

# Using AutoQSAR to select the most predictive modeling methods

**9<sup>th</sup> ICCS**

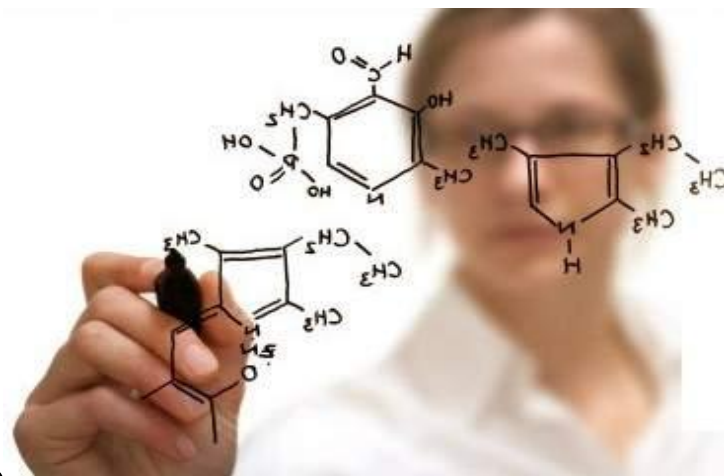
**6<sup>th</sup> June 2011**

**Sarah Rodgers (Aaron)  
Frank Brown  
David Wood  
Andy Davis**

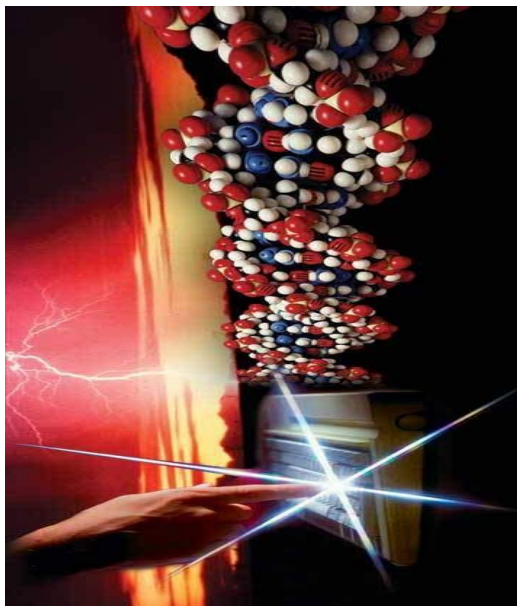
**AstraZeneca** 

 **accelrys**<sup>®</sup>

- A leading enterprise scientific informatics company
- Offering in
  - Modelling & Simulation
  - Lab execution analysis
  - Electronic Notebook
  - Data management & decision support
- Deep expertise in
  - Chemistry
  - Biology
  - Materials Science



# About Contract Research Services



Delivering Solutions for  
Real-World Scientific  
Problems

- 15+ years of successful contract research services
- Provides a framework to approach industrially relevant problem
- Uncompromising commitment to the quality of research
- Apply the correct resources - Wide range of research projects accomplished with Accelrys' full range of software products
- Experts ask the right questions - Scope of work, deliverables, timelines, cost defined and agreed in advance
- On time delivery of results
- Customer retains IP

# Global Brand Customer Portfolio Excellence Delivered to Leading Clients



Energy  
& Petroleum

Chemicals

Aerospace  
& Automotive

FMCG

Pharmaceuticals



MERCK



NOVARTIS



Genentech  
IN BUSINESS FOR LIFE



sanofi aventis

AstraZeneca



3M



Johnson & Johnson



BAE SYSTEMS

gsk  
GlaxoSmithKline



Bayer



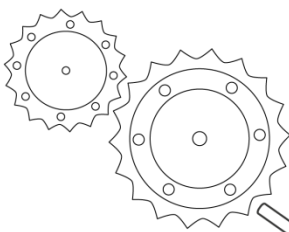
MOTOROLA

Lilly

AstraZeneca

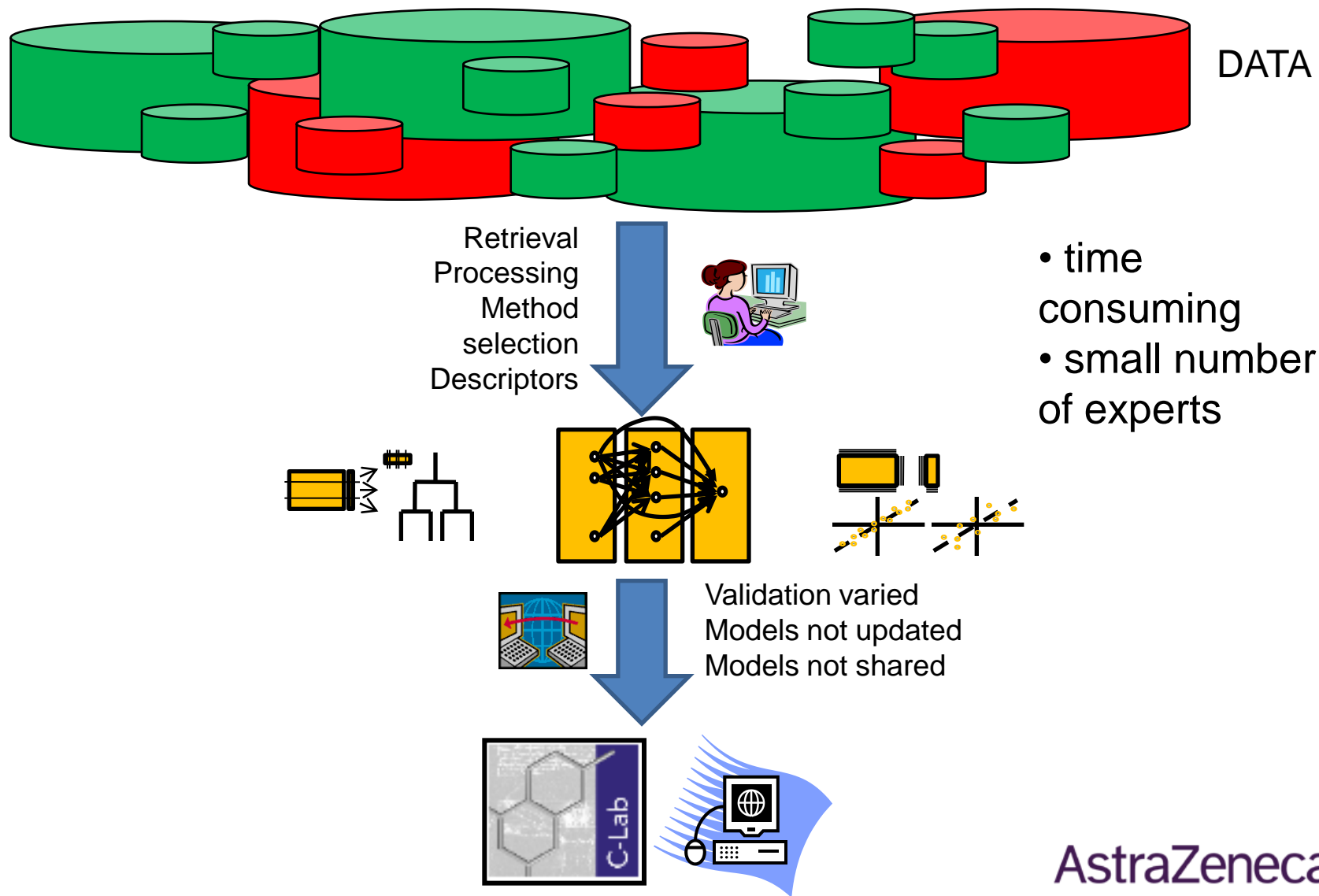
# What is AutoQSAR?

- A proprietary system developed at AstraZeneca in collaboration with Accelrys for the automatic creation, evaluation and maintenance of QSAR models

Aut  SAR

- AutoQSAR System
  - Input/output
  - Statistical methods
  - Competitive workflow
- Global Models
  - Statistical methods comparisons
- Local Models
  - Statistical methods comparisons
  - Project/series level
- Summary

# QSAR Modeling Previously



## Science

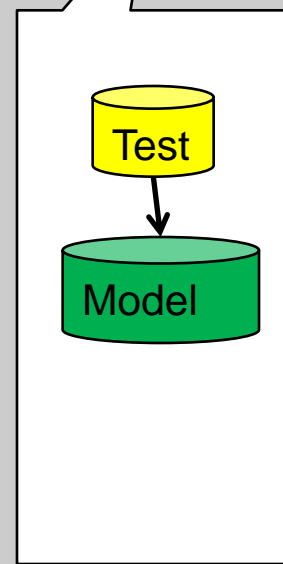
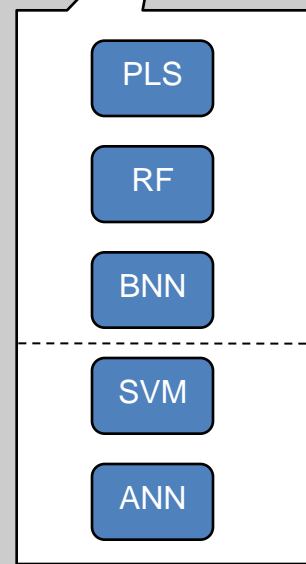
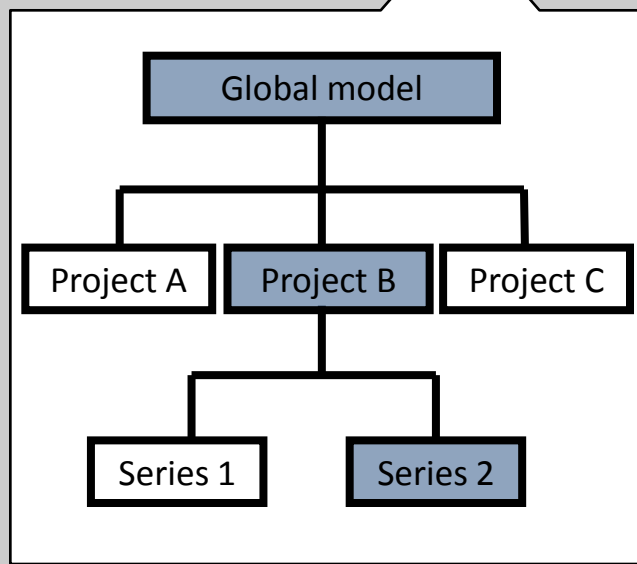
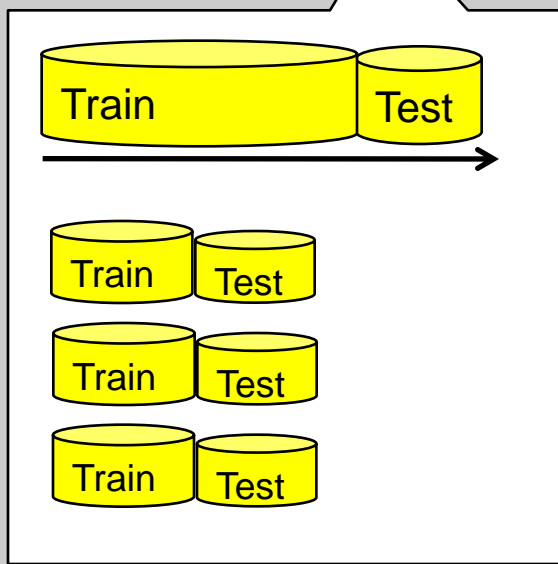
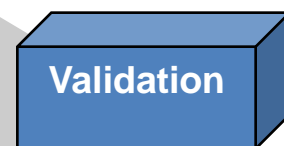
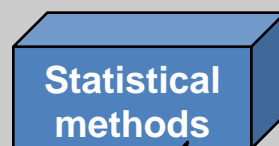
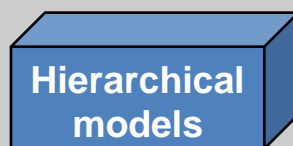
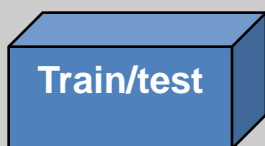
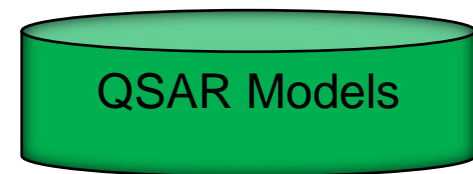
- Keep models up to date
- Correct biases (project models correct global model biases)
- Automate model selection – best predictions
- Development environment for testing new methods

## Business

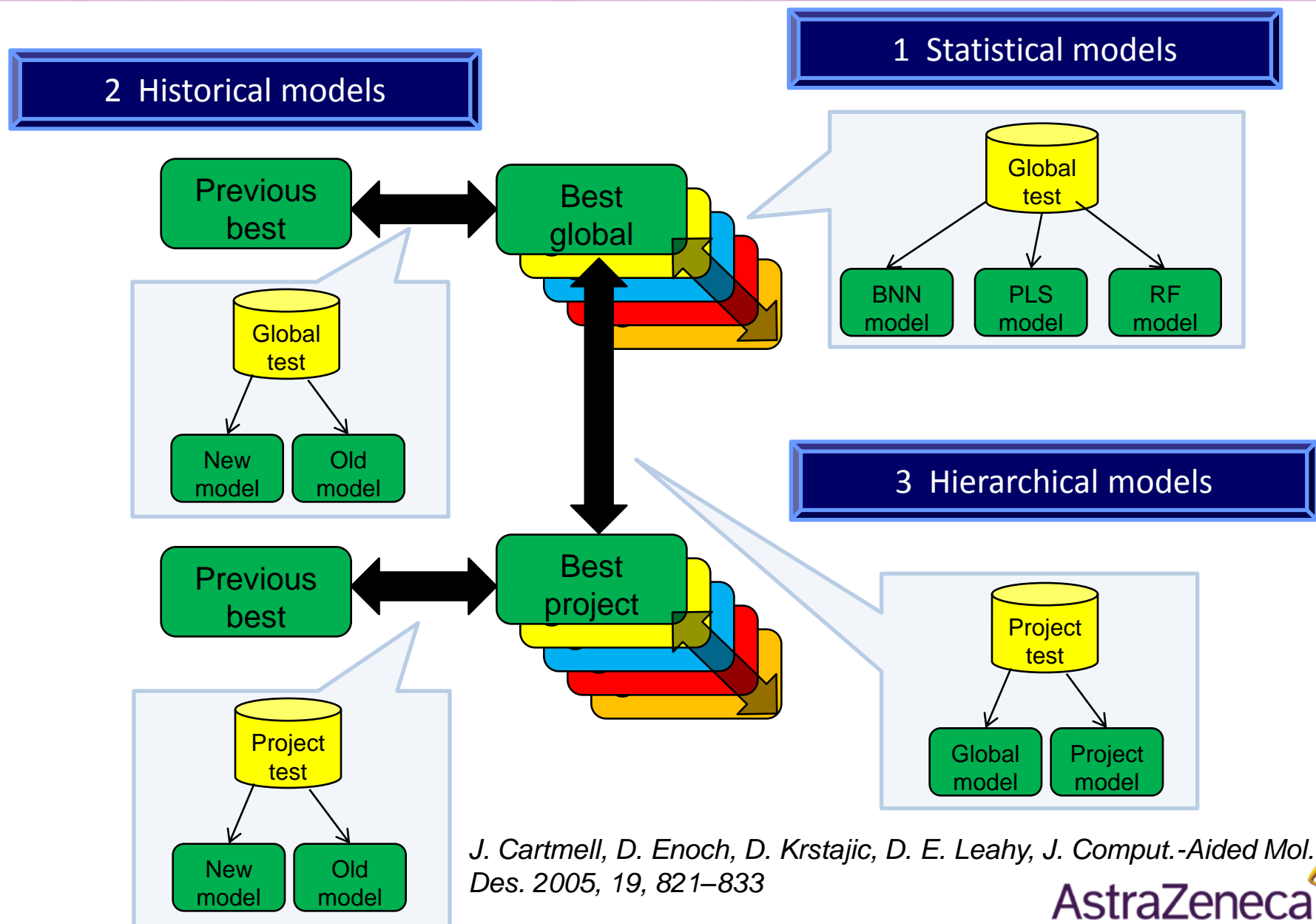
- Apply good QSAR practise
- Exploit measurement database
- Free-up computational chemists' time
- Track performance



Aut  SAR

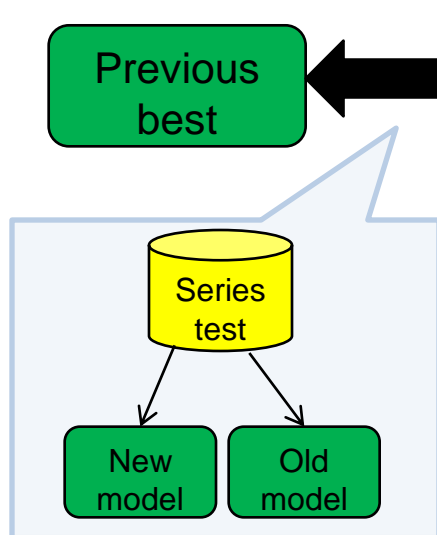
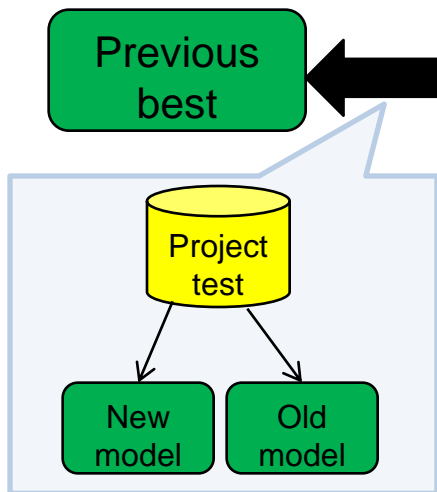


# Competitive Workflow: Global

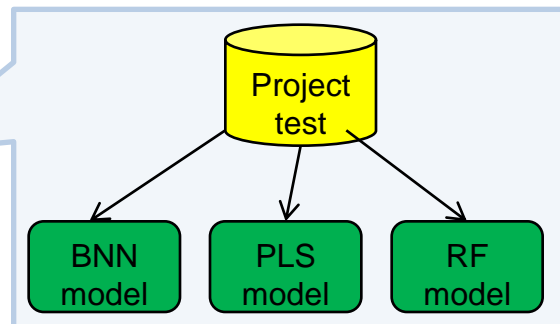


# Competitive Workflow: Project

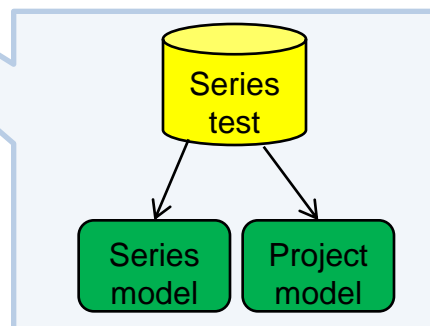
## 2 Historical models



## 1 Statistical models



## 3 Hierarchical models



*J. Cartmell, D. Enoch, D. Krstajic, D. E. Leahy, J. Comput.-Aided Mol. Des. 2005, 19, 821-833*

## Navigation

- Protocols
  - 1 Register IBIS Search
  - 2 Build Models
    - 1a Build a Global or Local Model for a Global Search
    - 1b Build All Local Models Associated with a Global Search
    - 2a Build a Single Project or Series Model for a Project
    - 2b Build All Models Associated With Project
  - 3 Prediction
  - 4 Reporting
  - 5 Administration
    - Build Monitor
    - Delete Model
    - Job Monitor
    - Model Mount
    - Model Viewer
  - 6 Scheduler tasks
    - Create, Modify or Delete Scheduled Job

Model Viewer [1] ✕

Build Monitor [1] ✕

## Job List

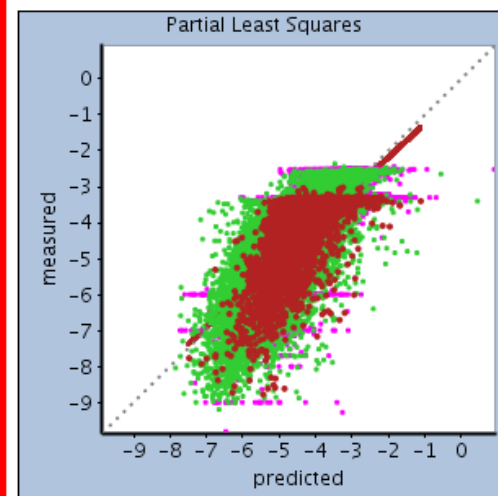
Build Id	Model Name	Stats Method	Status	Start Date	Finish Date
11573	AlderleyPark_CIRA_PL...	Random Forest	Complete	3/3/11 5:56 PM	3/3/11 6:53 PM
11572	AlderleyPark_CIRA_PL...	Random Forest	Complete	3/3/11 5:55 PM	3/3/11 6:52 PM

## Model Build Description Summary

Model Name: Sol74\_SA\_DriedDMSO  
 Model Id: 2159  
 Build Method: Partial Least Squares  
 IBIS query id: 2aa8a4c3-855b-47a9-9d51-df85d5244f74  
 Project:  
 Model Description: Sarah's version of the Dried DMSO Solubility model

[Shortcut to QSAR Prediction](#)

## Graph plotting predicted value against actual values

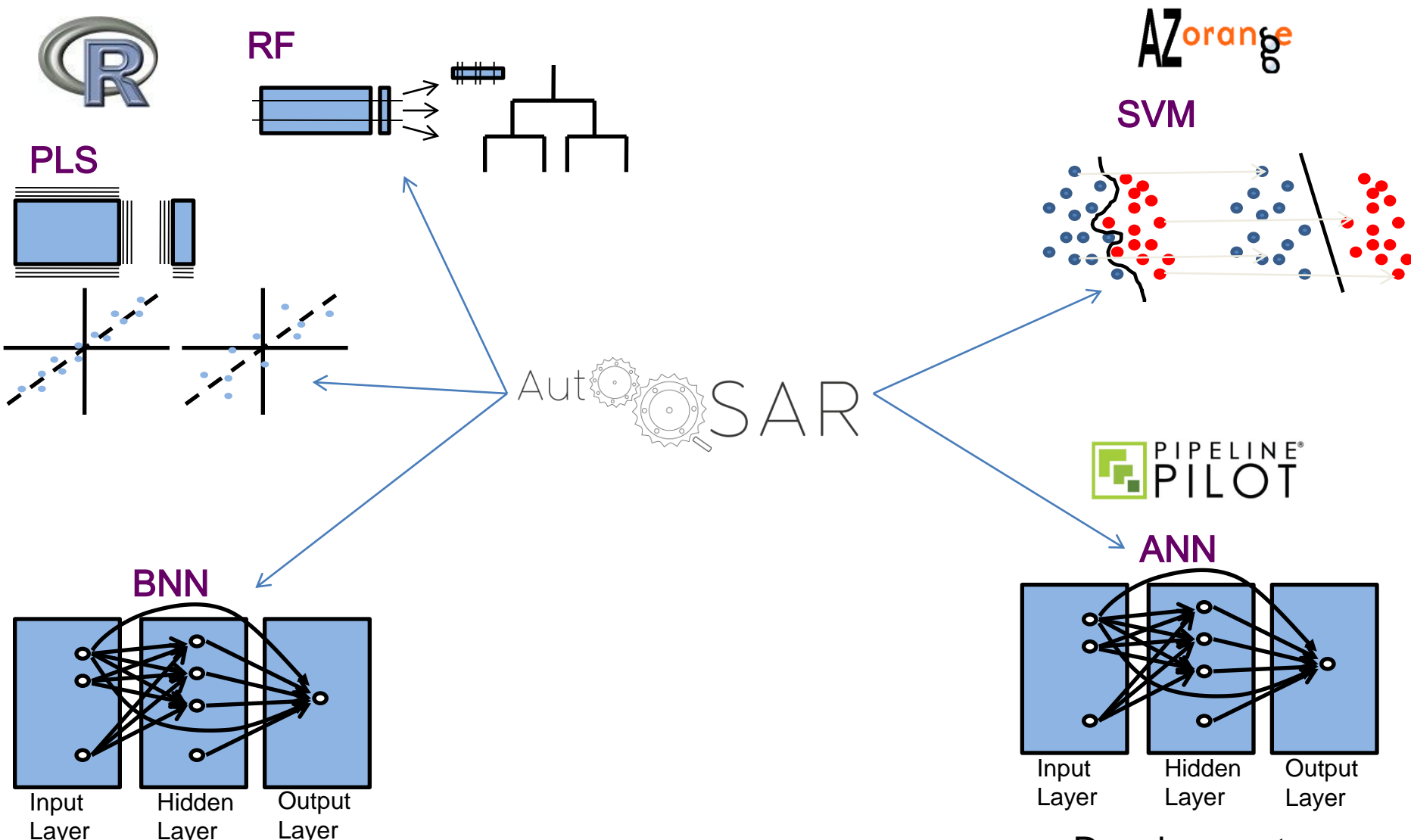


Creation date = Tue May 3 11:23:44 2011

Build Id for model = 12614

11550	AlderleyPark_CIRA_PL...	Random Forest	Complete	3/3/11 2:22 PM	3/3/11 2:22 PM
11549	AlderleyPark_CIRA_PL...	Random Forest	Complete	3/3/11 2:21 PM	3/3/11 2:21 PM

# Integration of Statistical Methods



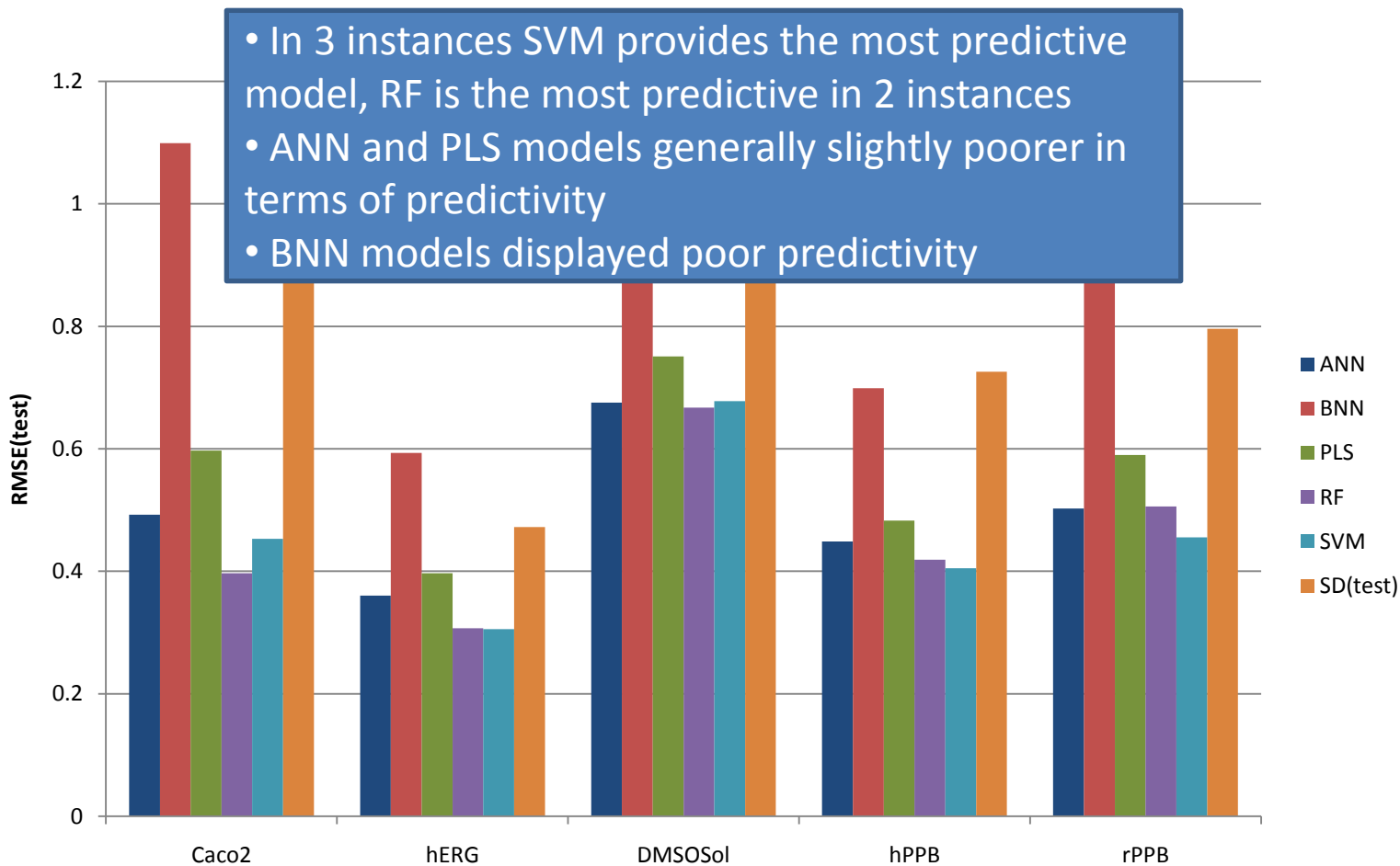
Production

Development

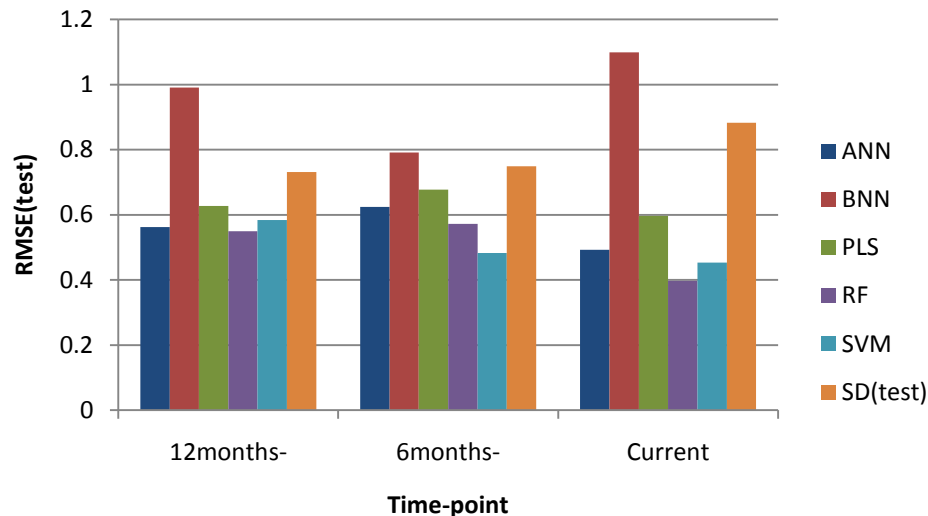
- Large, cross-site models (can contain 10,000s+ compounds) including multiple projects
- Useful in the absence of a predictive project-level model, or for compounds that do not have a project association
- Considered 5 global models:
  - Human plasma protein binding
  - Rat plasma protein binding
  - Human ether-a-go-go (hERG)
  - Caco2
  - Dried DMSO Solubility
- Temporal test set
- 193 in-house bulk property descriptors



For each endpoint, the temporal test set is used to compare the different statistical methods

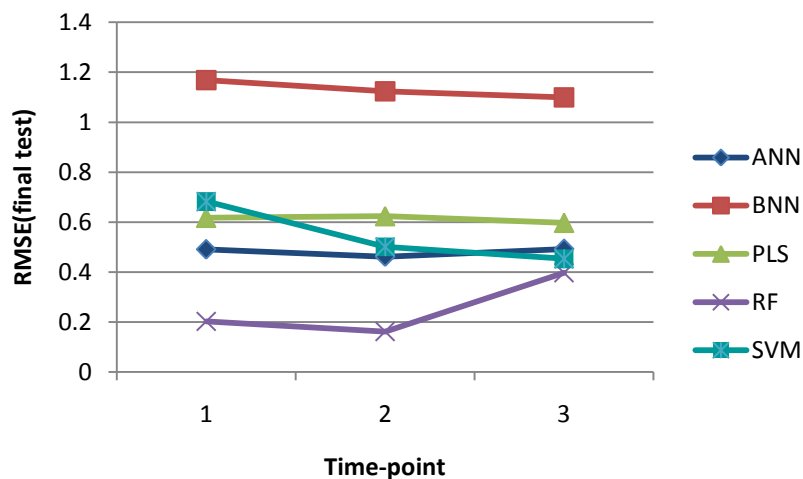


## Comparison of statistical methods at time of model build



- RF provides most predictive model at beginning and end of time-series
- SVM most predictive at 6 months
- BNN models poor at each time-point

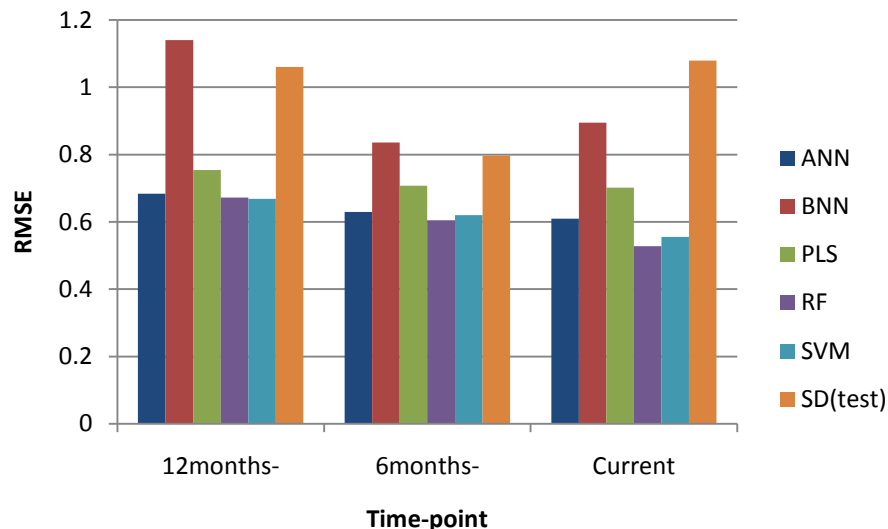
## Retrospective analysis of model performance



- PLS and ANN models provide stable predictivity across the time-series
- BNN and SVM models improve over time (retrospectively)
- Final RF model less predictive than previous models (Competitive workflow compares all new models with previous 'best', RF model at 6 months would be missed here)

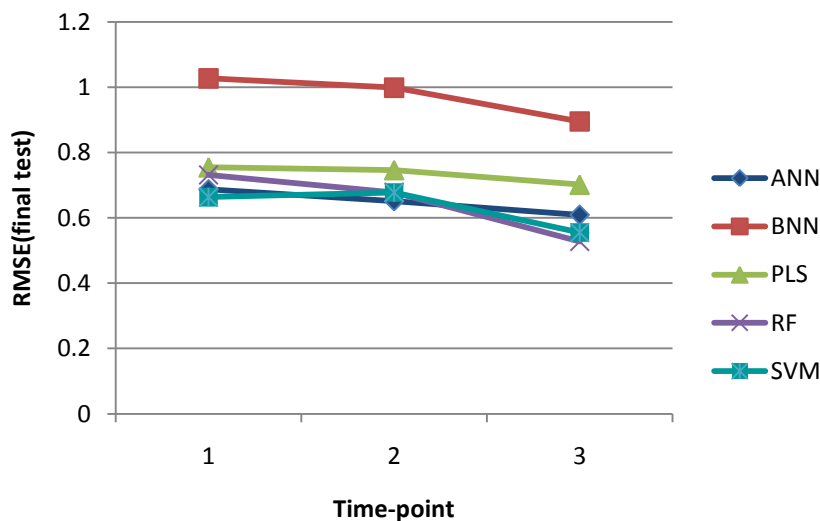
# Dried DMSO Solubility Historical

## Comparison of statistical methods at time of model build



- SVM statistically most predictive initially, then RF at 6 months old and current model
- ANN, RF and SVM very similar predictivity at each time point
- PLS less predictive
- BNN models poor

## Retrospective analysis of model performance

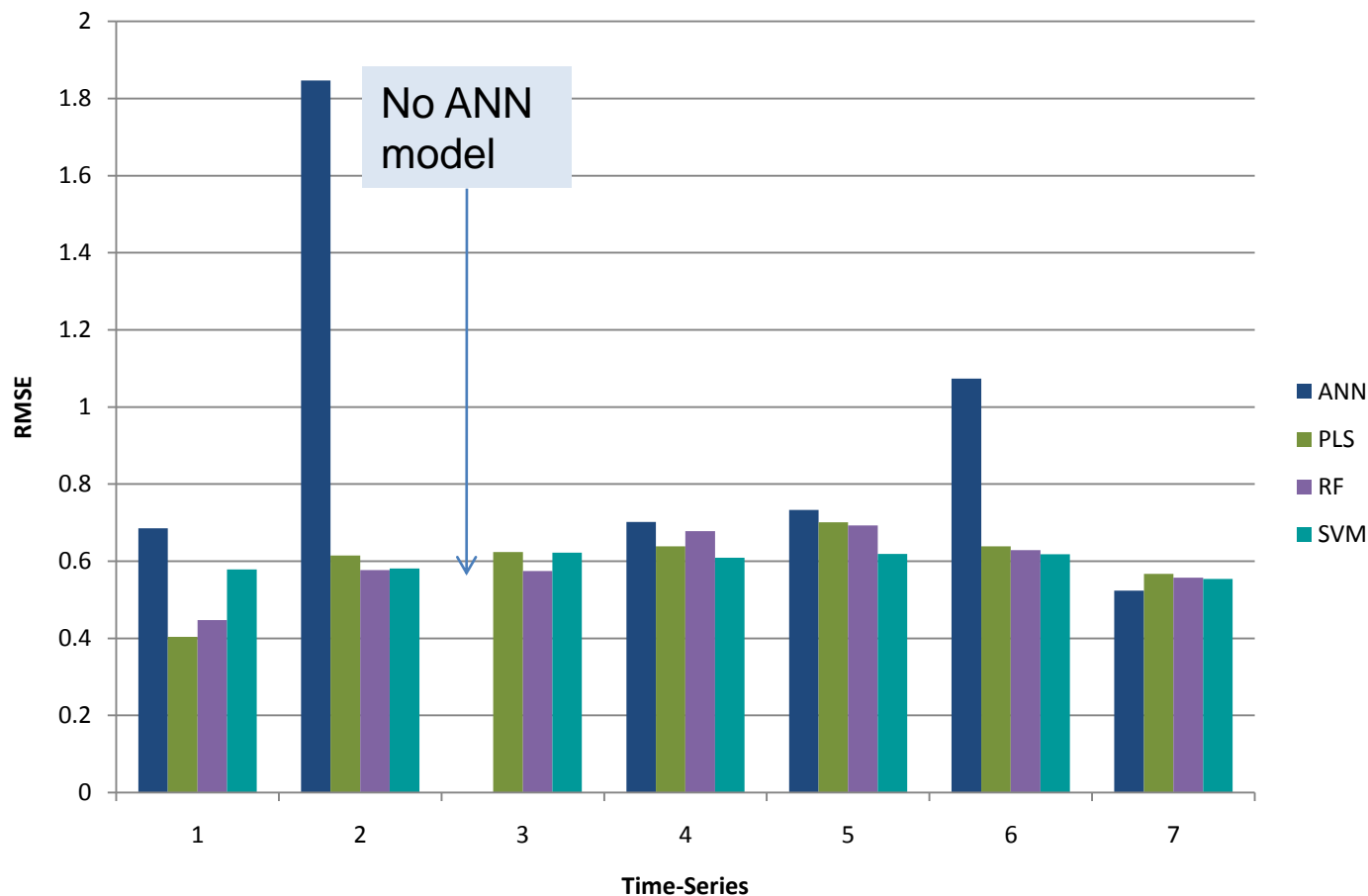


- All models tend to improve in predictivity across the time-series
- RF and SVM most predictive at the end of the time period
- PLS the most stable across the period

- For the endpoints tested, RF and SVM provide the most predictive models
- Findings reflect more general conclusions from global model building within AstraZeneca
- SVM
  - Finds a global solution (not local minima)
  - Less prone to overfitting than NN
- RF
  - building of multiple trees akin to building multiple local models – so all compounds are well represented
  - Not affected by irrelevant descriptors
  - Including multiple trees reduces the effect of overfitting
- PLS
  - Linear, unable to model more complex relationships
- NN methods tend to overfit, even BNN which is designed to prevent overfitting. Although internal validation statistics are promising, the models perform less well on external data
- BNN
  - More difficult to parameterise
  - Possible issues matching the query to the sub-model

- 12 endpoints (**logD**, human PPB, hERG, CYPS, rat HEPS etc)
- 14 series
- Models updated weekly
- 4 statistical methods
  - PLS
  - RF
  - SVM
  - ANN

# Model predictivity across the time series



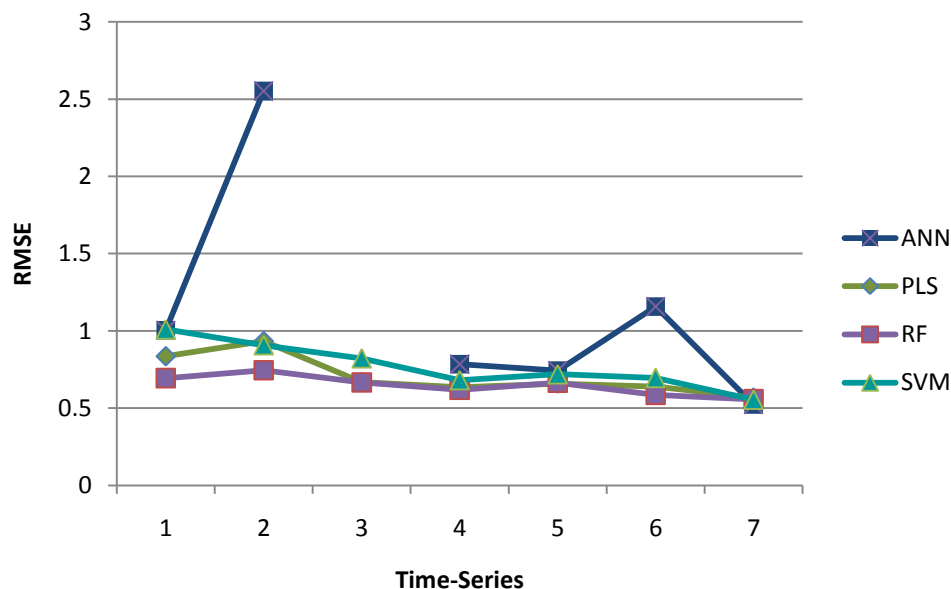
- Predictivity of each model as it is built, across the time-series (using the temporal test set from that build)
- At each time-point, the training set and test set are identical for each method

- Across the time-series, the most predictive method at each time-point varies

Time-point	1	2	3	4	5	6	7
Most Predictive	PLS	RF	RF	SVM	SVM	SVM	ANN

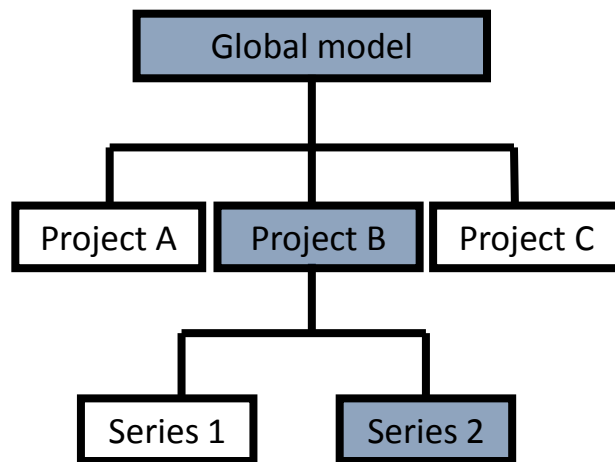
- Each method is most predictive at at least 1 time-point
- Initially, the difference between the methods is greater than towards the end
  - Number of compounds in the project is increasing
- ANN and SVM methods less well suited to smaller datasets at the beginning of the time-series

- Take test set from end of time-series (from final model) and make predictions from historical models
- Provides a guide to whether competitive workflow is selecting the correct models



All models are more predictive towards the end of the time series, when training set compounds are more likely to share greater similarity with the test compounds

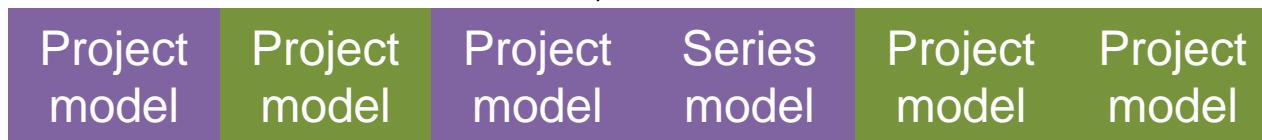
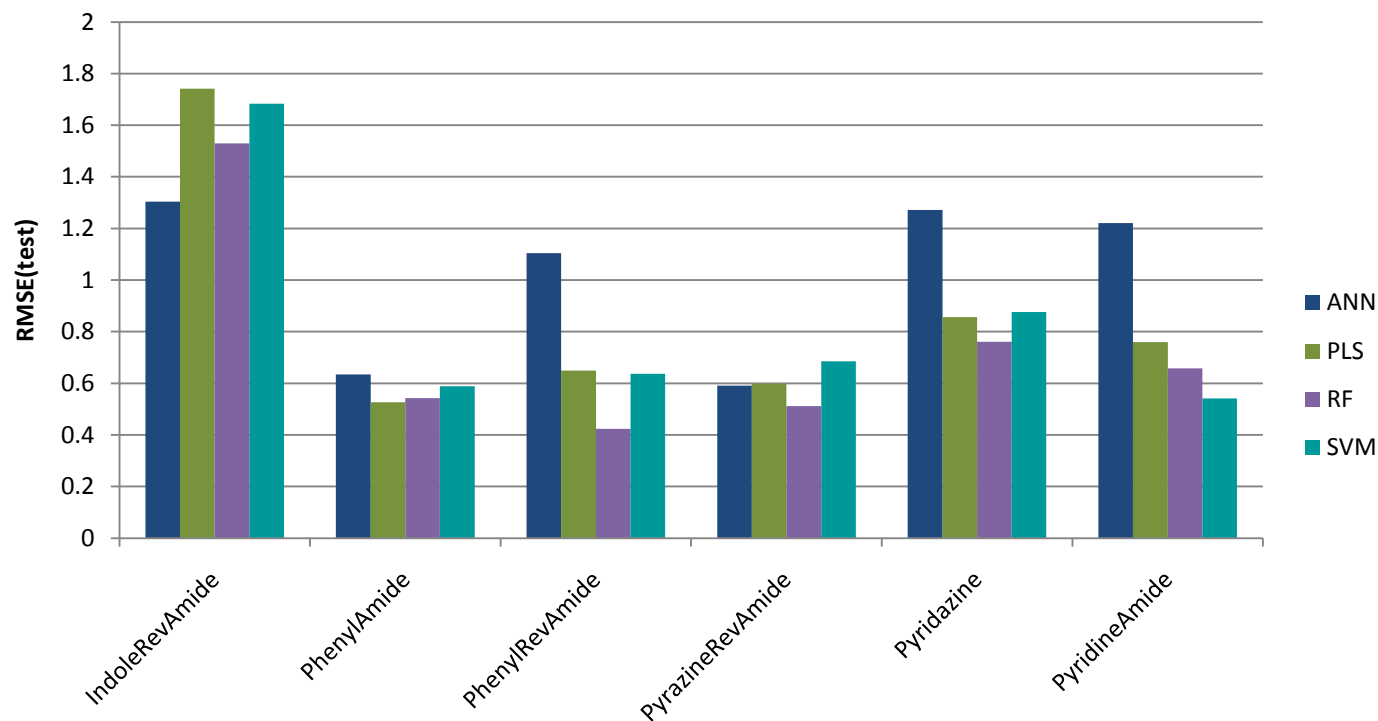
- In addition to building project-level models, AutoQSAR will build a model for each series
- Competitive workflow will then select the most predictive model for each series
- For each series, the project or series model may be the most predictive
- To evaluate which is the most predictive model, the series test set is used for comparison:
  - Is  $RMSE(test)_{series} < RMSE(test)_{project}$  ?



# Hierarchical model time series



For each series, the series test set is used to compare the different statistical methods



RMSE = 0.45

RMSE = 0.48

RMSE = 0.34

RMSE = 0.32

RMSE = 0.43

- In most cases, the project level model is more predictive than series models
  - Series training sets smaller: possibly too small to extract a meaningful relationship with these methods
  - Series test sets, used to compare the models are very small
  - Comparison is a very simple  $RMSE(test)_{series} < RMSE(test)_{project}$ , in some cases the difference between the two is small (and with the smaller test sets, probably not significant)
  - Many of the methods used work better with larger training sets
- PLS and RF models selected most often for the series
- ANN was found to be the overall most predictive project model at this time point (though difference in predictivity between methods was small)
  - Competitive analysis of the hierarchical models identifies that the statistical method selected as 'best' at project level, may not be 'best' for each series

- AZ want to get the best predictions possible so that they make the right decisions with regards to which compounds to progress
- AutoQSAR focusses on the bulk properties to build predictive models
- Wizepairz is concerned with the substructural features of compounds and how they contribute to activity
- Each approach has been developed separately within AZ, plan going forwards is that they will be used in conjunction with each other by projects to provide the best possible information on our lead compounds

Wednesday June 8	
8:30 - 13:00	Analysis Of Large Chemistry Spaces Pat Walters, Presiding
11:00 - 11:30	D-5 : <i>WizePairZ: auto-curation of matched molecular pairs</i> David Wood, AstraZeneca

- AutoQSAR speeds up the process of QSAR model generation
  - Project/series-level models more widely available
  - Models are kept up-to-date
  - Data collection, descriptors, statistical methods, validation methods are standardised
  - Provides a benchmark against which new descriptors/methods can be measured
- AutoQSAR explores model space, selecting the most predictive model using competitive workflow
- No ‘winning’ statistical method, from our study:
  - Global models: SVM and RF
  - Project models: PLS and RF
- Many different factors to test
  - Different endpoints
  - Descriptor sets, fingerprints
  - Classification models.....all of which we can do within AutoQSAR

- Cartmell, J., Enoch, D., Krstajic, D., Leahy, D.E. (2005) Automated QSPR through Competitive Workflow. *J. Comput.-Aided Mol. Des.* 19, 821–833.
- Wold, Herman. (1966). *Estimation of principal components and related models by iterative least squares*. In P.R. Krishnaiah (Ed.). *Multivariate Analysis*. (pp. 391–420) New York: Academic Press.
- Breiman, Leo (2001). *Machine Learning* 45(1), 5-32.
- Cortes, C. and Vapnik, V. (1995) *Support-Vector Networks*, *Machine Learning*, 20.
- Venables, W. N. & Ripley, B. D. (2002) *Modern Applied Statistics with S*. Fourth Edition. Springer, New York.
- R Development Core Team 2006 Version 2.7.1 (Windows), Available from: <http://www.r-project.org/index.html>
- R.M. Neal, Software for Flexible Bayesian Modeling, Version of 06-12-1999. <http://www.cs.utoronto.ca/~radford>
- Stålring, J., Carlsson, L.A., Almeida, P. and Boyer, S. (2011) *AZOrange - High performance open source machine learning for QSAR modeling in a graphical programming environment*. To be published

# Acknowledgements



- Mark Faller
- Stephen Blatch
- Michael Kanz
- Dana Honeycutt



- Jonna Stålring
- Andreas Loong
- Pierre Bruneau

# Thank You



*For more information contact ...*

Sarah Rodgers (Aaron)  
*Contract Research Scientist*  
Accelrys, Inc.  
[saaron@accelrys.com](mailto:saaron@accelrys.com)

