WORKSHOP ON QSAR MODELS FOR REACH

Mario Negri Institute, Milan, Italy - March 10-11, 2009

SIXTH FRAMEWORK PROGRAMME

Frank Lemke

KnowledgeMiner Software

Enhancing CAESAR Models

http://www.caesar-project.eu/

High Value Properties of CAESAR Models

- High quality of data
- Out-of-sample validation of models
- Reproducibility
- Transparency
- Application domain
- Ready- and Easy-to-use

Visions for CAESAR Models

Implementation of
Hybrid models from existing models
Prediction interval and uncertainty
Optimisation according to FN and FP costs

Hybrid QSAR Models: Motivation

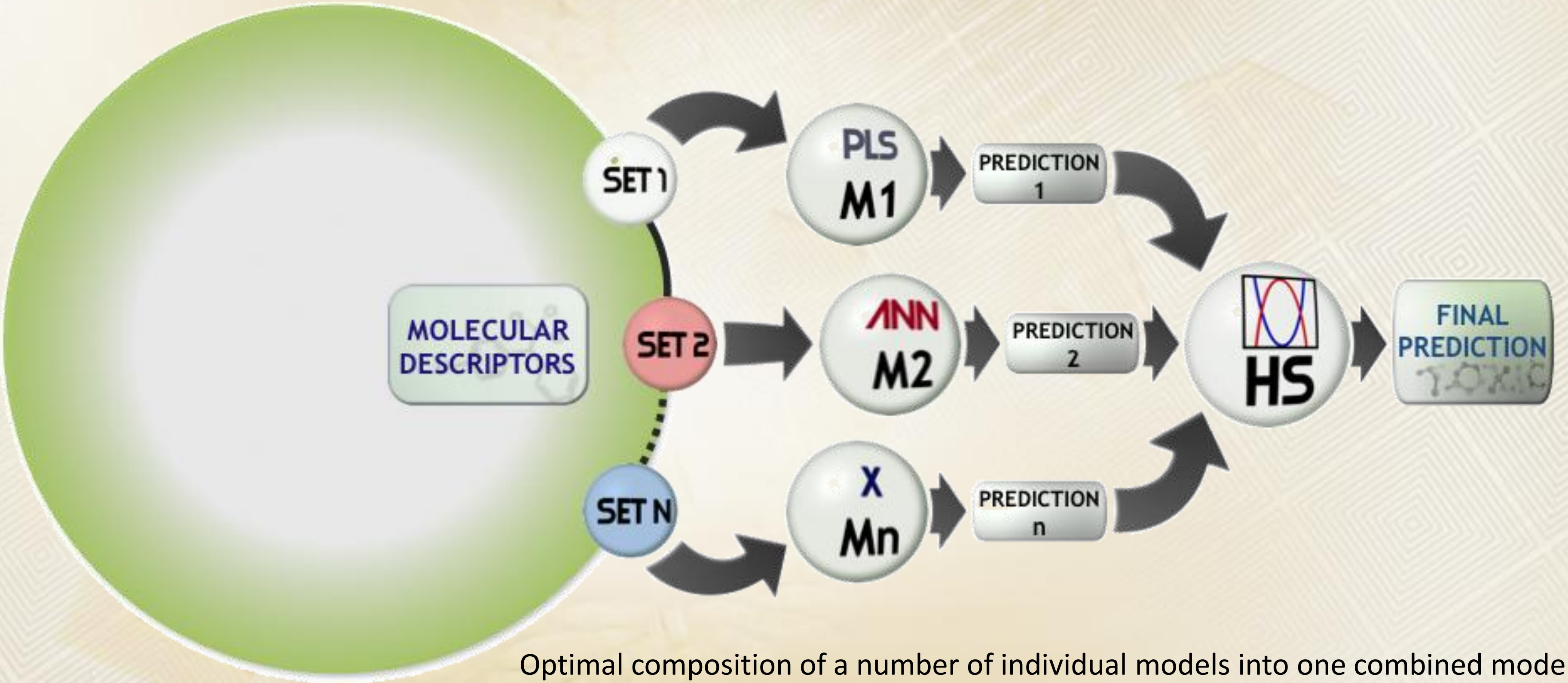
- On noisy, uncertain data sets a number of models can be built, which are comparable with respect to prediction accuracy. (in CAESAR: ≈ 25 / endpoint)
- Commonly, a model is a simplified reflection of the complex reality, only. It describes a specific part of the object's behavior.

So why only use one model?

Hybrid QSAR Models: Motivation

- A *more complete reflection of the reality* can be obtained when combining several models:
  - Different modeling approaches
  - Different input data
  - Different parameters
- *Increased prediction accuracy* of up to about 10% is possible.

Hybrid QSAR Models: Principle

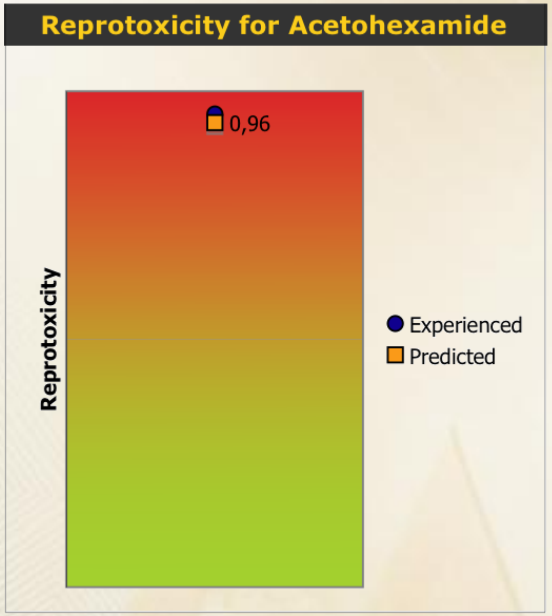


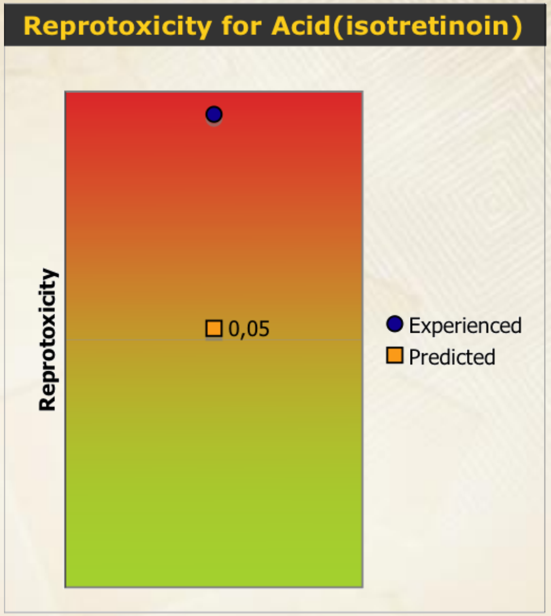
Optimal composition of a number of individual models into one combined model

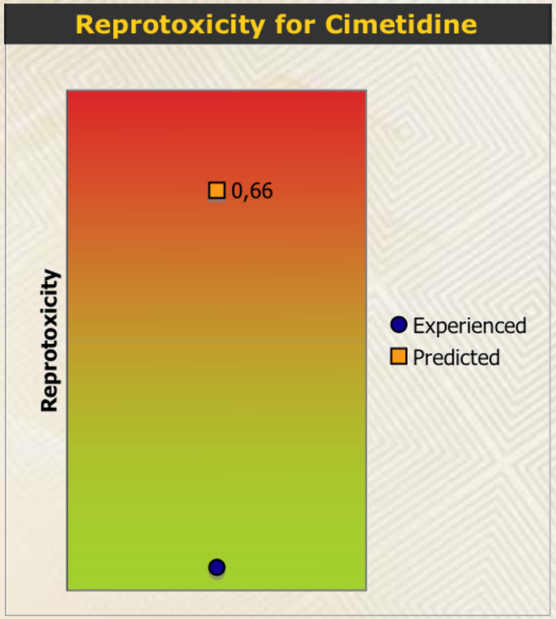
Visions for CAESAR Models

Implementation of
Hybrid models from existing models
Prediction interval and uncertainty
Optimisation according to FN and FP costs

Prediction: Commonly







Regression models

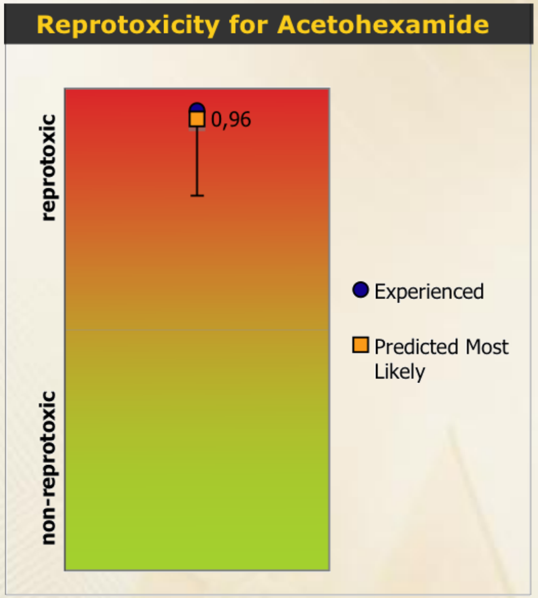
Predicted Class Value	Predicted Class Value	Predicted Class Value
reprotoxic	reprotoxic	reprotoxic

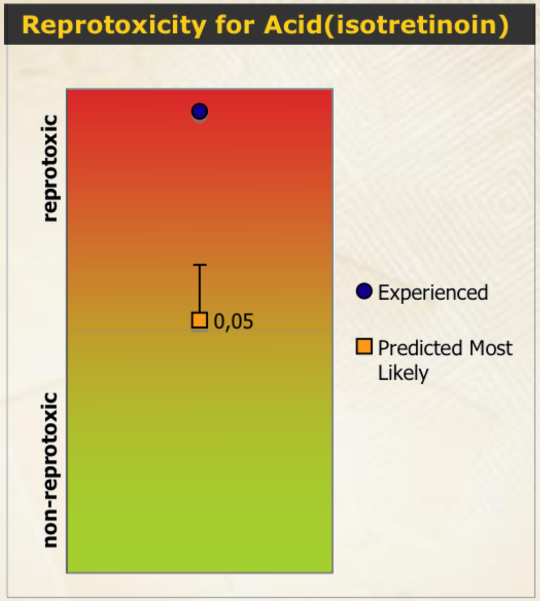
Class. models

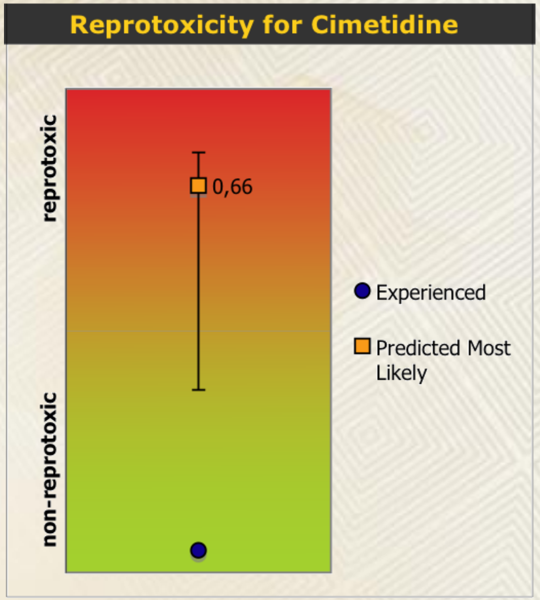
WORKSHOP ON QSAR MODELS FOR REACH
Mario Negri Institute, Milan, Italy - March 10-11, 2009

*) the values shown do not necessarily correspond to the final model for developmental toxicity.

Prediction Interval







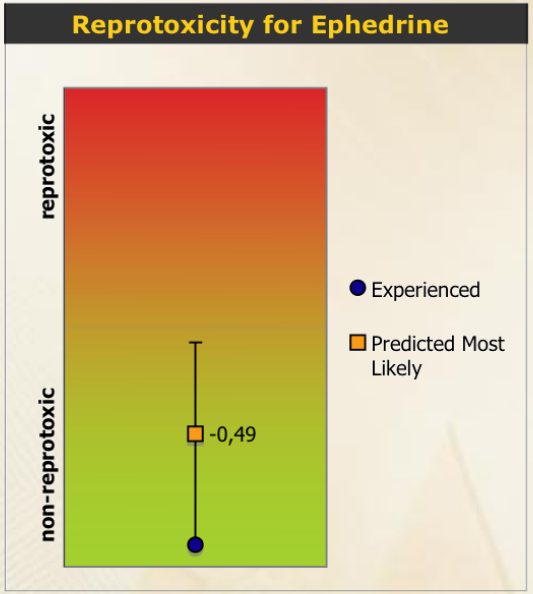
Classification and regression models

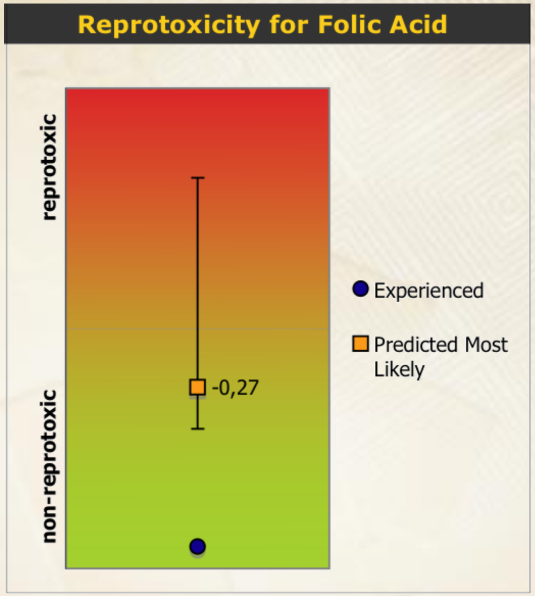
Per compound prediction uncertainty available for decision-making

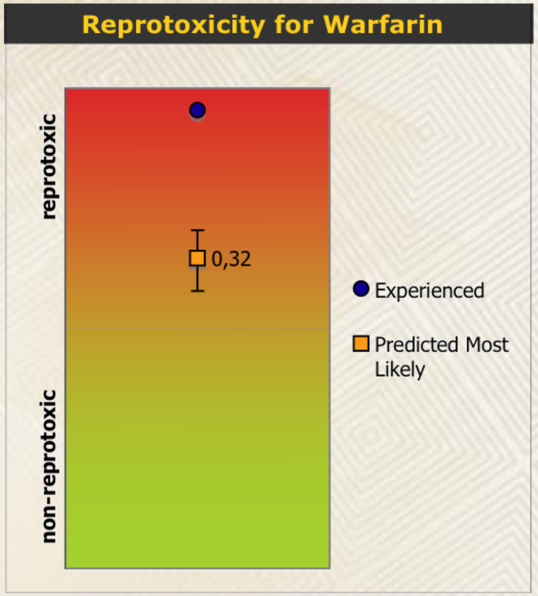
Freedom of choice

*) the values shown do not necessarily correspond to the final model for developmental toxicity.

Prediction Interval







Classification and regression models

Uncertainty is huge for experimental data, already.
We cannot expect QSAR models built on this data *being less uncertain* than the original information is.

*) the values shown do not necessarily correspond to the final model for developmental toxicity.

Visions for CAESAR Models

Implementation of
Hybrid models from existing models
Prediction interval and uncertainty
Optimisation according to FN and FP costs

Classification: Current Praxis

Given: Data set of experimental values about carcinogenicity (the „Truth“)
100 compounds are carcinogenic (Positive)
100 compounds are not carcinogenic (Negative)

Balanced classifier		
Confusion Matrix	**Truth: Positive**	**Truth: Negative**
Predicted: Positive	74	25
Predicted: Negative	26	75

→

Accuracy	74,5 %
Sensitivity	74 %
Specificity	75 %

Balanced sensitivity and specificity

Cost-sensitive Models

What if there are *different costs* for misclassified compounds (FP/FN) and/or *different benefits* for correctly classified compounds (TP/TN)? → **Real-world scenario**

High relative False Negative costs		
Cost-Benefit Matrix	**Truth: Positive**	**Truth: Negative**
Predicted: Positive	0	1
Predicted: Negative	9	-3

&

Balanced classifier		
Confusion Matrix	**Truth: Positive**	**Truth: Negative**
Predicted: Positive	74	25
Predicted: Negative	26	75

Cost/compound	0,09	**Relative cost**	3,2%

Cost-sensitive Models

Using a *cost-sensitive approach* to find the **optimal classifier** for cost-benefit matrix:

False Negative Optimisation

High relative False Negative costs		
Cost-Benefit Matrix	**Truth: Positive**	**Truth: Negative**
Predicted: Positive	0	1
Predicted: Negative	9	-3

&

False Negative optimised classifier		
Confusion Matrix	**Truth: Positive**	**Truth: Negative**
Predicted: Positive	89	42
Predicted: Negative	11	58

Benefit/compound	0,22	**Relative benefit**	11,8%

Cost-sensitive Models

How does the balanced classifier perform in the **inverse situation**?

False Positive Optimisation

High relative False Positive costs		
Cost-Benefit Matrix	**Truth: Positive**	**Truth: Negative**
Predicted: Positive	-3	9
Predicted: Negative	1	0

&

Balanced classifier		
Confusion Matrix	**Truth: Positive**	**Truth: Negative**
Predicted: Positive	74	25
Predicted: Negative	26	75

Cost/compound	0,14	**Relative cost**	5,6%

Cost-sensitive Models

Using a *cost-sensitive approach* to find the **optimal classifier** for cost-benefit matrix:
False Positive Optimisation

High relative False Positive costs		
Cost-Benefit Matrix	**Truth: Positive**	**Truth: Negative**
Predicted: Positive	-3	9
Predicted: Negative	1	0

&

False Positive optimised classifier		
Confusion Matrix	**Truth: Positive**	**Truth: Negative**
Predicted: Positive	70	21
Predicted: Negative	30	79

Benefit/compound	0,02	**Relative benefit**	1,8%

Cost-sensitive Models

One Example QSAR Model		
Summary Benefits	**Balanced Classifier**	**Optimised Classifier**
FN Minimisation	-3,2 %	11,8 %
FP Minimisation	-5,6 %	1,8 %
Balanced	24,1 %	24,1 %

Values in one column are not comparable
since based on different cost-benefit matrices.

Cost-sensitive Models

- Apparently, there is an **optimal classifier** for given cost-benefit matrix and model; balanced classifier optimal only for balanced costs/benefits
- **Objective** *accuracy- and cost-driven optimisation* of FP or FN
- **Live** optimisation according to given costs by the user at runtime

Visions: Summary

Hybrid Models	• More complete reflection of the complexity of the problem • Increasing prediction accuracy
Prediction Interval	• Per-compound prediction uncertainty available • Freedom-of-choice for decision making • Individual selection of prediction value based on purpose
Cost-sensitive Models	• Live, objective accuracy- and cost-driven optimisation of a model for minimising FN or FP • Finally, the purpose of a QSAR prediction, the evaluation task it is used for, is driving the model result • Dealing with uncertainty of results