Nathan Brown

nathan.brown@novartis.com

The Application of Consensus Modelling and Genetic Algorithms to Interpretable Discriminant Analysis

Workshop 'Chemoinformatics in Europe: Research and Teaching'

30th May 2006

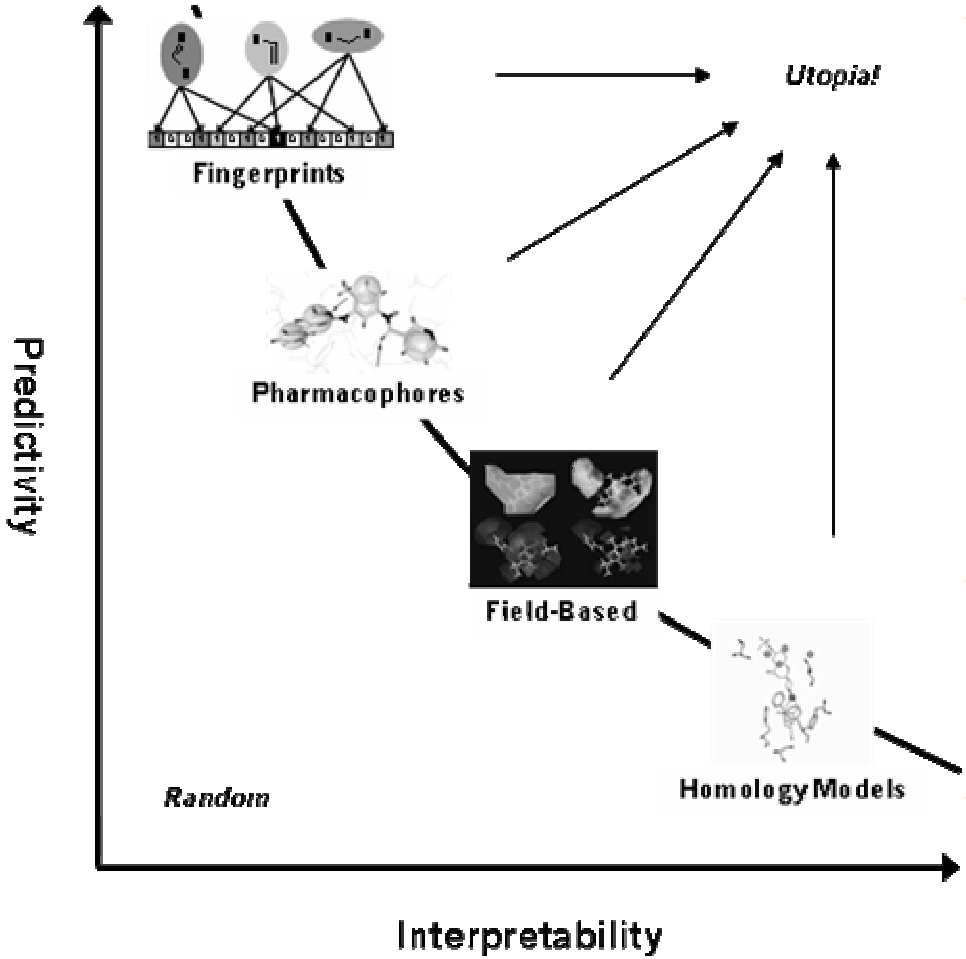
NOVARTIS

Discriminant Analysis Using a GA

- Predictive *vs.* Diagnostic Modelling
- Discriminant Analysis with a Genetic Algorithm
- Consensus and Splice Modelling
- Experimental Studies
  - MDDR: 1130 renin and 636 COX inhibitors[1]
  - Oral drugs: 1082 FDA-approved drugs[2]

1. Hert, J.; Willet, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A.; Comparison of Fingerprint-Based Methods for Virtual Screening Using Multiple Bioactive Reference Structures. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1177-1185.
2. Vieth, M.; Siegel, M. G.; Higgs, R. E.; Watson, I. A.; Robertson, D. H.; Savin, K. A.; Durst, G. L.; Hipskind, P. A. Characteristic Physical Properties and Structural Fragments of Marketed Oral Drugs. *J. Med. Chem.* **2004**, *47*, 224-232.

Predictive *versus* Diagnostic Models



- Highly predictive models tend to obfuscate what is important for the property being modelled
- Highly interpretable models tend to be less effective in prediction power
- However, both objectives are very important
- We want highly predictive models that can also guide our decision-making processes

* Adapted from a diagram by Richard Lewis

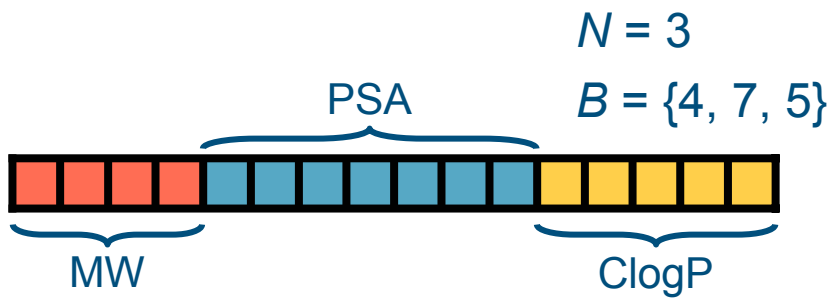
Discriminant Analysis

- Supervised learning
  - Dependent variable is known for dataset and used in training the model with the independent variables
- Optimize separation of classes
  - Evolve weights for binned descriptors
  - Score solutions according to ability to separate objects
- Discover descriptor ranges that are important for discrimination which can then be applied to make informed decisions

1. Gillet, V. J.; Willett, P.; Bradshaw, J. Identification of Biological Activity Profiles Using Substructural Analysis and Genetic Algorithms. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 165 – 179.

Chromosome Encoding

- Selection of N descriptors
- Each descriptor partitioned into B_i bins
- Each bin can take any value in the range $\{0 \ldots W\}$
- Chromosome length is then $(N \cdot \sum B_i)$



NOVARTIS
NOVARTIS INSTITUTES FOR BIOMEDICAL RESEARCH

Descriptor Selection

- Calculate physicochemical descriptors
- Cluster descriptors (not objects)
- Select descriptors that are:
  - more orthogonal, and
  - more interpretable for the medicinal chemist
- Some dataset dependency

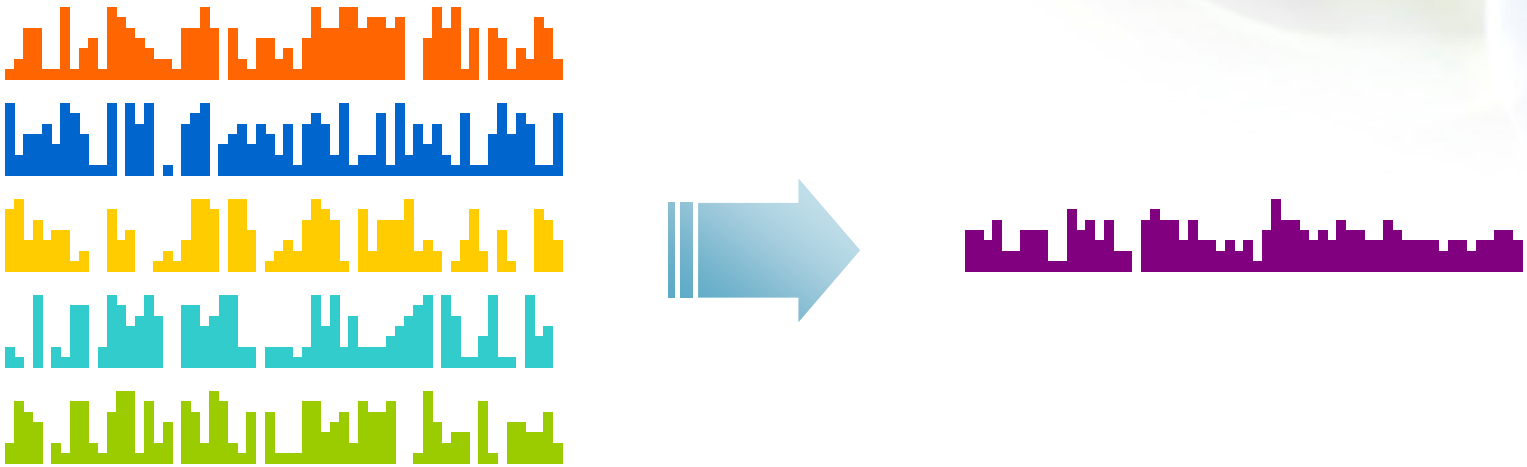
Fitness Functions 1

- Initial Enhancement (IE)
  - Emphasises enrichment in top *NACT*% of recalled molecules
  - *i.e.* mean rank of all actives recalled after *NACT*
- Global Enhancement (GE)
  - Emphasises enrichment of all actives in recalled molecules
  - *i.e.* mean rank of all actives
- Maximum Difference Enhancement (MDE)
  - Emphasises maximum difference in scores between the two classes

Fitness Functions 2

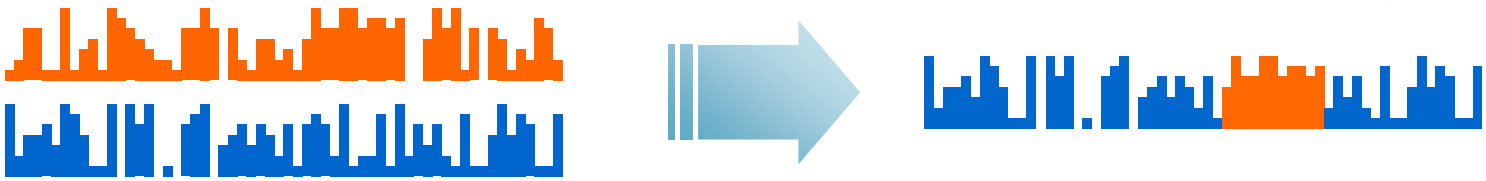
- Existing fitness function used a combination of evaluations:
  - Number of actives in the top *N*%
  - Average rank of actives over entire rank
- Maximised Difference Enhancement (MDE)
  - The difference of the average rank of the two classes being discriminated
- MDE will tend to result in molecules where the separation between the two classes is maximised globally and rewarding ranks with more interesting molecules in the initial part of the rank

Consensus Models



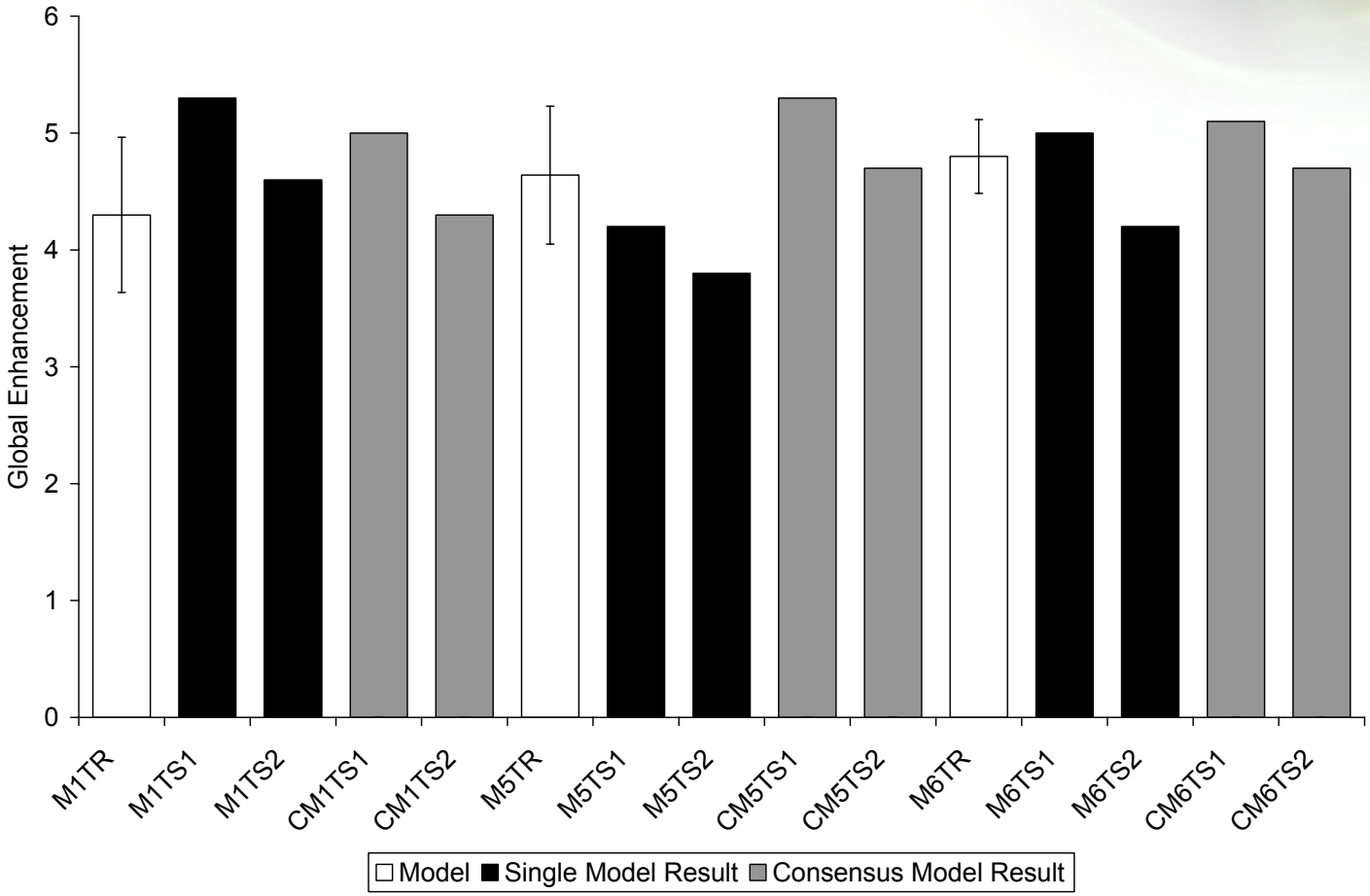
- Aim to reduce stochastic effects of using a single chromosome

Splice Models



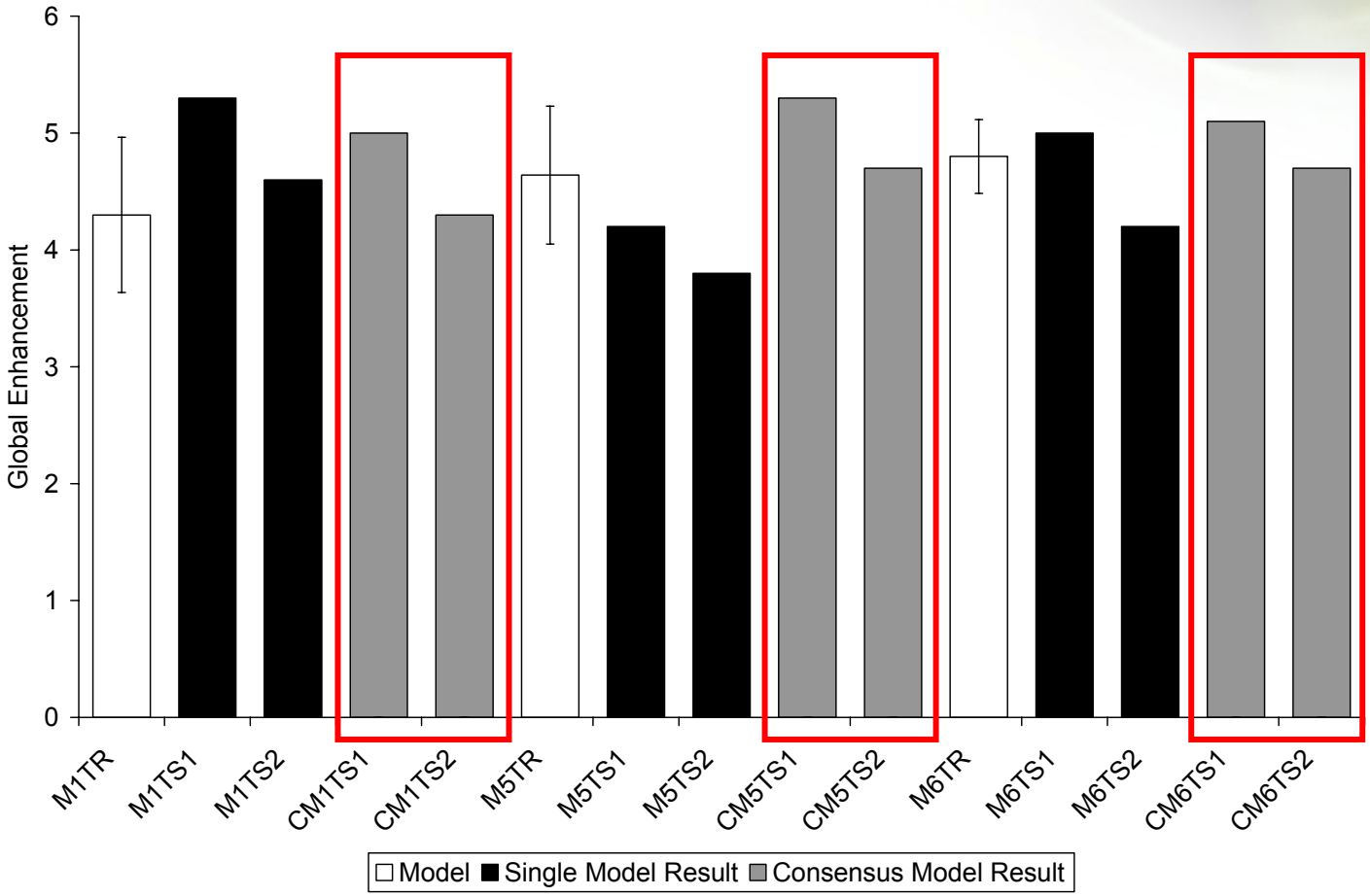
- Essentially a manual recombination operator to effect a more optimal solution model based on feedback and intuition

Renin Consensus Discrimination Model

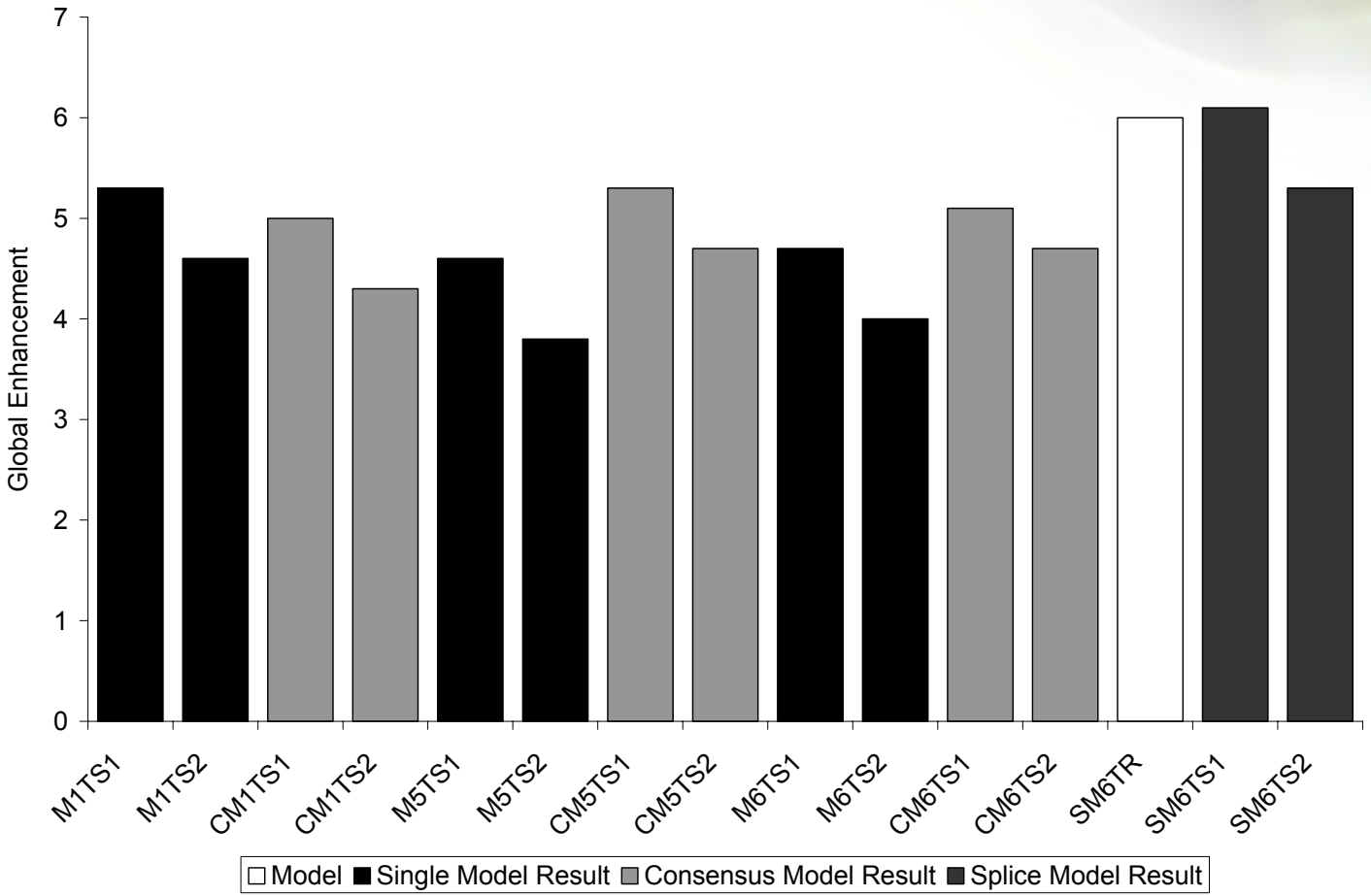


NOVARTIS
NOVARTIS INSTITUTES
FOR BIOMEDICAL RESEARCH

Renin Consensus Discrimination Model

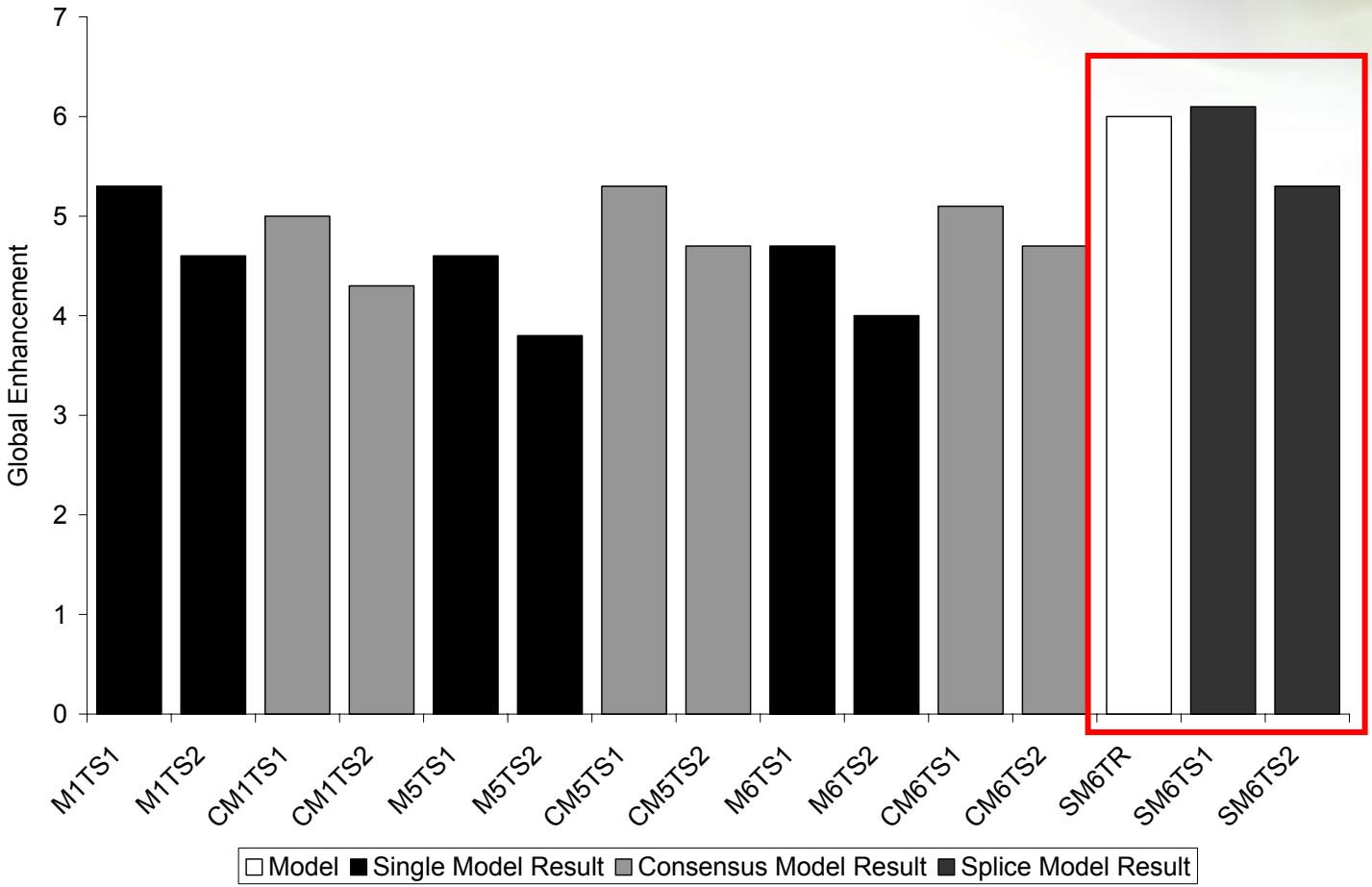


Renin Splice Discrimination Model



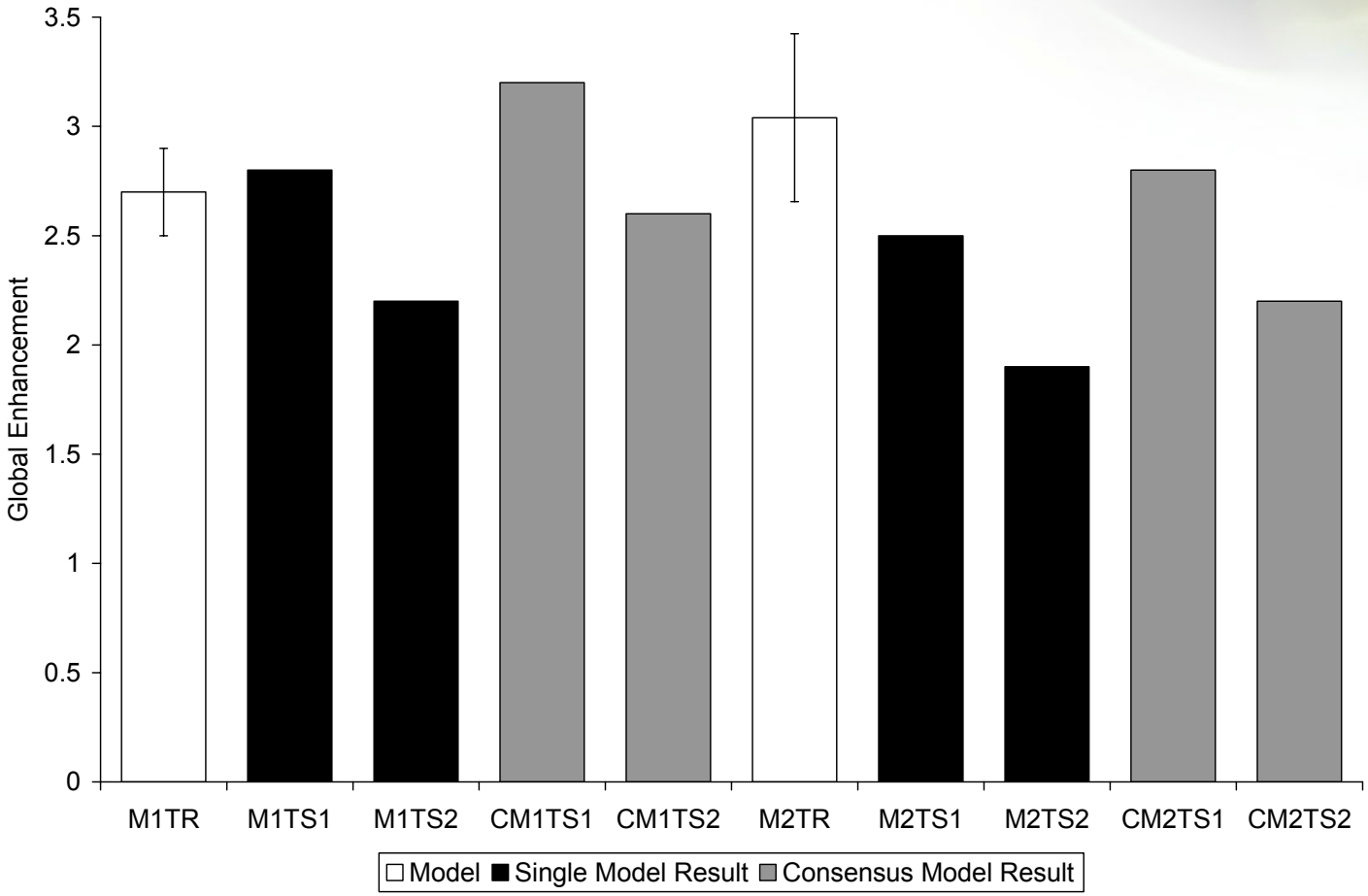
NOVARTIS
NOVARTIS INSTITUTES
FOR BIOMEDICAL RESEARCH

Renin Splice Discrimination Model

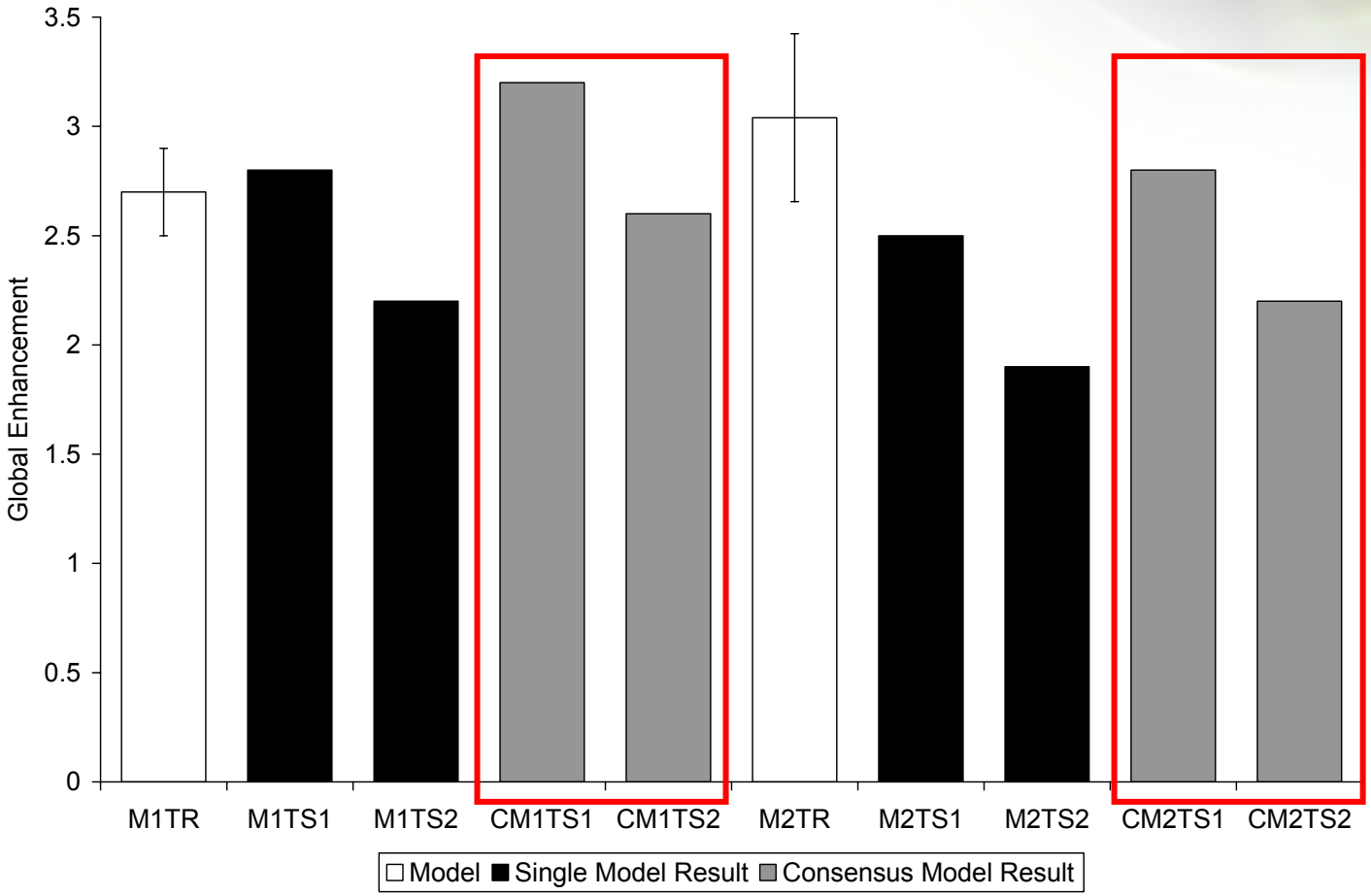


NOVARTIS
NOVARTIS INSTITUTES
FOR BIOMEDICAL RESEARCH

COX Splice Discrimination Model



COX Splice Discrimination Model

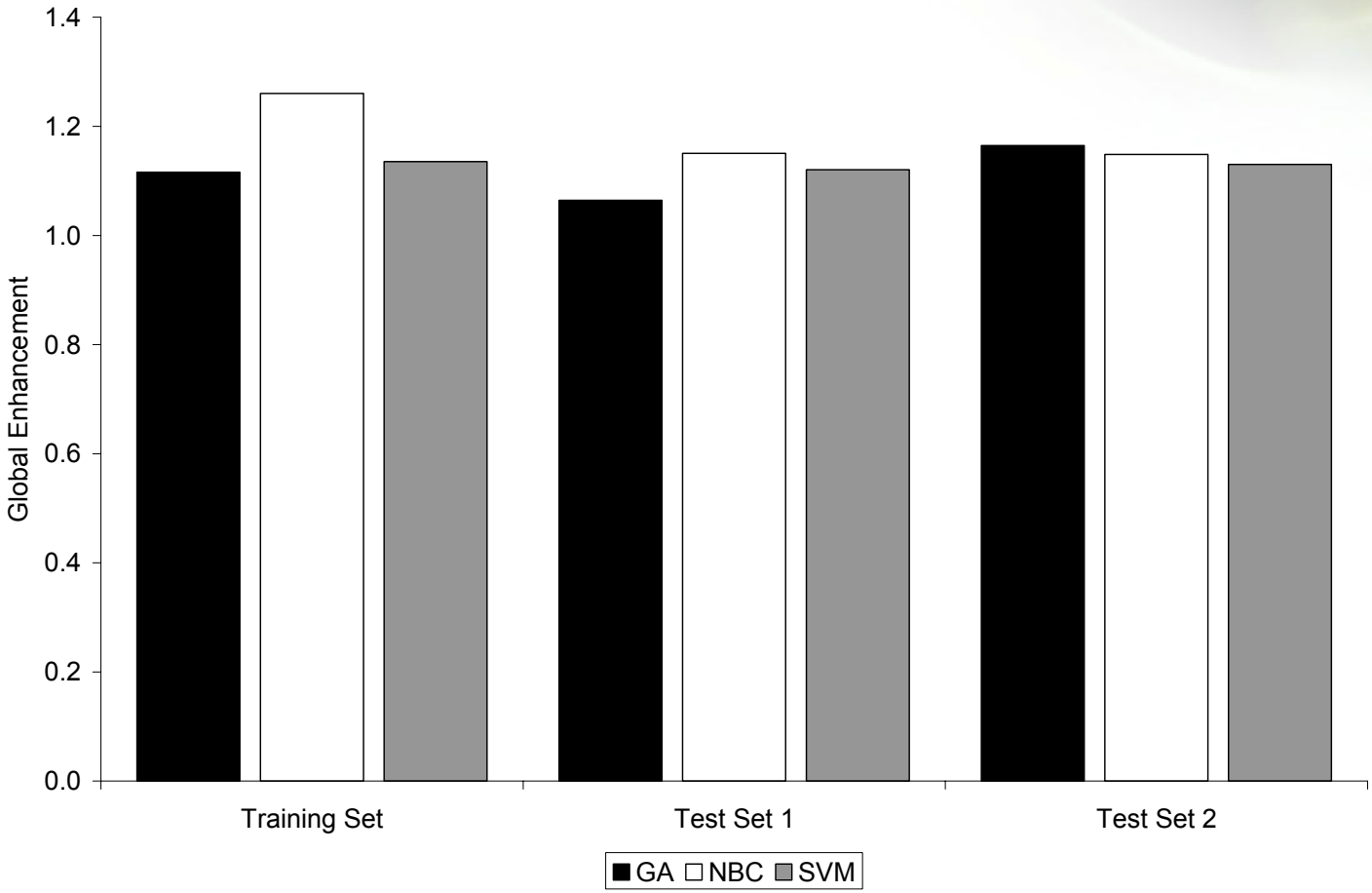


Comparative Study: Oral *vs.* Non-Oral Drugs

- Oral *vs.* non-oral drugs dataset[1]
- GA model compared with models generated with
  - Naïve Bayes Classifier (NBC)
  - Support Vector Machines (SVM)
- Investigating:
  - Consistency of results
  - Interpretation of models

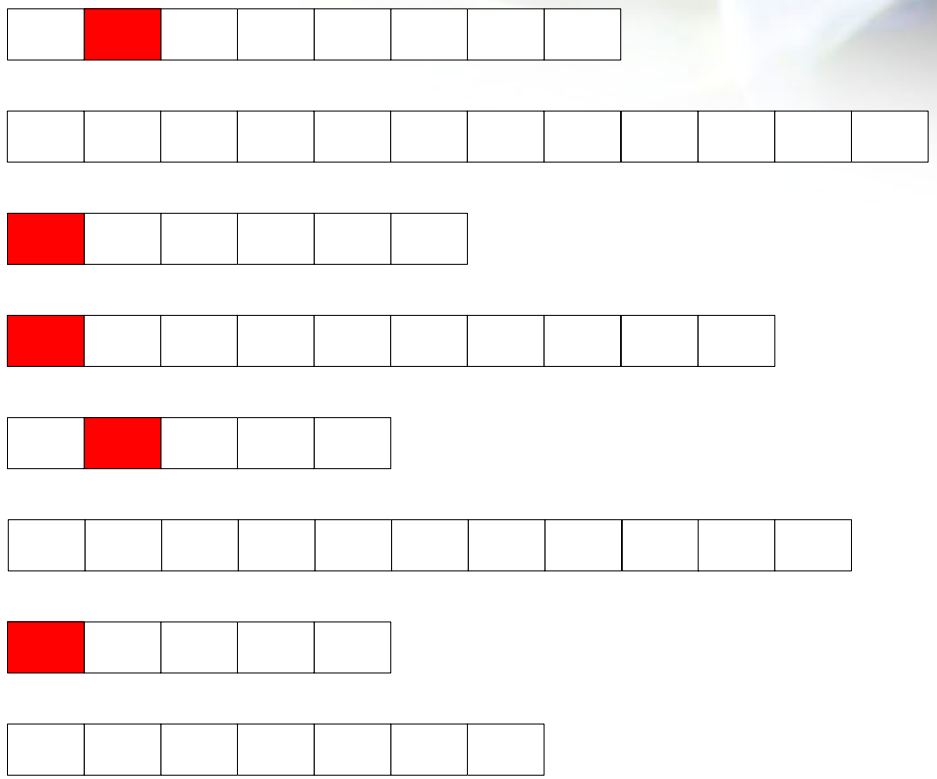
1. Vieth, M.; Siegel, M. G.; Higgs, R. E.; Watson, I. A.; Robertson, D. H.; Savin, K. A.; Durst, G. L.; Hipskind, P. A. Characteristic Physical Properties and Structural Fragments of Marketed Oral Drugs. *J. Med. Chem.* **2004**, *47*, 224-232.

Oral Drug Discrimination Model



Model Interpretability

- Model weights indicate
  - Important descriptors
  - Important ranges
- Used to guide decision-making processes
  - Similarity searching
  - Filtering rules
- Rules are focused on domain of interest



Conclusions

- Consensus and splice models provide consistently improved results
- GA models provide greater or similar interpretability than other methods applied here
  - Models are transparent as to which descriptors and their ranges are of greatest importance in discriminating
- Indications that the GA and NBC methods could be applied in combination
  - Investigation of complementarity

1. Ganguly, M.; Brown, N.; Schuffenhauer, A.; Ertl, P.; Gillet, V. J.; Greenidge, P. A. Introducing the Consensus Modeling Concept in Genetic Algorithms: Application to Interpretable Discrimination Analysis. Submitted to *J. Chem. Inf. Mod.*

Areas the Student Covered

- Cluster analysis
- Druglikeness
- Discriminant analysis
- Variable selection
- Genetic algorithms
- Statistical learning methods
- Java programming
- Method development

What does the student gain?

- Coding and adapting software
- Tackling everyday challenges of research
- Performing research in industry
- Application-context drug research
- Empowered to pursue their own research

What do the mentors gain?

- Freedom to pursue an avenue of interest
- Developing skills in student mentoring
- A new viewpoint with new ideas

- *Assisting in training the next generation of scientists*

Acknowledgements

- University of Sheffield
  - ***Milan Ganguly***
  - ***Val Gillet***
  - Peter Willett
- UCSF
  - Jérôme Hert
- Cheminformatics
  - ***Peter Ertl***
  - Stephen Jelfs
- Computer-Aided Drug Discovery
  - ***Paulette Greenidge***
  - Richard Lewis
  - Nikolaus Stiefl
- Molecular & Library Informatics
  - Kamal Azzaoui
  - Edgar Jacoby
  - ***Ansgar Schuffenhauer***

NOVARTIS
NOVARTIS INSTITUTES FOR BIOMEDICAL RESEARCH