Automated Predictive Modelling: Modeller's Utopia, or Fool's Gold?

Darren Green

Computational Chemistry and Informatics

Introduction to QSAR

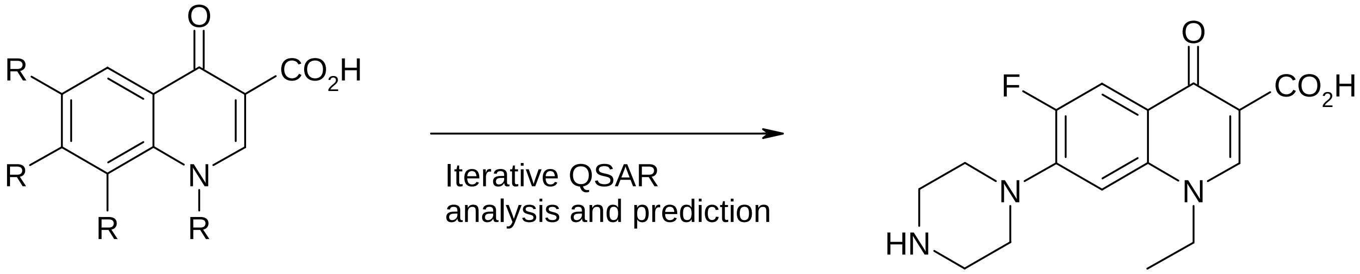
Predictive modelling

- A predictive model quantitatively relates a number of *descriptors* (variable factors that are likely to influence future behaviour or results) to an *outcome*.
  - In marketing, for example, a customer's gender, age, and purchase history (*descriptors*) might predict the likelihood of a future sale (*outcome*).

- In drug discovery, descriptors tend to be derived from chemical structure, and outcomes are *in vitro* or *in vivo* phenomena
  - the goal is to predict behaviour before synthesis
  - models can be built from experimental data too:
    - e.g. prediction of %F from solubility, permeability and clearance data
  - "QSAR" has been practiced since the 1960s
    - and can trace its ancestry back to 1863!

Early QSAR in drug discovery

- Hansch & Fujita, Free & Wilson
- "R-group" style correlative analysis and prediction to aid optimisation



Norfloxacin

QSAR: dead or alive?
Arthur M. Doweyko. J Comput Aided Mol Des (2008) 22:81–89

Modern Drug Discovery

High throughput sciences

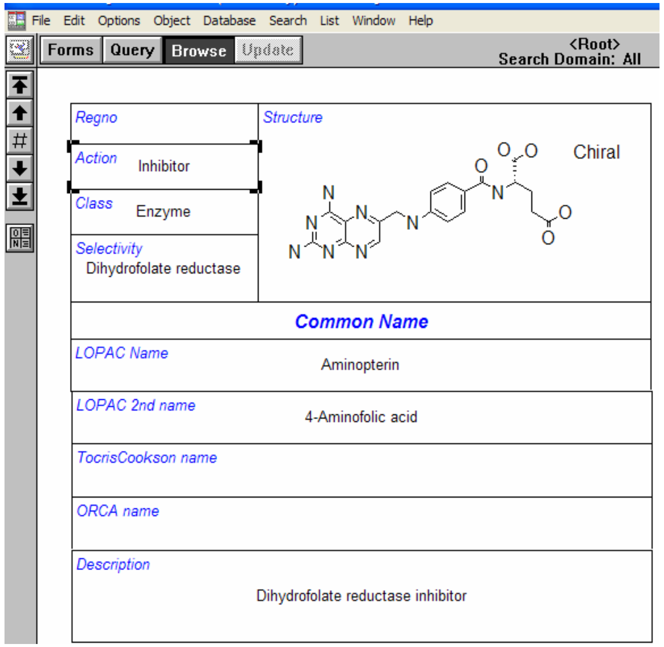
- Massive investments in the ‘90s
  - and still ongoing
- Automation
  - synthesis
  - screening
  - protein expression
  - crystallisation
  - gene expression/screening
    - ‘omics in general

Progress

- It has taken 10-15 years to achieve:
  - 100Ks of purified single compounds per year
  - >100K screening wells per day
  - protein crystallisation/structure solution inside a week
- And not just throughput
  - high content screening
  - ‘omics data

gsk GlaxoSmithKline

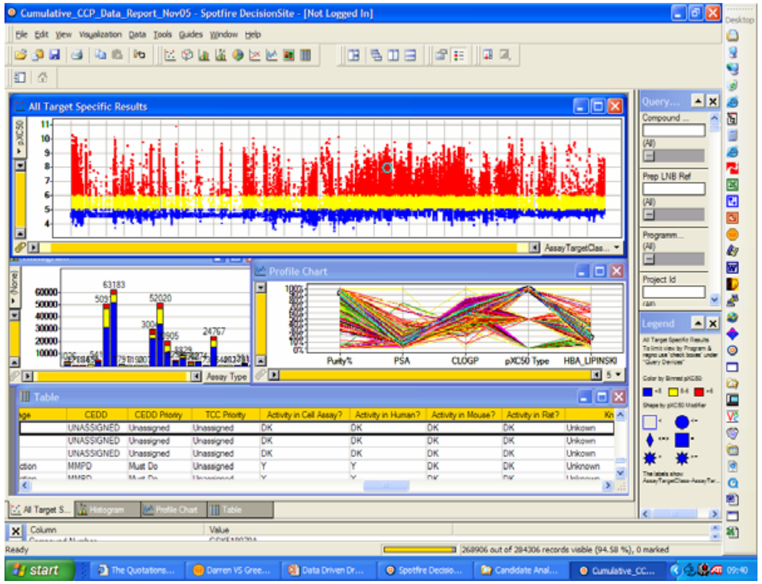
The effect of automation on decision & design complexity in medicinal chemistry



Regno	Structure
Action: Inhibitor	
Class: Enzyme	
Selectivity: Dihydrofolate reductase	

Common Name

LOPAC Name	Aminopterin
LOPAC 2nd name	4-Aminofolic acid
TocrisCookson name	
ORCA name	
Description	Dihydrofolate reductase inhibitor



...ge	CEDD	CEDD Priority	TCC Priority	Activity in Cell Assay?	Activity in Human?	Activity in Mouse?	Activity in Rat?	Kn...
	UNASSIGNED	Unassigned	Unassigned	DK	DK	DK	DK	Unkown
	UNASSIGNED	Unassigned	Unassigned	DK	DK	DK	DK	Unkown
	UNASSIGNED	Unassigned	Unassigned	DK	DK	DK	DK	Unkown
...ction	MMPD	Must Do	Unassigned	Y	Y	DK	DK	Unknown

All Target Specific Results
To limit view by Program & regno use 'check boxes' under "Query Devices"

The labels show AssayTargetClass-AssayTar...

Column	Value

Ready

268906 out of 284306 records visible (94.58 %), 0 marked

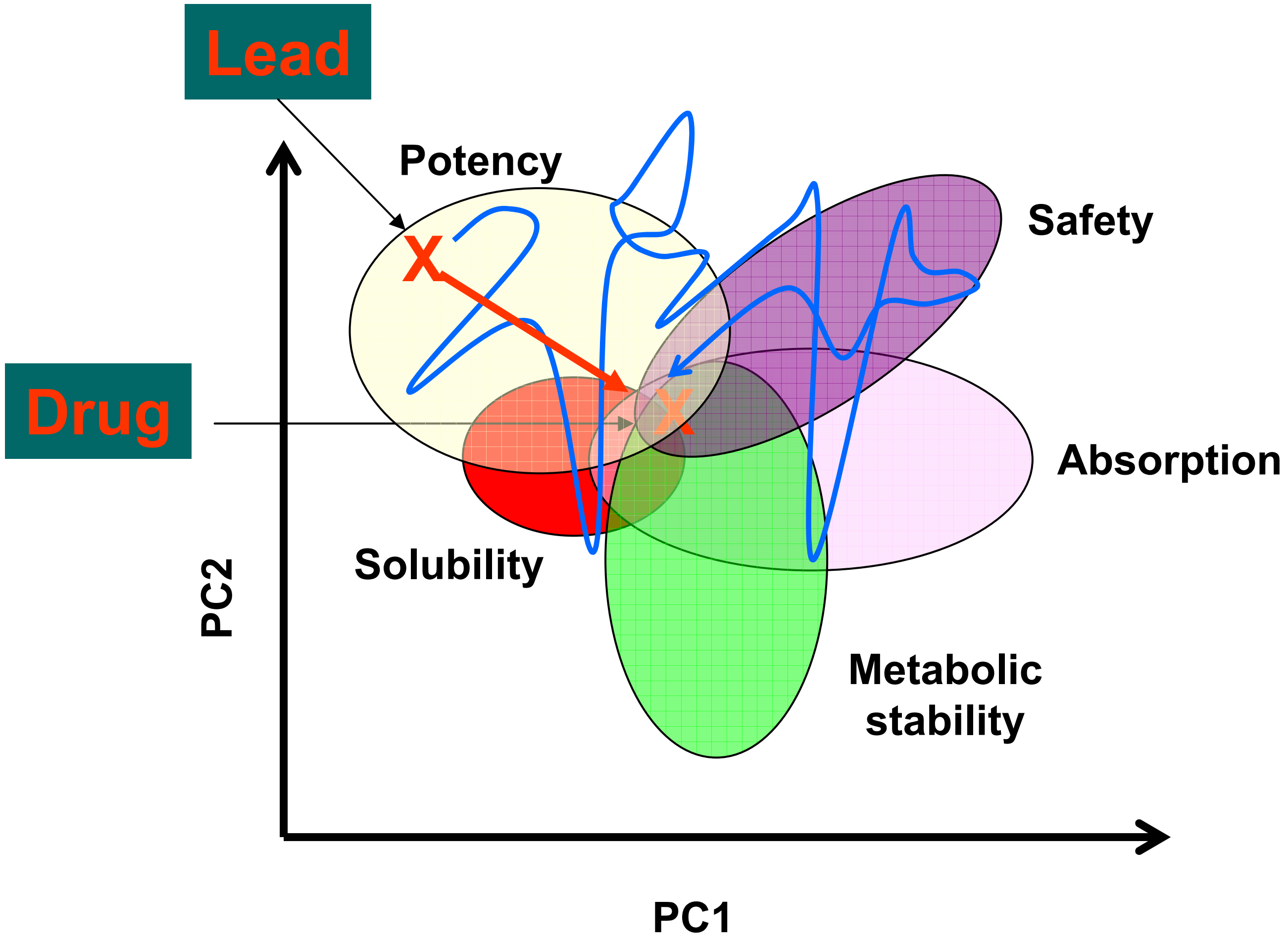
Machine Learning and Chemoinformatics

Rejuvenated QSAR

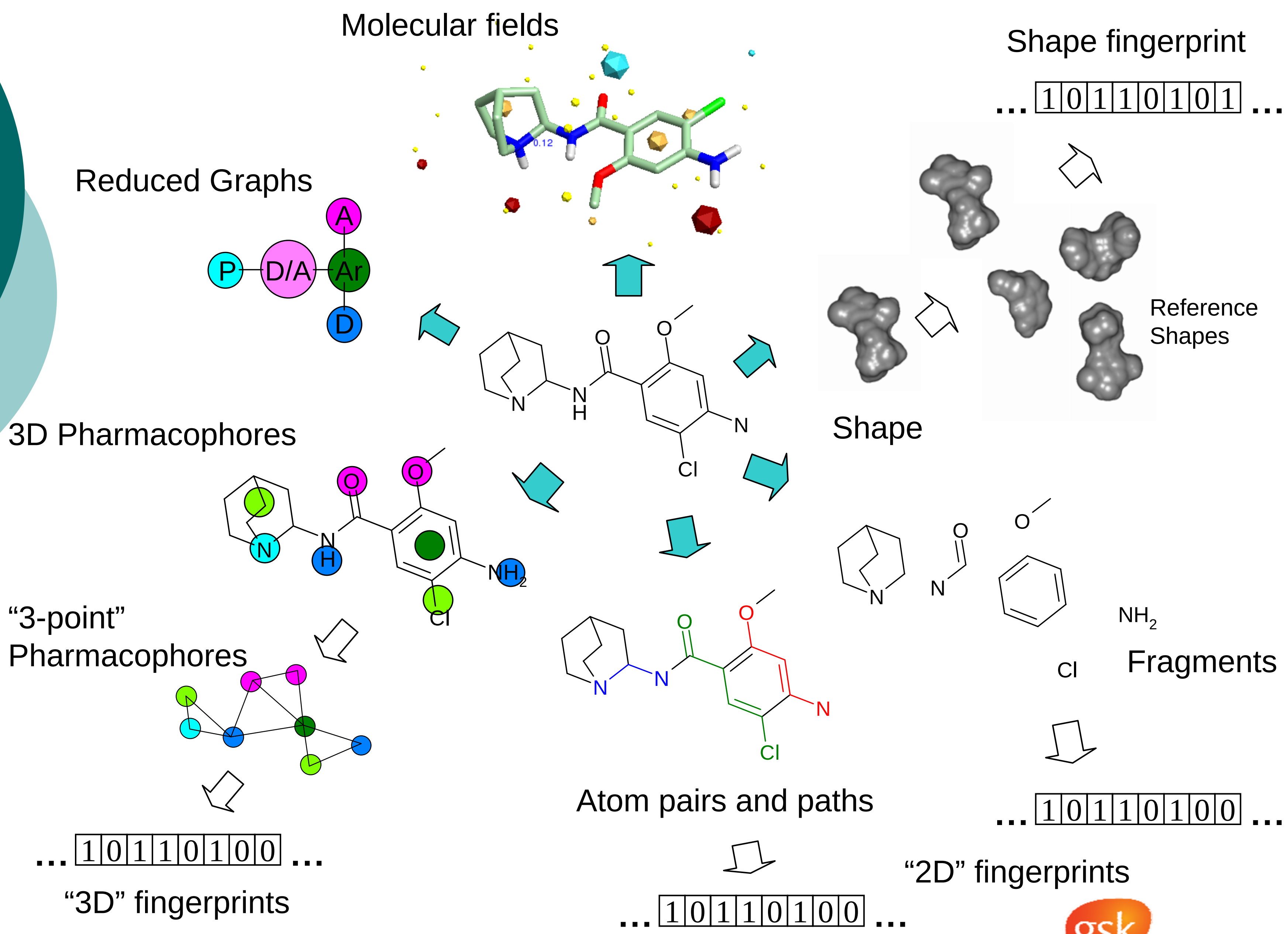
- Renewed interest in “QSAR” modelling
  - More distinct end points measured
  - With higher throughput => more data
  - Greater access to chemical descriptors
  - New statistical methods
  - End user tools for access to predictions
- A different type of modelling to that practiced in the ’70s
  - better?

Why Predict?

A multi-objective optimisation process



Predictive QSAR: Faster way to navigate the route via prediction and knowledge



Molecular fields

Shape fingerprint

... 1 0 1 1 0 1 0 1 ...

Reduced Graphs

Reference Shapes

Shape

3D Pharmacophores

"3-point" Pharmacophores

Fragments

... 1 0 1 1 0 1 0 0 ...

Atom pairs and paths

... 1 0 1 1 0 1 0 0 ...

"3D" fingerprints

"2D" fingerprints

... 1 0 1 1 0 1 0 0 ...

GlaxoSmithKline

Statistics

- A bewildering variety of statistical methods are available
  - Mainly driven by data mining & prediction needs in other industries: marketing, finance, telecom, insurance etc
  - multiple linear regression, logistic regression, K-nearest neighbours, PLS, linear discriminant analysis, decision trees, random forests, neural networks, Support Vector Machines and many, many more

The “new” QSAR

- Major difference from Free-Wilson
  - Less obviously interpretable descriptors
  - Non-linear relationships in many dimensions
  - BUT
    - Can model multiple chemotypes
    - Can extrapolate to new chemotypes?

Drivers for Change

Drivers for change in predictive modelling

- **Too many models, too little time…**
  - Many new parameters being added to address “attrition” in drug discovery
  - Data volumes and rate of production ↑
  - QSAR specialists ~ constant (or declining!)
  - **Modelling cannot produce up to date, hand crafted models for every required end point**

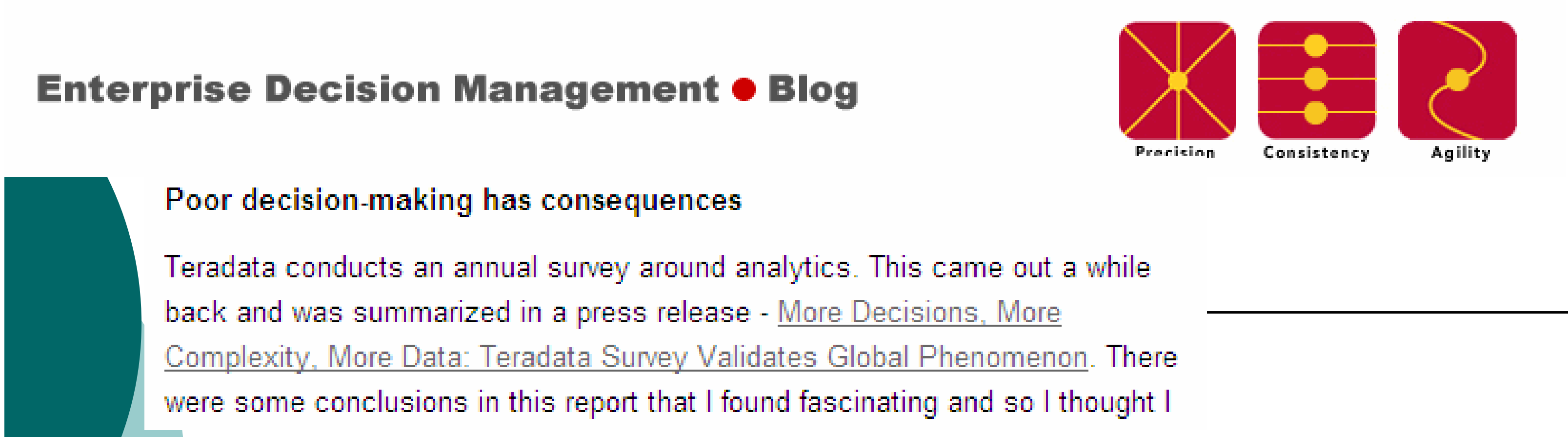
Drivers for change in predictive modelling

- **Modelling tools come from different sources**
  - Internal, external, scripts, etc.
  - Data types, descriptors, file formats, etc.
- There is no “holy grail” method
  - Model and descriptor performance is variable
- Individual expertise typically in one or two tools/techniques
- Data preparation typically takes more time than model building
- **Results in local expertise, global questions**

Drivers for change in predictive modelling

- Consumption by non-specialists
  - profiling of virtual compounds
  - design tools etc
- Build->deployment needs to be low effort
- “Blind” use of models
  - model being used by someone who does not have “mothers love”

AME Vision and Implementation



Enterprise Decision Management ● Blog

Poor decision-making has consequences

Teradata conducts an annual survey around analytics. This came out a while back and was summarized in a press release - More Decisions, More Complexity, More Data: Teradata Survey Validates Global Phenomenon. There were some conclusions in this report that I found fascinating and so I thought I

Data volumes are increasing massively - over half of respondents saying that data volume is doubling or tripling over the previous year

To me this means **automate** or die. Data volumes are going nowhere but up and it's going to get harder and harder for people to effectively use all this data. One of the great things about computers is a much greater capacity for managing volume so why not turn all this data into insight/ actions that can be executed **automatically**?

One common response to this extra data is to add staff

I suspect most of this was to handle the data management itself but if companies are adding staff to try and process the extra data, that's not going to scale. This data is only going to be useful if it can be applied to business problems and the volume and volume growth mean that this application will have to be **automated**

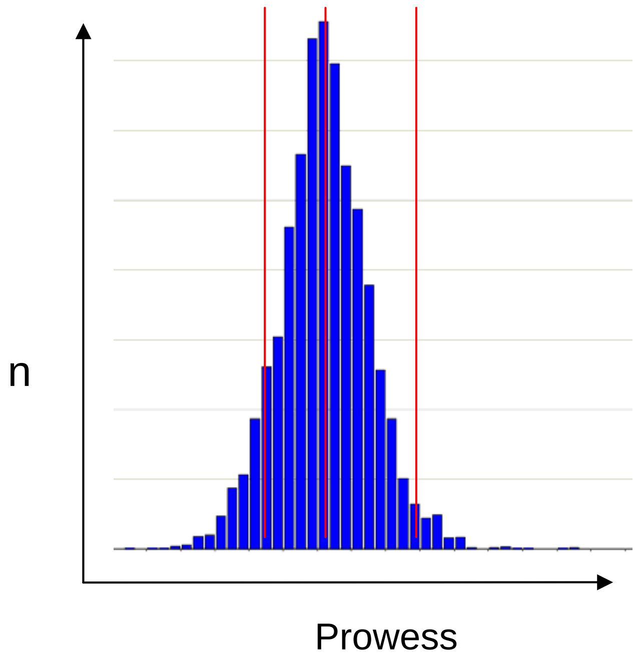
The importance of real-time information is a five-year trend, and this year, eighty-five percent of respondents said that decision-makers need more up-to-date information than in the past

So more and more people need to take more and more decisions with more and more data in less and less time. Hmm, sounds like these people need to **automate** some of this...

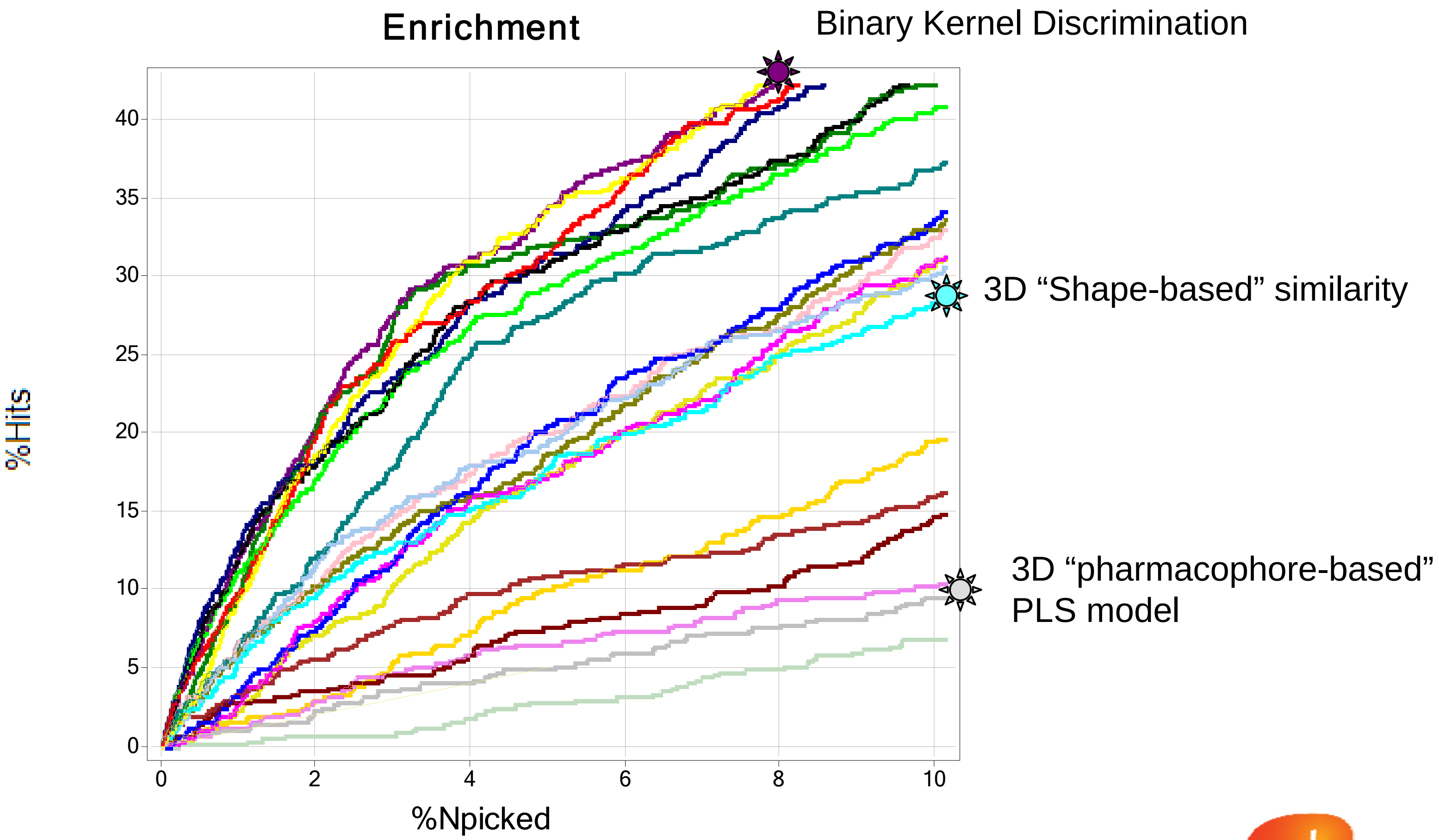
http://www.edmblog.com/weblog/2006/10/poor_decisionma.html

Automation in a scientific environment

- The automated science must work
- What does *work* mean for analysis/design methods?



Virtual Screening Experiment



Original slide provided by Paul Bamborough

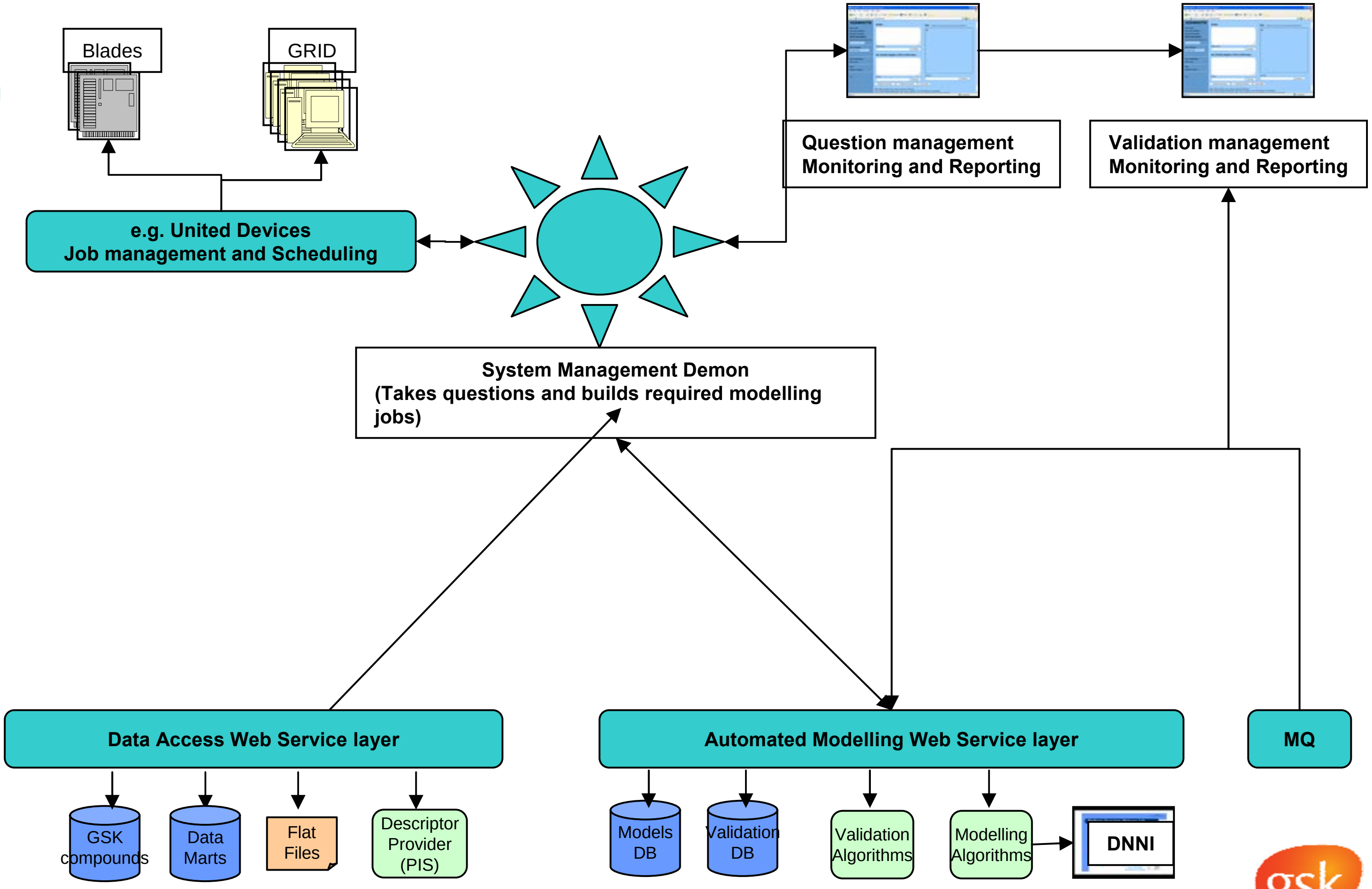
Strategy

- People
  - Automate what machines do best
    - Data QC, processing
    - Repetitive tasks
- Good modelling practice/method parameterisation
  - encode in the system
- Implement
  - Handling multiple models and distributed processing
  - Robust automated modelling methodology
- Automated model validation/refinement
  - Let the modeller know when the model is not performing and why

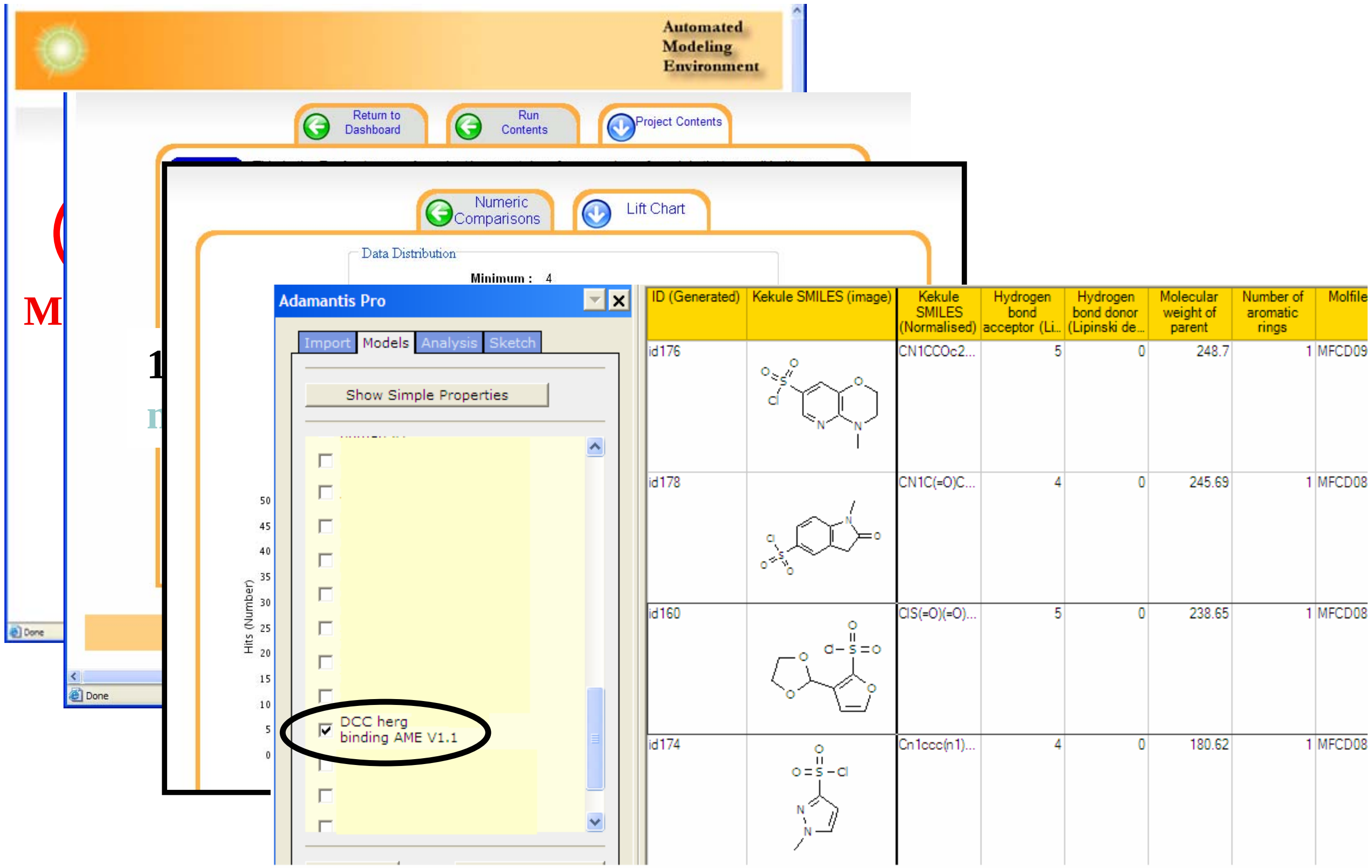
Strategy

- **People**
  - **Automate what machines do best**
    - **Data QC, processing**
    - **Repetitive tasks**
- **Good modelling practice / method parameterisation**
  - **encode in the system**
- **Implement**
  - **Handling multiple models and distributed processing**
  - **Robust automated modelling methodology**
- Automated model validation/refinement
  - Let the modeller know when the model is not performing and why

AME Vision


Blades
GRID
e.g. United Devices
Job management and Scheduling
Question management
Monitoring and Reporting
Validation management
Monitoring and Reporting
System Management Demon
(Takes questions and builds required modelling jobs)
Data Access Web Service layer
Automated Modelling Web Service layer
MQ
GSK compounds
Data Marts
Flat Files
Descriptor Provider (PIS)
Models DB
Validation DB
Validation Algorithms
Modelling Algorithms
DNNI

GSK Automated Modelling Environment (AME)



Data Distribution

Minimum : 4

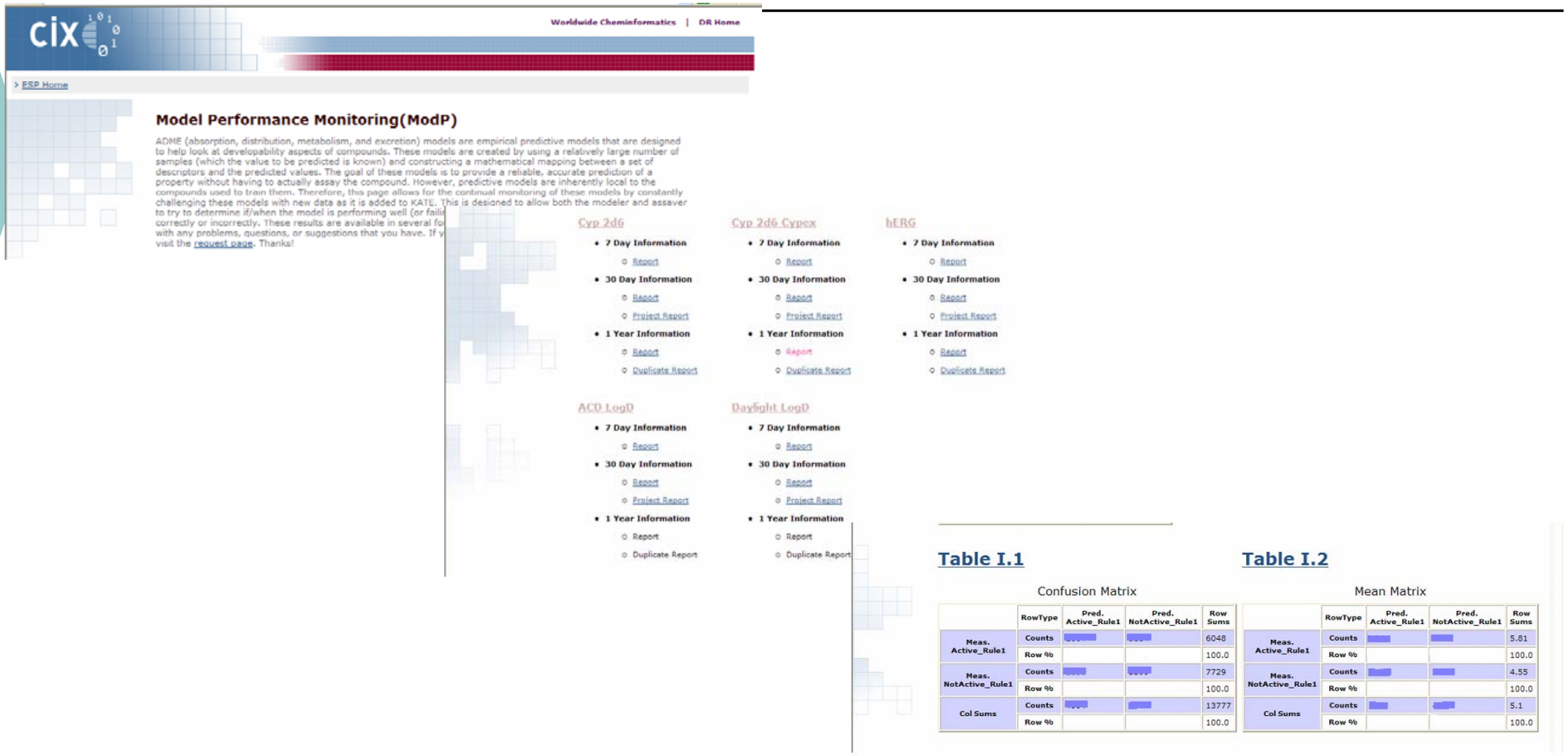
☑ DCC herg binding AME V1.1

ID (Generated)	Kekule SMILES (image)	Kekule SMILES (Normalised)	Hydrogen bond acceptor (Li...	Hydrogen bond donor (Lipinski de...	Molecular weight of parent	Number of aromatic rings	Molfile
id176		CN1CCOc2...	5	0	248.7	1	MFCD09
id178		CN1C(=O)C...	4	0	245.69	1	MFCD08
id160		ClS(=O)(=O)...	5	0	238.65	1	MFCD08
id174		Cn1ccc(n1)...	4	0	180.62	1	MFCD08

Strategy

- People
  - Automate what machines do best
    - Data QC, processing
    - Repetitive tasks
- Good modelling practice/method parameterisation
  - encode in the system
- Implement
  - Handling multiple models and distributed processing
  - Robust automated modelling methodology
- **Automated model validation/refinement**
  - **Let the modeller know when the model is not performing and why**

Automated Validation



Model Performance Monitoring(ModP)

ADME (absorption, distribution, metabolism, and excretion) models are empirical predictive models that are designed to help look at developability aspects of compounds. These models are created by using a relatively large number of samples (which the value to be predicted is known) and constructing a mathematical mapping between a set of descriptors and the predicted values. The goal of these models is to provide a reliable, accurate prediction of a property without having to actually assay the compound. However, predictive models are inherently local to the compounds used to train them. Therefore, this page allows for the continual monitoring of these models by constantly challenging these models with new data as it is added to KATE. This is designed to allow both the modeler and assayer to try to determine if/when the model is performing well (or faili correctly or incorrectly. These results are available in several fo with any problems, questions, or suggestions that you have. If y visit the request page. Thanks!

Cyp 2d6

- 7 Day Information
  - ○ Report
- 30 Day Information
  - ○ Report
  - ○ Project Report
- 1 Year Information
  - ○ Report
  - ○ Duplicate Report

Cyp 2d6 Cypex

- 7 Day Information
  - ○ Report
- 30 Day Information
  - ○ Report
  - ○ Project Report
- 1 Year Information
  - ○ Report
  - ○ Duplicate Report

hERG

- 7 Day Information
  - ○ Report
- 30 Day Information
  - ○ Report
  - ○ Project Report
- 1 Year Information
  - ○ Report
  - ○ Duplicate Report

ACD LogD

- 7 Day Information
  - ○ Report
- 30 Day Information
  - ○ Report
  - ○ Project Report
- 1 Year Information
  - ○ Report
  - ○ Duplicate Report

Daylight LogD

- 7 Day Information
  - ○ Report
- 30 Day Information
  - ○ Report
  - ○ Project Report
- 1 Year Information
  - ○ Report
  - ○ Duplicate Report

Table I.1

Confusion Matrix

	RowType	Pred. Active_Rule1	Pred. NotActive_Rule1	Row Sums
Meas. Active_Rule1	Counts	[redacted]	[redacted]	6048
	Row %			100.0
Meas. NotActive_Rule1	Counts	[redacted]	[redacted]	7729
	Row %			100.0
Col Sums	Counts	[redacted]	[redacted]	13777
	Row %			100.0

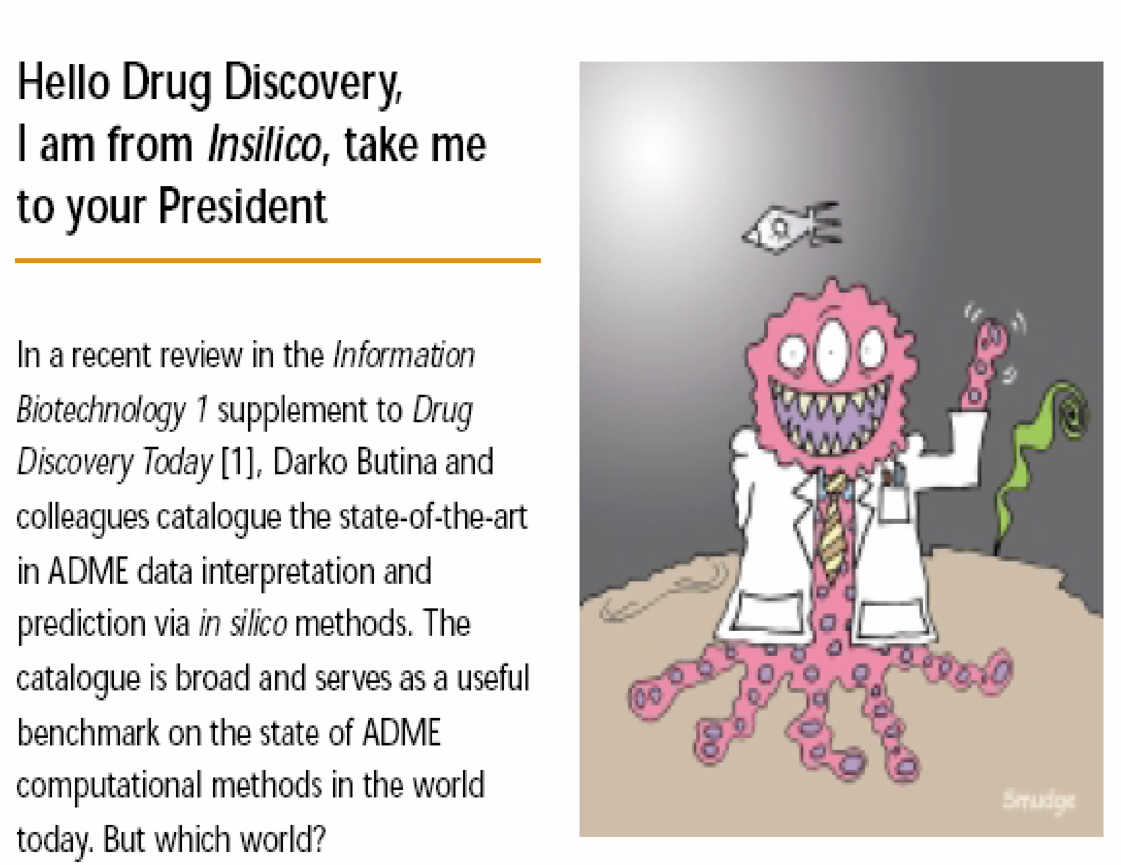
Table I.2

Mean Matrix

	RowType	Pred. Active_Rule1	Pred. NotActive_Rule1	Row Sums
Meas. Active_Rule1	Counts	[redacted]	[redacted]	5.81
	Row %			100.0
Meas. NotActive_Rule1	Counts	[redacted]	[redacted]	4.55
	Row %			100.0
Col Sums	Counts	[redacted]	[redacted]	5.1
	Row %			100.0

Change Management Challenges

Culture

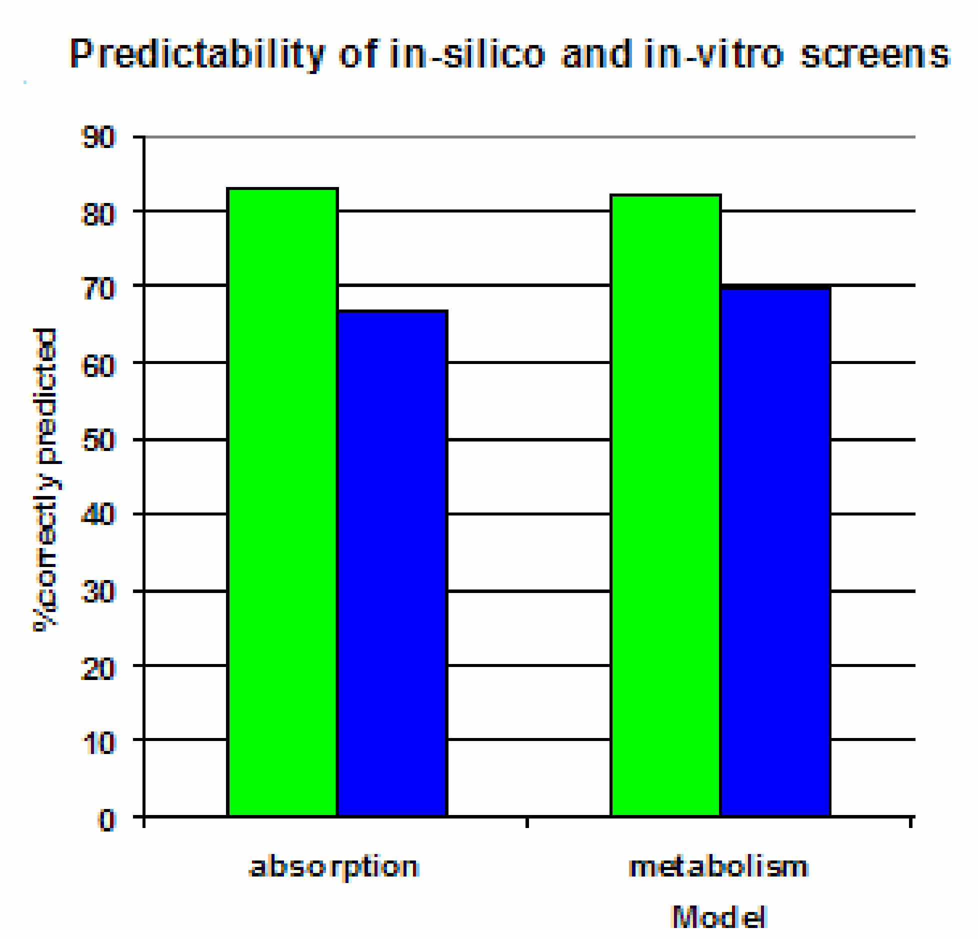


Hello Drug Discovery, I am from *Insilico*, take me to your President

In a recent review in the *Information Biotechnology 1* supplement to *Drug Discovery Today* [1], Darko Butina and colleagues catalogue the state-of-the-art in ADME data interpretation and prediction via *in silico* methods. The catalogue is broad and serves as a useful benchmark on the state of ADME computational methods in the world today. But which world?

Drug Discovery Today, Volume 7, Issue 21, 1 November 2002

in vitro screens are models too….



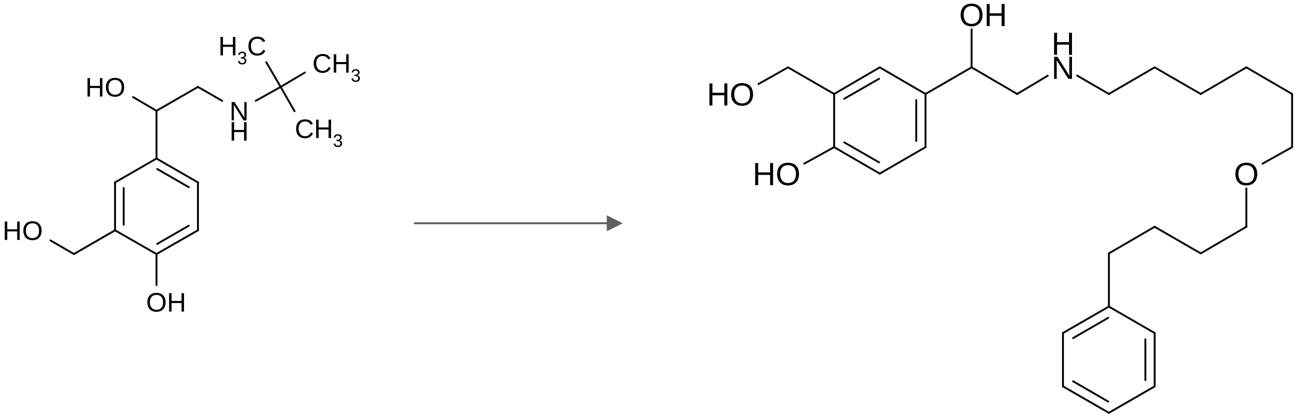
Science

- Modellers
  - Trusting automation
  - Letting go of old practices
  - Intellectual emphasis shifts from model *building* to model *choosing*
  - Processing the deluge of information from model and validation services

- Chemists
  - Using multiple predictive models
    - Multi Objective analysis
      - Not using as filters in XL!
  - Using continually changing models (as more information becomes available)
    - What was correctly predicted last week might not be this week, but there should be fewer false predictions overall…

Embracing Uncertainty

- Salmeterol
  - QSAR analysis => *hypothesis*
    - med. chem. testing of hypothesis



http://www.chemsoc.org/chembytes/ezine/2001/newton_sep01.htm

Summary

Opportunities and fantasies

- A shift from rich GUIs to consumption of modelling results, run online as the data is generated
- Automated processing of data and “push” of interesting results/diagnostics
  - “the latest results show that this compound has the profile you have been looking for…”
  - “there is something unexpected about this result, you might want to review the experiment…”
  - “this molecule doesn’t match your potency cut-off, but is predicted to meet your solubility criteria, so has been routed to AC for measurement….”
  - “based on the latest screening results, AME has been able to build a predictive QSAR model for your selectivity assay X. Click here to see Chemonaut suggestions….”

Summary

- The automated generation of predictive QSAR models is now possible
  - this capability will enable a whole range of new applications
- The key challenges are cultural
  - Exploiting the automation
  - Using models wisely, not blindly

Acknowledgements

- Chris Keefer
- Chris Bizon
- Nate Woody
- Chris Luscombe