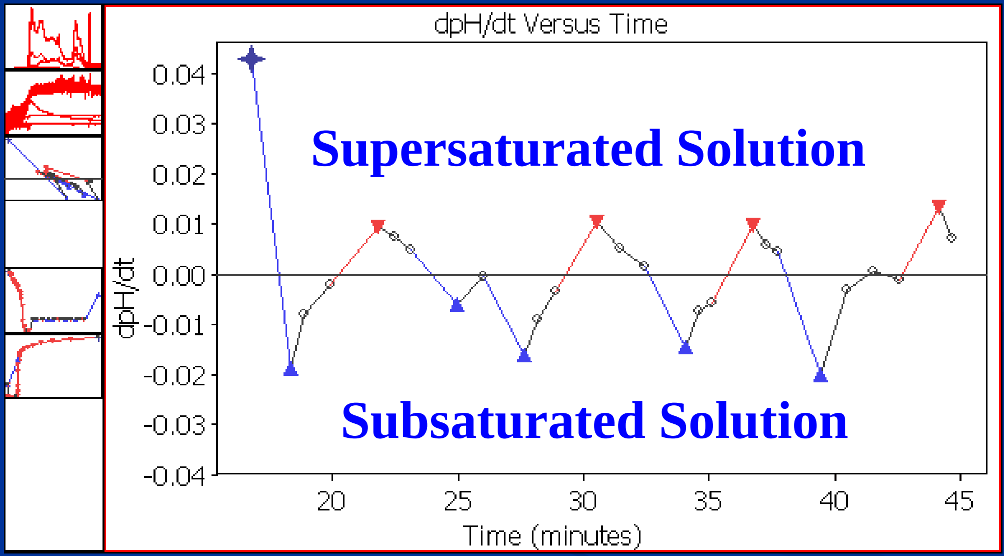
In this case, we have decided to create our own data and not to combine it with other literature data.

Potentiometric cycling method for very accurate and controlled measurement of solubility

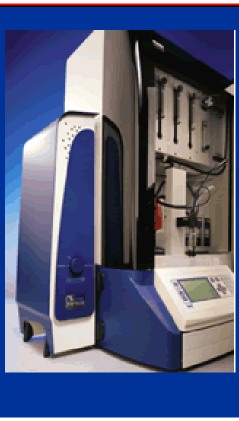


A Na → A^- ·········· Na^+ ⇌ AH ⇌ AH





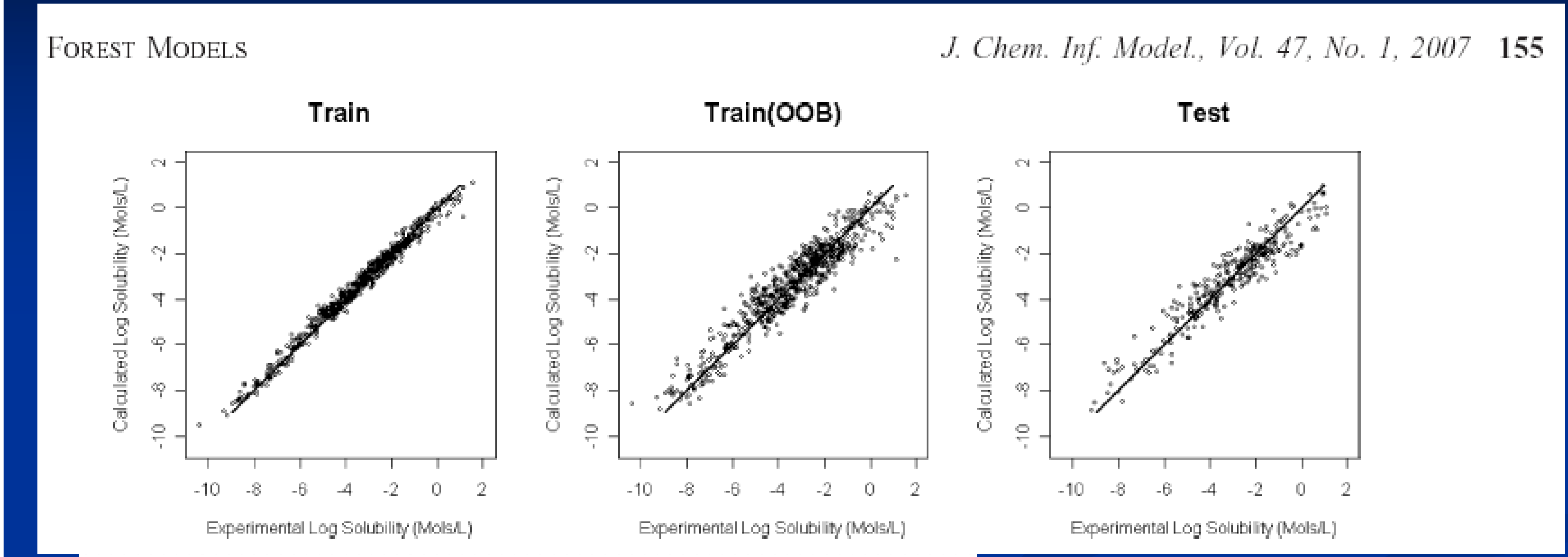
Stuart, M., Box, K. Chasing equilibrium: measuring the intrinsic solubility of weak acids and bases. Anal. Chem. **2005**, 77(4), 983-990.



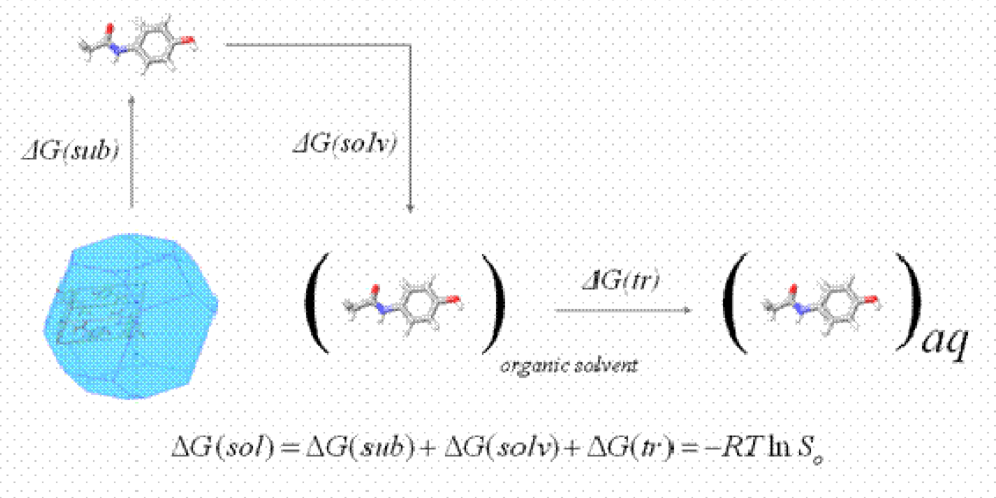
We use a Sirius glpKa instrument
With a DPAS detector

Random Forest Models To Predict Aqueous Solubility. Palmer, David S.; O'Boyle, Noel M.; Glen, Robert C.; Mitchell, John B. O.. Journal of Chemical Information and Modeling (2007), 47(1), 150-158.

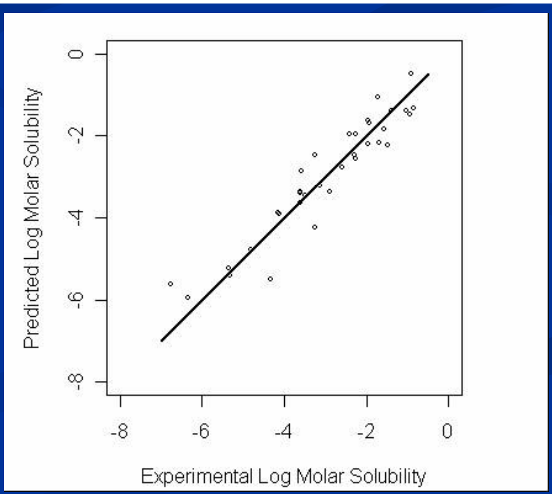
Random Forest models from solubility data : interestingly, minor improvement using QSPR methods even using the accurate data. A full thermodynamic cycle also gives minor improvement.



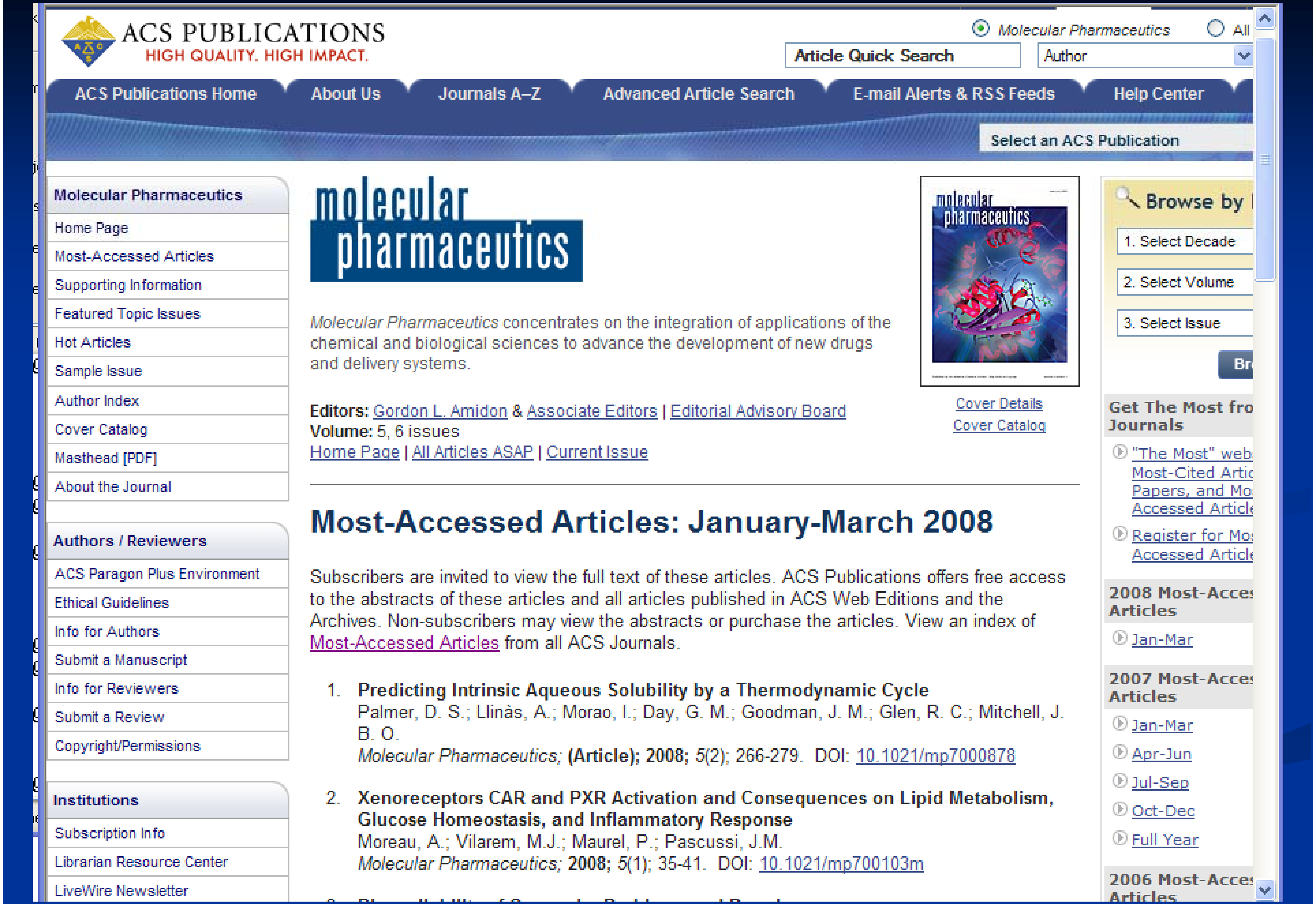
FOREST MODELS *J. Chem. Inf. Model., Vol. 47, No. 1, 2007* **155**



$$\Delta G(sol) = \Delta G(sub) + \Delta G(solv) + \Delta G(tr) = -RT \ln S_o$$



Predicting Intrinsic Aqueous Solubility by a Thermodynamic Cycle (published in Molecular Pharmaceutics)
David S. Palmer, Antonio Llinàs, Iñaki Morao, Graeme M. Day, Jonathan M. Goodman, Robert C. Glen, John B. O. Mitchell



Molecular Pharmaceutics concentrates on the integration of applications of the chemical and biological sciences to advance the development of new drugs and delivery systems.

Editors: Gordon L. Amidon & Associate Editors | Editorial Advisory Board
Volume: 5, 6 issues
Home Page | All Articles ASAP | Current Issue

Most-Accessed Articles: January-March 2008

Subscribers are invited to view the full text of these articles. ACS Publications offers free access to the abstracts of these articles and all articles published in ACS Web Editions and the Archives. Non-subscribers may view the abstracts or purchase the articles. View an index of Most-Accessed Articles from all ACS Journals.

1. **Predicting Intrinsic Aqueous Solubility by a Thermodynamic Cycle**
   Palmer, D. S.; Llinàs, A.; Morao, I.; Day, G. M.; Goodman, J. M.; Glen, R. C.; Mitchell, J. B. O.
   Molecular Pharmaceutics; **(Article); 2008;** *5*(2); 266-279. DOI: 10.1021/mp7000878

2. **Xenoreceptors CAR and PXR Activation and Consequences on Lipid Metabolism, Glucose Homeostasis, and Inflammatory Response**
   Moreau, A.; Vilarem, M.J.; Maurel, P.; Pascussi, J.M.
   Molecular Pharmaceutics; **2008;** *5*(1); 35-41. DOI: 10.1021/mp700103m

So, we have created a Solubility Challenge

- *JCIM* Solubility Challenge: Coming Soon! Please check Journal of Chemical Information and Modeling for more details.

- We deposited 100 accurate measurements of intrinsic solubility of drug-like molecules.
- You predict 40 unknowns
- JCIM publish the ‘best’ attempts.

Before beginning

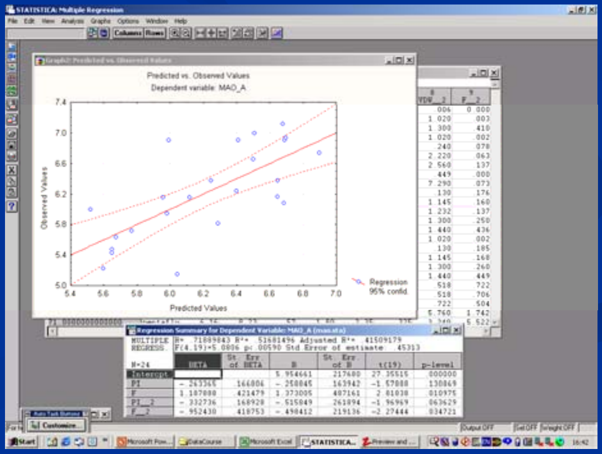
- This is obvious, but….Even if you use a pre-computed model, check the data sources
  - Are data compatible, and can they be combined
    - It often the case that non-compatible data are merged to create a database 'large enough' to do statistics on
  - Is there sufficient background information to determine the model's relevance
    - The 'ontology' of the information can be vital – what were the units of measurement ? (in the solubility example, some have mixed up ug/ml and umol/ml
  - Do they cover the 'chemical property space' required
    - Are my compounds very different from those used in the model ?

So, if we have accurate data, what's in a model ?

Molecular database

Calculate/measure molecular parameters

This is the most common 'paradigm for molecular analysis and prediction



VDW__2	F__2
.006	0.000
1.020	.003
1.300	.410
1.020	.002
.240	.078
2.220	.063
2.560	.137
.449	.000
7.290	.073
.130	.176
1.145	.160
1.232	.137
1.300	.250
1.440	.436
1.020	.002
.130	.185
1.145	.168
1.300	.260
1.440	.449
.518	.722
.518	.706
.722	.504
5.760	1.742
3.240	5.522

Regression Summary for Dependent Variable: MAO_A (mao.sta)

MULTIPLE REGRESS. R= .71889843 R²= .51681496 Adjusted R²= .41509179
F(4,19)=5.0806 p<.00590 Std.Error of estimate: .45313

N=24	BETA	St. Err. of BETA	B	St. Err. of B	t(19)	p-level
Intercept			5.954661	.217680	27.35515	.000000
PI	-.263365	.166806	-.258045	.163942	-1.57088	.130869
F	1.187080	.421479	1.373005	.487161	2.81838	.010975
PI__2	-.332736	.168928	-.515849	.261894	-1.96969	.063629
F__2	-.952430	.418753	-.498412	.219136	-2.27444	.034721

Analysis

$$\text{LogP} = \sum_{i=1,N} a_i f_i + \sum_{j=1,M} b_j F_j$$

Prediction

What's in a model ?

- The objective is usually (in drug discovery) to select a molecule (e.g. molecular similarity) or predict a property of a molecule and even explain the properties observed in another experiment.

- ***All models rely on the variance of the data***
- ***All models are susceptible to database bias***

- That is, the range of data values **and** their distribution.
  - If the points all had the same value, they would be easy to look up, there would be no model and one prediction for everything
  - The point is to extract a relationship between **calculable** parameters and the property of interest
  - The design of the experiment to obtain the data is therefore very important (and often ignored) – experimental design (Chemometrics can help)

Methods to discover models

- Models are generated using statistical or machine learning methods
  - Statistical methods usually rely on a normal distribution of the data and provide a fit to the data while minimising the error in the fit.
  - Are either supervised (e.g. regression) or unsupervised (e.g. principal components)
  - Machine learning methods are usually heuristic based and nearly all depend on local clustering (classification) – There are lots of flavours…..

Methods for Machine Learning….there are many……

Modeling conditional probability density functions: regression and classification

- Artificial neural networks
- Decision trees
- Gene expression programming
- Genetic algorithms
- Genetic programming
- Dynamic programming
- Gaussian process regression
- Linear discriminant analysis
- K-nearest neighbor
- Minimum message length
- Perceptron
- Quadratic classifier
- Radial basis function networks
- Support vector machines

Modeling probability density functions through generative models

- Expectation-maximization algorithm
- Graphical models including Bayesian networks and Markov Random Fields
- Generative Topographic Mapping

Approximate inference techniques

- Markov chain
- Monte Carlo method
- Variational Bayes
- Variable-order Markov models
- Variable-order Bayesian networks

Optimization

- Most of methods listed above either use optimization or are instances of optimization algorithms

Meta-learning (ensemble methods)

- Boosting
- Bootstrap aggregating aka bagging
- Random forest
- Weighted majority algorithm

Inductive transfer and learning to learn

- Inductive transfer
- Reinforcement learning
- Temporal difference
- Monte-Carlo method

They can be traced back to the ID3 method of Ross Quinlan – worth a look

Some comments about making models (includes QSAR, SAR, QSPR…)

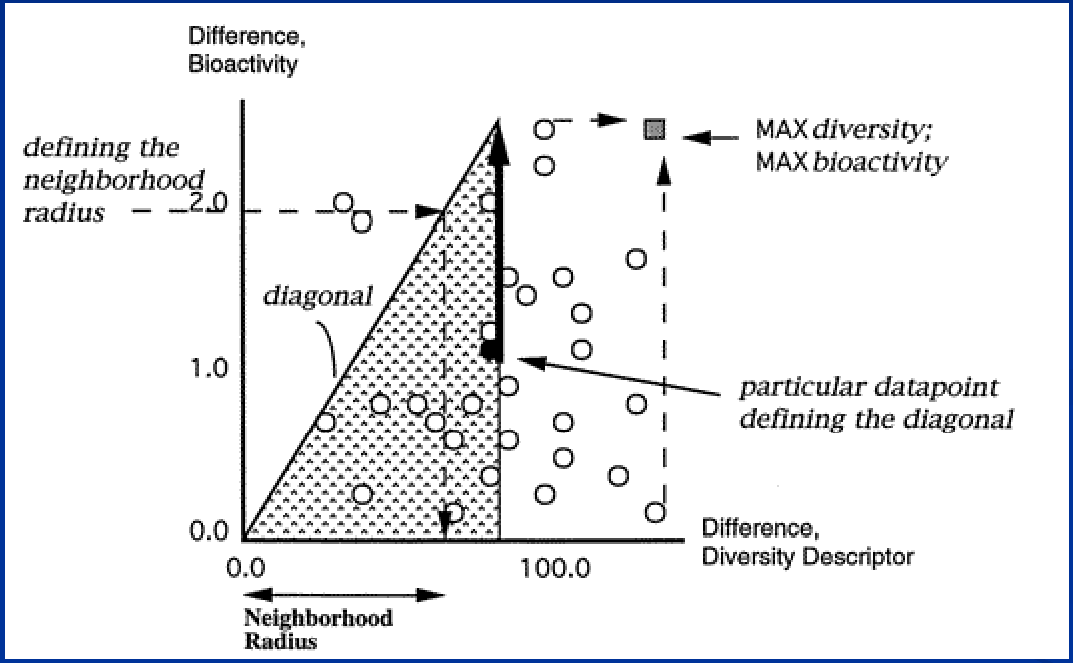
- The parameters used to predict a physical property (like solubility and logP) compared to e.g. a binding affinity must often behave in a fundamentally different way.
- Reason: a property like logP in octanol/water is consistent in that the medium doesn't change. However, both the medium (the receptor) and the ligand change upon binding and different ligand/receptor combinations really require different models!

Property behaviour

- So, in property space, we should expect behaviour that was consistent in that it was : linear, exponential, parabolic – i.e. predictable
- However, in SAR space – it's disjointed and, if we're lucky, **clustered** e.g. depending on the mode of binding (if you look at SAR predicted/measured plots in the literature, many join clusters and not compounds)
- So, parameters must have the following 'property'
  - Small changes in the parameter should produce small changes in the bio-activity (e.g. affinity)
  - Large changes in the parameter can produce large or small changes in the affinity
  - This is exactly how medicinal chemists optimise compounds

  - Neighborhood Behavior: A Useful Concept for Validation of "Molecular Diversity" Descriptors
    Patterson, D. E.; Cramer, R. D.; Ferguson, A. M.; Clark, R. D.; Weinberger, L. E.
    J. Med. Chem.; (Expedited Article); 1996; 39(16); 3049-3059. DOI: 10.1021/jm960290n

This is neatly summed up in this paper, which analysed diversity and similarity



- Neighborhood Behavior: A Useful Concept for Validation of "Molecular Diversity" Descriptors
Patterson, D. E.; Cramer, R. D.; Ferguson, A. M.; Clark, R. D.; Weinberger, L. E.
J. Med. Chem.; (Expedited Article); 1996; 39(16); 3049-3059. DOI: 10.1021/jm960290n

So – does (Q)SAR work ?

- Yes, for localised sets of compounds – often simple parameters, if spatially localised and functionally dependant, will e.g. provide a useful regression
- A mistake is often to use a dataset of molecules and their activities that actually requires **multiple models (see our 5-HT example earlier)**

- Another is to rely on vast numbers of parameters and model selection such as cross validation. I'm not a great fan of ' lets use all the available parameters and cross-validation will save the day' – the variance of a large number of parameters will often match the variance of the data – just put in enough variables.
- Cross validation can be tricky. We need enough 'similar' molecules to perform cross validation, which again raises the problem of memorising subsets.

Overfitting and cross validation
- three papers to read by Douglas Hawkins

- The Problem of Overfitting
  Hawkins, D. M.
  J. Chem. Inf. Comput. Sci.; (Perspective); 2004; 44(1); 1-12. DOI: 10.1021/ci0342472
- Assessing Model Fit by Cross-Validation
  Hawkins, D. M.; Basak, S. C.; Mills, D.
  J. Chem. Inf. Comput. Sci.; (Article); 2003; 43(2); 579-586. DOI: 10.1021/ci025626i
- QSAR with Few Compounds and Many Features
  Hawkins, D. M.; Basak, S. C.; Shi, X.
  J. Chem. Inf. Comput. Sci.; (Article); 2001; 41(3); 663-670. DOI: 10.1021/ci0001177

Which leads on to another problem, database bias…a particular problem with
The kind of limited molecular diversity we typically deal with

Database bias. ‘Sophisticated models’ are sometimes little better than simple models. The ‘Database bias’ in activity databases is simply that the active molecules are generally very similar classes and are memorised !

Put another way, the information content of many common structure-based descriptors for virtual screening purposes is, in some cases, not higher than the nonstructural information about the number of atoms per element in the structure.

Below, is an example using only **atom counts** compared to more complex similarity descriptors. Note the high performance of the ‘dumb descriptors’

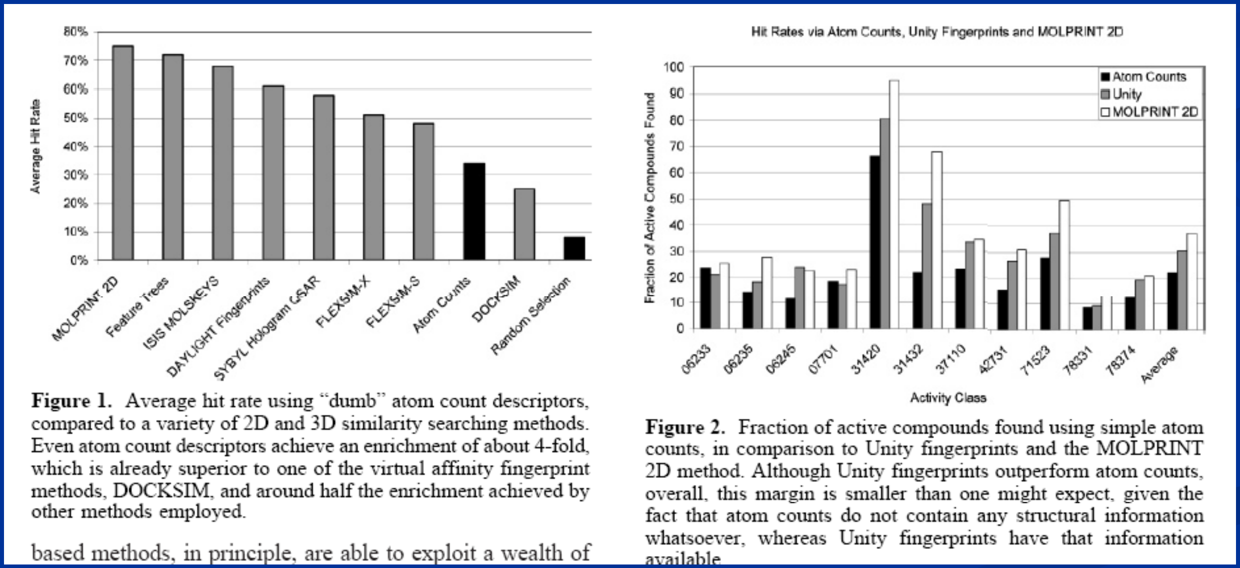


Figure 1. Average hit rate using “dumb” atom count descriptors, compared to a variety of 2D and 3D similarity searching methods. Even atom count descriptors achieve an enrichment of about 4-fold, which is already superior to one of the virtual affinity fingerprint methods, DOCKSIM, and around half the enrichment achieved by other methods employed.

based methods, in principle, are able to exploit a wealth of

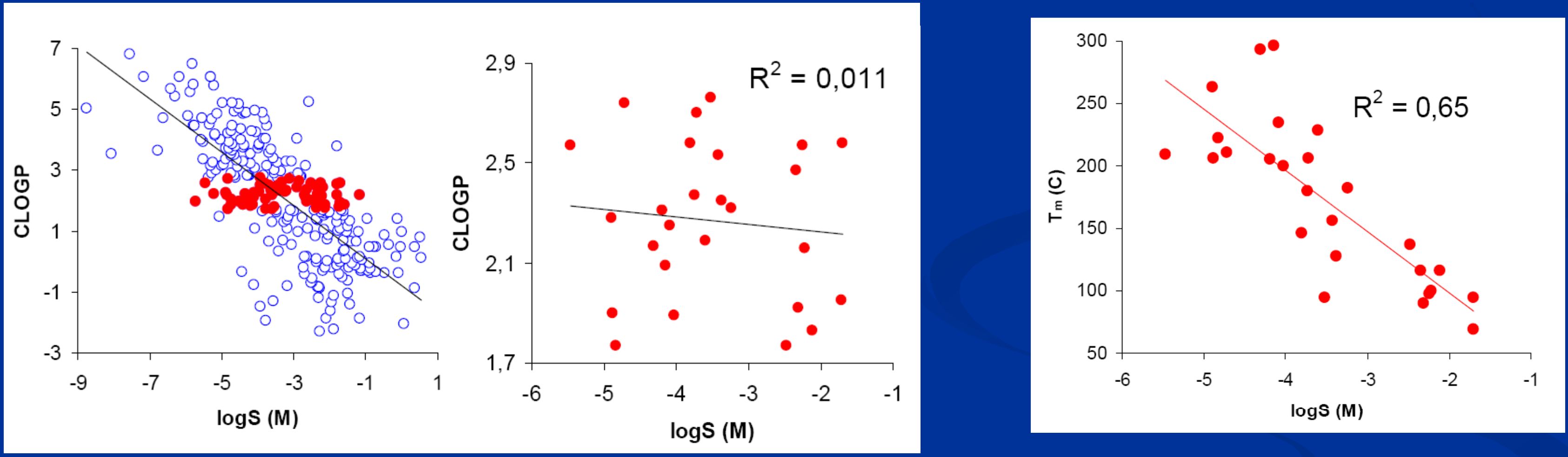
Figure 2. Fraction of active compounds found using simple atom counts, in comparison to Unity fingerprints and the MOLPRINT 2D method. Although Unity fingerprints outperform atom counts, overall, this margin is smaller than one might expect, given the fact that atom counts do not contain any structural information whatsoever, whereas Unity fingerprints have that information available.

Bender et al. J. Chem. Inf. Model. ***2005, 45, 1369-1375 1369***

A simple model example, again using solubility – putting in parameters that relate to the phenomenon

- The failure to account for the influence of the solid state on solubility
- The General Solubility Equation is a rare examples that does.

$$\text{LogS} = 0.8 - \log P - 0.01(\text{MP}-25)$$



(from Wassvik, C. Uppsala Pharmaceutical Profiling Conference)

But, accounting for the solid state requires an understanding of the dissolution process, and our earlier slide showed that this is still missing a fundamental property (or two)

Conclusions

- The data is king – comprehensive, in an extensible format is best (XML)
- Parameters in a model should relate to the phenomenon being studied. If not, smell a rat.
- Machine learning methods have the property of local models – best for discontinuous SAR data
- Combining pattern recognition and phenomenological modelling with experiment is best
- Design the testing regime for the model before creating it – can it be properly tested ?

Acknowledgements

Jonathan Goodman, John Mitchell, Peter Murray-Rust, Tony Llinas,

James Bell, Max Macaluso, Andreas Bender, David Palmer

EPSRC, BBSRC, Gates Trust, Tripos, Unilever