

# Evolving Predictors of Molecule-Cytochrome P450 Interaction for Drug Discovery

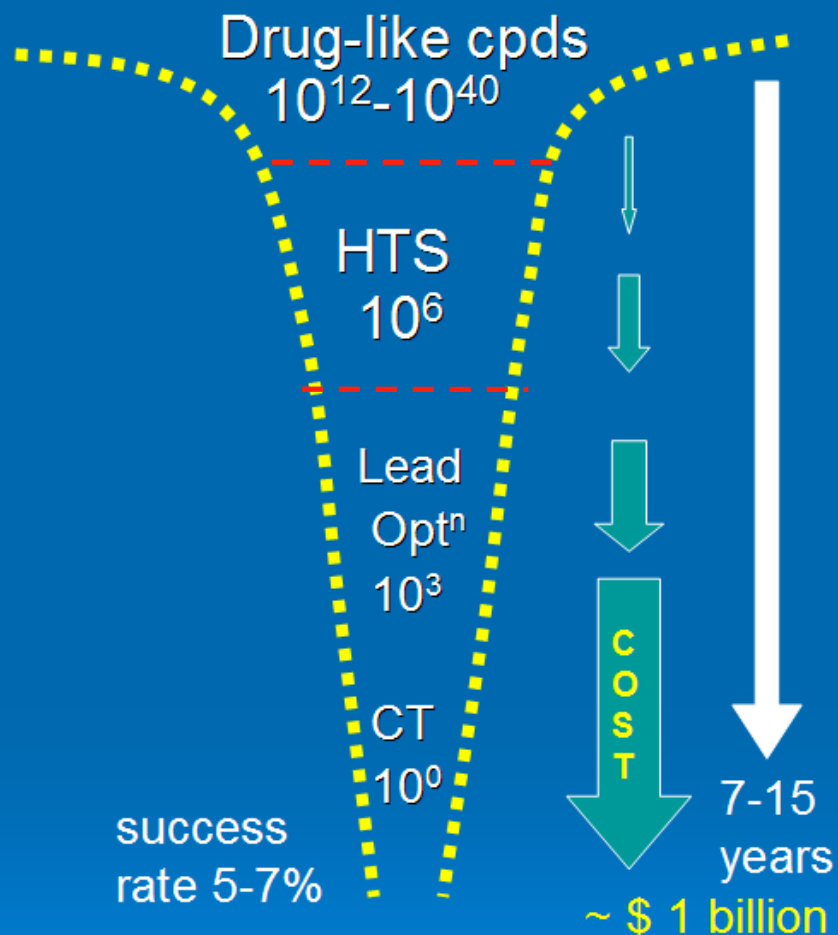
**Dr Steven Barrett, GlaxoSmithKline**

*Cercia Workshop on Computational Intelligence in Cheminformatics,  
Conference Park, University of Birmingham, 2nd March 2006*

# Summary

- Drug discovery process
- Adverse drug reactions
  - Importance of Cytochrome P450
- Predicting molecular P450 interaction
  - Approaches and problems
- Evolutionary '*in silico*' prediction exp'ts
  - Brief evolutionary computing and GP
  - GP classification/regression
  - Ensemble

# Early Drug Discovery - Summarised !



## ➤ Initial screening for molecules effecting therapeutic targets

- target disruption hypothesised to form the basis of treatment
- est. search space in region of  $10^{12}$  -  $10^{40}$  drug-like chemicals.
- high throughput screening of the order of  $10^6$  measured.

## ➤ Computer models

- **filters** used in early stages to help focus discovery efforts
- guide scientists where to look first
  - which existing sets of molecules to test next for specific activities
  - 'virtual screening', predicting for vast numbers of molecules that do not physically exist
    - interesting chemicals made and tested

# Adverse Events, Cytochrome P450 and Drug Discovery

- **As candidate drugs proceed from discovery to development research costs escalate**
  - greater losses the later a potential drug is abandoned
  - major cause of late failure is 'off-target' interactions with adverse effects
- **Cytochrome P450 enzymes** [ <http://www.icgeb.org/~p450srv/> ]
  - Heme containing enzymes, mainly in the liver, key to catabolism
  - They not only metabolise endogenous substrates, **but also exogenous ones, incl. drugs**
  - 26 Human Cyt p450 families distinguished
  - Isoenzymes of families 3A-, 2C-, 2D- and 1A- most important to pharma
    - responsible for metabolising some 70% of drugs on the market.
    - **2D6, 3A4, 2C9 account for ~70% of p450 drug metabolism**
      - most studied/screened against

# Early Study of P450 interaction

- P450 metabolism can influence the whole *pharmacological and ADME / toxicological profile* of drugs which are **substrates**
- Importantly, some drugs can also **inhibit or enhance the normal function of cytochrome p450s** leading to *Adverse Drug Reactions*
- *In vitro* screening finds worst molecules - before clinical trials
  - Don't screen-out substrates therapeutically active at lower concentration
- Companies use various methods for computer-based studies
  - **Protein Studies** – active site information from 3D re-construction – needs amino-acid sequence and ideally the actual protein crystal structure (lacking for eukaryotic membrane-bound p450s)
  - **Pharmacophore models** – require active site information plus template molecules – OK for substrates assuming similar active-site orientation, difficult for inhibitors (specific site of reaction is lacking)
  - **QSAR** – quantitative structure-activity relationship modelling
    - predict biological effect from computed chemical features, calculated/measured physico-chemical properties

# Issues for *in silico* Prediction

- Difficult to obtain p450 models with good generalisation
  - P450s present a **multi-mechanism situation**
    - substrates – metabolised in different ways
    - inhibitors – competitive/non-competitive and for different underlying reasons
  - P450 **measurements from limited compound space**
    - previous published efforts limited by this, esp. wrt **model validation**
- Models have to be used very cautiously – restricts utility
  - ***alert flags rather than a decisions***
- Can't easily extrapolate *in vitro* data to clinical world !
  - Many P450s with different activities/substrate selectivities and overlaps
  - Individual isoenzyme differences across human populations
    - continuum of effects of polymorphisms
    - differing outcomes in different polymorphic individuals/sub-populations
      - adverse Drug Reactions, lack of drug efficacy, etc.

# Evolutionary Computing

- Family of techniques based on Darwinian natural selection

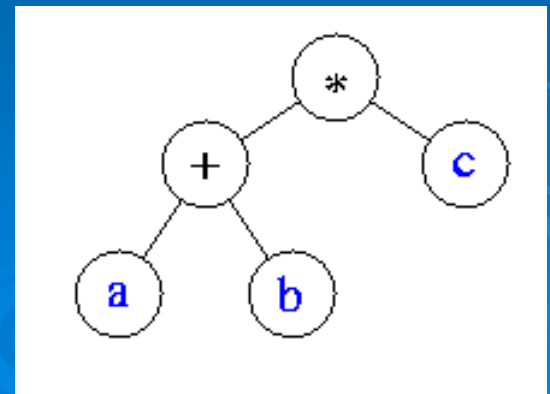
Biological Metaphor :

EVOLUTION		PROBLEM
Environment	<----->	Problem
Individual	<----->	Candidate Solution
Fitness	<----->	Quality of Solution
Recombination/mutation	<----->	Adaptation of Solution

- Problems posed as optimisation of parameters via a **'fitness' function**
- Includes Genetic Algorithms, **Genetic Programming**

# What is Genetic Programming?

- Genetic Programming creates candidate **computer programs as the solution**
  - unlike Genetic Algorithms which usually create a string of numbers that represent the solution
- Model represented as tree, i.e.  $\text{eval}=(a+b)*c$ 
  - $a, b, c$  inputs to model, leaves of tree
  - $+*$  functions, internal nodes of tree
  - result from root node of tree



# Why is GP Appropriate here?

- **Well proven in optimisation and function fitting**
  - incl. classification & regression tasks
- **Heuristic search of vast, discontinuous solution spaces possible**
  - solutions of unknown (or controlled) size and shape
  - minimal 'intervention' by user,
    - but can be interactive, incorporating user-proposed partial solutions
- **Custom fitness functions** can be created for specific problems
- Can give **interpretable and robust models**
- **Multiple solutions** can be found

# P450 2D6 GP Modelling Expt.

- GSK formed by merger of Glaxo Wellcome + SmithKline Beecham in 2000
  - GW and SB had large libraries of molecules for research in *different* therapeutic areas
  - post merger effort to measure f-GW molecules in f-SB Cyp 2D6 assay system
  - **Opportunity : better test of generalisation performance than previously possible**
- GP effort was part of a wider blind trial to better assess the true feasibility of predictive technologies for *In Silico* modelling
  - Train on f-SB molecules
  - Test on f-SB molecules
  - Validation - extrapolate to f-GW molecules

# *In vitro* p450 Assay - 50% Inhibition Concentrations

Molecules split into 3 classes on IC50 measure

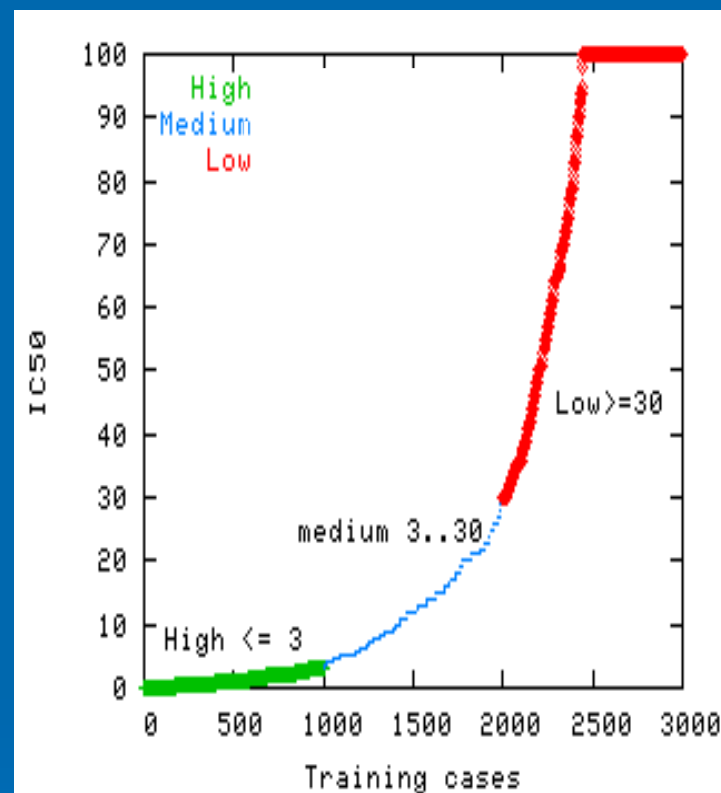
- inhibitor ( $\leq 3\mu\text{M}$ )
- substrate ( $>3$  to  $<30\mu\text{M}$ )
- inactive ( $\geq 30\mu\text{M}$ )

Quiet large compound data sets:

- f-SB Training – random - 3000 cpds  
equally ‘balanced’ across classes
- f-SB Test - 4570 cpds
- f-GW Extrapolation - 1932 cpds

But:

- Test / extrapolation sets – *unequal* class splits



Total Inhibitor Substrate Inactive

1000	1000	1000	3000	Training SB	Value given
4570	91	1562	2916	Testing SB	no value
1932	114	446	1372	Validn GW	no value

# Variables used in Predicting Compound Interaction Class

- 121 “features” calculated from chemical structure
  - Boolean, categorical or continuous
  - Selection mostly previously found useful in predicting p450 interactions
    - Structural
      - i.e. presence/absence of basic N atom, aromatic groups, etc
    - Physicochemical properties
      - i.e. lipophilicity, charge, acid/base, etc
- No feature transformations applied – used ‘as is’

# Predicting P450 inhibition

- Set Goal: Categorisation to the 3 classes
- Two GP approaches :
  - **Regression**
    - single GP tree individual to predict continuous IC50 value
    - then threshold to give class
  - **Classification**
    - individual of 3 GP sub-trees, using 3-way 'winner-take-all' strategy
    - 1 tree per class not hard coded - GP *evolved* this !
      - inhibitor vs substrate/inactive
      - substrate vs inhibitor/inactive
      - inactive vs substrate/ inhibitor
    - class given by GP sub-tree returning highest value

# Basic GP set-up

- Function set - kept fairly simple
  - + - \* / IFLTE min max minA maxA
- Terminal set
  - 121 features
  - numbers 0 - 9
  - 90 unique constants
- Population size = 5000 individuals
  - Selection – non-elitist, generational 7-tournament
  - Mutation - various ...
- 5 runs each for classification and regression
  - Termination: 50 generations

# Mutation – previous experience

➤ Boosting degree of variation for evolution to work on enhances possibilities

- **high mutation rate**
- **several types of mutation used**

Select part of a fit program tree, then either

- replace sub-tree with randomly created new one
- shrink: replace sub-tree with part of itself
- replace function with another with same number of inputs
- replace constant with Gaussian value near it
- replace an input by another

# GP Regression

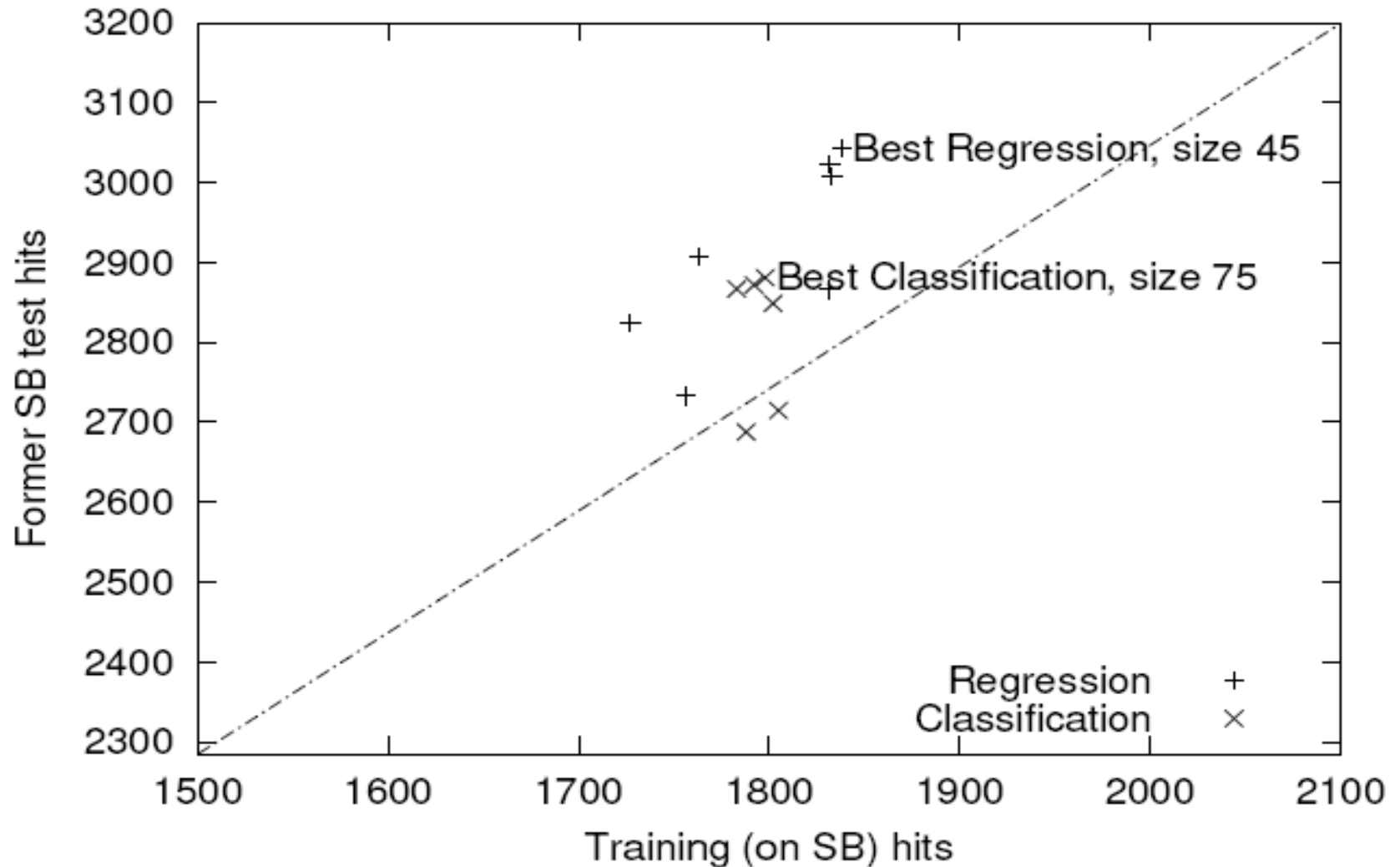
- Fitness =  $(20,000 * \text{\#hits}) - \sum |f_i - \text{IC50}_i|^2$ 
  - $f_i$  (floating pt ) output of GP tree for one input example  $i$
  - a hit is a correctly classified example after thresholding
  - Hits vs  $\sum \text{error}^2$  weighting value of 20,000 found by experiment
    - trying to roughly balance of terms in fitness
- Individual : Single tree
- Best individual: generalisation across the 3 classes
  - Simplified by further GP run (altering max tree size param) and then by hand (size reduced from 65 to 45)

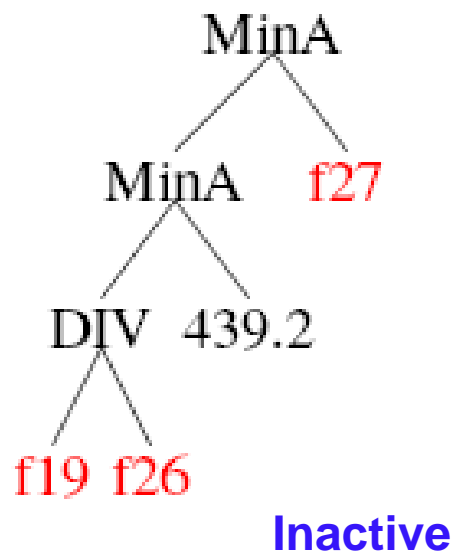
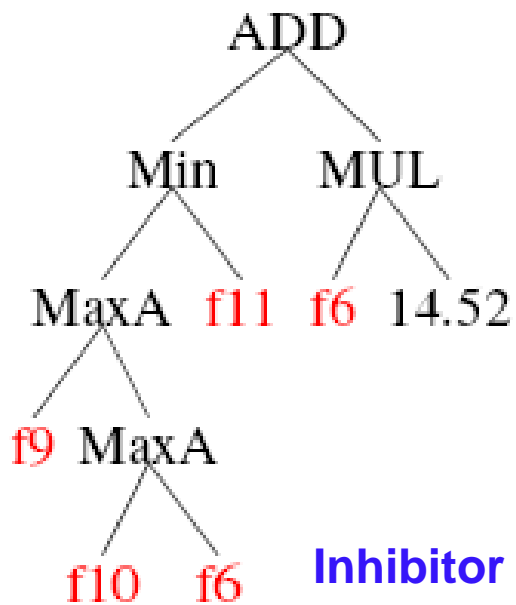
# Classification

- Fitness = 'hits' (= # correctly classified examples)
- Individual : 3 trees (...evolved as 1 per class)
- Best individual: generalising across f-SB data
  - Simplified by further GP run, altering max tree size param (size reduced from 93 to 75)

# Generalisation to f-SB Training vs Test

Classification vs Regression: 5 runs ea.+ simplified



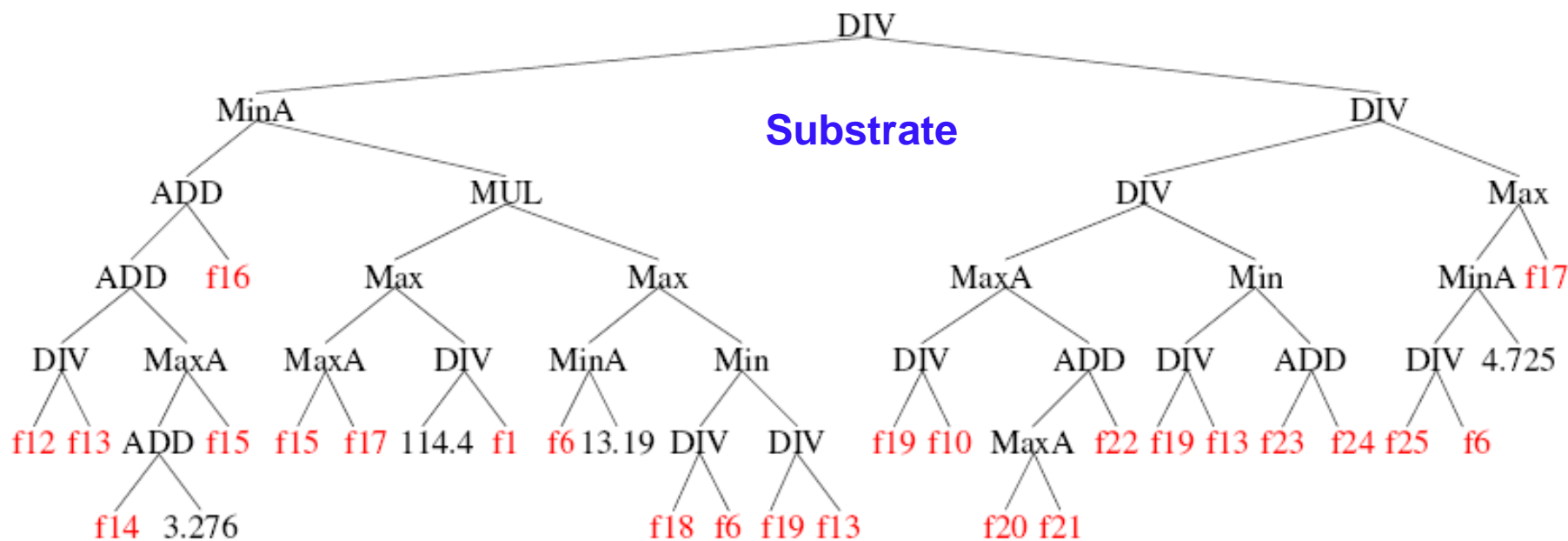


# Classification

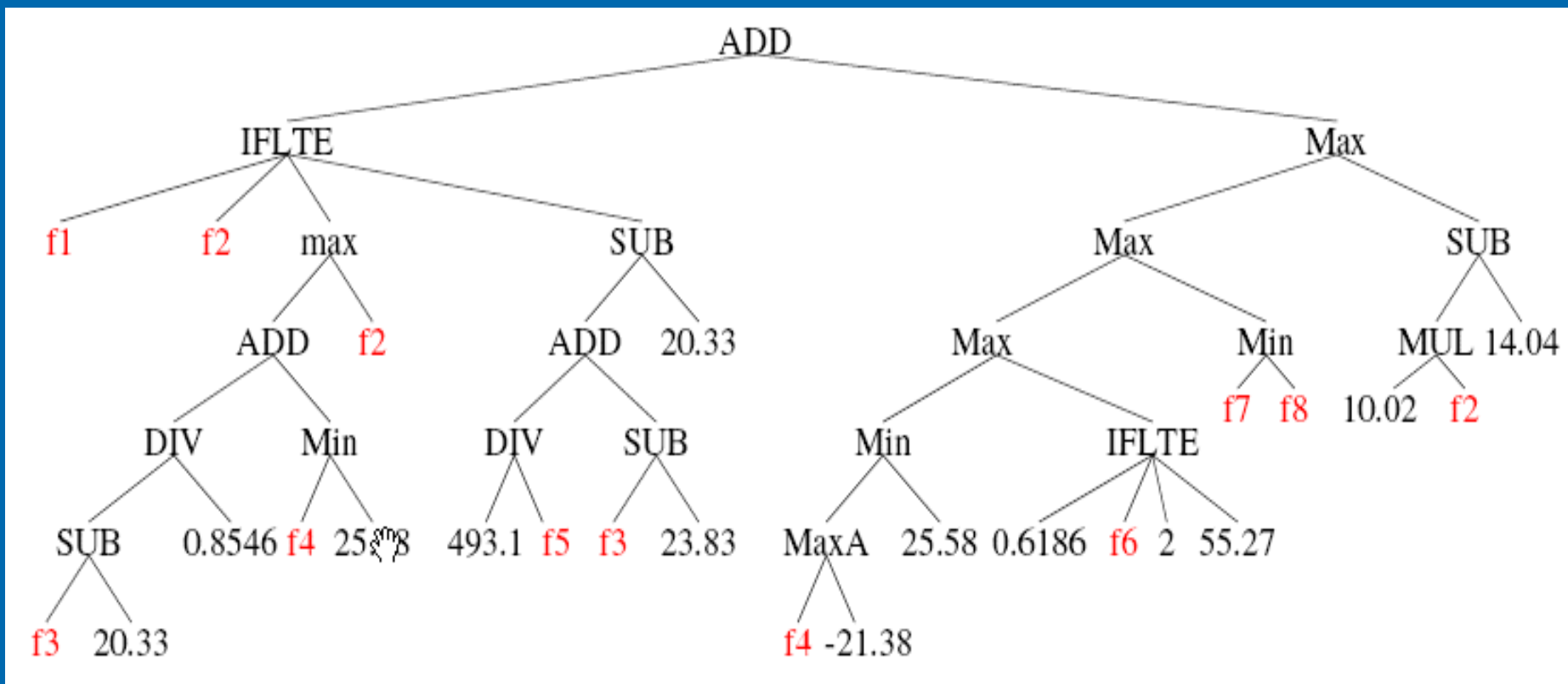
*evolved*

## Class-specific GP Sub-Trees

*f#s* = chemical features



# Simpler Evolved Tree – Regression



- The 8 features were chosen by GP direct from the 121 available
  - **f1, f2, ... f8** - fewer chemical features used overall
  - ....only 2 features used by both approaches?

# Regression << Results >> Classification

TRAINv2 fitness 1838 hits size 45				
	1	2	3	%Acc
1	<u>579</u>	294	127	<u>58</u>
2	188	<b>504</b>	308	<b>50</b>
3	45	200	<b>755</b>	<b>76</b>
Overall 61%				

TRAINv2 : fitness 1798 hits, size 75				
	1	2	3	%Acc
1	<u>568</u>	352	80	<u>57</u>
2	187	<b>569</b>	244	<b>57</b>
3	64	275	<b>661</b>	<b>66</b>
Overall 60%				

fSB_test1v2 fitness hits 3043 size 45				
	1	2	3	%Acc
1	<u>69</u>	19	3	<u>76</u>
2	306	<b>778</b>	479	<b>50</b>
3	153	566	<b>2197</b>	<b>75</b>
Overall 67%				

fSB_test1v2 : fitness 2881, size 75				
	1	2	3	%Acc
1	<u>56</u>	31	4	<u>62</u>
2	342	813	393	<b>53</b>
3	179	740	2012	<b>69</b>
Overall 63%				

fGW_test2v2 fitness hits 1187 size 45				
	1	2	3	%Acc
1	<u>41</u>	32	41	<u>36</u>
2	92	119	235	<b>27</b>
3	103	242	1027	<b>75</b>
Overall 61%				

fGW_test2_v2 : fitness 1109 hits, size 75				
	1	2	3	%Acc
1	<u>45</u>	26	43	<u>39</u>
2	62	148	236	<b>33</b>
3	96	360	916	<b>67</b>
Overall 57%				

Regression better overall vs f-SB, ...niether good vs f-GW !

# Overall Conclusions

- GP produced a relatively **compact, low complexity models**
  - small # features
  - understandable, made some sense to P450 modelling experts
- GP could show **reasonable generalisation *within* company compound library**
- GP was **less able to generalise *across* company compound libraries**

**Hard problem, but ..**

- not enough training compound diversity ?
  - was goal self-compromising – aiming for ‘excessive generality’ ?
- ....how to improve generalisation ?**

# How to Improve Generalisation?

- If '*uncontrolled generality*' compromises QSAR  
....can we improve classification by pooling predictions of multiple 'niche' classifiers ?
- "Committee of experts" could benefit from
  - 'expert' classifiers for multi-chemical series / multi-mechanisms
  - complementary expertise provided by 'niche' classifiers, one classifier performs well on examples where others are poor
  - different classifier types, neural nets, decision trees, etc
- How to determine ...
  - which classifier is best on which examples?
  - ...the best thresholds for multiple classifiers?
  - .....and how to combine classifiers?

# Experiment : p450 Inhibitor Classification

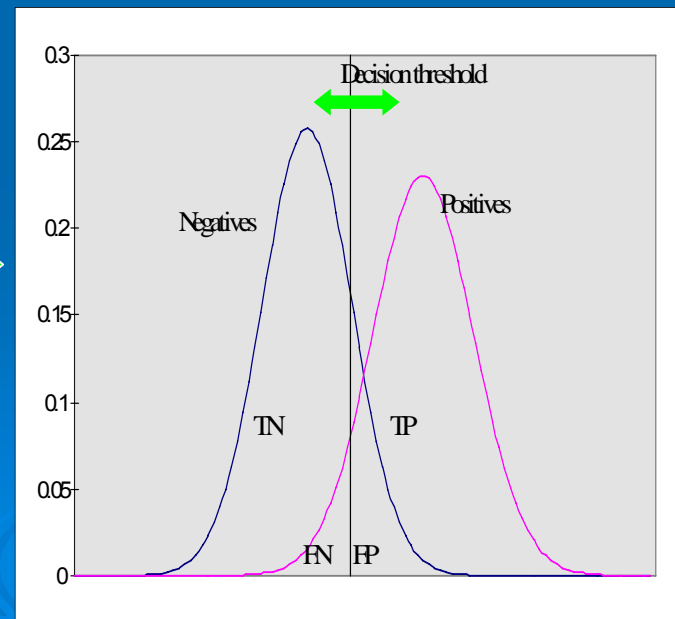
2 stage process to evolve GP-ANN ensemble model

## 1. Train 75 Artificial Neural Net - binary classifiers

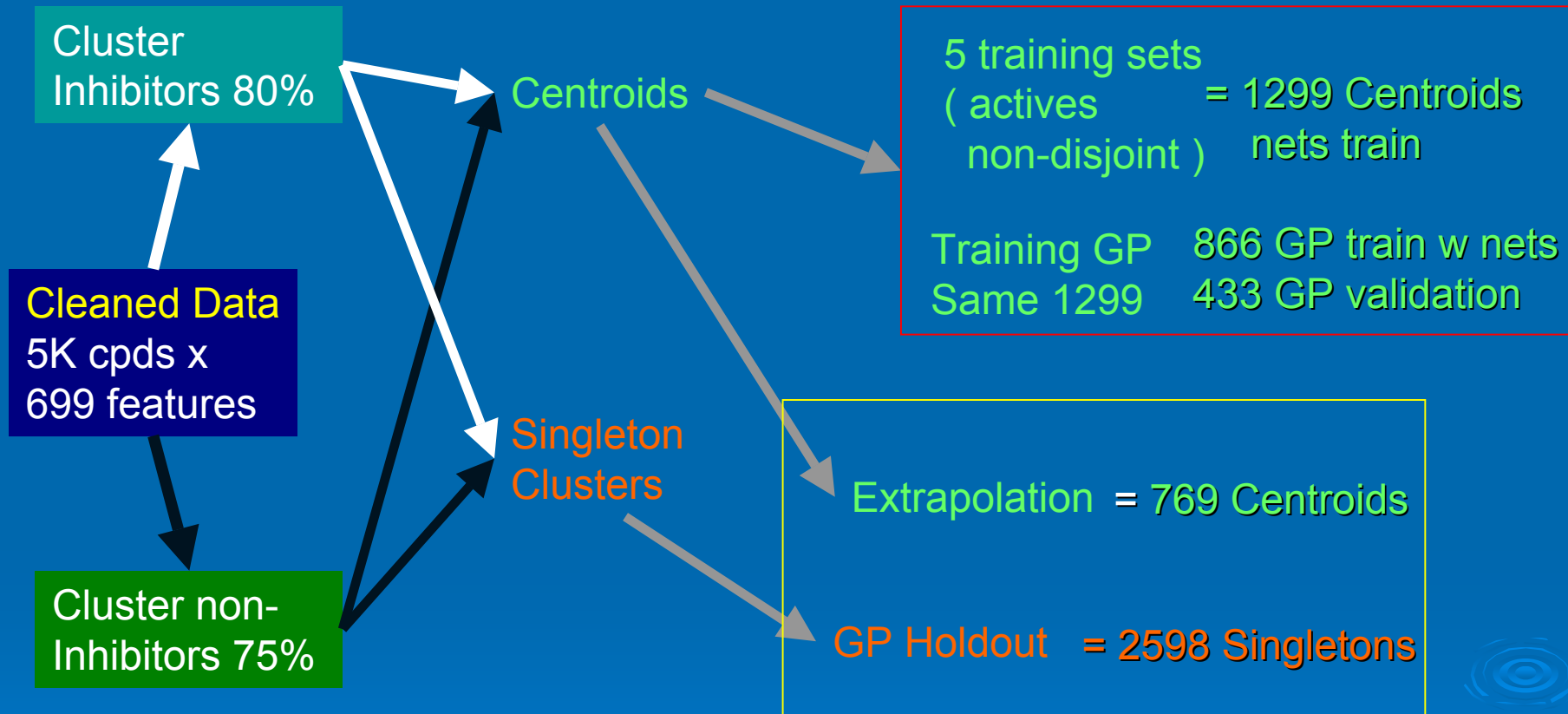
- 75 ANNs trained within SPSS Clementine

## 2. GP selects from 75 ANNs fusing them into a composite classifier

- *Optimises overall ensemble and individual ANN class separation cut-points*

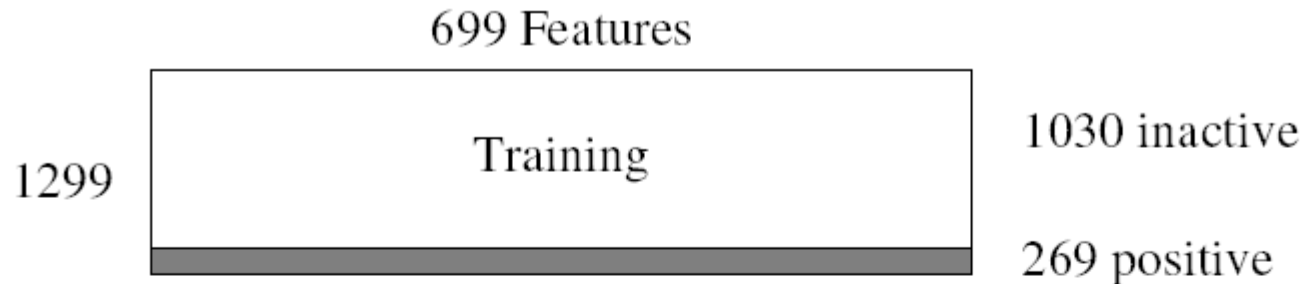


# P450 HTS Data sets

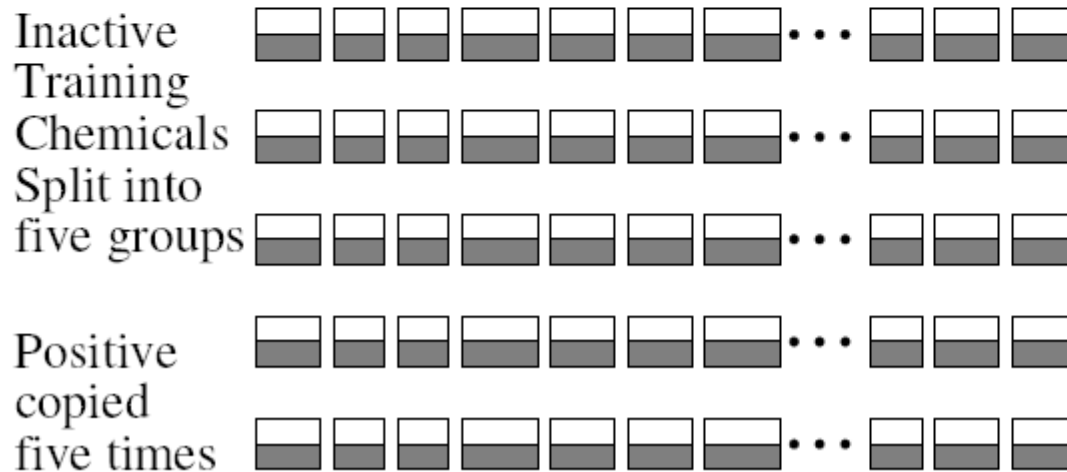


# Base-Classifier Training : ANNs

...**feature-specific**, early stopping to avoid over-fitting



699 features split into 15 related groups



75 training sets  
each used by Clementine  
to train one neural network

# Clementine BP nets: 'Packaged' for use by GP

- Nets Output prediction as confidence value, range 0-1
  - confidence nearness to 0 or 1 indicates greater certainty
    - Clementine thresholds at 0.5,  $\geq 0.5$  is an inhibitor
- -0.5 subtracted from ANN confidences for use in GP – larger abs value, greater confidence in *prediction for a specific case*    *zero is threshold for class*
  - Non-inhibitors – negative, -0.5 to  $< 0$
  - Inhibitors – positive, 0 to 0.5

(GP DOESN'T SEE ACTUAL COMPOUND FEATURES)

# GP Set-up: Classifier fusion

- Function set: + - \* / if Min Max Frac Int
  - plus* 75 'packaged' ANN classifier confidences
    - each a unary function accepting/returning a float
    - confidences implicit to individual training cases
- Terminal set: 100 unique random constants  $\{-1 \dots 1\}$ 
  - plus* special terminals accepting thresholding values  $\{0, 0.1, 0.2, \dots 1\}$
- Fitness = Area Under ROC
  - 11 thresholding cut-points (0, 0.1, 0.2, ... 1 ) applied to output of GP individual to compute
- Population 500, 50 generations (max)
  - selection / mutation - as on earlier slides

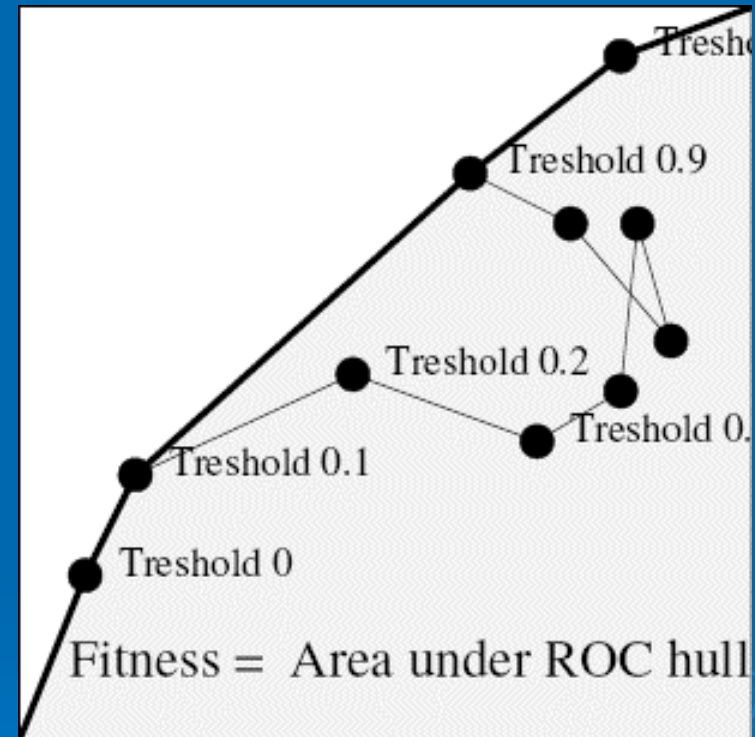
# Evolutionary Ensemble-Classifier Learning Cycle

{ ANN classifiers trained + packaged for GP }

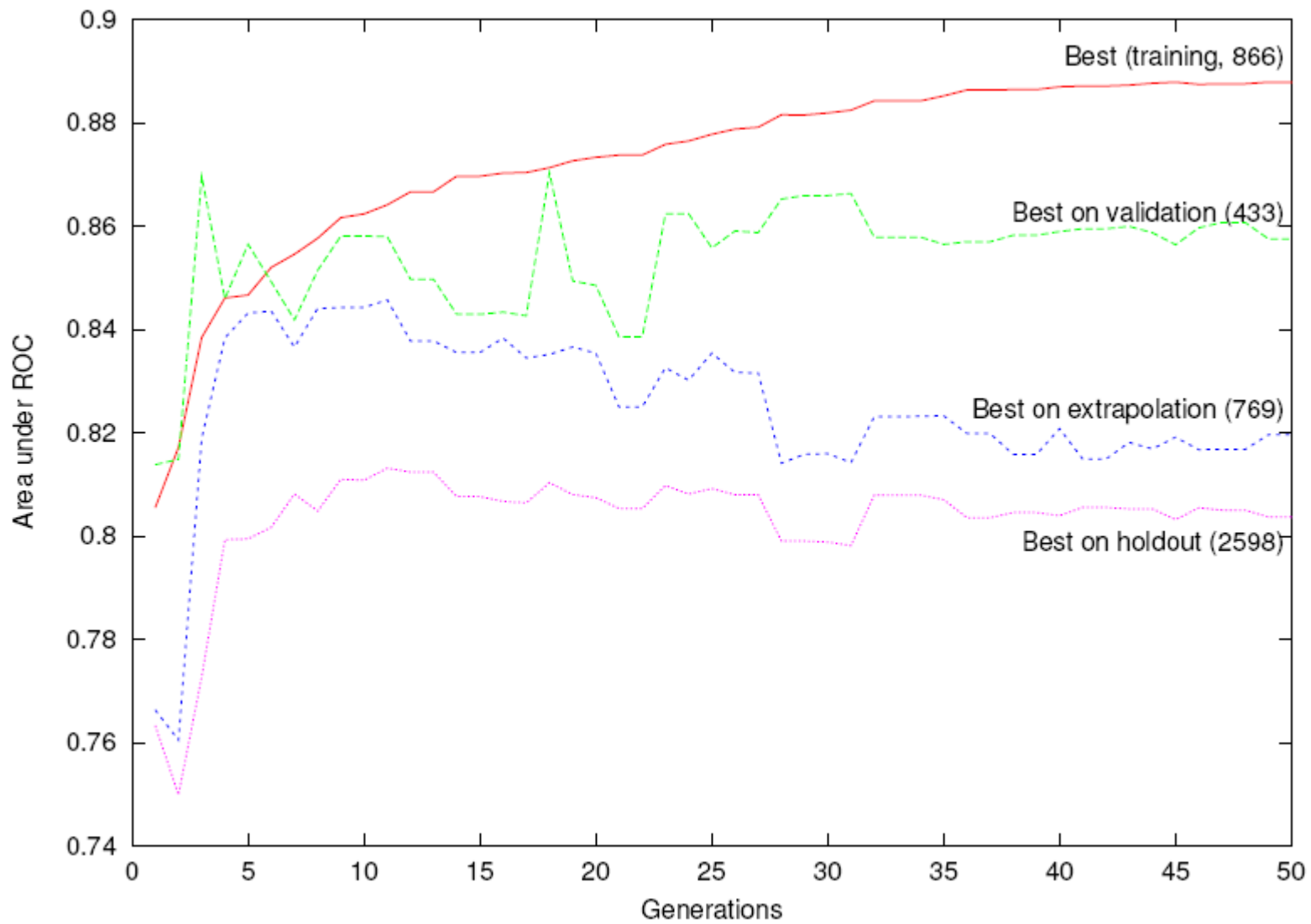
- 0 Random initial classifier combinations=GP individuals
- 1 Test each GP individuals' fitness
- 2 Select fittest (generation  $n$ )
  - Use of AUROC fitness optimises sub-tree (incl. classifiers) re-combination wrt decision threshold
- 3 Breed new generation of combinations with crossover and mutation (generation  $n+1$ )
- 4 Generation < 50? Y = Iterate 2-3  
N = Stop and Assess vs validation

# AUROC from Applied Multi-Thresholding

- Performance at a range of thresholds are calculated per GP classifier individual
- Only those on 'convex hull' of thresholding points are used to compute the AUROC
  - can always create an 'interpolated' classifier
  - deterministic, polynomial time alg.

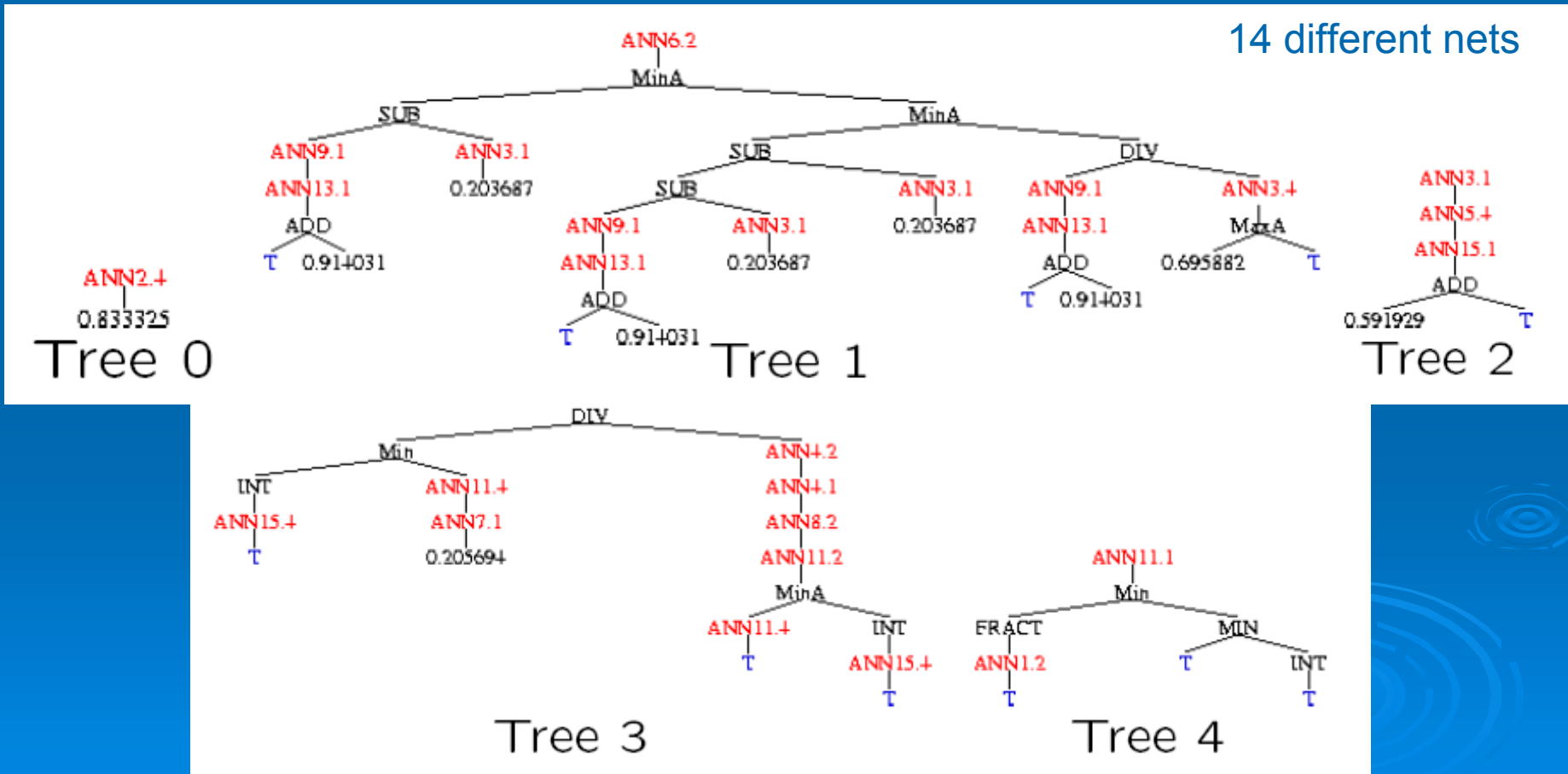


# Evolution of Fitness

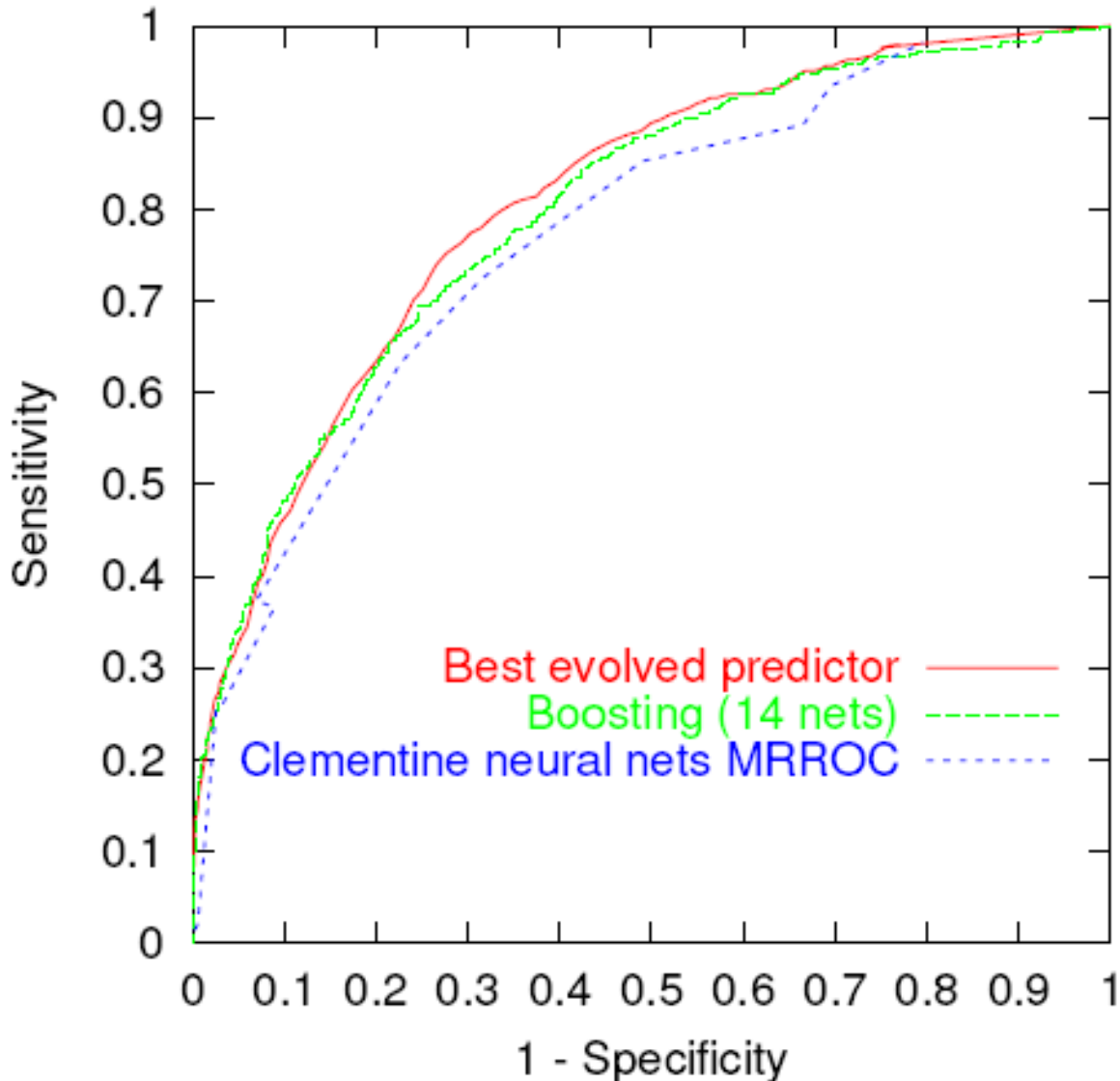


# GP Representation and 'Best Individual'

- Each GP individual consists of 5 trees
  - sum (values from individual trees 0-4)  $\geq 0$   $\rightarrow$  inhibitor



# ROC Best Individual vs Holdout (2598 singletons)



## Holdout AUROCs

**0.8096 GP gen50**

**0.7994 14net AdaBoost**

**0.7768 MRROC 75 nets**

# References

- W. B. Langdon , S. J. Barrett and B. F. Buxton (2003)  
Predicting Biochemical Interactions – Human P450  
2D6 Enzyme Inhibition. Proceedings of the 2003  
Congress on Evolutionary Computation CEC2003, pp.  
807-814, IEEE Press, 8-12 December 2003.
- W. B. Langdon and S. J. Barrett and B. F. Buxton (2001)  
Genetic Programming for Combining Neural Networks  
for Drug Discovery. Soft Computing and Industry  
Recent Applications, pp. 597-608, Springer-Verlag, 10-  
24 September 2001.


# Acknowledgements

- **Bill Langdon !**
- **EPSRC RAIS + Faraday INTERSECT**  
funding

# Back-ups



# Generic Evolutionary Cycle

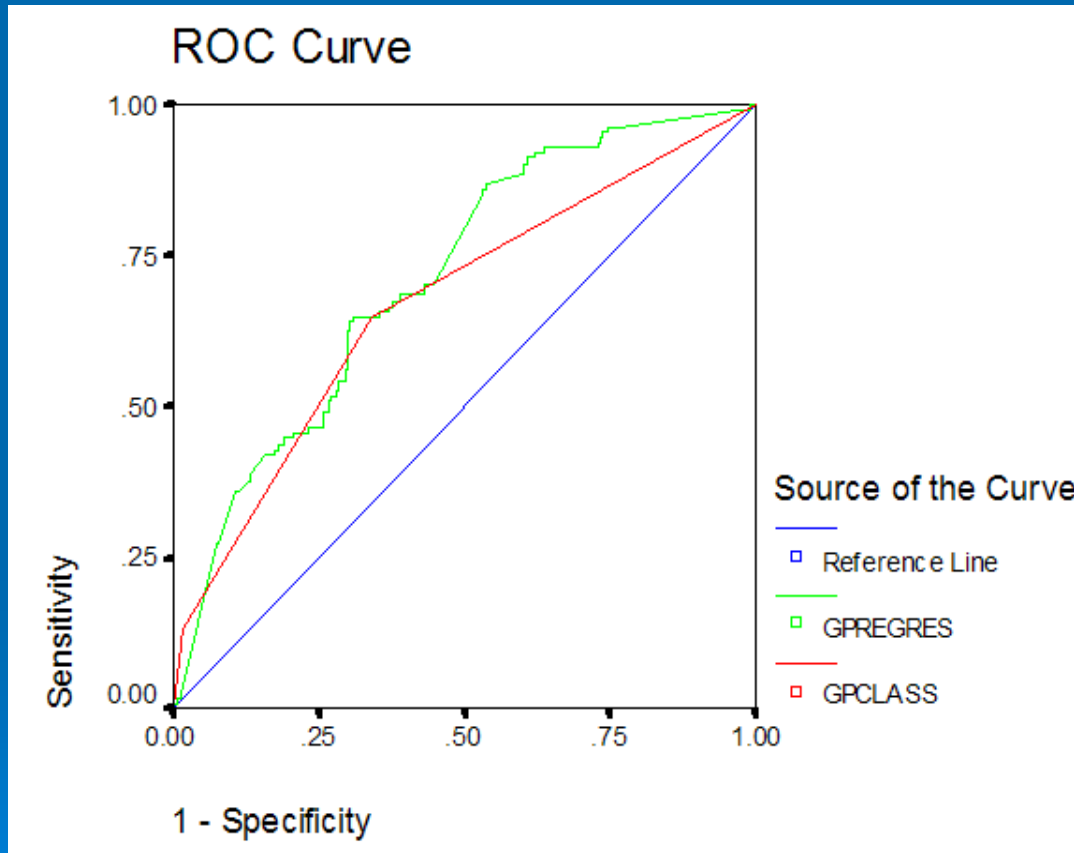
- 0 Initial random populations of trial solutions
  - 1 Test individuals
  - 2 Good enough? **Stop**
  - 3 Select better
  - 4 Breed new generation with crossover and mutation
  - 5 Goto 1
- 
- The background of the slide is a solid blue color. In the lower right quadrant, there are several faint, concentric circles that resemble ripples in water, creating a decorative pattern.

# P450 HTS Data

- High Throughput Screening
  - volume data
  - single concentration (no curve info.), very noisy
- To counter noise effects
  - took subset of molecules screened in triplicate on 2 separate HTS runs
  - then selected molecules having consistent measurements (within 15% variation, across 6 readings)
    - averaged the values
    - binary thresholded to determine p450 inhibition class
- To help counter 4:1 class imbalance
  - separated classes
  - hierarchical clustering
    - Wards' linkage, Tanimoto similarity Daylight fingerprints
    - Inhibitor groups selected at 80%, Non-inhibitors at 75%

# Generalisation (wrt f-GW)

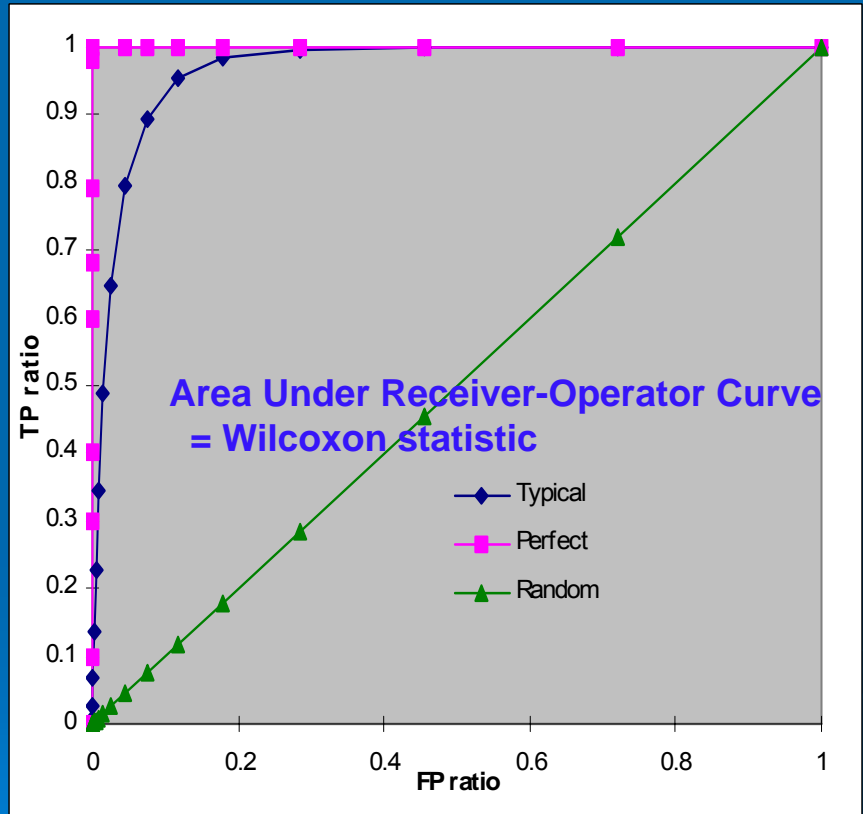
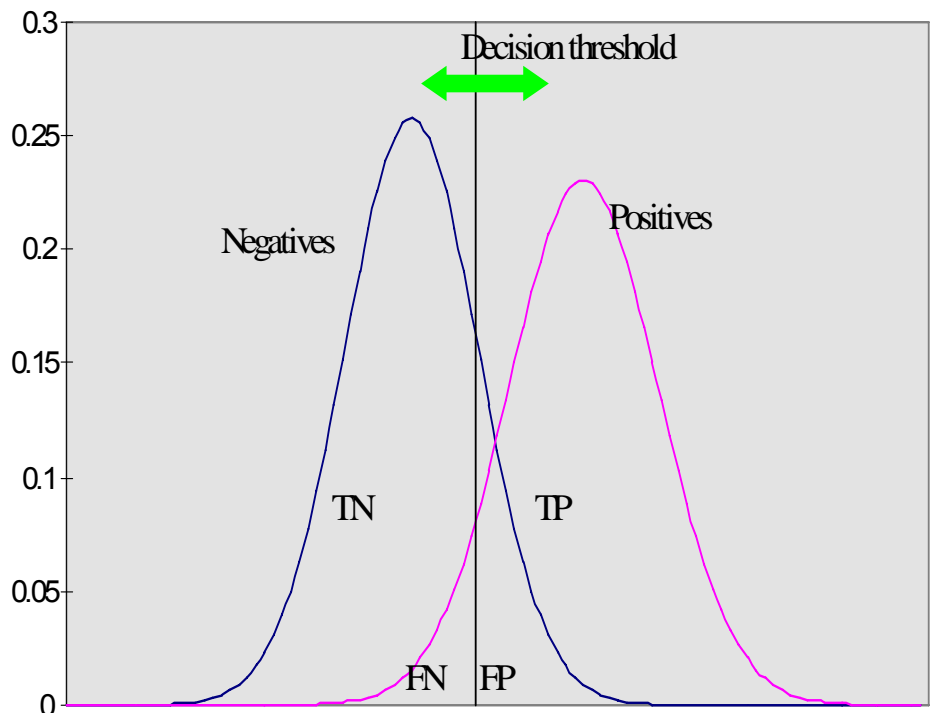
## Binary Performance - 2D6 inhibitor rejection



# Binary Classifier Cut-points?

Two classes with overlapping distributions  
- where to place decision threshold?

...trade-off errors :  
false positives vs false negatives



...can we simultaneously optimise the use of multiple classifiers together ?