Presented by Giuseppina Gini

DEI, Politecnico di Milano

The CAESAR mutagenicity model: a multistep model

Mutagenicity

Predict the results of the AMES test

- Mutagenicity was introduced as a support for carcinogenicity studies
- Mutagenicity tests, when introduced, were criticized by regulators
- Today regulators request them
- … more info we have, the better

Mutagenicity in REACH

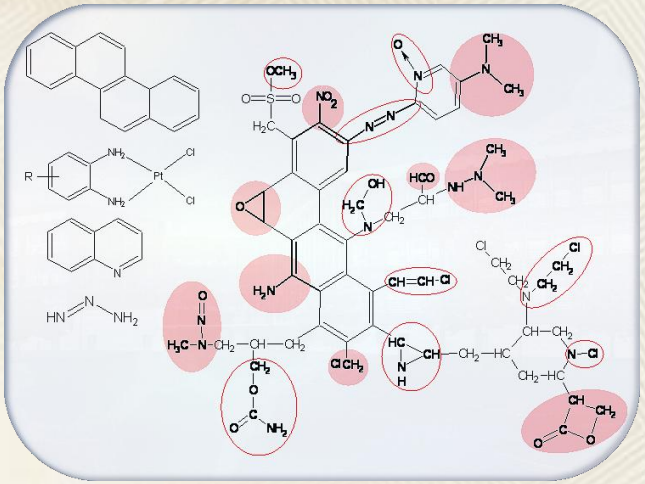
- Mutagenicity (and carcinogenicity) are requested
- REACH wants to minimize mutagens / carcinogens and the exposure to them
- It may be useful to have models to identify mutagens/carcinogens (YES/NO models), or to take into account their potency (dose models)

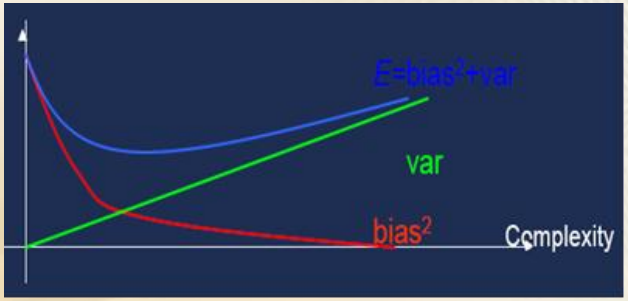
Mutagenicity classification models

QSAR models of noncongeneric compounds to predict mutagenicity can use two approaches:

1) *structural alerts*

2) statistics





Data for the CAESAR model

1) The Kazius-Bursi Mutagenicity Dataset, originally containing *4337* chemical compounds, supplied by R. Bursi

2) Data are categorical

3) Following quality checks (IRFMN and CSL) the database has been pruned and modified to *4225* compounds: *2358* classified as mutagens and *1867* classified as non-mutagens by Ames test

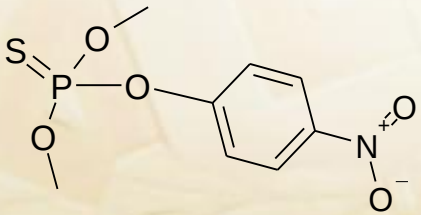
4) For validation, the dataset has been divided into training *(80%)* and test *(20%)* sets (UFZ)

Statistical model for Mutagenicity

1) Method: advanced statistical tools available in open source software

2) Descriptors: 4 global descriptors + 23 simple substructural counts (i.e. a SMARTS matching task) calculated with MDL software

2D structures



MDL descriptors

Selected descriptors

23 simple substructural counts

Symbol	Definition		
SsCH3_acnt	Count of all (– CH3) groups in molecule	*StN_acnt*	Count of all (≡ N) groups in molecule
SdCH2_acnt	Count of all (= CH2) groups in molecule	*SdsN_acnt*	Count of all (= N –)groups in molecule
SssCH2_acnt	Count of all (– CH2 –) groups in molecule	*SaaN_acnt*	Count of all (⋯ N ⋯)groups in molecule
SdsCH_acnt	Count of all (= CH –) groups in molecule	*SsssN_acnt*	Count of all (> N –)groups in molecule
SaaCH_acnt	Count of all (CH) groups in molecule	*SdaaN_acnt*	Count of all (⋯ N ⋯) groups in molecule
SssssCH_acnt	Count of all (> CH –) groups in molecule	*SsOH_acnt*	Count of all (– OH) groups in molecule
SdssC_acnt	Count of all (= C <) groups in molecule	*SdO_acnt*	Count of all (= O) groups in molecule
SaasC_acnt	Count of all (⋯ CH ⋯) groups in molecule	*SssO_acnt*	Count of all (– O –) groups in molecule
SaaaC_acnt	Count of all (⋯ CH ⋯) groups in molecule	*SaaO_acnt*	Count of all (⋯ O ⋯) groups in molecule
SsssssC_acnt	Count of all (> C <) groups in molecule	*SHsOH_Acnt*	Count of all [– OH] groups in molecule
SsNH2_acnt	Count of all (– NH2) groups in molecule	*SHother_Acnt*	Count of all [other] groups in molecule
		SHCHnX_Acnt	Count of all Halogen on C with 1 or 2 H atoms

4 global descriptors

MDL code	Definition
MDL187	Smallest atom E-State value in molecule
MDL198	Bonchev-Trinajsti mean information content
MDL226	Calculated value of LogP
MDL230	Number of rings (cyclomatic number) in a molecular graph

Local
and
Global

Modelling method and software

Algorithm:

***C*-SVC** - a *Support Vector Machine* classifier

Software tools:

- **Weka 3.5.7** - the popular open source suite of machine learning software, written in Java and developed at the University of Waikato

- **LibSVM 2.85** - an open source Java library for SVM by Chang and Lin, wrapped inside Weka

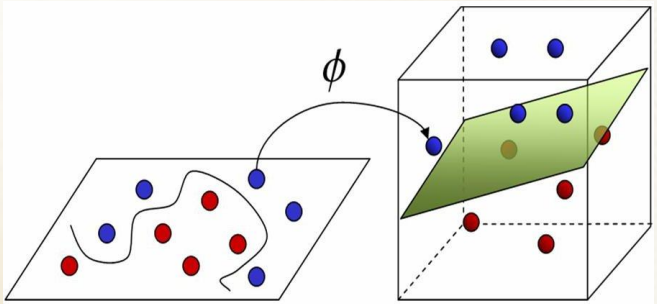
Modelling method

The chosen algorithm is a classification method from the **Support Vector Machines** (**SVM**) family.

SVM can perform a *non linear* classification using *classical linear regression techniques*

1. input space is **mapped** into a higher (maybe infinite) dimensional space with a non linear mapping (the ***kernel* function**)
2. a linear classification in the new space is not linear in the original one

separation may be easier in higher dimensions



Modelling steps

1) **Normalization of the training set:**
   by scaling (dividing) each descriptor column by its maximum absolute value
2) **Choice of the kernel function:**
   Radial Basis Function
3) **Parameterization of the learning algorithm:**
   best parameters assignment (C, γ) = (8, 8) found by an automated search in the space of parameters
4) **Training of the classifier:**
   on the training set

- **Validation of the resulting model:**
  fitting of the training set
  predictive ability for the test set
  robustness by cross-validation on the training set

Confusion matrix

- Given a binary classification problem, there are *4* possible outcomes:

	Prediction	
True class	**T**rue **P**ositives	**F**alse **N**egatives
	False **P**ositives	**T**rue **N**egatives

✓ **TP**: the chemical is *positive* and classified as *positive*
✗ **FN**: the chemical is *positive* but classified as *negative*
✓ **TN**: the chemical is *negative* and classified as *negative*
✗ **FP**: the chemical is *negative* but classified as *positive*

Performance metrics

False Negative rate = $\frac{FN}{TP + FN}$

False Positive rate = $\frac{FP}{FP + TN}$

Sensitivity = 1 – *FN rate* **=** $\frac{TP}{TP + FN}$

Specificity = 1 – *FP rate* **=** $\frac{TN}{FP + TN}$

Accuracy = $\frac{TP + TN}{TP + FN + TN + FP}$

SVM model: *confusion matrices*

Model validation:

TRAINING SET *3380 chemicals*	**Predicted *mutagen***	**Predicted *non-mutagen***
mutagen	***1766***	***122***
non-mutagen	*137*	***1355***
Correct classification rate: ***92.3%***		

TRAINING SET

TEST SET

TEST SET *845 chemicals*	**Predicted *mutagen***	**Predicted *non-mutagen***
mutagen	***407***	***63***
non-mutagen	*79*	***296***
Correct classification rate: ***83.2%***		

SVM model: *statistics*

CAESAR dataset	training set	**test set**	total
accuracy:	*92%*	***83%***	*91%*
sensitivity:	*94%*	***87%***	*92%*
specificity:	*91%*	***79%***	*88%*

(Ames test average interlaboratory reproducibility: 85%)

SVM model: *10-folds cross-validation*

The **10 fold cross-validated** statistics are the average performances of 10 models obtained **without employing the test set**.

The training set is split in 10 parts, and each model is validated on a different 1/10 of the training set, and trained on the remaining 9/10.

CAESAR dataset	10CV on the training set
accuracy:	*82%*
sensitivity:	*84%*
specificity:	*79%*

- *It is a **measure of the robustness** of the original model, since every one of these 10 models is obtained by a slight perturbation (reduction) in the training data, and tested without using the test set.*

next step?

- Problem to address:
  AVOID FALSE NEGATIVES *(FN)*

- How:
  integrated system to avoid *FN* using also ***structural alerts (SA)***

- The tool:
  Toxtree *1.50*, an open source software developed at ECB, Ispra, encoding the Benigni/Bossa rulebase for mutagenicity

Toxtree: *confusion matrix*

The **Benigni/Bossa rulebase** evaluated on the **CAESAR test set**:

TEST SET *845 chemicals*	Predicted *mutagen*	Predicted *non-mutagen*
mutagen	***404***	***66***
non-mutagen	*117*	***258***
Correct classification rate: ***78.3%***		

- ✘ *+ 48% FP*
- ✘ *+ 5% FN*

(compared to SVM model)

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%***

Combining models

- Even though **Toxtree** shows an overall lower accuracy than the **SVM model** in the test set , its good *sensitivity* can still be opportunely integrated into a combined model.

➢*Extracting only the complementary knowledge to the SVM, it is possible to purify the SVM predictions reducing the FN rate*

The idea

- *Structural alerts (SA)* may provide the knowledge lacking from the SVM
  After an analysis of the predictions obtained cross-validating the SVM model, ***2 different subsets of SA*** *were showed for their capacity to remove mutagens incorrectly predicted as safe from the* ***training set****.*

➢ *Will these SA improve also the predictive ability of the* ***test set*** *?*

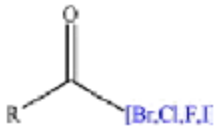
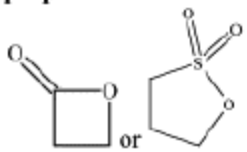
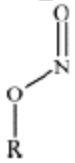
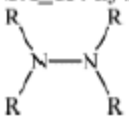
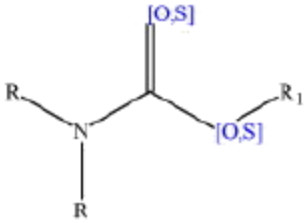
Structural Alerts

The *2 SA* subsets extracted:

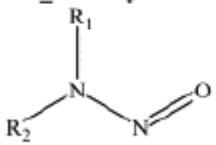
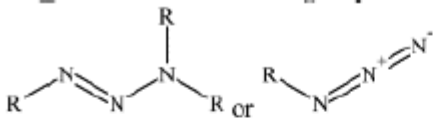
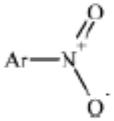
1. **The "*sharp*" rules set (*10 SA*)**
   enhance the accuracy to identify misclassified *FN*

2. **The "*suspicious*" rules set (*5 SA*)**
   remarkable *FN* removal power, but higher misclassification rate

1° rules set: *10 SA*

SA_1: acyl halides 	R= any atom/group, except OH, SH
SA_6 propiolactones or propiosultones 	
SA_9: alkyl nitrite 	R= any alkyl group
SA_13: hydrazine 	R=any atom/group
SA_16: alkyl carbamate and thiocarbamate 	R= Aliphatic carbon or hydrogen R1 = Aliphatic carbon

1° rules set: *10 SA*

SA_18: Polycyclic Aromatic Hydrocarbons	Three or more fused rings, not heteroaromatic
SA_19: Heterocyclic Polycyclic Aromatic	Three or more fused rings, heteroaromatic
SA_21: alkyl and aryl *N*-nitroso groups 	R1= Aliphatic or aromatic carbon, R2= Any atom/group
SA_22: azide and triazene groups 	R=any atom/group
SA_27: nitro-aromatic 	Ar = Any aromatic/heteroaromatic ring Chemicals with ortho-disubstitution, or with an ortho carboxylic acid substituent are excluded. Chemicals with a sulfonic acid group ($-SO_3H$) on the same ring of the nitro group are excluded.

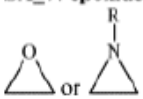
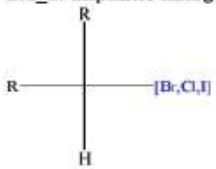
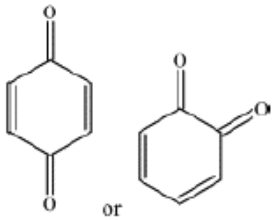
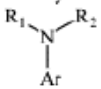
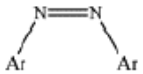
SVM + 1° checkpoint

Confusion matrix on the test set:

TEST SET *845 chemicals*	Predicted *mutagen*	Predicted *non-mutagen*
mutagen	***415***	$(\leftarrow 8)$ ***55***
non-mutagen	*86*	$(\leftarrow 7)$ ***289***
Correct classification rate: ***83.3%***		

- ✓ + 0.1% accuracy
- ✓ ***- 13%*** ***FN !***

2° rules set: *5 SA*

SA_7: epoxides and aziridines 	R= any atom/group
SA_8: aliphatic halogens 	R= any atom/group
SA_12: quinones 	
SA_28 bis: aromatic mono- and dialkylamine 	Ar = Any aromatic/heteroaromatic ring R1= Hydrogen, methyl, ethyl R2 = Methyl, ethyl Chemicals with ortho-disubstitution, or with an ortho carboxylic acid substituent are excluded. Chemicals with a sulfonic acid group ($-SO_3H$) on the same ring of the nitro group are excluded.
SA_29: aromatic diazo 	Ar = Any aromatic/heteroaromatic ring Chemicals with a sulfonic acid group ($-SO_3H$) on both rings linked to the diazo group are excluded

SVM + 1° & 2° checkpoints

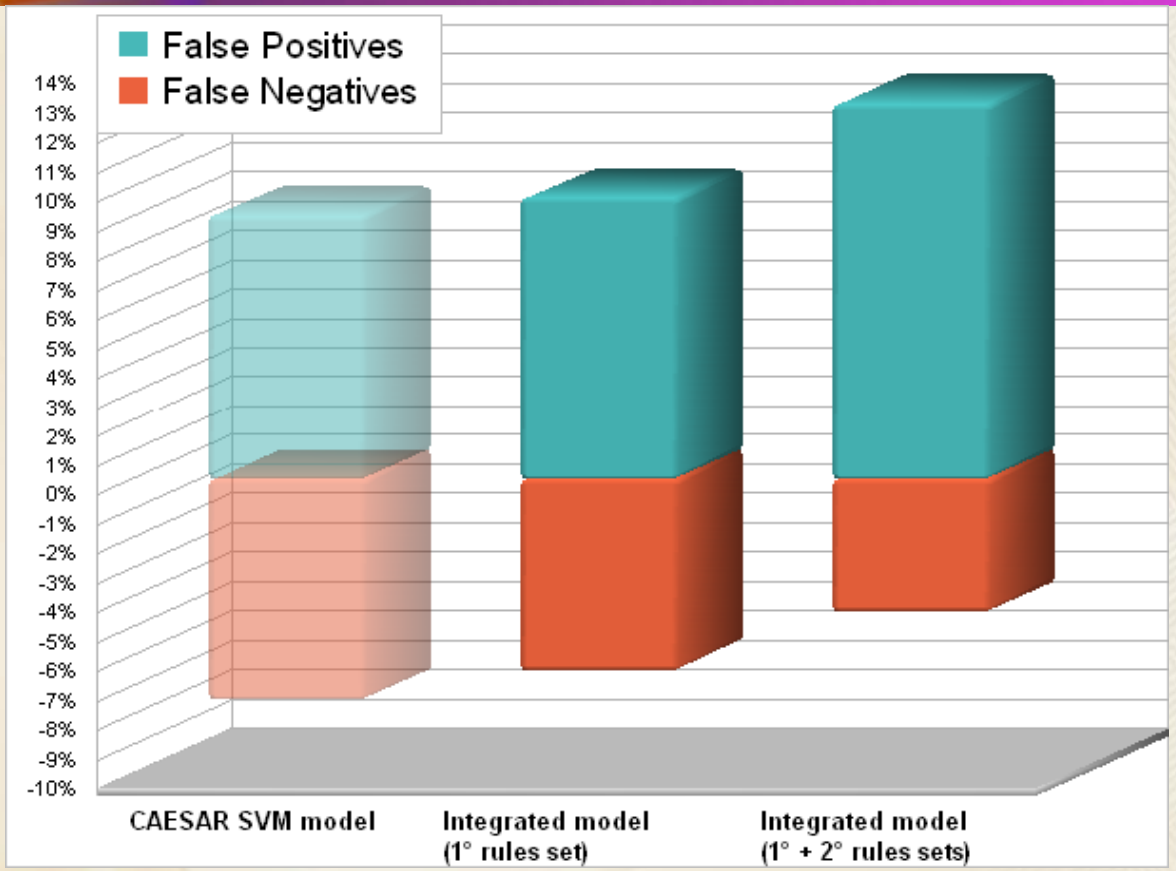
Confusion matrix for the test set:

TEST SET *845 chemicals*	Predicted *mutagen*	Predicted *non-mutagen*
mutagen	***427***	(←20) ***43***
non-mutagen	*108*	(←29) ***267***
Correct classification rate: ***82.1%***		

✘ - 1.1% accuracy

✓ ***- 32%*** ***FN !***

Error shifting analysis

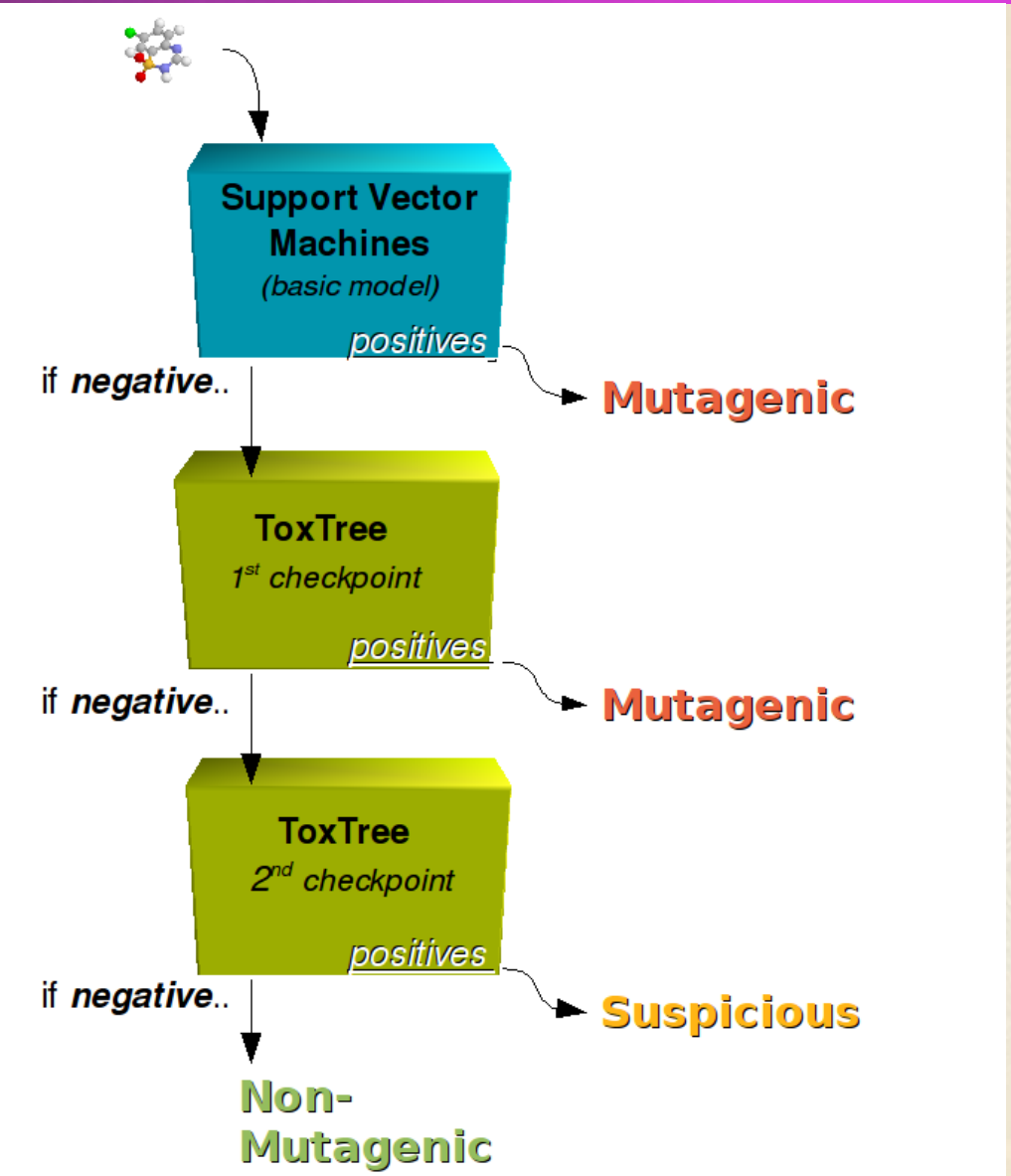


Accuracy:	good	✓ **maximized**	slight decrease
FN:	balanced	✓ reduced	✓ **1/3 off!**
Overall:	ok	**MAX PERFORMANCE**	**HIGH PRUDENCE**

Architecture of the integrated system

3 steps in cascade:

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



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

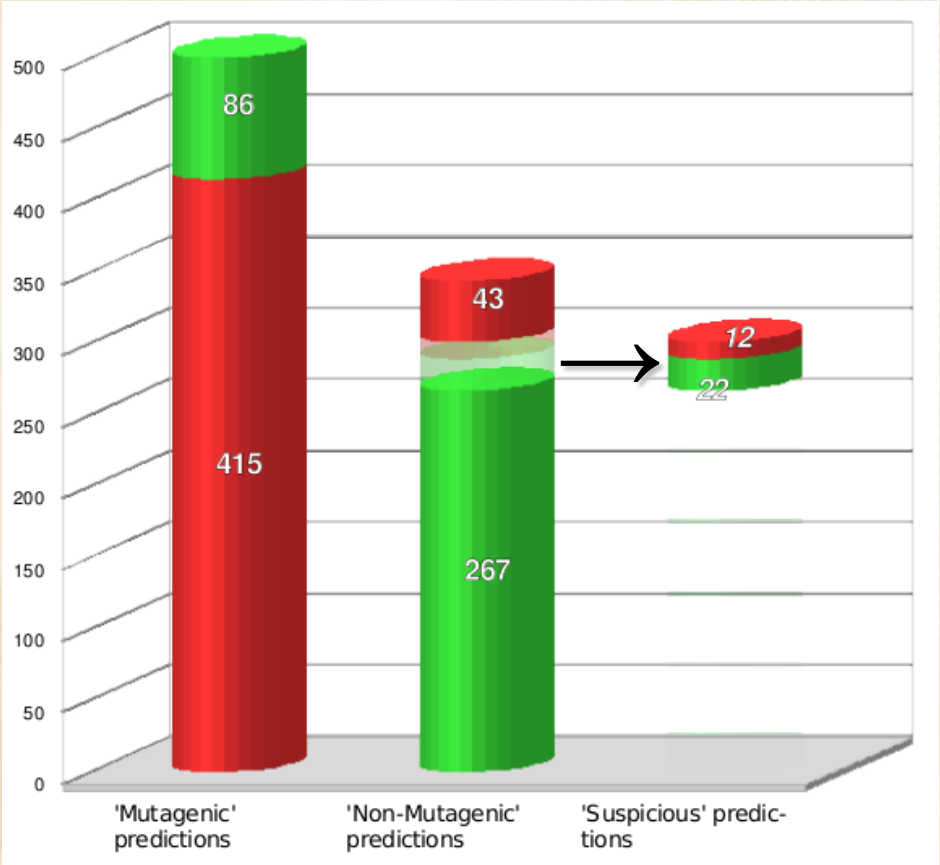
✓ *Accuracy* close to the reliability of the experimental test (85%)

PRUDENT CHOICE

✓ *Sensitivity* boosted over 90%

Integrated model overlook

- The 2 checkpoint *extracts the suspicious compounds* from the presumed safe ones with an impressive *specificity*, if related to the very low number of real mutagens still present.



…concluding

1. 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;
2. The experimental error is a major bottleneck;
3. This gives evidence that very good performance is possible with machine learning software from public domain;
4. Selected structural alerts can discover FN (but can moderatly increase FP as well);
5. The CAESAR model has been checked against commercial systems (Multicase, Derek); it gave always not worse results.

References

Data:

J. Kazius, R. McGuire, and R. Bursi. *Derivation and Validation of Toxicophores for Mutagenicity Prediction*. J. Med. Chem. 2005, *48: 312-320*.

MDL: http://www.mdl.com/
Toxtree: http://ecb.jrc.ec.europa.eu/qsar/
SVM: http://www.csie. ntu.edu.tw/~cjlin/libsvm
WEKA: http://www.cs.waikato.ac.nz/ml/weka