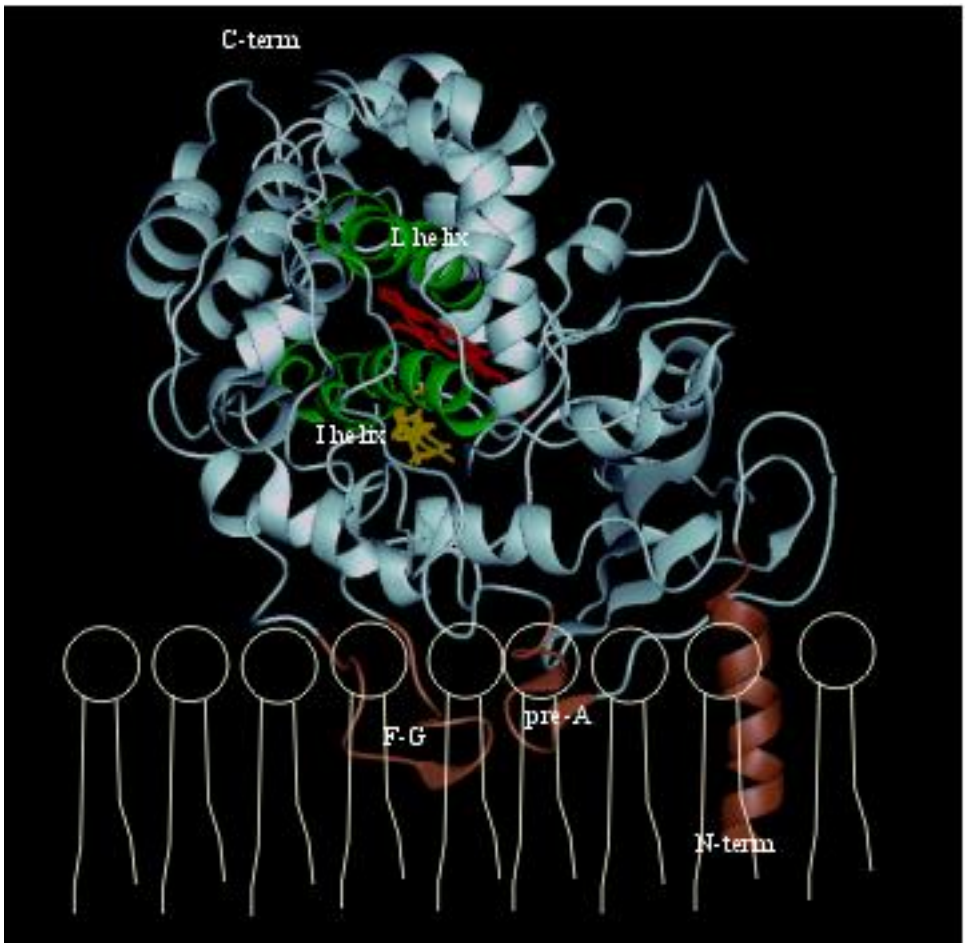
Combining Machine Learning techniques to Predict Compounds' Cytochrome P450 High Throughput Screening Inhibition

W. B. Langdon, B. F. Buxton and S. J. Barrett

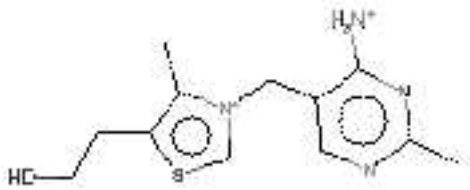
University College, London
W.Langdon@cs.ucl.ac.uk

GlaxoSmithKline
Steven.J.Barrett@gsk.com



Model of P450 showing membrane attachment

Compute Electro/Chemical/Structural Features

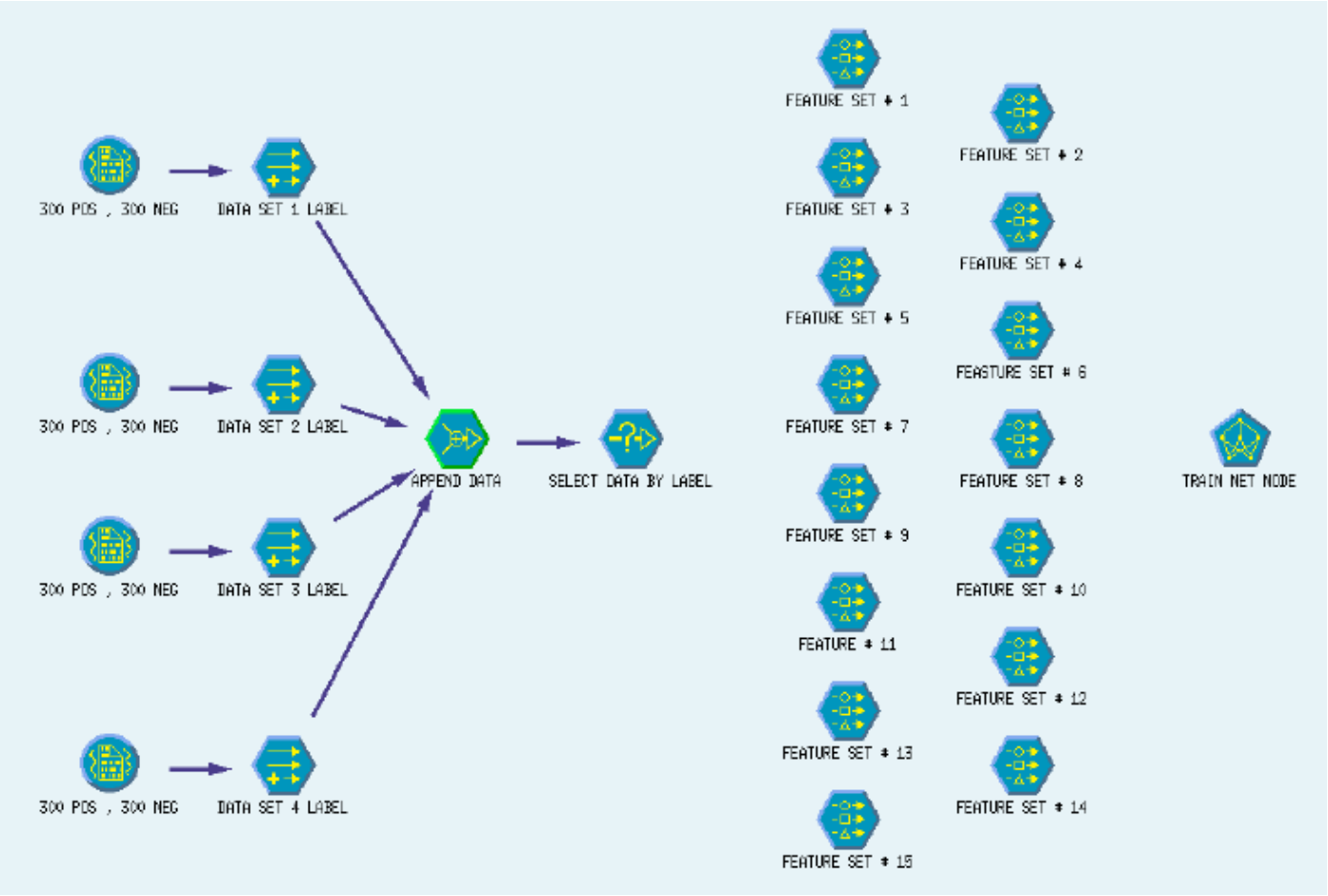


Example Compound (vitamin B1)

The primary chemical structure is used to calculate 699 features for use in predicting activity of compounds.

Individually none of these features are able to predict activity with P450.

Combined Classifier via 2 Stages: Clementine. Genetic Programming or Boosting



Data mining tool Clementine used to train Neural Networks

699 features split into 15 related groups (vertical)

Inactive records split into 5 groups (horizontal)

5 "balanced" training sets produced by reusing inhibiting compounds with each inactive set.

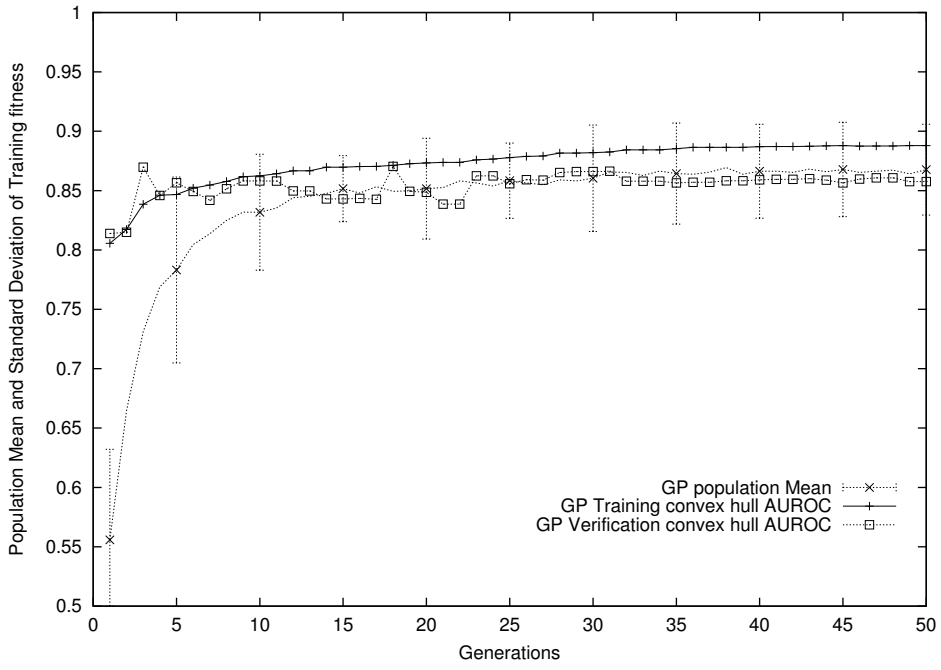
Total of $5 \times 15 = 75$ neural networks produced by Clementine.

Composite of 75 ANN: Genetic Programming

Data randomly split into training (866 compounds) and verification (433).

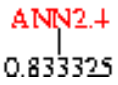
Clementine Neural Networks used a functions within genetic programming trees.

Performance of individual trees within GP population given by the area under the convex hull of their Receiver Operating Characteristics (Wilcox' statistic).

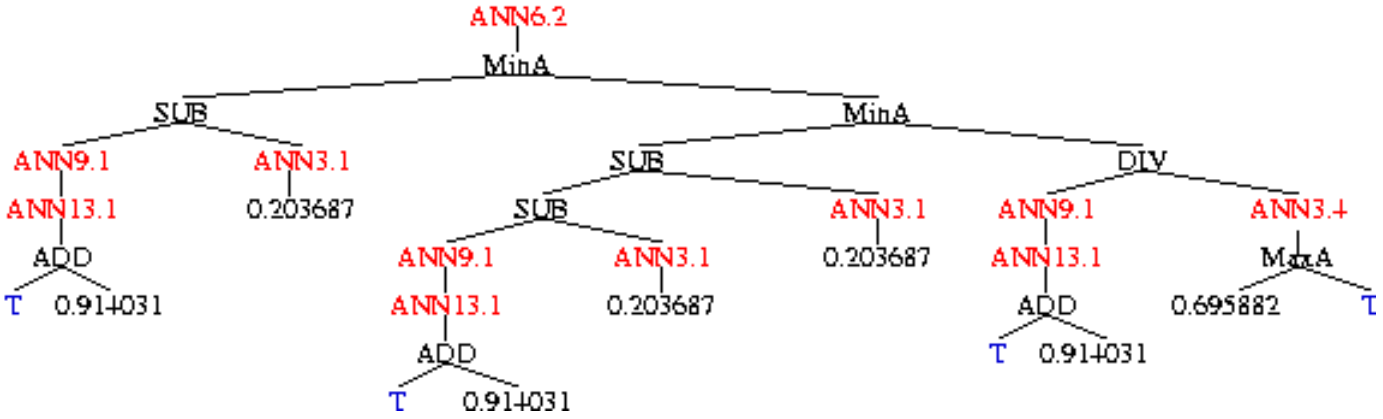


Genetic Programming Evolution of Area Under ROC

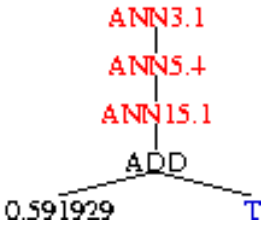
Example Genetic Programming P450 Classifier



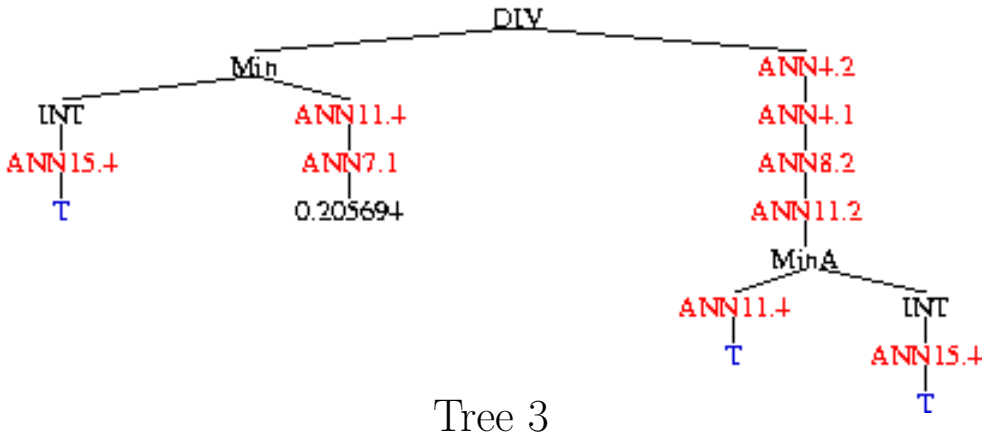
Tree 0



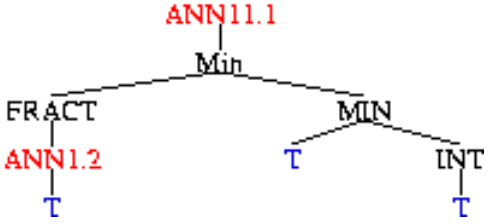
Tree 1



Tree 2



Tree 3



Tree 4

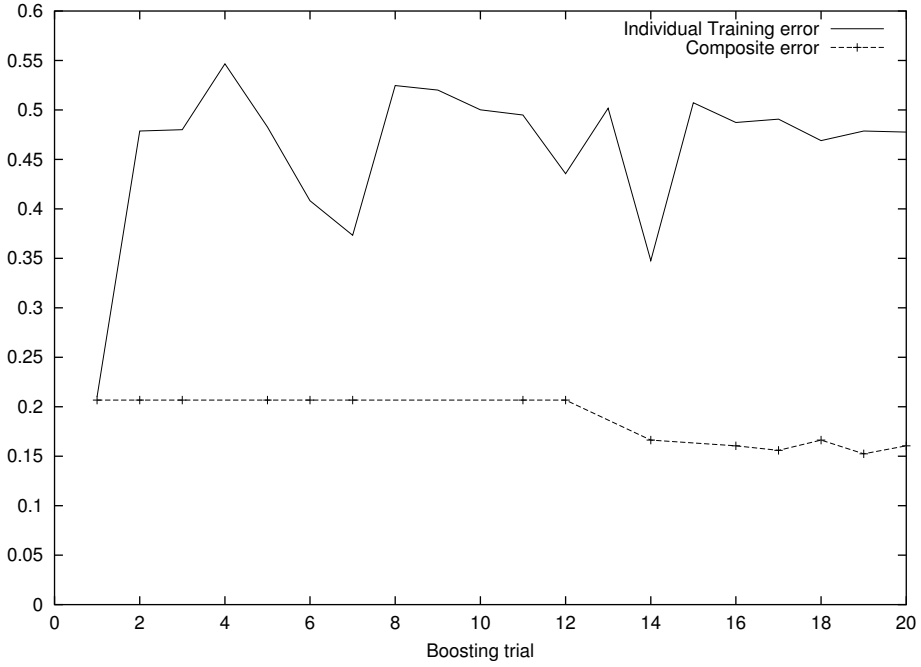
Composite of 75 Networks: Boosting

GP and Matlab boosting code use identical data.

Unlike GP, combination rule is Matlab feed forward neural net with over fitting stopping rule.

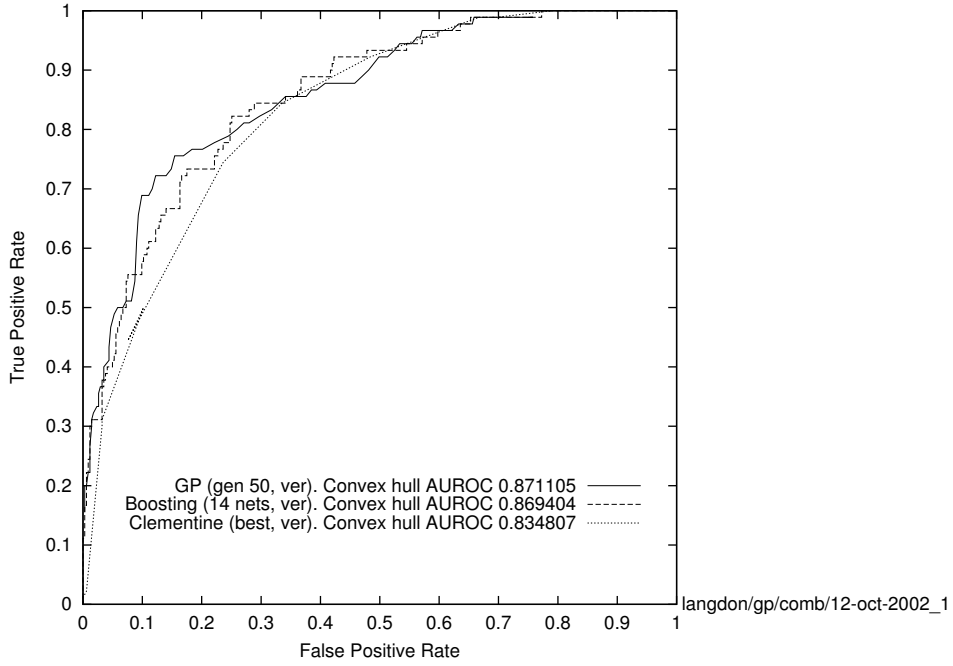
Experiments (without boosting) indicated little performance difference between 2 and 20 hidden units, so smallest neural network was used (75 inputs, 2 hidden units, 2 output neurons).

AdaBoost.M2 using training error and resampling from re-weighted training set.



Error on Boosting run

Comparison GP and AdaBoost.M2



Receiver Operating Characteristics (ROC) of Combined Features

Conclusions

70% of inhibiting compounds (HTS) can be predicted (at the expense of misclassifying 12% of inactives) using Genetic Programming composite classifier based on readily computed features.

For a boosted combination, at 70% true positive rate, 16% of inactives are incorrectly predicted.

The best Clementine network, at 70% true positive rate, wrongly suggests 21% of inactives inhibit P450.

None of the features give adequate performance if used singularly.