MARIO NEGRI
ISTITUTO DI RICERCHE FARMACOLOGICHE
Emilio Benfenati
Professor
CAESAR Project Coordinator
Mario Negri Institute
Integrating Infrastructure Requirements for Chemistry Research in Predictive Toxicology
Pre-conference OpenTox workshop of the AXLR8 2010 meeting
May 30th, 2010 : Potsdam, Germany
OpenTox

(Q)SAR

=

(quantitative) structure-activity relationship

IN SILICO

MARIO NEGRI
ISTITUTO DI RICERCHE FARMACOLOGICHE

Partial view in QSAR

- Errors – poor statistical meaning; wrong chemical info; wrong tox basis
- False belief – skepticism / affected by personal background
- Opportunities missed

Explicit and implicit knowledge

Probabilistic and deterministic approach

THE CAESAR ENDPOINTS

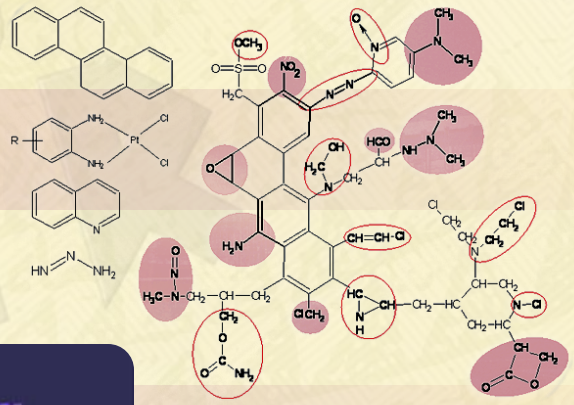
- BIOCONCENTRATION FACTOR
- SKIN SENSITIZATION
- MUTAGENICITY
- CARCINOGENICITY
- DEVELOPMENTAL TOXICITY

Mutagenicity

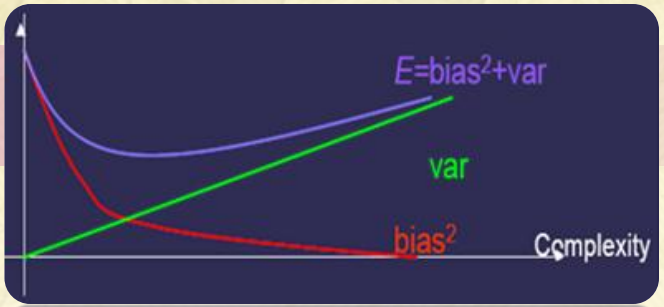
Classification Models

QSAR models of noncongeneric compounds to predict mutagenicity can use TWO APPROACHES:

1 : STRUCTURAL ALERTS



2 : STATISTICS



CAESAR MODELLING FOR MUTAGENICITY

Dataset

- *Kazius-Bursi Mutagenicity Dataset (Kazius et al. J Med Chem, 2005),* originally containing 4337 chemical compounds, supplied by R. Bursi
- Data are *categorical*
- Following quality checks the database has been pruned and modified to *4225 compounds*: 2358 classified as *mutagens* and 1867 classified as *non-mutagens* by *Ames test*
- For validation, the dataset has been divided into training (80%) and test (20%) sets

CAESAR MODELLING FOR MUTAGENICITY

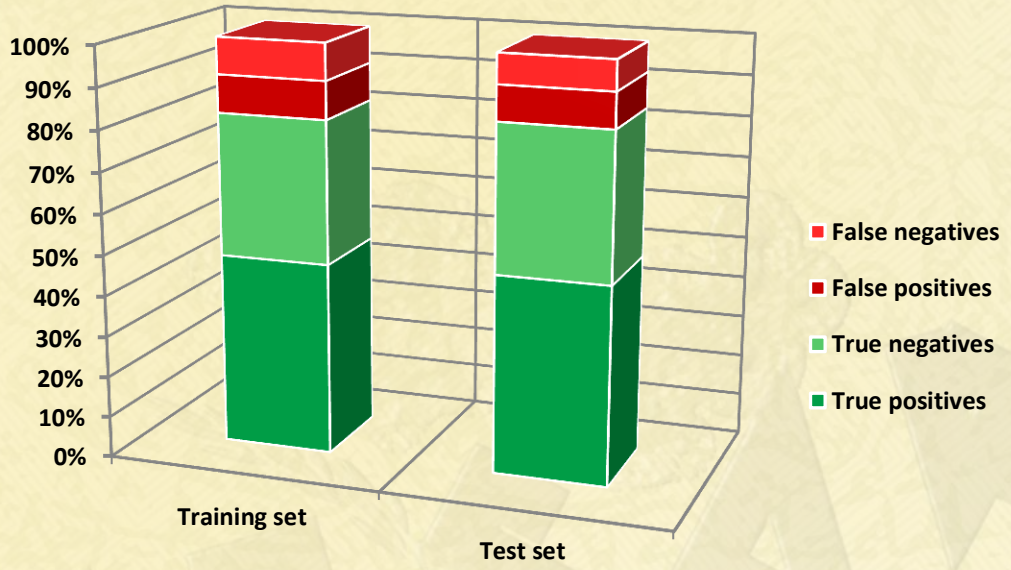
Descriptors

- 2D descriptors: MDL software

Models

- Classification: SVM (Support Vector Machines)
- 10 fold cross-validation

Results of CAESAR Modelling

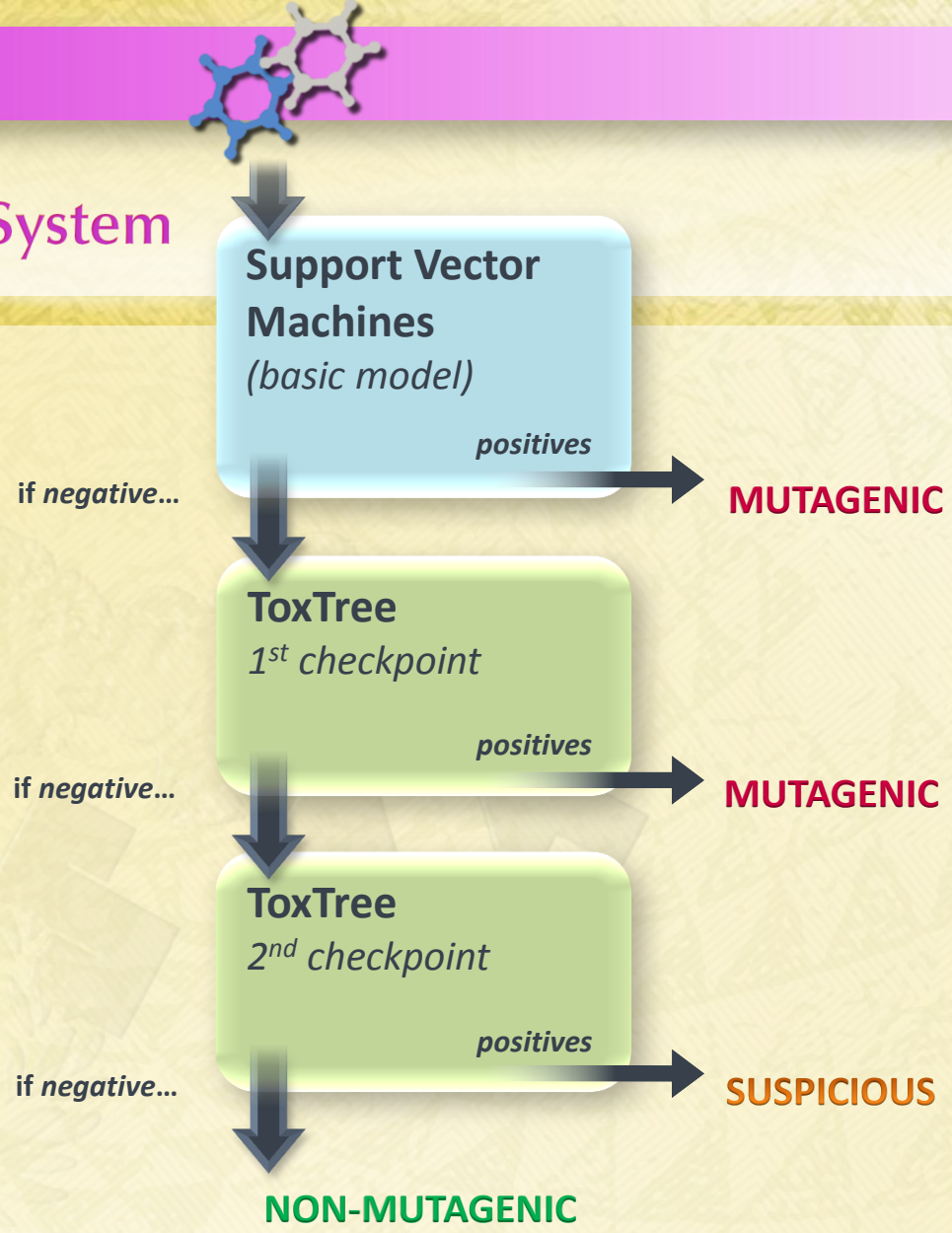


- Good accuracy (considering reproducibility of the experimental data about 85%)
- A cost-sensitive model was also evaluated to reduce *FN*

Architecture of the Integrated System

3 STEPS IN CASCADE:

- **statistical model** (based on chemical descriptors)
- **knowledge-based filter** (based on structural alerts)



ToxTree vs SVM

COMPARISON OF PERFORMANCE (*on the same data*):

CAESAR Test Set	Toxtree	SVM model
accuracy:	78%	✓ **83%**
sensitivity:	86%	✓ **87%**
specificity:	69%	✓ **79%**

Integrated Model Statistics

CAESAR Test Set	SUSPICIOUS taken as NON-MUTAGENIC	SUSPICIOUS taken as MUTAGENIC
accuracy:	**83.3%**	82.1%
sensitivity:	88.3%	**90.9%**
specificity:	**77.1%**	71.2%

CONFIDENT CHOICE

Accuracy close to the reliability of the experimental test (85%)

PRUDENT CHOICE

Sensitivity boosted over 90%

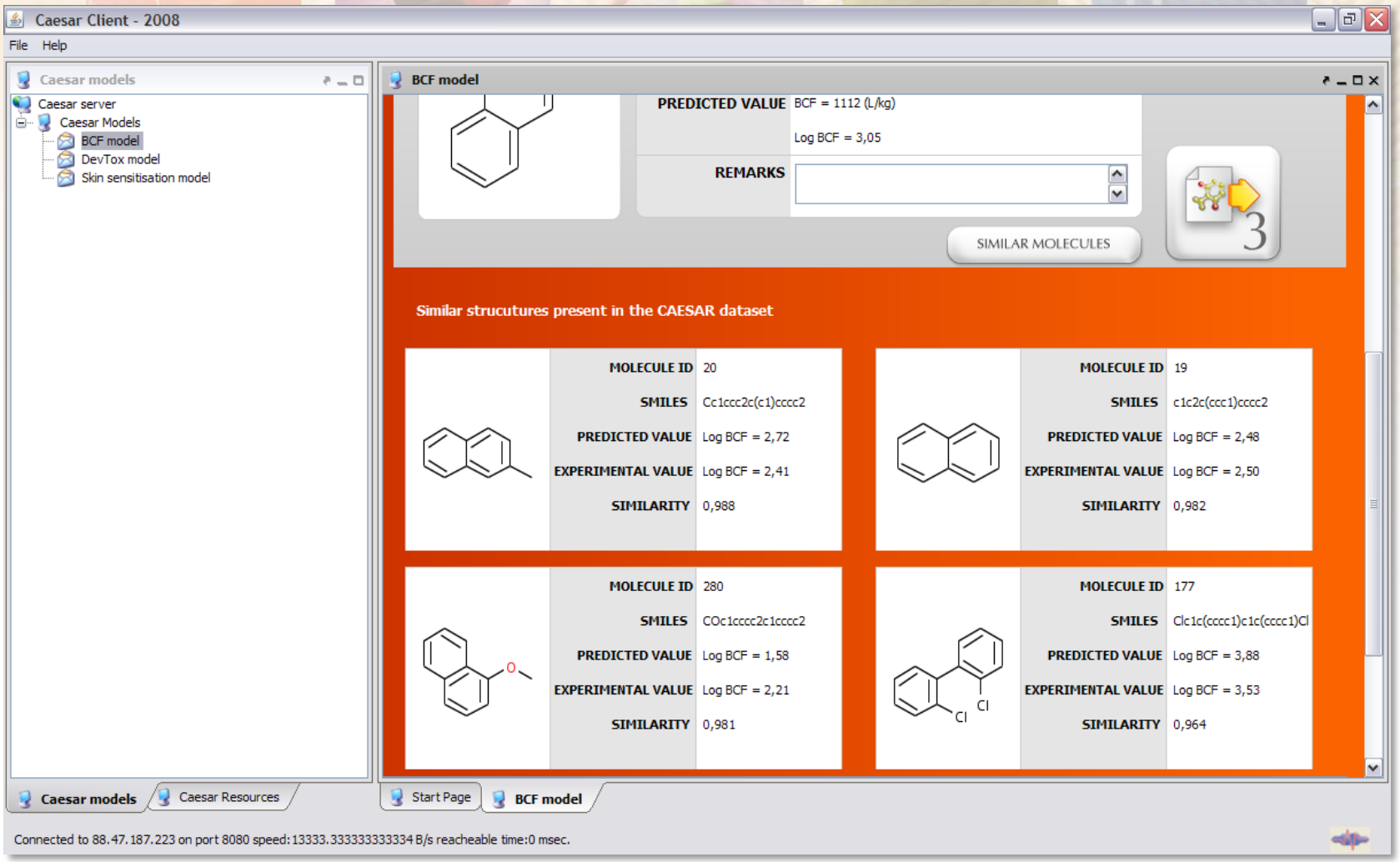
Mutagenicity: Conclusions

- The cascade model has achieved a classification accuracy close to the reliability of the *Ames test* data (*average interlaboratory reproducibility error of 15%*) used to train and validate the model;
- The experimental error is a major bottleneck;
- This gives evidence that very good performance is possible with machine learning software from public domain;
- Selected structural alerts can discover FN (but can moderately increase FP as well);
- The CAESAR model has been checked against commercial systems (Multicase, Derek); it gave always not worse results.

CAESAR
CAESAR Remote Desktop Client
QSAR

Viewing the Results

SIMILAR MOLECULES



MARIO NEGRI
ISTITUTO DI RICERCHE FARMACOLOGICHE

http://www.caesar-project.eu/software/index.htm

APPLICABILITY DOMAIN: CHEM, TOX, MATH

Current methods:

on chemical info / a priori

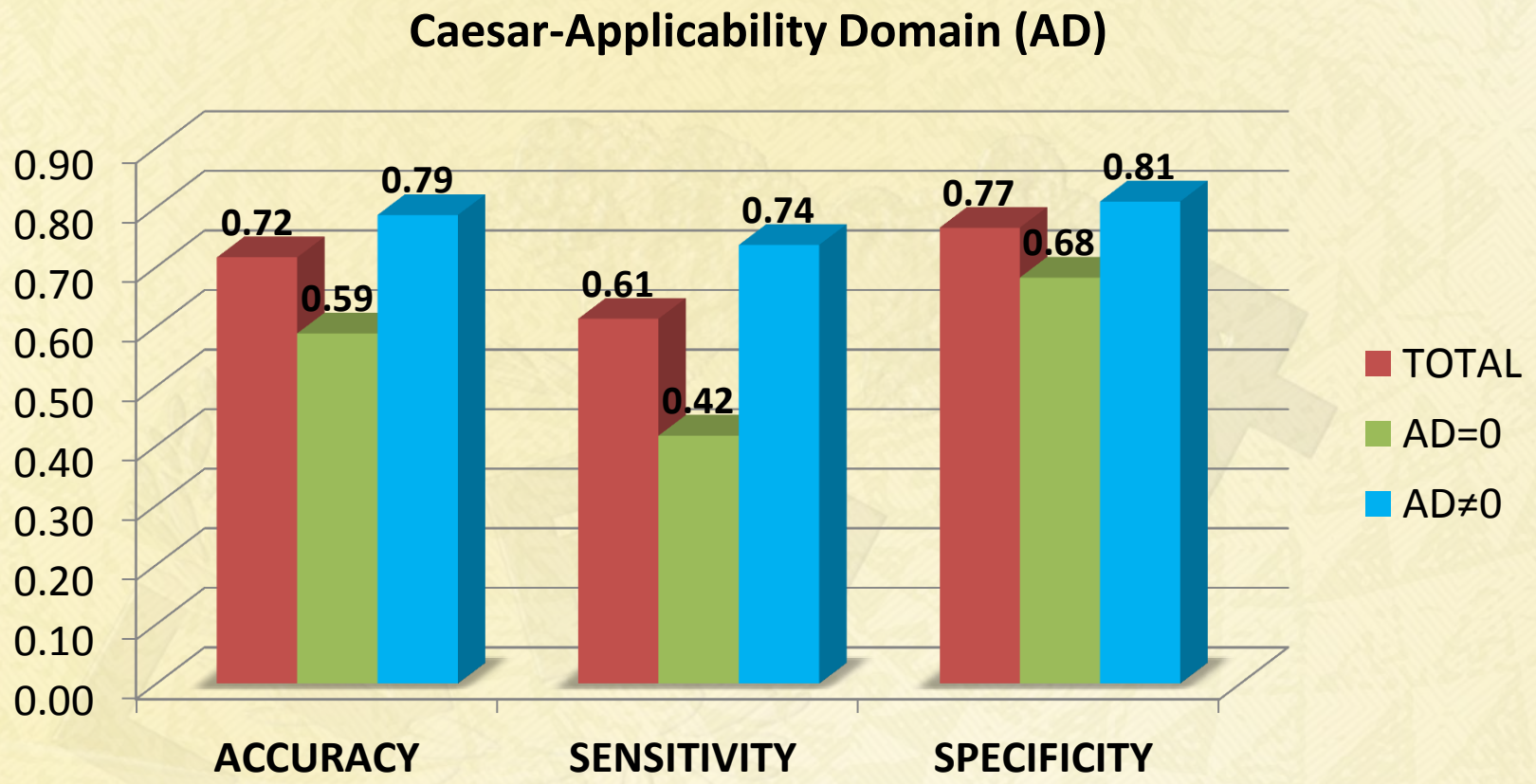
CAESAR approach:

on chem; tox; math

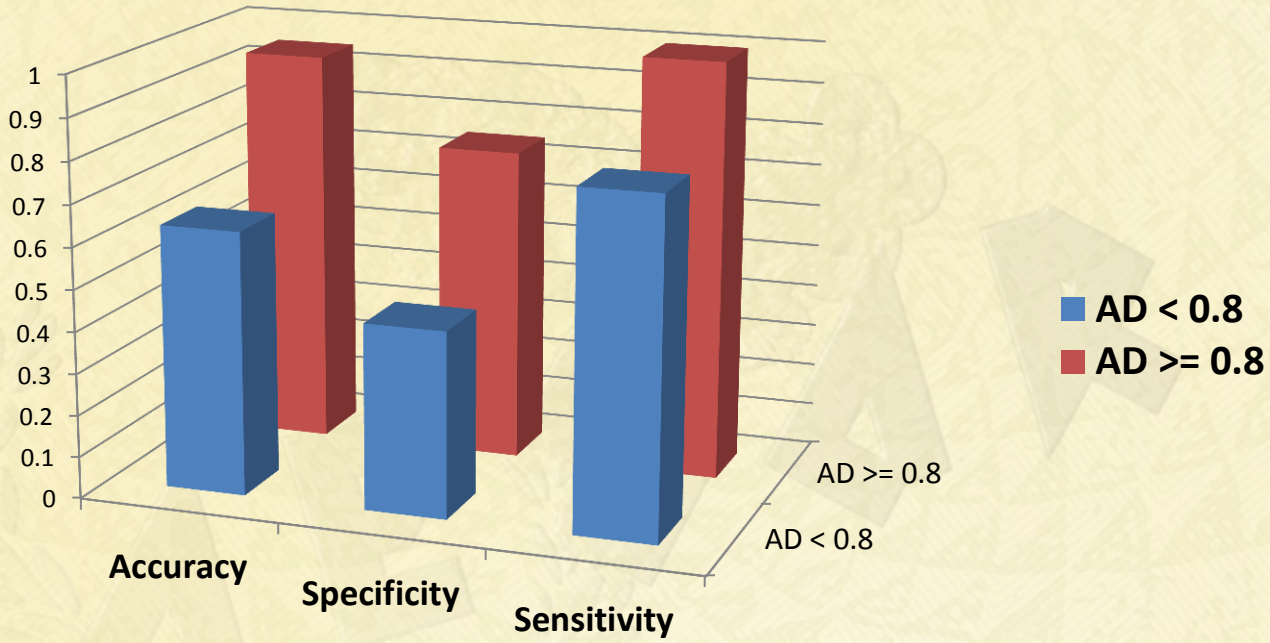
a priori and a posteriori

based on input and output space

THE CAESAR APPLICABILITY DOMAIN



THE CAESAR APPLICABILITY DOMAIN



Pre-conference OpenTox workshop of the AXLR8 2010 meeting

May 30th, 2010 : Potsdam, Germany

OpenTox