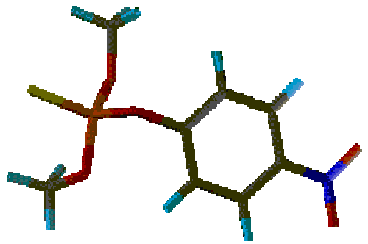
Ensemble Methods for Data Mining and Knowledge Extraction in Scientific Data Bases

Giuseppina Gini

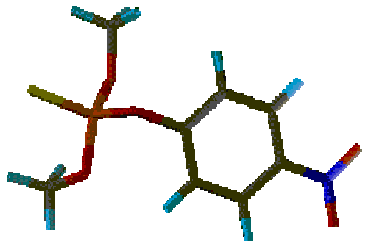
Politecnico di Milano, DEI



index

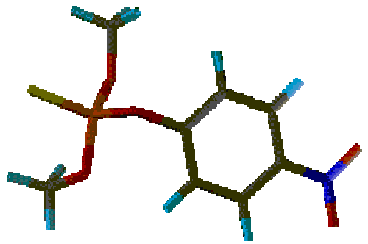
- Data mining and Scientific Data sets – problems
- Case study: chemometrics, QSAR
- Data production and analysis
- Model construction: regression, classification, hybrid
- experiments
  - Carcinogenicity (graphs and NN)
  - Pesticide evaluation
  - acquatic toxicity and MOA
- Validation and interpretability

conclusion



What is KDD

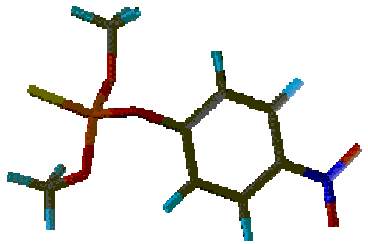
- *automated discovery of patterns and the development of predictive and explanatory models*
- *It is based on Data mining* selection and processing of data for the identification of novel, accurate, and useful patterns, and the modeling of real-world phenomena.



KDD => MODELS

- a. Theory-driven approach
- For complex ill-defined systems we have insufficient a priori knowledge about the relevant theory, uncertain a priori information with regard to the selection of the model structure as well as insufficient knowledge about interference factors
- b. Data-driven approach
- usually we have no a priori knowledge about the structure of the mathematical model.

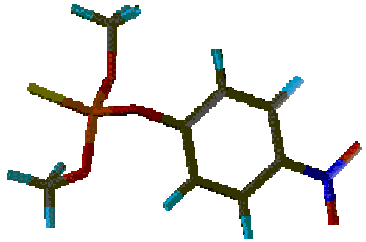
Problems in scientific prediction



- a large collection of data (more variables than cases) has problems dimensionality problem;
- Most of the reported classifiers and regression models are so bad in prediction power that cannot be used for real problems/ most of the systems are intended for DSS
- So far no relevant knowledge extracted

COST 282

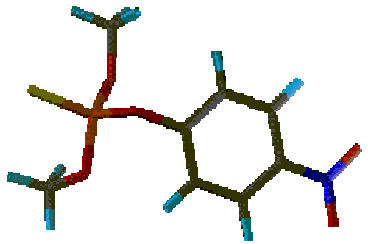
KNOWLEST

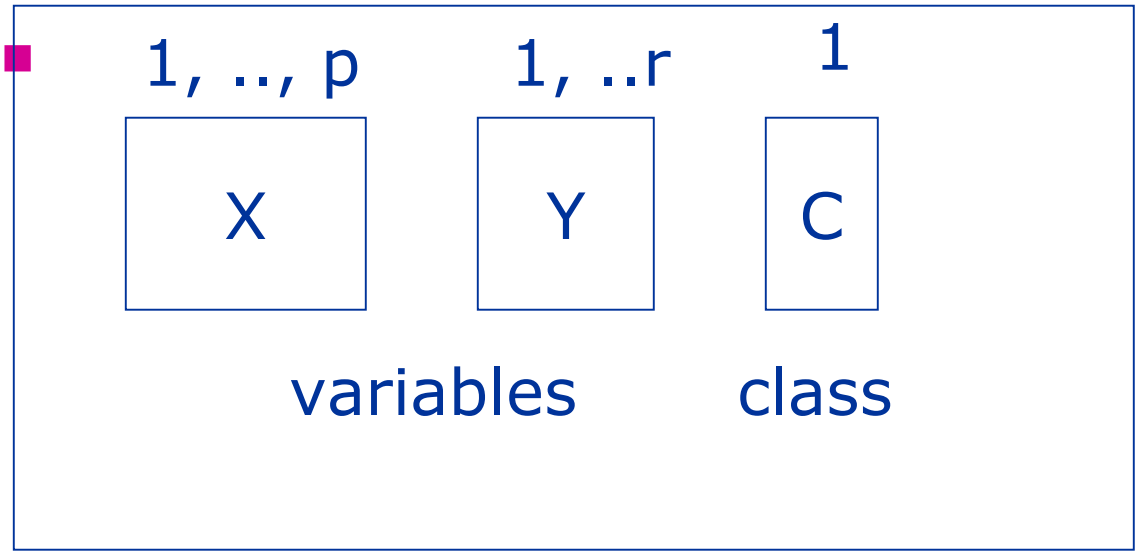


Knowlege Exploration in Science and Technology

- "… extracting previously unknown, non-trivial, and potentially useful knowledge from structurally complex, high-volume, distributed, and fast-changing scientific and R&D databases within the context of global computing and data infrastructures such as the GRID".
- *incorporating general background knowledge and user experience into the knowledge discovery process ….*
- *Non text, non relational data (molecular data mining…)*

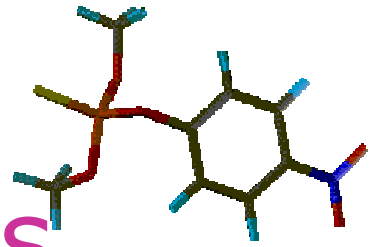
Chemometrics - *the information aspects of chemistry*

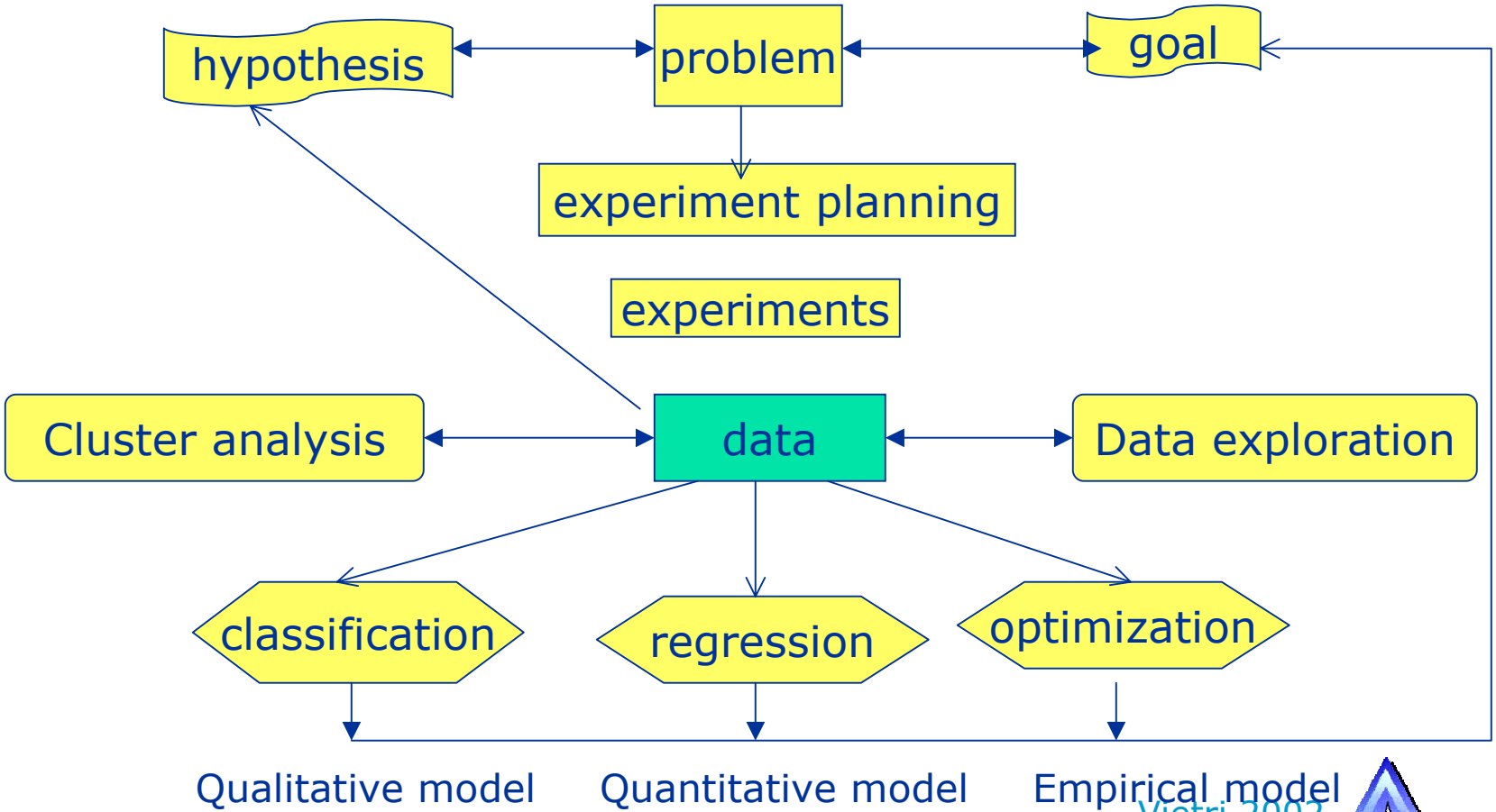


- Extracting information from chemical data **Data analysis**
- Making chemical data have information **Experimental design**
- Investigating complicated relationships **Modelling**
- 

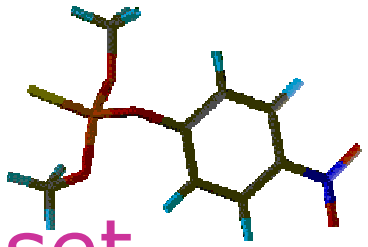
  - multivariate data obtained from experiments

Chemometrics strategies



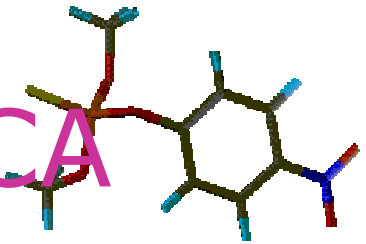


Methodology and Statistical experimental design => data set



- In 1925 Fisher started the development of methods of statistical experimental design [DoE]
- Generate a set of examples
- Reduce attribute dimensionality
- Reduce attribute value ranges
- Transform data
  - simplify the response function by linearizing;
  - stabilize the variance;
  - make the distribution more normal
- *A GOOD METHODOLOGY IS FOLLOWED BY THE PRODUCERS OF DATA?*

Feature selection and PCA (Pearson 1901, Hotelling 1933)



Select a minimum set of features such that the probability distribution of different classes given the values for those features is as close as possible to the original distribution given the values of all features

- Why: evaluate variable correlation, relevance, for data reduction

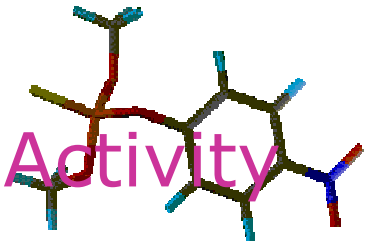
Build matrix **A** with eigenvectors as rows

$=>\mathbf{y}= \mathbf{A}(\mathbf{x} - \mu_x)$

we choose the first k eigenvectors (k?)

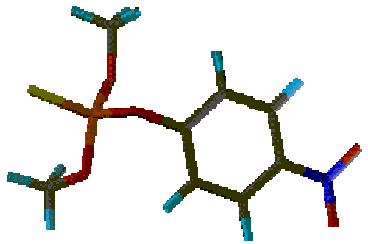
- $\mathbf{y}= \mathbf{A}_k(\mathbf{x} - \mu_x)$

QSAR ((Quantitative Structure Activity Relationships)

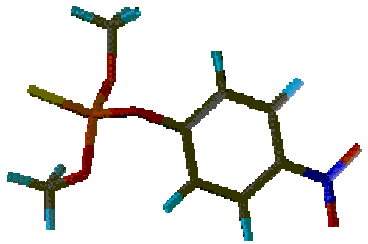


- Since 40 years is the way to assess the value of drugs
- Since 10 years
- => a way to assess toxicity? As a way to obtain new knowledge

QSARs as regression or classification



- For drug activity and toxicity, most of the QSAR models are regressions, referring to the dose giving the toxic effect in 50% of the animals
- Classification systems for QSAR or SAR refer to regulatory bodies (NTP, EU plans to use predictive methods for priority setting and for risk assessment)

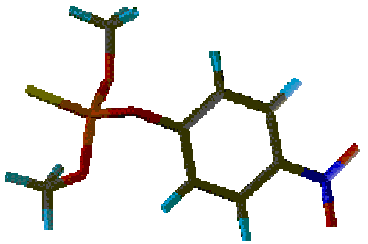


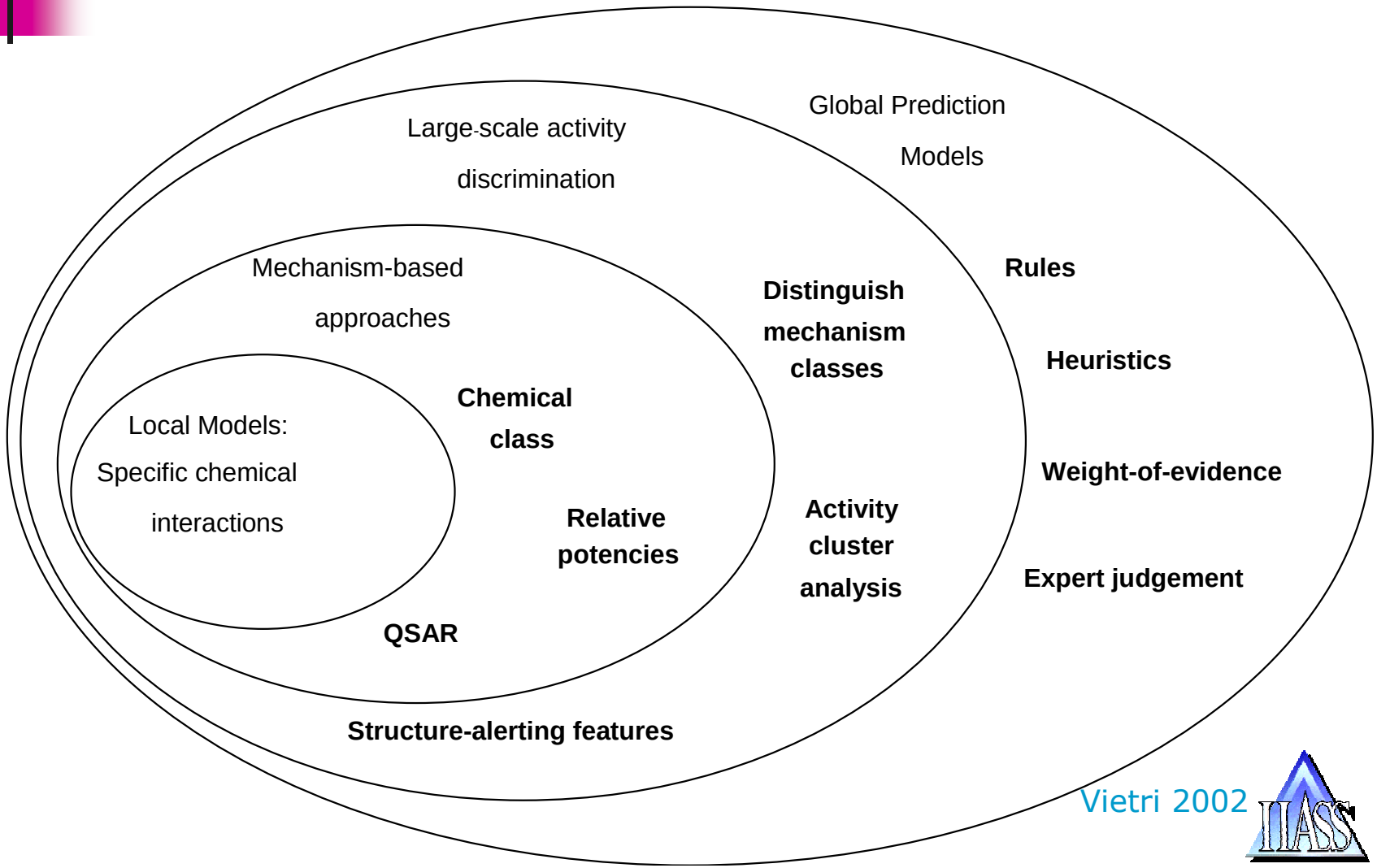
QSAR "postulates"

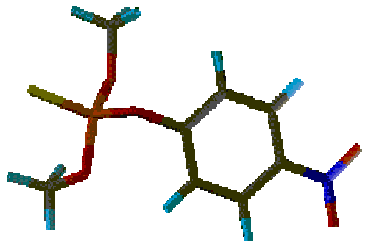
the molecular structure is responsible of all the activities shown

- Similar compounds have similar biological and chemico-physical properties (Meyer 1899)
- Hansch (1963) postulate:
- *biological system* + *compound* gives answer = f_1(Lipolificity) + f_2(Electronics) +f_3(Steric) +f_4(Molecular-prop)
- **Congenericity postulate**: QSAR is applicable only to similar compounds

Locality of the model



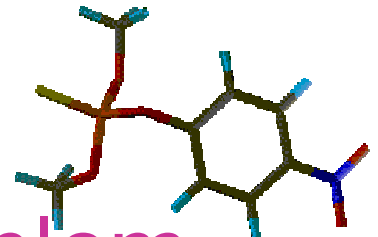




Is locality a problem?

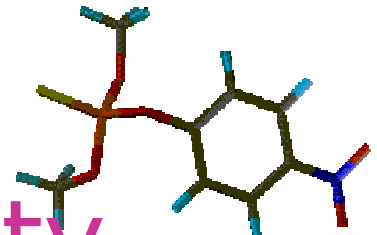
- *NUMBER PROBLEM: 20 millions registered CAS against 2 thousand studied*

- ONTOLOGY PROBLEM: how we subdivide the compounds to have homogeneous? What is toxicology?
- REPRESENTATION PROBLEM
  - (quantum similarity, spectral, descriptors, …)



The Predictive Toxicology Problem -

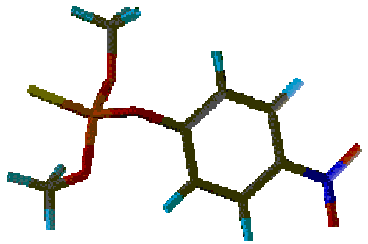
- to develop predictive models, in order to obtain improved applicability of these systems
- to get knowledge from data to speed up scientific discovery
- Needs:
  - large and peer reviewed data sets
  - Ideas how to combine toxicity for different organisms
- Target: To work in silico, not in vivo
- *Example: challenge (IJCAI 1997)*



The virtual lab for toxicity

- All chemistry is computer chemistry (descriptors…)
- All chemistry is a model => the model is good if it gives an explanation to the experimental results
- A virtual lab is a set of tools to compute descriptors, input and output scaling, molecular properties, toxicity …..

Where are data to mine

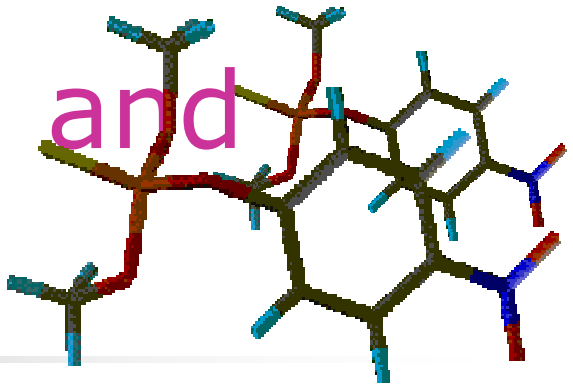


- Standard data set as in UCI have

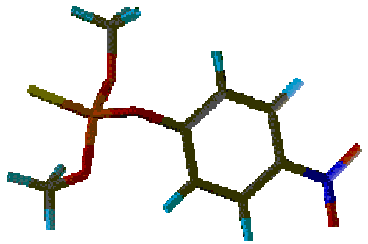
shorthcomings:

- Not apt to extract knowlege
- *Good properties*:
- Number , comparison…
- WHY NO TOXICOLOGY DATA THERE?

Data sets developed and studied



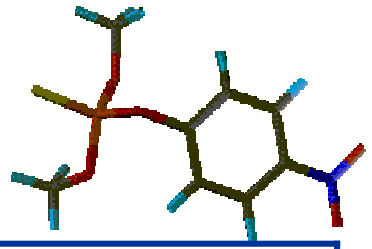
- Carcinogenicity data set – to predict TD50
- EPA data set – to predict lethal concentration for 50% of the test animals (LC_{50}), towards the fish fathead minnow (*pimephales promelas*).
- Pesticides data set – to predict toxicity LC_{50} for different species



experiments

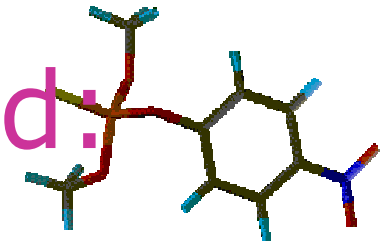
	REGRESSION	CLASSIFICATION
	PLS, statistics	CART
	ANN, FNN	SIMCA, statistics
	NIKE	NIKE
	AFP	AFP
	WEKA	WEKA
Feature sel	*PCA*	*PCA*
	GA - wrapper	*GA - wrapper*
ensemble	Hybrid and fuzzy	Hybrid and fuzzy

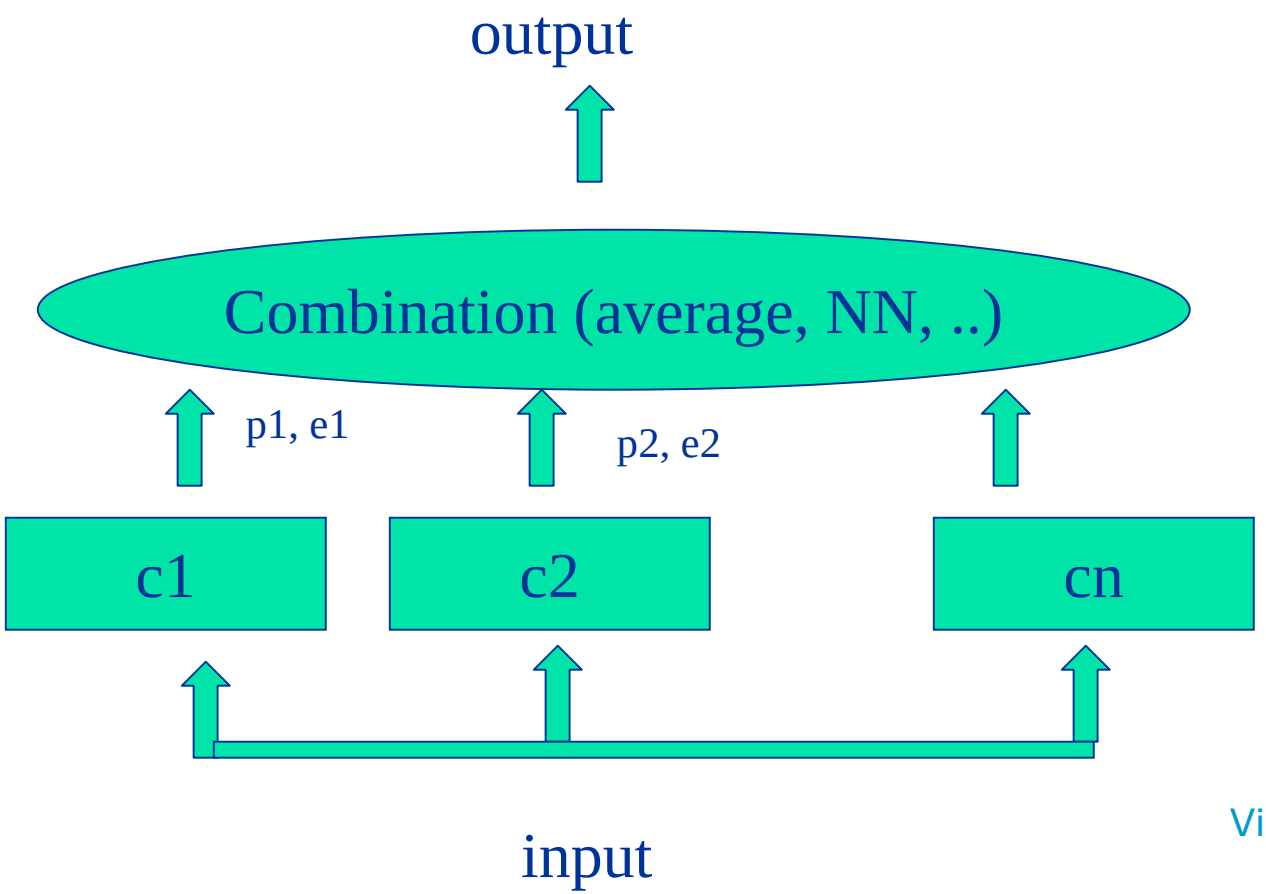
Data analysed/method



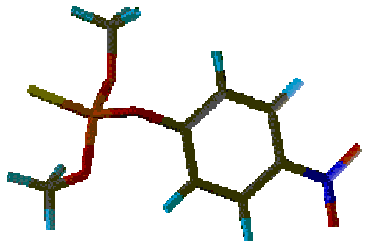
method	aromatic	pesticides	EPA fish
ANN, FNN			
ensemble			
graphs			
trees			
stat			
GA			

The combination method: ensembles, mixture, …



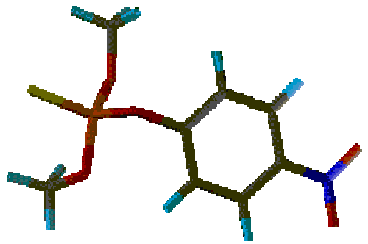


The origin of combining models



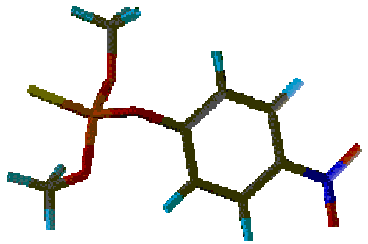
a simple averaging of the predictors generates a very good composite model -

- => generate highly correct classifiers that disagree as much as possible (with dissimilar learning parameters, different classifier architectures, various initial neural-network weight settings, or separate partitions of the training set.



Mixture of experts

- train individual networks on a subtask, and then combine these predictions with a ``gating'' function that depends on the input. The key idea is that a decomposition of the problem into specific subtasks might lead to more efficient representations and training.
- gating function can be a network that *learns* how to allocate examples to the experts.

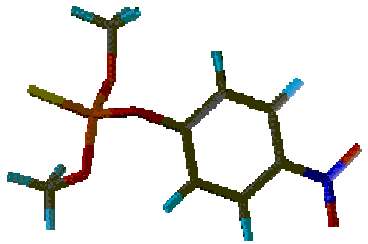


Connectionism /symbolic

- **translating the domain knowledge into a neural network,** then modifying the weights of this resulting network.

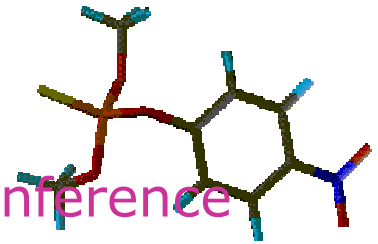
.

- **Rule extraction from NN** Gallant [1988]
- Architecture-analysis based
- Causal index (for a net with h hidden neurons)
  - CI = Sum wkj*wji all the patways from input i to j and from j to output k
- function-analysis based (learning)



Neuro/fuzzy integration

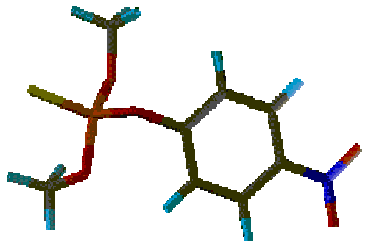
- Any rule based fuzzy system may be approximated by a neural net
- Any neural net may be approximated by a fuzzy system
  - Mandami or Sugeno type
- Neuro-fuzzy hybridization



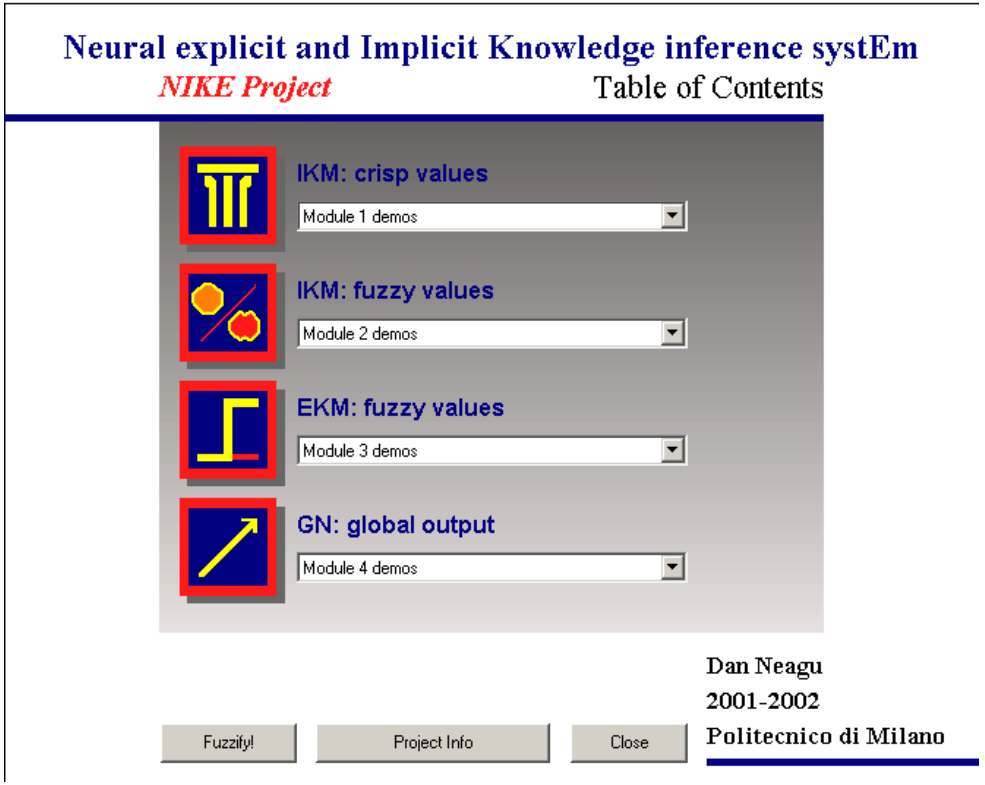
•NIKE (Neural explicit&Implicit Knowledge inference system)

- **NIKE** is a *hybrid intelligent system* shell based on *modular neural networks*, supporting different strategies to build assemblies of *neural, neuro-fuzzy, and fuzzy inference systems* implemented in Matlab. It combines:
- *implicit knowledge* (**IKM**), represented by neural/ neuro-fuzzy networks, created and adapted by a *learning algorithm*.
- *explicit knowledge* (**EKM**), a collection of connectionist structures, which are computationally identical to the I/O relations set, and are created by mapping existing fuzzy rules into hybrid neural networks.

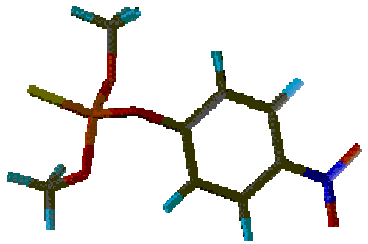
Major functions of NIKE

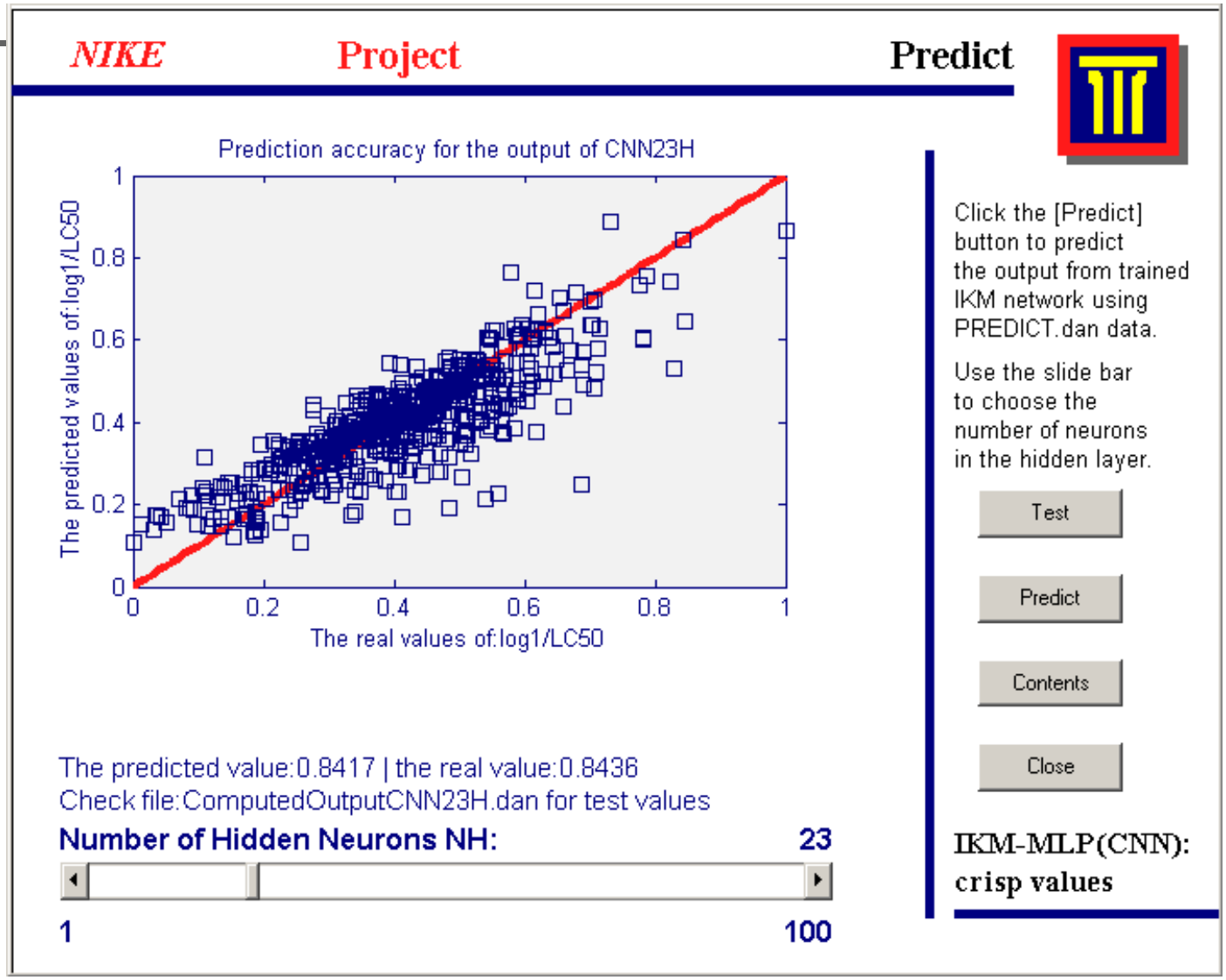


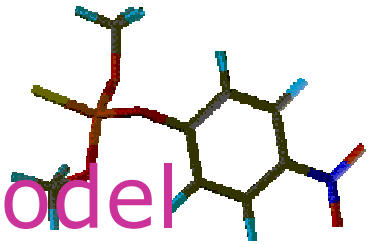
- Defining, training, using ANNs.
- Knowledge refinement from neural networks.
- Using connectionist fuzzy systems.
- Integrating neural nets with fuzzy inference systems.
- QSAR representation as fuzzy inference systems.
- Knowledge modules integration (modular nets)
- Data mining



IKM-CNN representation

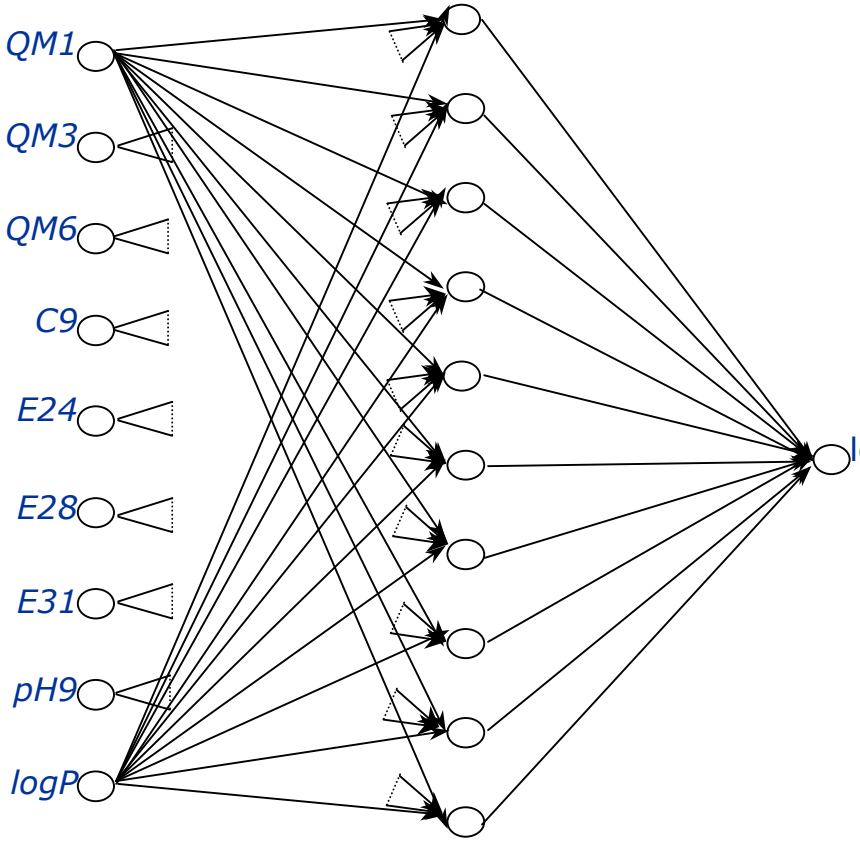






Example: MLP (IKM-CNN) model for toxicity of organic compounds

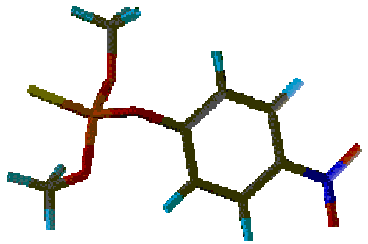
Acute toxicity 96 hours (LC_{50}), for fathead minnow (*Pimephales promelas*): 568 compounds.



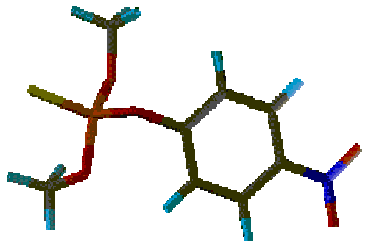
Descriptors	Code
Total Energy (kcal/mol):	*QM1*
Heat of Formation (kcal/mol):	*QM3*
LUMO (eV):	*QM6*
Relative number of N atoms:	*C9*
Relative number of single bonds:	*C24*
Molecular weight:	*C35*
Kier&Hall index (order 0):	*T6*
Average Information content (order 1):	*T22*
Moment of inertia B:	*G2*
Molecular volume:	*G10*
Molecular surface area:	*G12*
TMSA Total molecular surface area:	*E13*
FPSA-2 Fractional PPSA (PPSA-2/TMSA):	*E24*
PPSA-3 Atomic charge weighted PPSA:	*E28*
FPSA-3 Fractional PPSA (PPSA-3/TMSA):	*E31*
logD:	*pH9*
logP:	*logP*

Vietri 2002

Implicit Knowledge in FNN



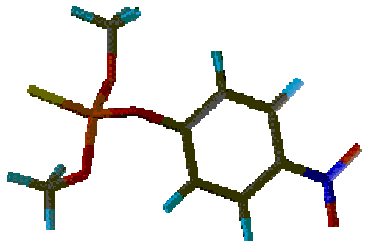
- The IKM-FNN: the input layer performing the membership degrees of the variables, a fully connected three-layered FNN2, and a defuzzification layer.
- A linguistic variable X_i is described by m_i fuzzy sets, A_{ij}, having the degrees of membership performed by the functions $\mu_{ij}(x_i)$, j=output number, i=input number
- as the output y_{defuz}).



Linguistic variables

- A numerical variable takes numerical values: LUMO=0.5572
- A linguistic variable takes linguistic values: *QM6* is Medium
- A linguistic value is a **fuzzy set**.
- The collection of all the linguistic values is a **term set**:

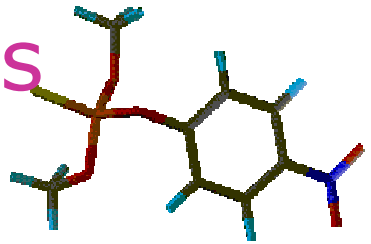
  QM6={Low,Medium,High}



Fuzzy IF-THEN Rules

- Mamdani fuzzy rule:
  - IF D_1 is Low AND D_2 is High THEN Tox is Medium
- zero-order Sugeno fuzzy rule:
  - IF D_1 is Low AND D_2 is High THEN Tox=k
- first order Sugeno fuzzy rule:
  - IF D_1 is Low AND D_2 is High THEN $Tox=0.72xD_1+0.12xD_2-0.11$

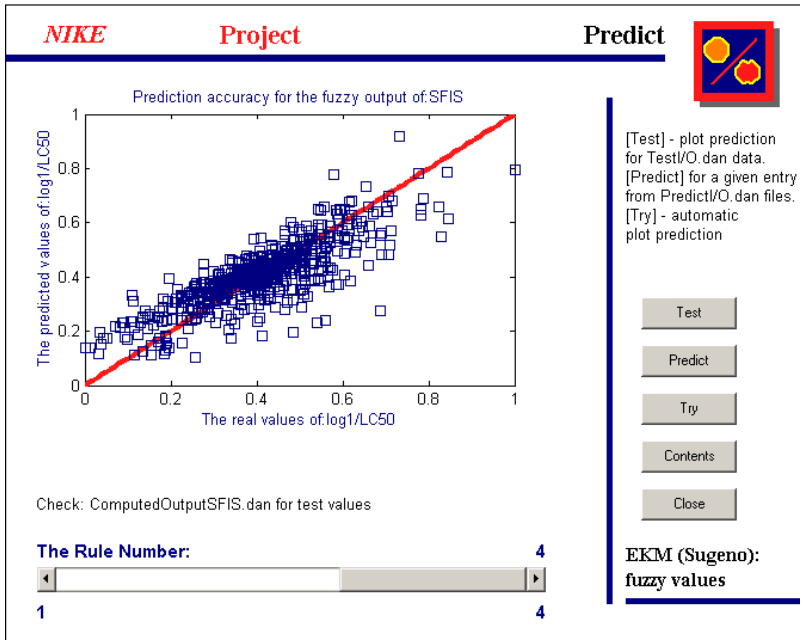
FIS representation for QSARs



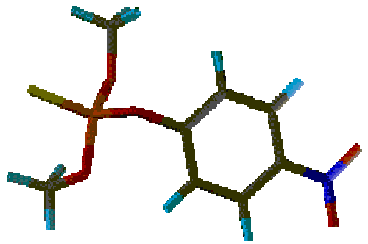
- Mamdani:
  - IF D_1 is Low AND D_2 is High THEN Tox is Medium
- zero-order Sugeno fuzzy rule:
  - IF D_1 is Low AND D_2 is High THEN Tox=k
- first order Sugeno fuzzy rule:
  - IF D_1 is Low AND D_2 is High THEN Tox=0.82+0.17**QM6*–0.79**logP*

Example:

- 1. If (logP is Low) then (log1/LC50 is QSAR2) (1)
- 2. If (logP is Med) then (log1/LC50 is QSAR2) (1)
- 3. If (logP is High) then (log1/LC50 is QSAR2) (1)

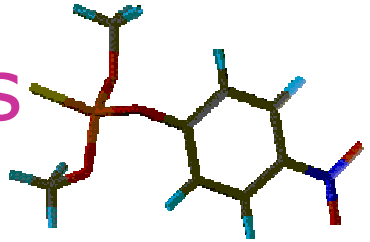


Extracted fuzzy rules

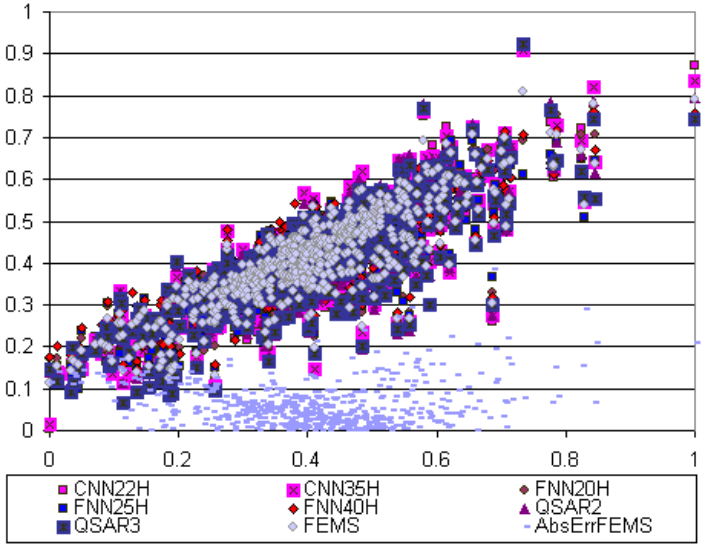


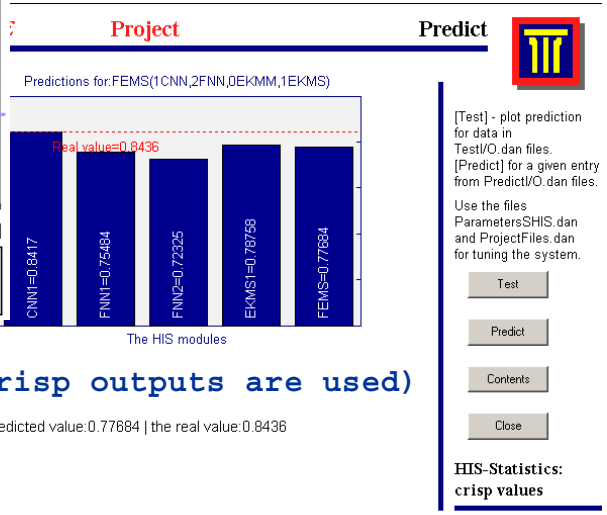
- -> from IKM-FNN using Effect Measure Method (EMM)
- pre-processing to delete the contradictory rules
  - (1) different output predictions than the same input class, and a small trust: `IF RdaFit1 is:`**`Medium`** `THEN class is:VeryLow(47.79%)`
    (2) big differences between the value of the input (the classification) and the output: IF KnnXFi1 is:High THEN class is:Low (78.70%)
  - WHAT IS THE PREDICTIVE POWER OF THE INDUCED FIS?

Statistical mixture of experts



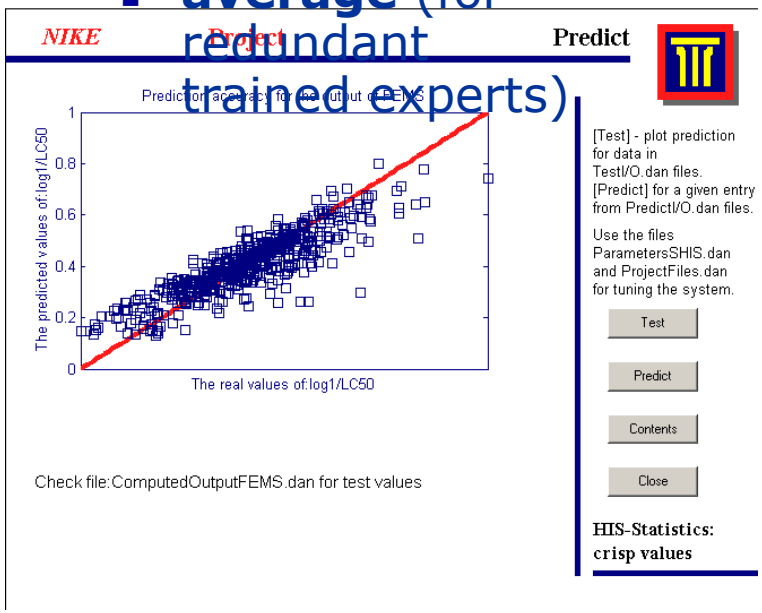
- The method of combining:
  - **max** (for disjunctive trained experts) and
  - **average** (for redundant trained experts)





ProjectFiles.dan (the crisp outputs are used)

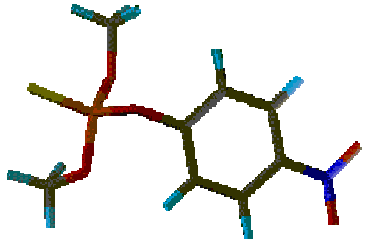
```
NumCNN=1
NumFNN=2
NumEKMMamdani=0
NumEKMSugeno=1
[CNN]
C:\IMAGETOX\DuluthMols\work\IKM\CNN\CNN23H\memvarCNN23Hnet.mat
[FNN]
C:\IMAGETOX\DuluthMols\work\IKM\FNN\FNN15H\memvarFNN15Hnet.mat
C:\IMAGETOX\DuluthMols\work\IKM\FNN\FNN25H\memvarFNN25Hnet.mat
[EKMMamdani]
[EKMSugeno]
C:\IMAGETOX\DuluthMols\work\Data\FuzzyIOS.fis
```

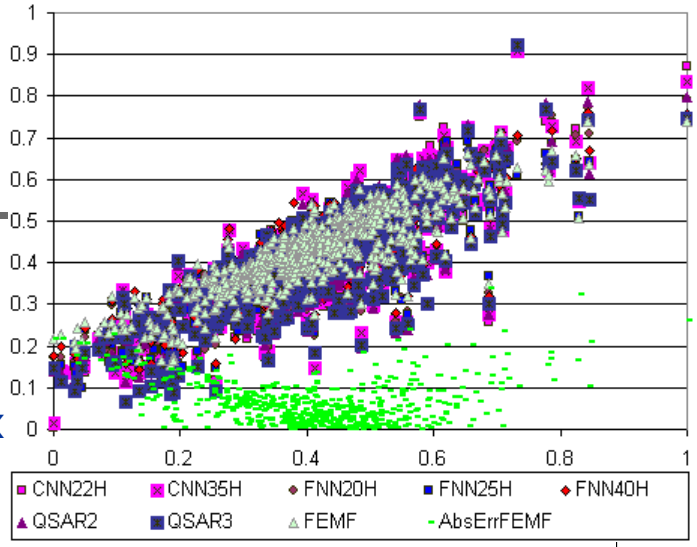


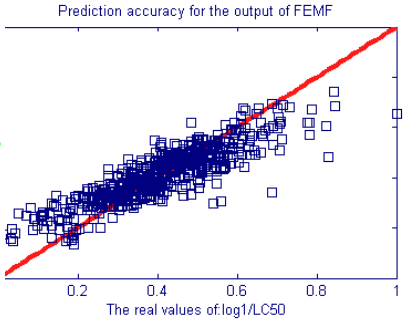
The predicted value:0.77684 | the real value:0.8436

Vietri 2002

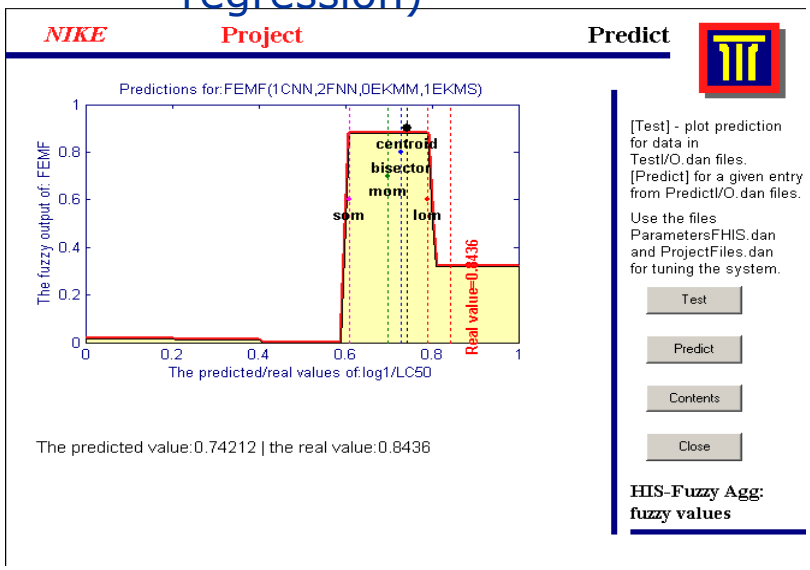
Fuzzy mixture of experts







- The method of combining:
  - Aggregation: **max** (for disjunctive trained experts) and
  - Defuzzification: **centroid** (for regression)

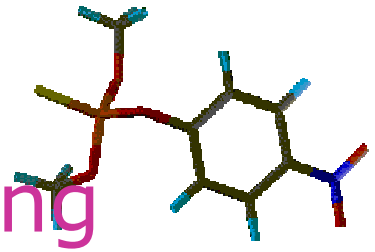


ProjectFiles.dan (just FIS (FNN, Mamdani) are used)

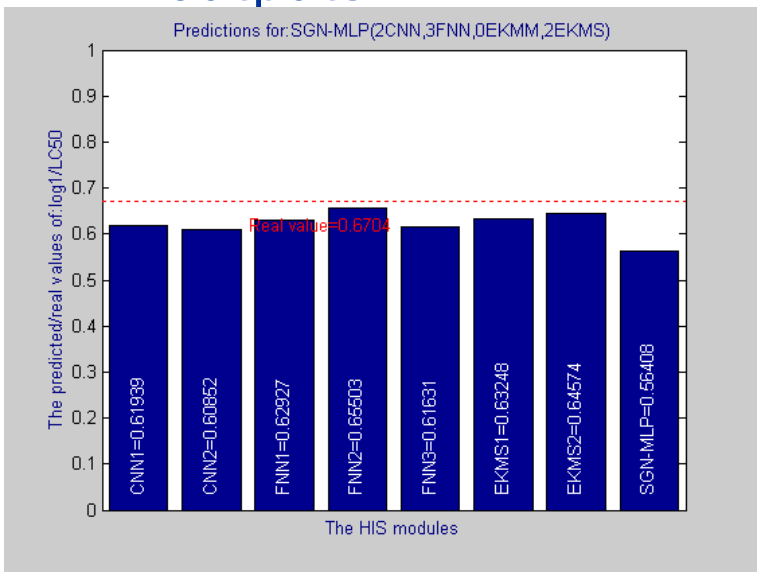
```
NumCNN=1
NumFNN=2
NumEKMMamdani=0
NumEKMSugeno=1
[CNN]
C:\IMAGETOX\DuluthMols\work\IKM\CNN\CNN23H\memvarCNN23Hnet.mat
[FNN]
C:\IMAGETOX\DuluthMols\work\IKM\FNN\FNN15H\memvarFNN15Hnet.mat
C:\IMAGETOX\DuluthMols\work\IKM\FNN\FNN25H\memvarFNN25Hnet.mat
[EKMMamdani]
[EKMSugeno]
C:\IMAGETOX\DuluthMols\work\Data\FuzzyIOS.fis
```

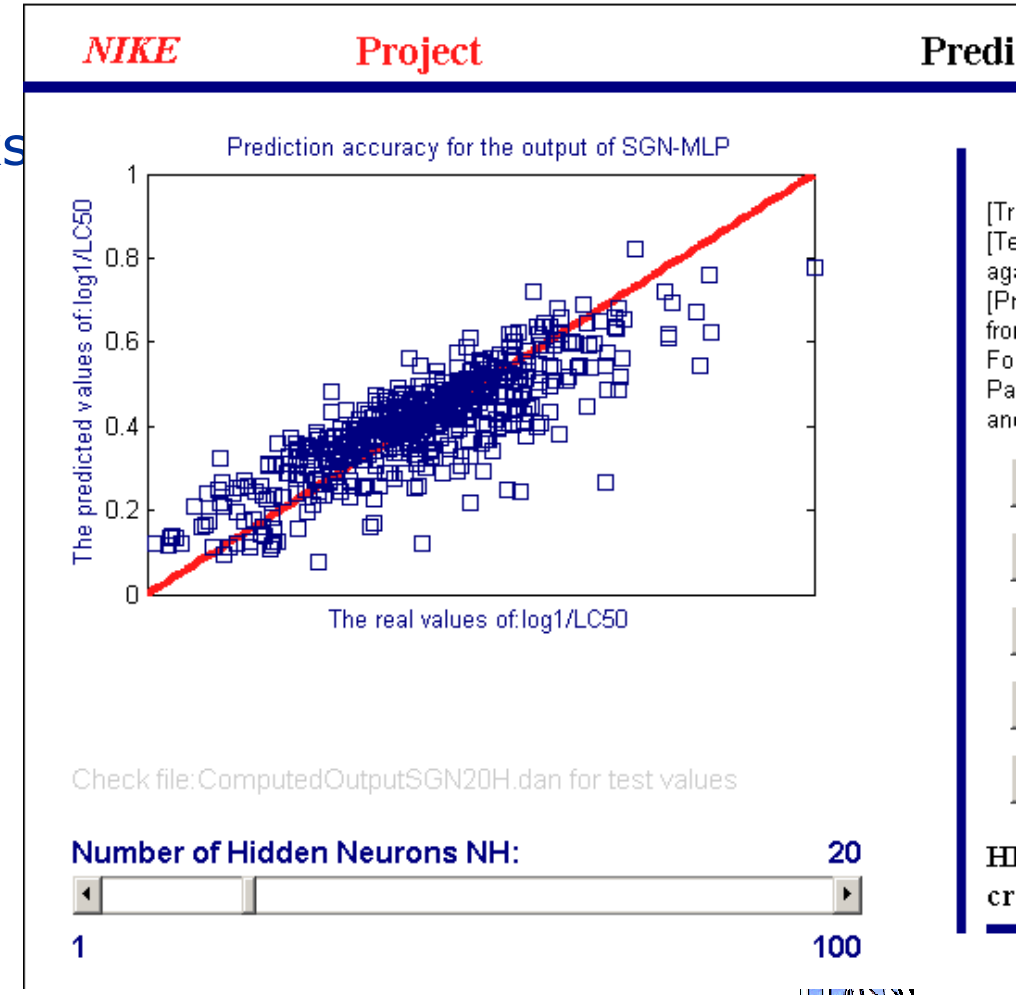
Vietri 2002

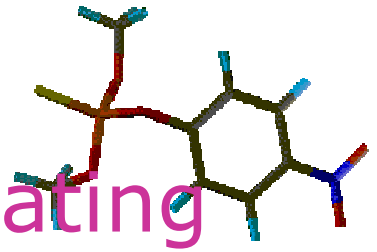
SGN (supervised-trained gating network) voting of experts



- SGN considers:
  - outputs of expert networks as inputs for GN
  - the gating network is trained with the experts opinions against the real outputs.

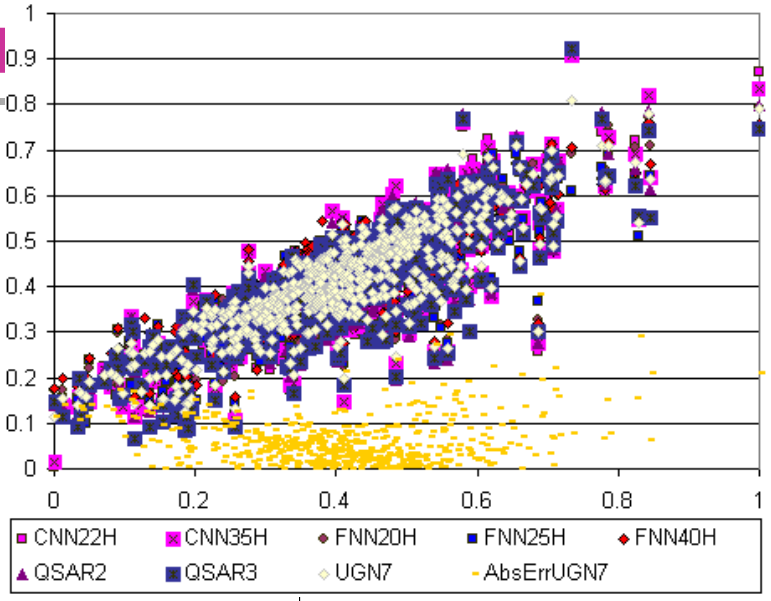


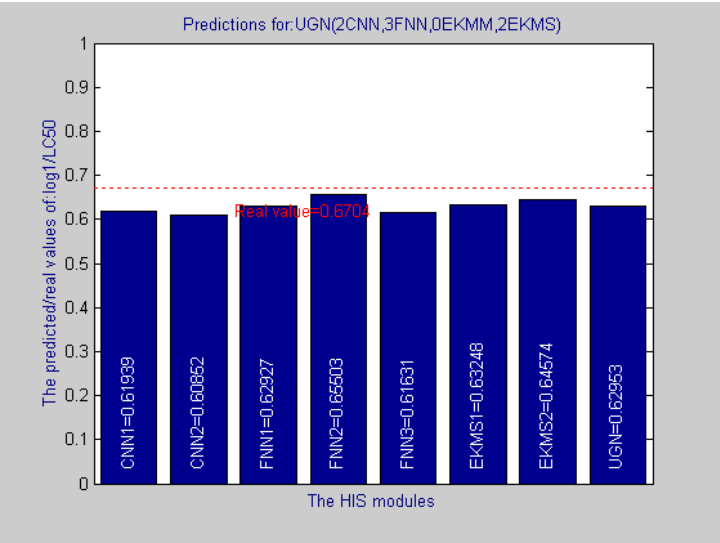


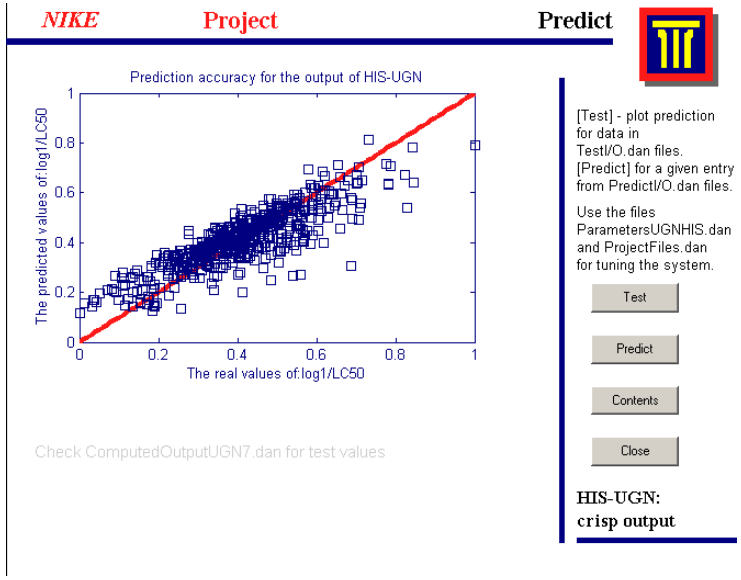


UGN (unsupervised-trained gating network) voting

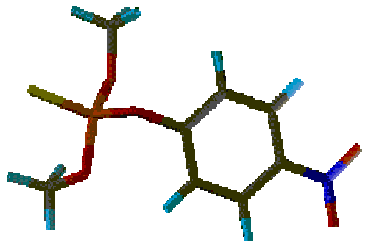
- UGN considers:
  - expert networks competing to learn the training patterns
  - the gating network mediating the competition between the







Regression models evaluation



RMSE
(root mean squared error)

• square root of the mean of the squared residuals obtained from a model

$RMSE = [(SUM [(y - \hat{y})^2]) / n]^{1/2}$

RSS residual sum of squares

• sum of the squared differences between the observed response and the response obtained from a model

$RSS = SUM [(y_i - \hat{y}_i)^2]$

MSS model sum of squares

•sum of the squared differences between the computed response and the average

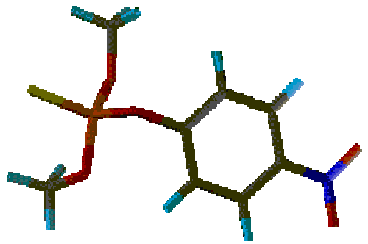
$MSS = SUM [(\hat{y}_i - y^a)^2]$

TSS total sum of squares

TSS= RSS+MSS

• sum of the squared differences between the observed response and the average

$TSS = SUM [(y_i - y^a)^2]$ *zero order model*



Model predictive value

MSS model sum of squares

• sum of the squared differences between the computed response and the average
MSS = SUM $[(\hat{y}_i - y^a)^2]$

R^2 determination coefficient

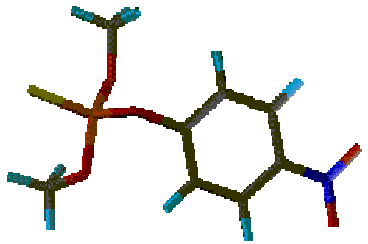
•**MSS/TSS = 1-RSS/TSS = R^2** ;
•R^2 * 100 the percentage variance expressed by the model,
•R is the coefficient of multiple correlation

PRESS
predicted error sum of squares

• sum of the squared differences between the observed response and the response obtained from the test set
PRESS = SUM $[(y_i - \hat{y}_i)^2]$

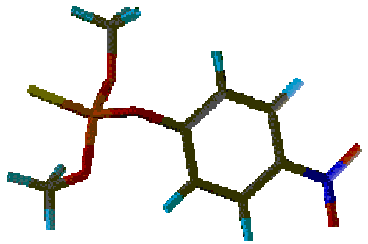
R^2_{cv} determination coefficient cross validated

1-PRESS/TSS = R^2_{cv}



For classification

- From the confusion matrix c we compute
- NER% = (Sum c_{dd})/n + 100
- ER% = 100 – NER%
- Using a loss matrix l
- MR% =[Sum (Sum $l_{dd'} * c_{dd}$)*p_g/n]*100



ENSEMBLE /MIXTURE

- To reach a reasonable good prediction by a single and combine a few

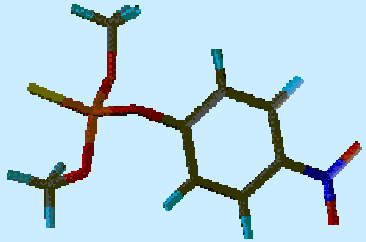
- **MOTIVATIONS: to exploit diversity**

  Carcinogenicity of aromatic compounds – ANN + graphs

  Letal dose of pesticides – gating network of classifiers

  Letal dose (EPA study) – the effect of scaling, symbolic rules

1. Carcinogenicity (of aromatic compounds)



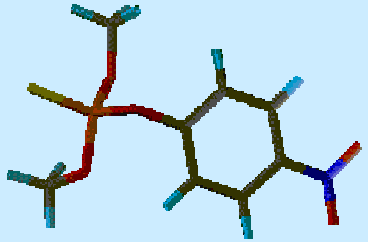
104 aromatic compounds

Predicting td50

finding residues

Carcinogenicity class

How to quantify carcinogenicity?



Classes or *doses*

Classes (IARC, EPA) assumption: one single molecule can produce cancer (no interest on the dose)

IARC (International Agency on Research on Cancer) *classes*:

1. Carcinogenic to man
2. carcinogenic to animals (2A: probable; 2B possible)
3. not classifiable
4. not carcinogenic

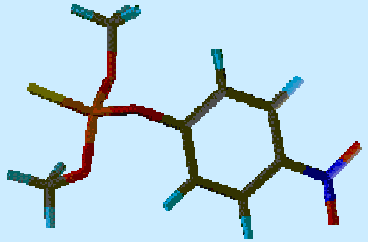
This classification combines, in the evaluation of carcinogenicity, the experimental evidences with the amount of epidemiological knowledge available.

TD50 (Gold) threshold dose:

- dose which kills 50% of animals
- it is a continuous value
- not for man toxicity

Gold and colleagues developed a numerical data set that contains standardized and reviewed results for carcinogenicity for more than 1200 chemicals. The cancerogenicity data on rat and mouse are expressed in term of the parameter TD50, which is the chronic dose rate, which would give half of the animals tumors within some standard experiment time.

Residues = substructures responsible of some activities

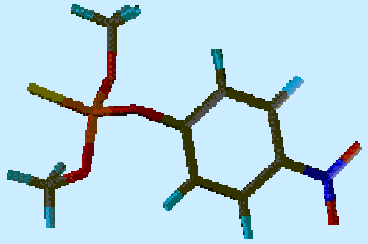


- *Activity:* carcinogenicity for aromatic compounds with at least a nitrogen linked to the aromatic ring (**Ar-N compounds**).
- The Ar-N group is divided into 10 chemical classes, defined by the presence of a chemical group characterizing the Ar-N bond.
- Subclasses splitting: same atom or substituent or structure in fixed position relative to Ar-N bond;

I convenience; affinity of chemicals.

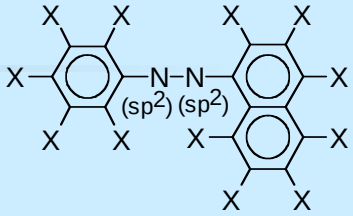
- **can be expressed as rules, but the graphic representation helps**

Residue search

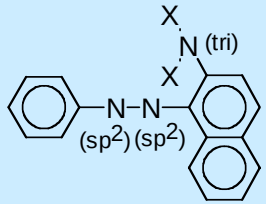


- For each subclass:
- - ***FIRST SEARCH***: search of the characterizing element of the subclass ("body" of the residue);
- - ***FIRST INHIBITION LEVEL***: is a negative condition, to exclude groups that are related to the structure of the subclass but not carcinogens.
- - ***SECOND INHIBITION LEVEL*****:** it excludes a specific compound (or a small group of compounds).
- As a result of the search, each fragment is associated with a category expressing the **level of toxicity (in 5 levels)**

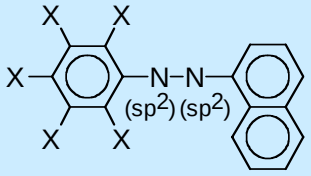
First Level Structure:
1-Naphtho azocompounds.



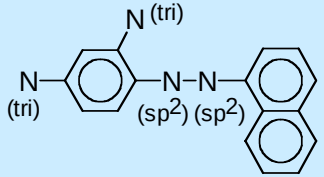
First Level Inhibition.

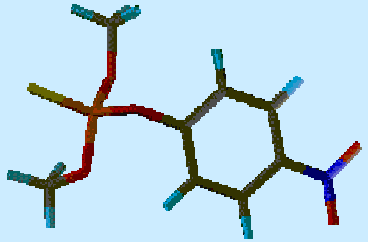


Second Level Structure:
Bensub-1NA residue



Second Level Inhibition



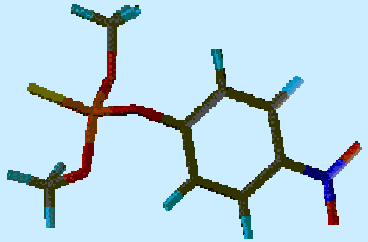


Chemical as graphs

- molecules and residues are represented by graphs - COSMIC format - atom hybridization instead of information on atomic bonds:

1. All bonds are equals
2. Hydrogens are left out.

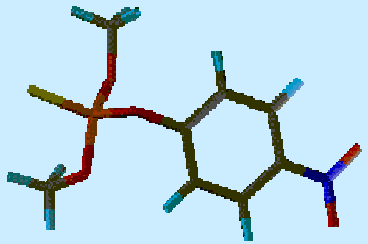
- structures are represented by adjacency lists.
- The **search of a fragment** in a molecule as a *subgraph isomorphism problem:* find *all* the isomorphisms between a graph and subgraphs of a given graph.



Graph isomorphism

- A graph is *isomorphic* to a subgraph iff there is a 1-to-1 correspondence between the node sets that preserves adjacency. The problem is, in general, NP-Complete.
- Ullmann's algorithm, modified to manage hydrogens and wildcards.
- The first search level: all isomorphisms between the structure considered and the molecule. When a first level structure is found, the second part checks positive and negative conditions.
- • If a second level structure and no inhibition, we have **one instance** of the residue in the molecule.

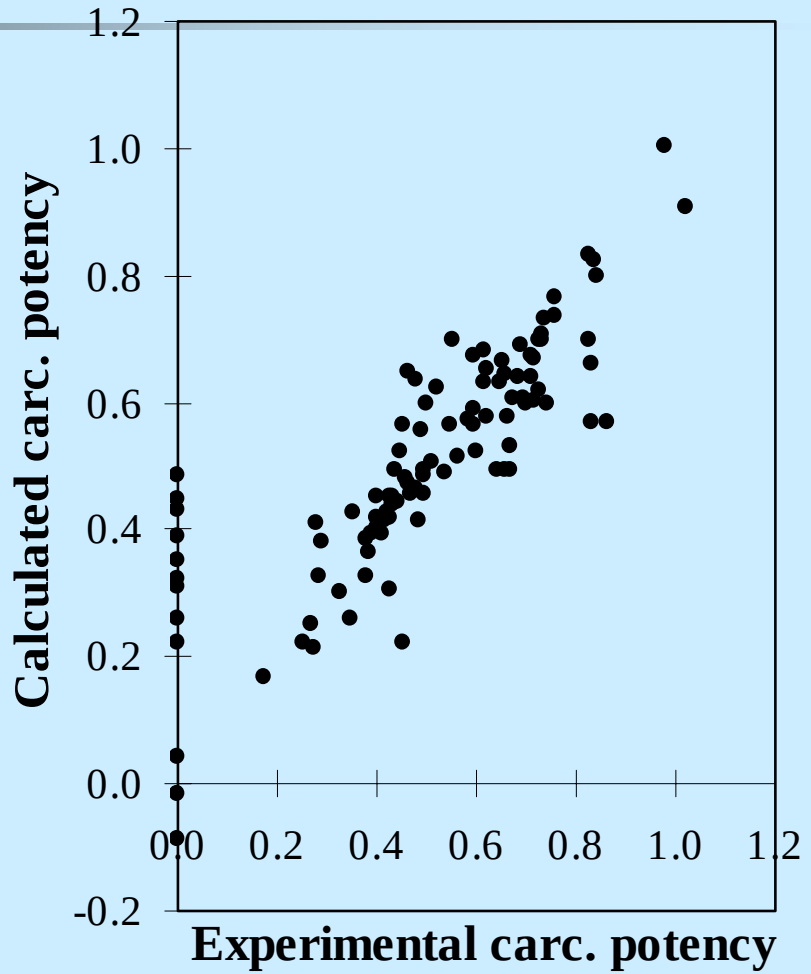
ANN prediction of TD50



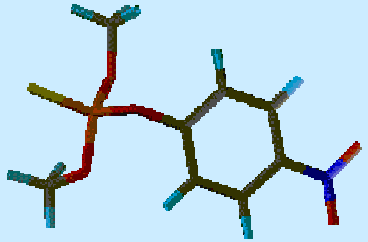
Input: 13 descriptors
Output:
Log(mw*1000/TD50)
Validation: N/2-fold-cross validation

Neurons	MSE	R^2_{cv}
3	0.0157	0.6752
4	**0.0146**	**0.6911**
5	0.0154	0.6756
6	0.0153	0.6758
7	**0.0146**	**0.6915**

$R^2_{cv} = 0.69$



Vietri 2002

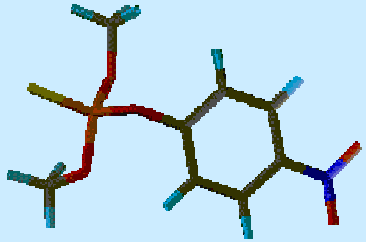


Best crossvalidated

- after removing 12 outliers. For 9 the experimental results were not statistically significant (arbitrary 10^{31})
- therefore a lower prediction for non carcinogenic compounds.

Neurons	MSE	R^2_{cv}
3	0.0062	0.7933
4	**0.0053**	**0.8237**
5	0.0053	0.8236
6	0.0057	0.8099
7	0.0061	0.7922
8	0.0073	0.7553

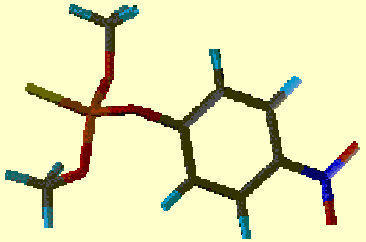
Hybrid system



	C4.5	CART	OC1
Training	93.3	88.5	90.2
Validation	81.9	85.5	82.8

- • 5 classes, from lower to higher risks
- • to each residue, a toxicity class as the mean of the toxicity of the molecules where found;
- • to assign to the molecule the maximum toxicity obtained from residues + ANN.
- classification :
- • **C4.5**, **CART**, **OC1**, accuracy % using the leave-one-out method

2. Pesticide toxicity

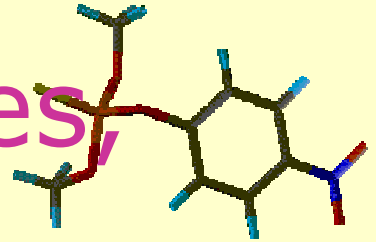


Species	Toxicity values	# compounds
Rainbow trout	LC_{50} 96h	233
Daphnia magna	LC_{50} 48h	217
Mallard duck	LD_{50}	110
Bodwhite quail	LD_{50}	133
Rat	LD_{50}	235

Toxicity values

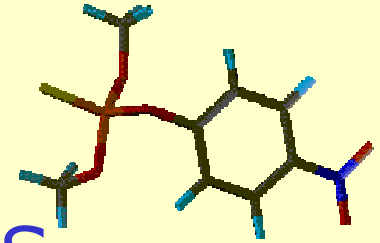
- ☞ Pesticide Manual
- ☞ RTECS
- ☞ HSDB
- ☞ Ecotox

Chemical classes, species, and r correlation



Chemical Class	Total	Training Set	Test Set
Anilines	39	21	18
Aromatic halogenated	83	57	26
Carbamates	26	23	3
Heterocycles	119	93	26
Organophosphorous	59	27	32
Ureas	31	24	7
Different Class	5	4	1
Total	**362**	**249**	**113**

	Quail	Trout	Daphnia
Trout	-0.02		
Daphnia	0.21	0.06	
Duck	0.55	0.44	0.14

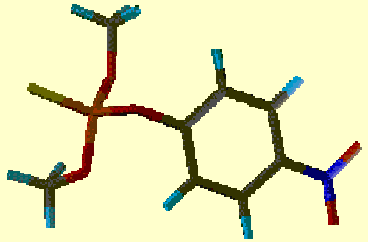


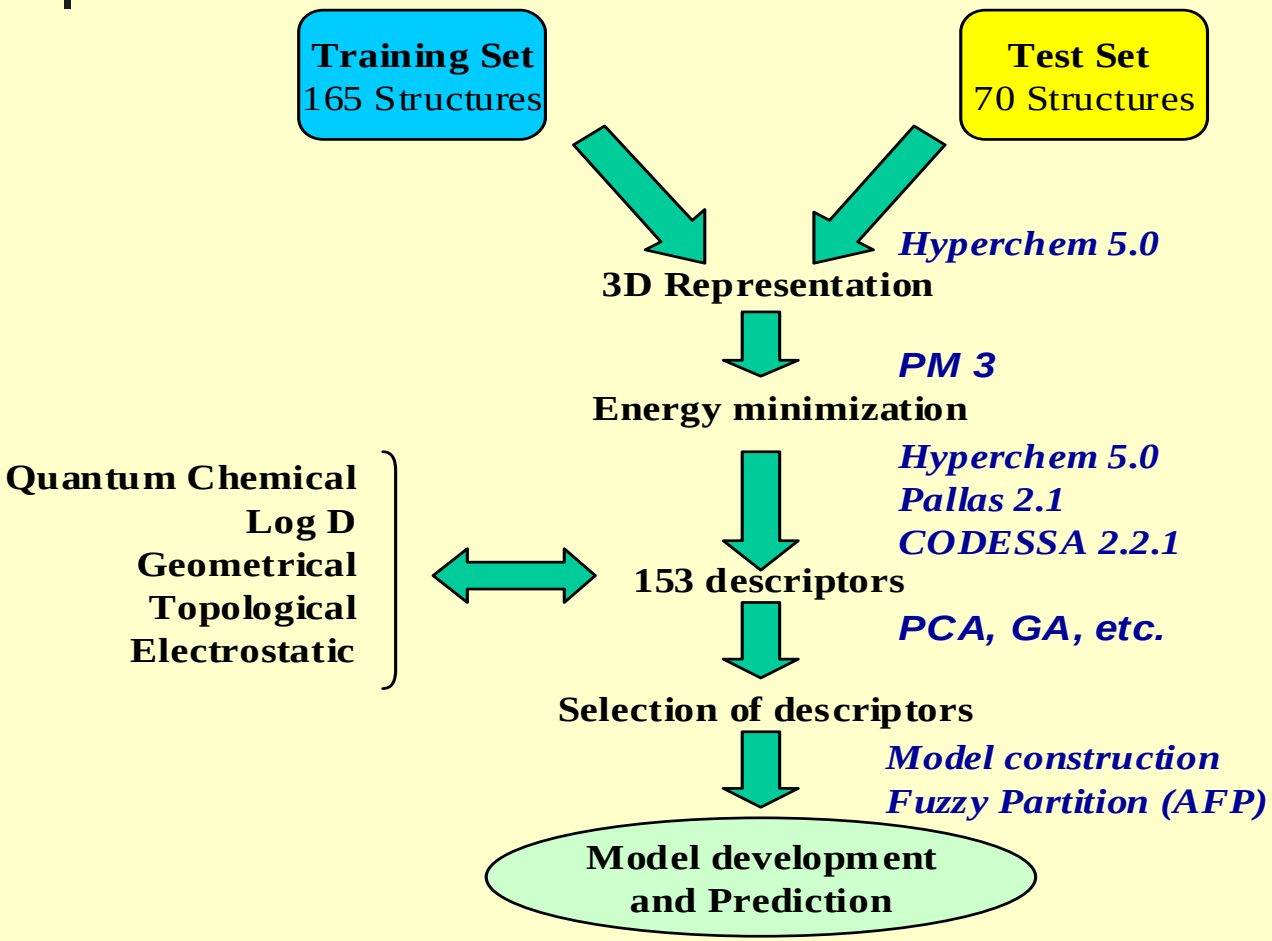
Linear regression for LC_{50}using PLS –

R^2_{cv} when > 0.5.

Chemical Class	rainbow trout	daphnia	rat	duck	quail
Aniline	0.78	0.72	No results	No results	No results
Carbamate	No results	No results	No results	No results	No results
Organophosphorus	No results	0.69	No results	No results	No results
Urea	0.78	0.85	0.59	No results	No results
Heterocyclic	No results	0.56	No results	0.55	No results
alogenated aromatic	No results	No results	No results	No results	0.55

Model development



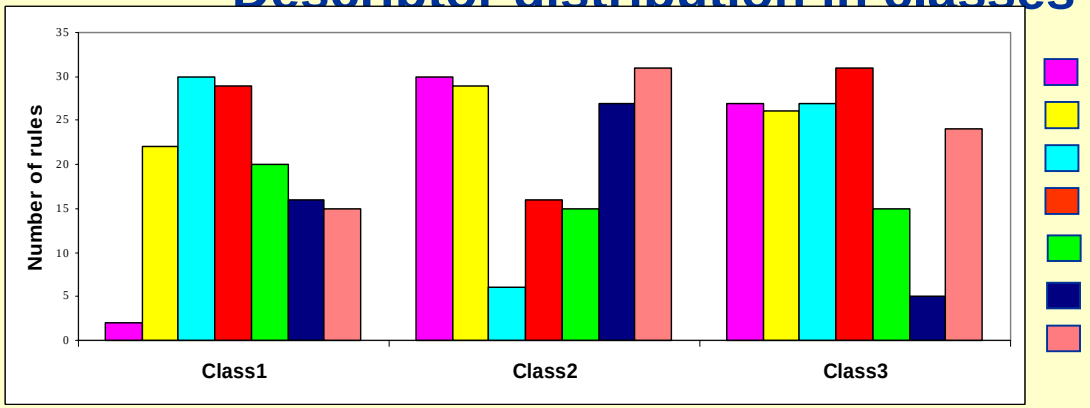


Toxicity against Rat (3 class)

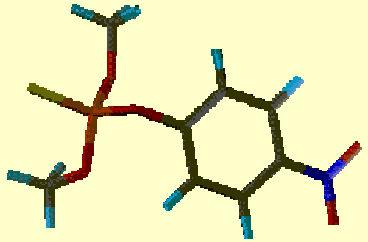
Classes	Intervals LD_{50} (mg/kg)	Training Set 165	Test Set 70
Class1	> 3000	56	16
Class2	700 - 3000	54	17
Class3	< 700	55	37

7 descriptors - Class1: 30 rules; Class2: 31 rules; Class3: 31 rules

Descriptor distribution in classes

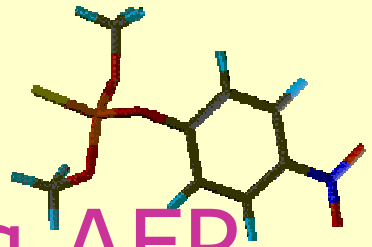


Vietri 2002



Validation

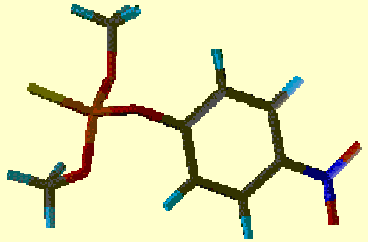
Classes	Intervals (mg/kg)	Training set validation (%)	Test set validation (%)
Class1	> 3000	**80**	**75**
Class2	700 - 3000	**68.5**	**53**
Class3	< 700	**82**	**86**
All classes		**77**	**76**



Adaptive Fuzzy Partitioning AFP

- Iteratively divide the descriptor hyperspace into fuzzy partitioned rectangular subspaces until :
  - # of molecular vectors within a subspace < threshold_{MIN};
  - the difference between two generated subspaces is negligible in terms of chemical activities;
  - # of subspaces > threshold_{MAX}.
- select the descriptor and the cut position to maximize the difference between the two fuzzy-rule scores generated by the new subspaces.

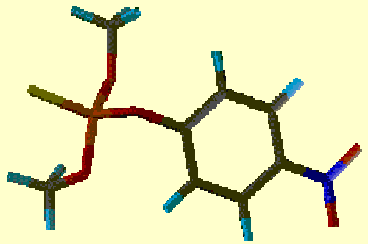
if x_1 is associated with $\mu_{1k}(x_1)$ ***and*** *x_2 is associated with $\mu_{2k}(x_2)$ …* ***and*** *x_N is associated with $\mu_{Nk}(x_N) \Rightarrow$ the score of the activity O for P is O_{kP},*



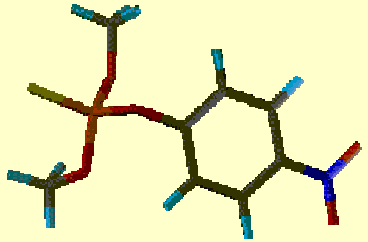
Classification results

- 4 classes (EU Directive 92/32/EEC); correct prediction 60% of the test set, 78% of the training set. The most toxic class better predicted (69%).
- 3 classes (in the training set a similar number of compounds). Correct 71% of the test set; class 3 (the most toxic) the best predicted(86%).
  - AFP builds up a scheme of the rules used for each toxicity class, as :
    - if $0 < x(\log D\text{-}pH5) < 0.26$ and $0 < x(\text{Balaban Index}) < 0.51$ and $x(\text{Randic Index}) > 0.81\ldots$. $\Rightarrow$ the membership degree of class 1, for the compound 34, is 0.5.

ensembling different classifiers



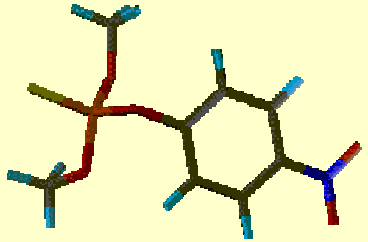
- 57 organophosphorous compounds.
- The toxicity value was $Log_{10}(1/LC_{50})$, scaled in the interval [-1..1].
- Class 1 [-1..-0.5],
- Class 2 [-0.5..0],
- Class 3 [0..0.5],
- Class 4 [0.5..1]



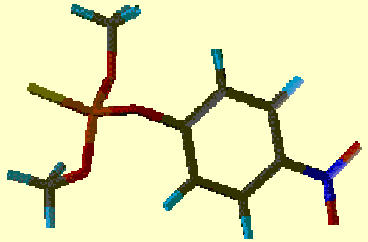
Single classifiers

- LDA (Linear Discriminant Analysis)
- RDA (Regularized Discriminant Analysis)
- SIMCA (Soft Independent Modeling of Class Analogy)
- KNN (K Nearest Neighbors classification)
- CART (Classification And Regression Tree)

results



	True Class	CART	LDA	KNN	SIMCA	RDA
[1]	[2]	[3]	[4]	[5]	[6]	[7]
Anilofos	2	2	2	1	2	2
Chlorpyrifos	1	2	2	1	2	2
Chlorpyryfos-methyl	2	2	2	1	2	2
Isazofos	1	1	1	2	1	1
Phosalone	2	2	2	2	2	2
Profenofos	1	2	2	1	2	2
Prothiofos	2	2	2	2	2	2
Azamethiphos	2	2	2	1	4	2
Azinphos methyl	1	1	1	2	1	1
Diazinon	3	3	1	1	4	1
Phosmet	2	2	2	1	2	2
Pirimiphos ethyl	1	1	1	1	1	1
Pirimiphos methyl	2	3	1	2	1	1
Pyrazophos	2	2	1	4	2	1
Quinalphos	1	1	1	2	1	1
Azinphos-ethyl	1	1	1	1	2	1
Etrimfos	1	1	1	3	3	1
Fosthiazate	4	2	2	2	4	2
Methidathion	1	1	1	1	1	1
Piperophos	3	3	3	2	2	3
Tebupirimfos	4	1	1	3	4	1
Triazophos	1	1	1	2	1	1
Dichlorvos	2	4	2	2	2	2
Disulfoton	3	3	3	1	3	3
Ethephon	4	4	4	4	4	4
Fenamiphos	1	1	3	2	1	1
Fenthion	2	2	3	2	2	3
Fonofos	1	1	3	2	1	3
Glyphosate	4	4	4	4	4	4
Isofenphos	3	3	3	1	3	3
Methamidophos	4	4	4	3	4	4
Omethoate	3	3	3	3	3	3
Oxydemeton-methyl	3	3	3	3	3	3
Parathion ethyl	2	2	2	3	1	3
Parathion methyl	3	3	3	3	3	3
Phoxim	2	2	1	1	1	1
Sulfotep	1	1	3	2	2	2
Tribufos	2	2	2	2	2	2
Trichlorfon	2	2	2	1	2	4

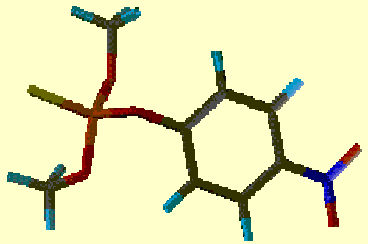


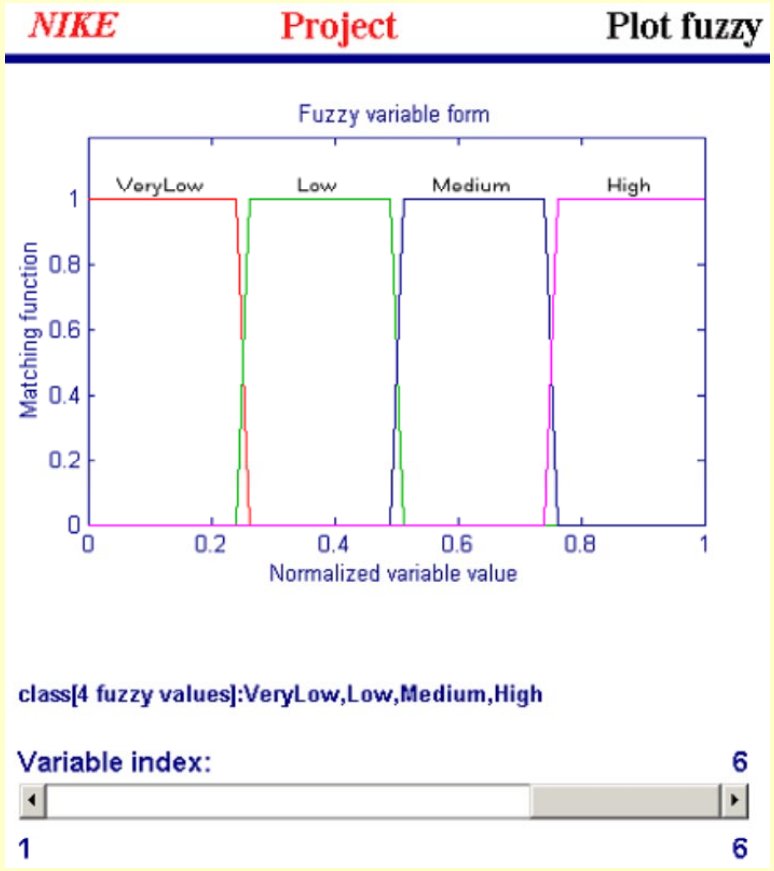
validation

	NER% fitting	NER% validation	Descriptors
LDA	64.91	61.40	D1,D2, D3, D4
RDA	84.21	71.93	D1, D2, D3, D4, D6, D7, D8, D11, D12, D13
SIMCA	92.98	**77.19**	D1, D2, D3, D4, D5, D6, D7, D8, D10, D11, D12
KNN	-	61.40	D1, D12
CART	85.96	77.19	D1, D2, D3, D4, D5, D9

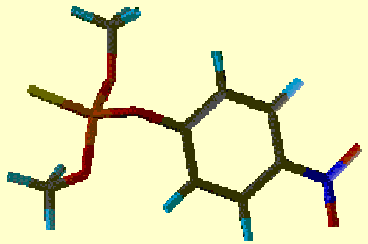
How to make an ensemble?
Maority vote 14 errors
Gating network?

ensemble learner

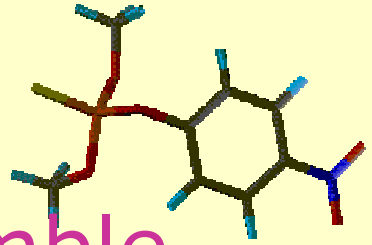




- a class represented by the centroid:
- 0.135 (class 1),
- 0.375 (class 2),
- 0.625 (class 3)
- 0.875 (class 4).
- trapezoidal:
- *VeryLow* (0..0.25),
- *Low* (0.25..0.5),
- *Medium* (0.5..0.75),
- *High* (0.75..1).



- For FNN, $p = 5$ inputs represent the answer of the classifiers for a given compound: x_1=output$_{CART}$, x_2=output$_{LDA}$, x_3=output$_{KNN}$, x_4=output$_{SIMCA}$, x_5=output$_{RDA}$.
- FNN trained on 40 cases (70%), with backpropagation. The neuro-fuzzy network was a multi-layered structure with the 5x4 above described fuzzy inputs and 4 fuzzy output neurons, the toxicity class linguistic variable. The best results obtained with 10, 12, 19 neurons.

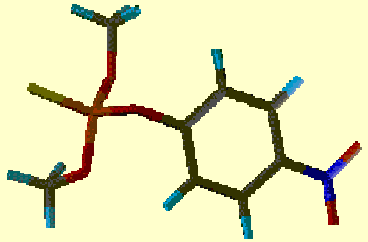


Confusion matrix of the ensemble

		Assigned Class				N° of objects
		1	**2**	**3**	**4**	
True Class	**1**	13	2			15
	2		20			20
	3		1	15		16
	4				6	6

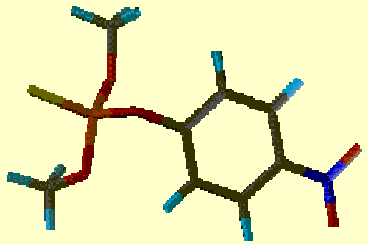
The error on the badly predicted

	True Class	CART	LDA	KNN	SIMCA	RDA	FNN
Chlorpyrifos	1	2	2	1	2	2	2
Profenofos	1	2	2	1	2	2	2
Fenitrothion	3	2	3	3	3	3	2



performances

	LDA	RDA	SIMCA	KNN	CART	FNN
NER% fitting	64.91	84.21	92.98	-	85.96	-
NER% validation	61.40	71.93	77.19	61.40	77.19	94.74



Extracted fuzzy rules

- *Same output for different opinions of classifier*

  IF CarFit1 is:VeryLow THEN class is:High (39.22%)
  IF CarFit1 is:Low THEN class is:High (82.30%)
  IF CarFit1 is:Medium THEN class is:High (48.74%)
  IF CarFit1 is:High THEN class is:High (39.04%)

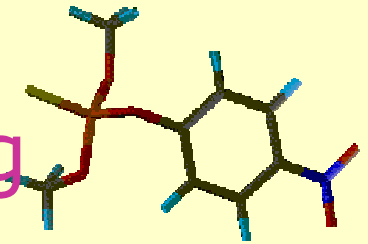
(for any answer of CART THEN class is High)

- *IF SimFit1 is:VeryLow THEN class is:Medium (61.25%)*
  IF SimFit1 is:Low THEN class is:Medium (36.04%)
  IF SimFit1 is:High THEN class is:Medium (43.72%)

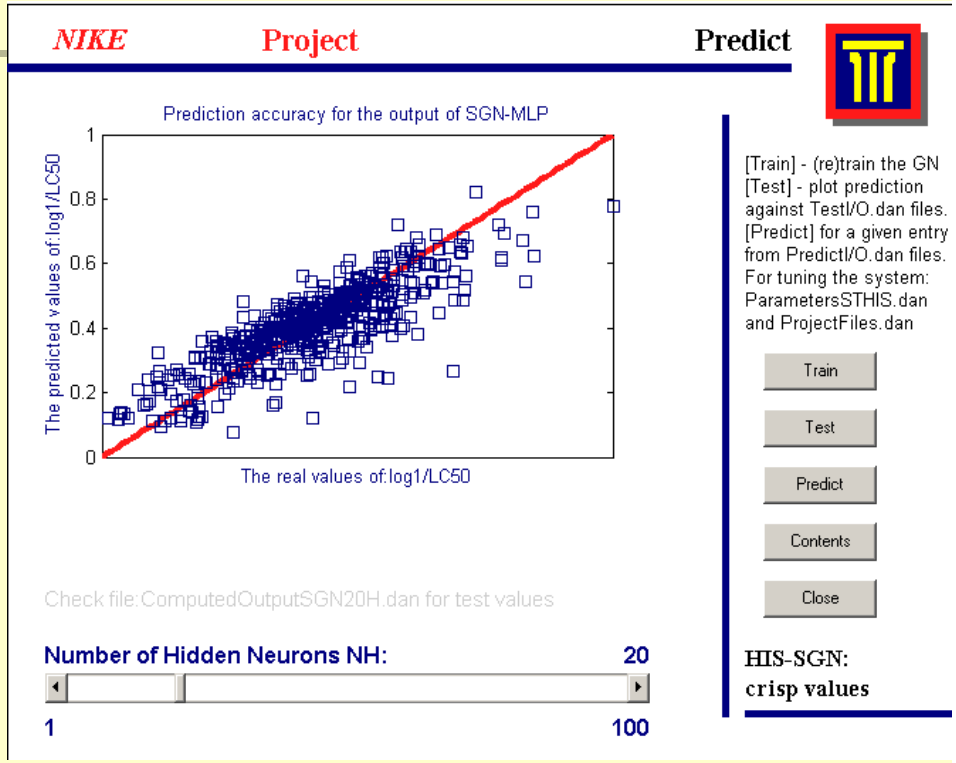
(for many answers of SIMCA THEN class is Medium)

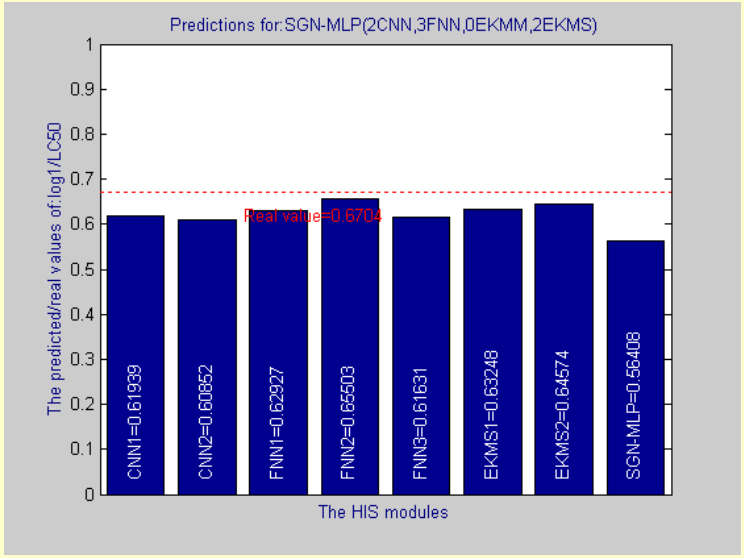
- *THE BEST CLASSIFIER :*
- **IF RdaFit1 is:VeryLow THEN class is:Low (75.65%)**
  IF RdaFit1 is:Low THEN class is:Low (100.00%)
  IF RdaFit1 is:High THEN class is:High (76.39%)

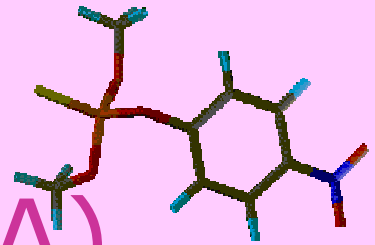
SGN (supervised-trained gating network) voting of experts



- SGN considers:
  - outputs of expert networks as inputs for GN
  - the gating network is trained with the experts opinions against the real outputs.



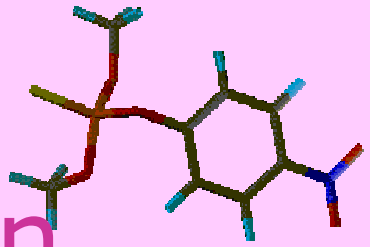




3. EPA (toxicity and MOA)

- 554 organic compounds, commonly used in industrial processes, with experimental data for acute toxicity 96 hours LC_{50}, for the fathead minnow (*Pimephales promelas*).
- Mechanism Of Action (MOA) to each compound.
- The data set was 70%-30% randomly partitioned between 388 training cases and 166 testing cases.

EPA Data set information



Maximum Value
75200.00

Minimum Value
0.00019

Range
7.5200e+004

Standard Deviation
5.7249e+003

Variance
3.2774e+007

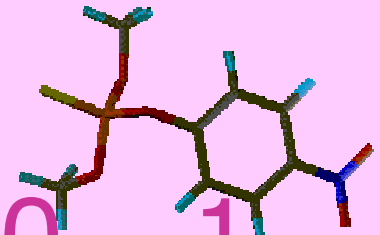
Geometrical Mean
24.1313

Arithmetic Average
1.0600e+003

Descriptors selected

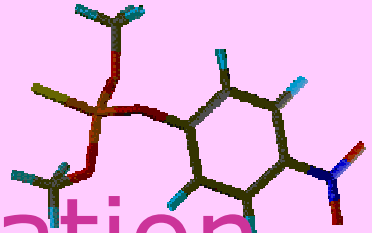
- Total Energy (kcal/mol) QM1
- Heat of Formation (kcal/mol) QM3
- LUMO (eV) QM6
- Relative# of N atoms C9
- Relative # of single bonds C24
- Molecular weight C35
- Kier&Hall index (order 0) T6
- Average Information (order 1) T22
- Moment of inertia B G2
- Molecular volume G10
- Molecular surface area G12
- Total molecular surface area E13
- FPSA-2 Fractional PPSA E24
- PPSA-3 Atomic charge weighted PPSA E28
- FPSA-3 Fractional PPSA E31
- LogD pH9 pH9
- LogP LogP

Vietri 2002



the effect of scaling in 0 - 1

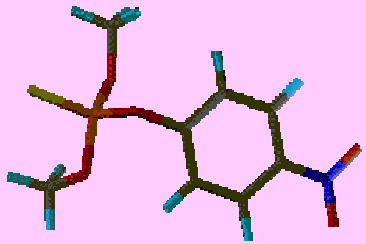
- To maintain the original distribution => range scaling
- For future integrations => the scaling must go beyond the limits of the data set. It exists a natural inferior limit (0 mg/L) but not a superior limit => a function defined between 0 and 1 with an asymptote to 1. The loss in knowledge about the highest values is acceptable (high values indicate less toxic, and on high values less precision is required).



EU directive for classification

LC_{50}	Dangerous for the environment
< 1 mg/L	Very toxic to aquatic organisms
1 mg/L – 10 mg/L	Toxic to aquatic organisms
10 mg/L – 100 mg/L	Harmful to aquatic organisms
> 100 mg/L	May cause long-term adverse effects in the aquatic environment

it is easily recognizable a logarithmic scale



scaling

1. Range scaling RS

$$y_i = \frac{x_i - min(x)}{max(x) - min(x)}.$$

2. Range logarithmic scaling RLS

$$y_i = \frac{log_{10}(x_i + 1) - min(log_{10}(x+1))}{max(log_{10}(x+1)) - min(log_{10}(x+1))}.$$

to co sider $log_{10}(x_i)$ when $x_i = 1$

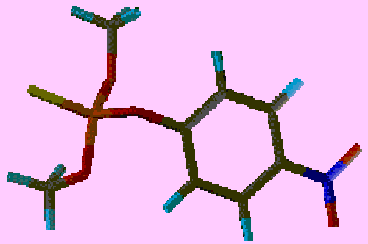
3. Tangent hyperbolic scaling THS

$$y_i = tanh(x_i).$$

4. Tangent hyperbolic logarithmic scaling THLS

$$y_i = tanh(log_{10}(x_i + 1)).$$

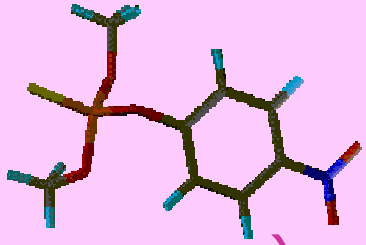
more specific scaling



5. Tangent hyperbolic logarithmic scaling modifiedTHLS

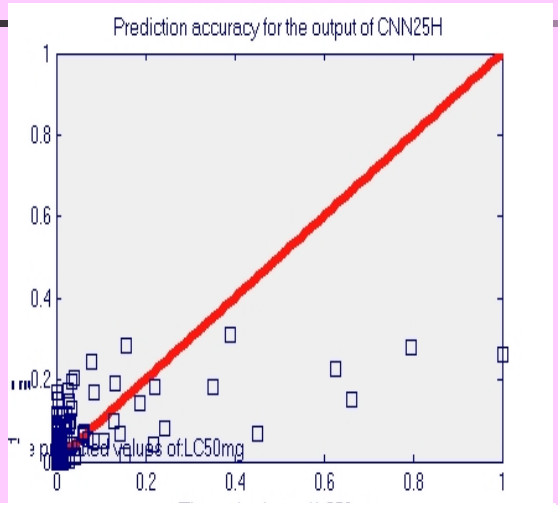
$$y_i = tanh\left(0.4903\, log_{10}\left(x_i + 1\right) + 0.0562\right) - 0.0095.$$

The *ideal* transformation succeeds in scaling the original toxic classes into classes of the same wideness. Thus, each transformed class has the same accuracy and the same original variance



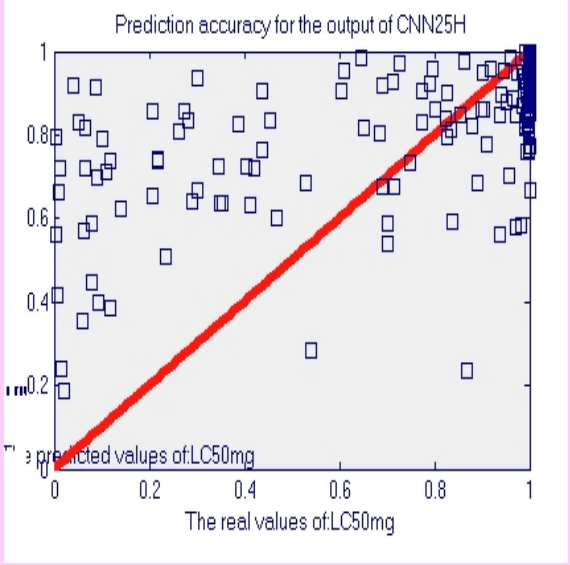
Prediction accuracy (NN with 25 hidden neurons)

RS



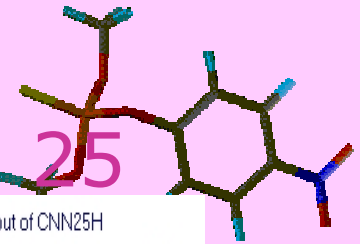
presence of few data with high value with respect to the others, it concentrates most of the data in a small interval; it loose information on the class of compounds more toxic

THS



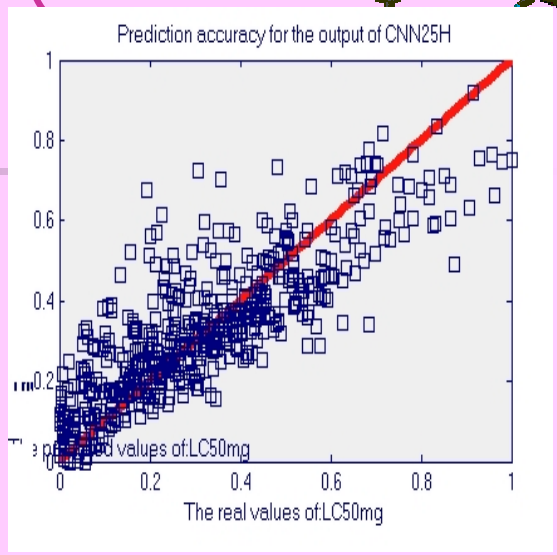
The object are well distributed.
The weakness, it needs a min and max value to be computed.

Prediction accuracy (NN with 25 hidden neurons)



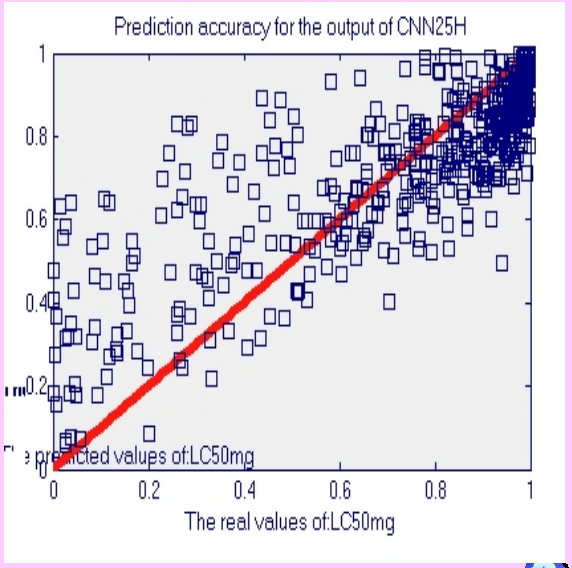
RLS

responds to our request to be a generalizable manipulation, but most of the data are compressed

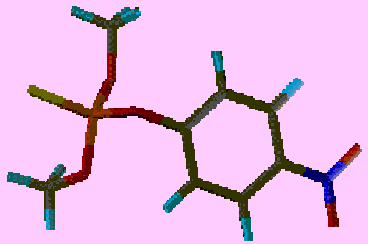


Doing first the logarithmic transformation in order to keep the guidelines of the EU Directive and then using a tangent hyperbolic in order to have a generalizable scaling we see a consistent improvement

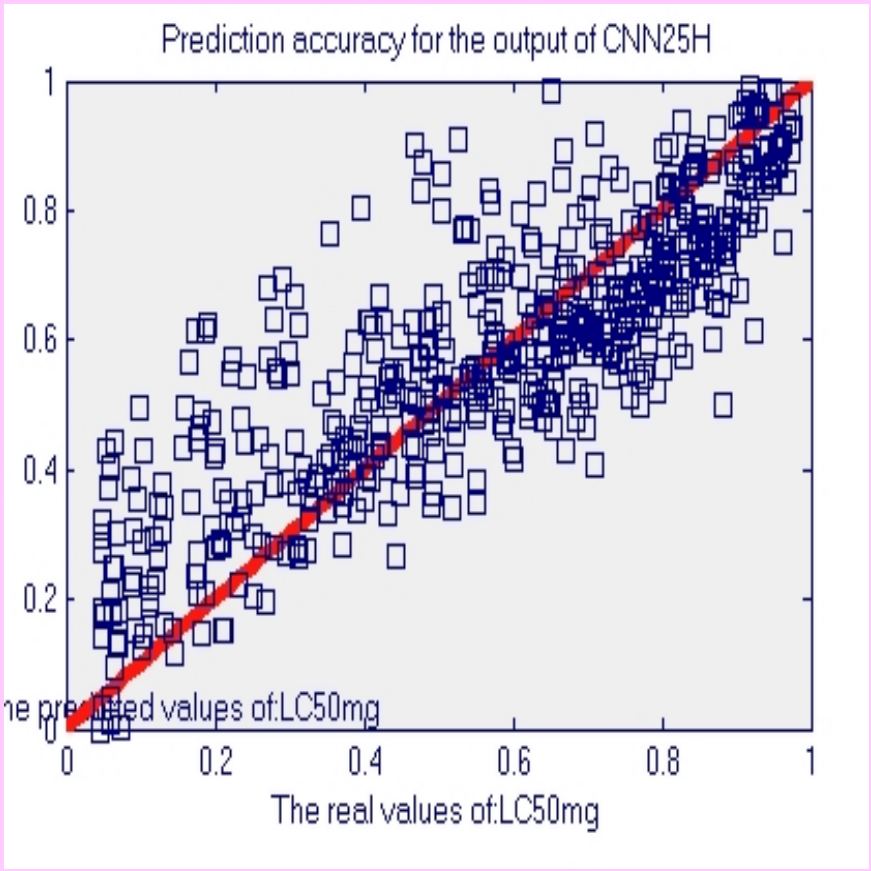
THLS



Transformed scaling



THLSM



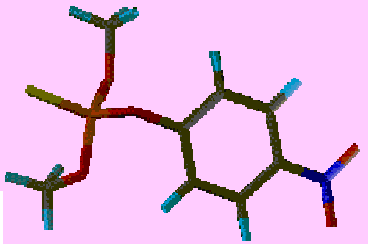
THLS needs only to be fit on the ideal distribution given by the Directive. We used a nonlinear curve-fitting solver in the least squares sense: find coefficients *x* that "best-fit" the equation *F(x, xdata)*:

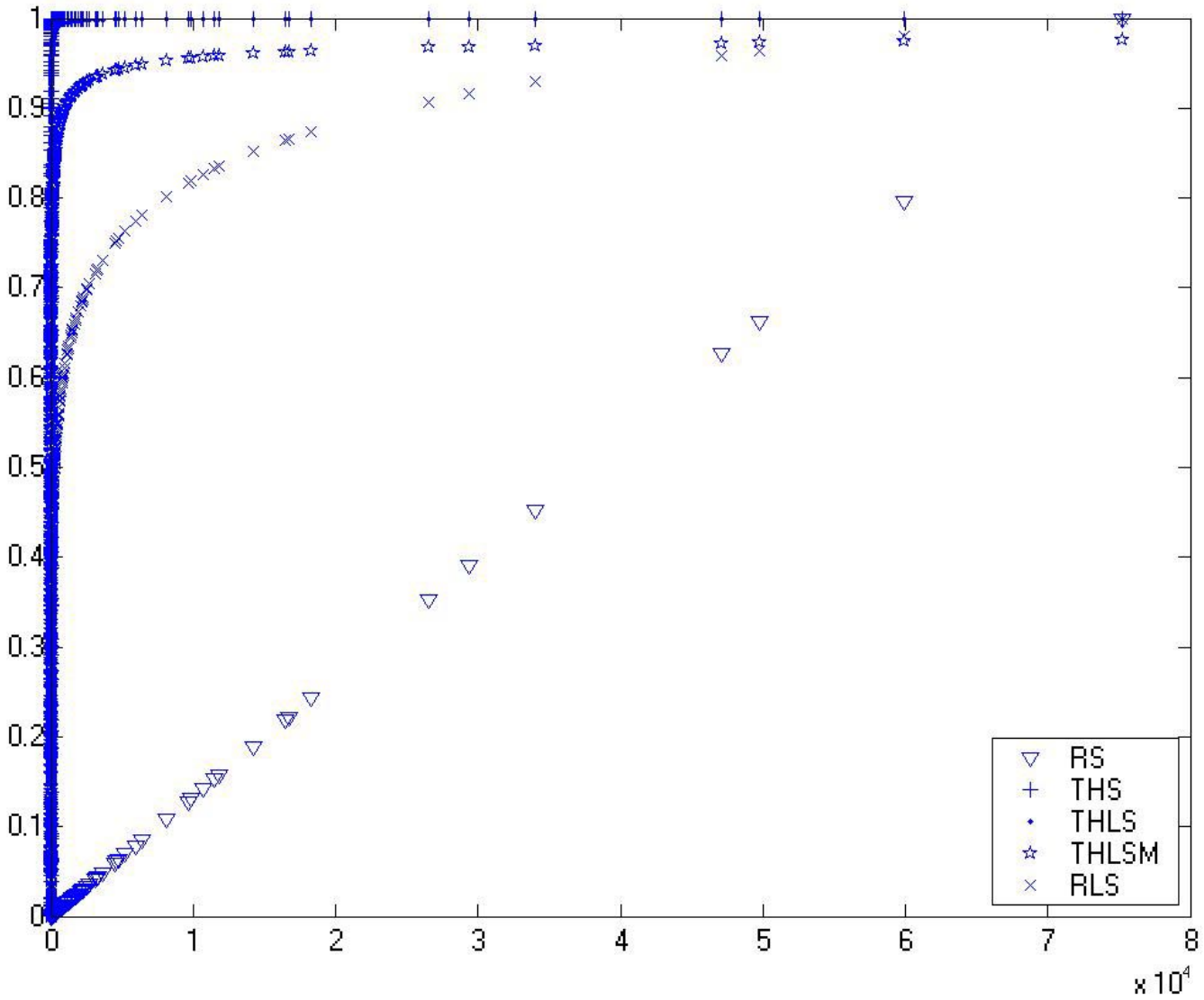
$$\min_x \frac{1}{2}\left|F(x,xdata)-ydata\right|_2^2 = \frac{1}{2}\sum_i \left(F(x,xdata_i)-ydata\right)^2.$$

xdata is the vector of the class limits given by the EC, *ydata* is the vector of the best ideal distribution and *F(x, xdata)* is the vector valued function:
xdata = [0 ; 1 ; 10 ; 100 ; inf]
ydata = [0 ; 0.25 ; 0.5 ; 0.75 ; 1]

$$F(x,xdata) = tanh\left(x_1\, log_{10}(xdata+1)+x_2\right)+x_3.$$

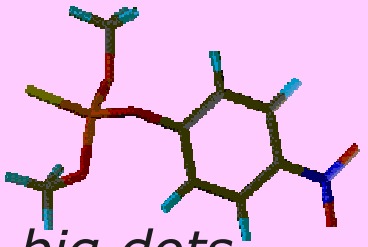
Ideal transformation

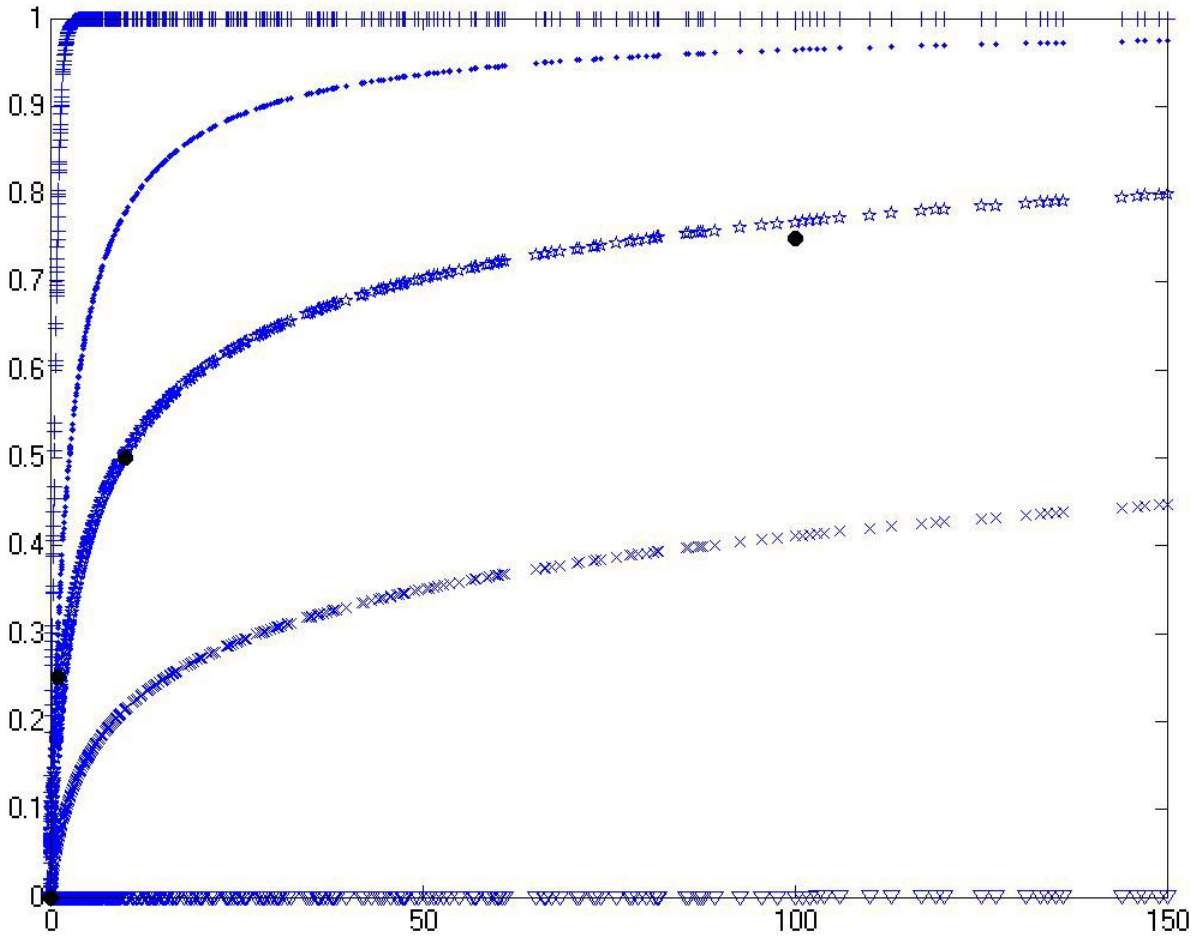




In the entire interval

Ideal transformation





In the significant interval [0-150 mg/L]

The big dots are land marks for the ideal transformation

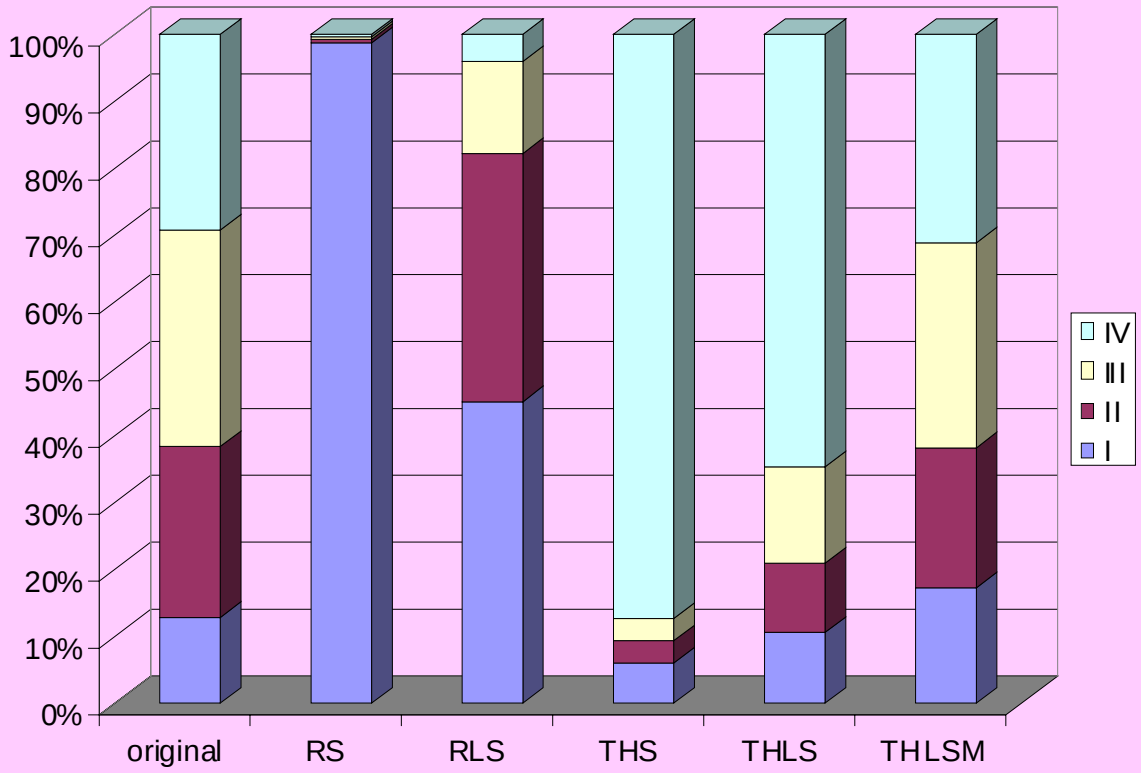
RS forces 99% in a very small interval (0 – 0.25).

Similarly THS, 87% of data in (0.75 – 1).

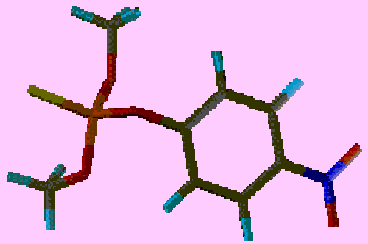
RLS and THLS have a better distribution THLSM best fits the characteristics of the *ideal* transformation.

Abundance of classes after each transformation.

- The THLSM keeps a better distribution

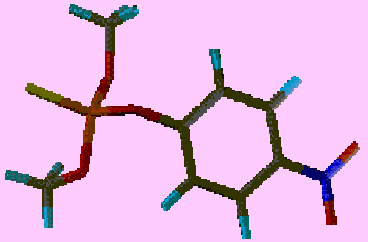


Vietri 2002



Models and Knowledge

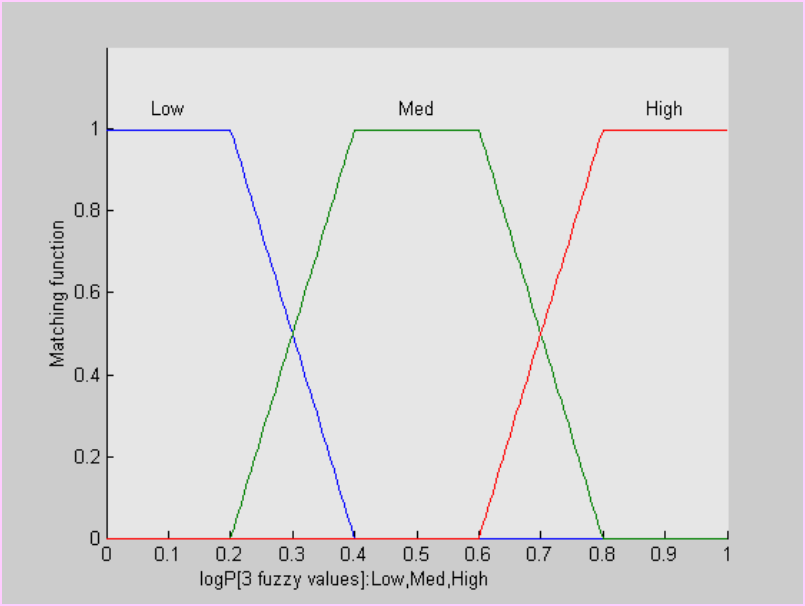
- analyse 568 organic compounds through neural/neuro-fuzzy nets.
- The most successful architectures are data mined, to obtain models, a reduced number of descriptors, to combine them with the explicit QSAR Finally, the models are integrated to develop the hybrid intelligent system



fuzzyfication

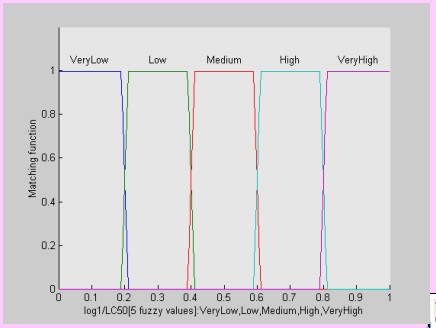
- Input: 17 descriptors output: $\log(1/LC_{50})$.
- the membership functions are trapezoidal. The linguistic variables for descriptors, and for toxicity, are characterized by the term sets

(a)

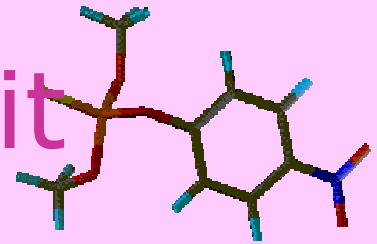


$$D_i = \{Low, Med, High\}, i = 1..17$$

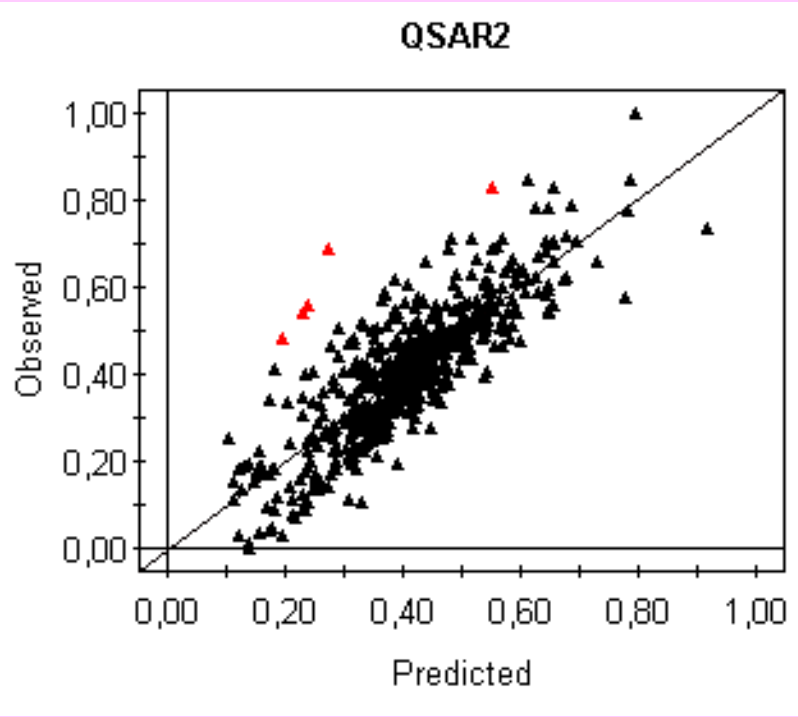
$$\log(1/LC50) = \{VeryLow, Low, Medium, High, VeryHigh\}$$

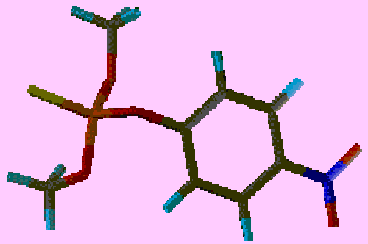


QSARs: Inserting explicit knowlege



- QSAR1: $\log(1/LC_{50}) = 0.7919 + 0.09772*QM6 - 0.2045*C35 + 0.1276*G2 - 0.3509*pH9 - 0.3879*logP$
- *QSAR2: $\log(1/LC_{50}) = 0.8779 + 0.1385*QM6 - 0.06703*C35 - 0.02937*T6 - 0.06165*G12 - 0.6854*logP$*
- *QSAR3: $\log(1/LC_{50}) = 0.8237 + 0.1711*QM6 - 0.7974*logP$*

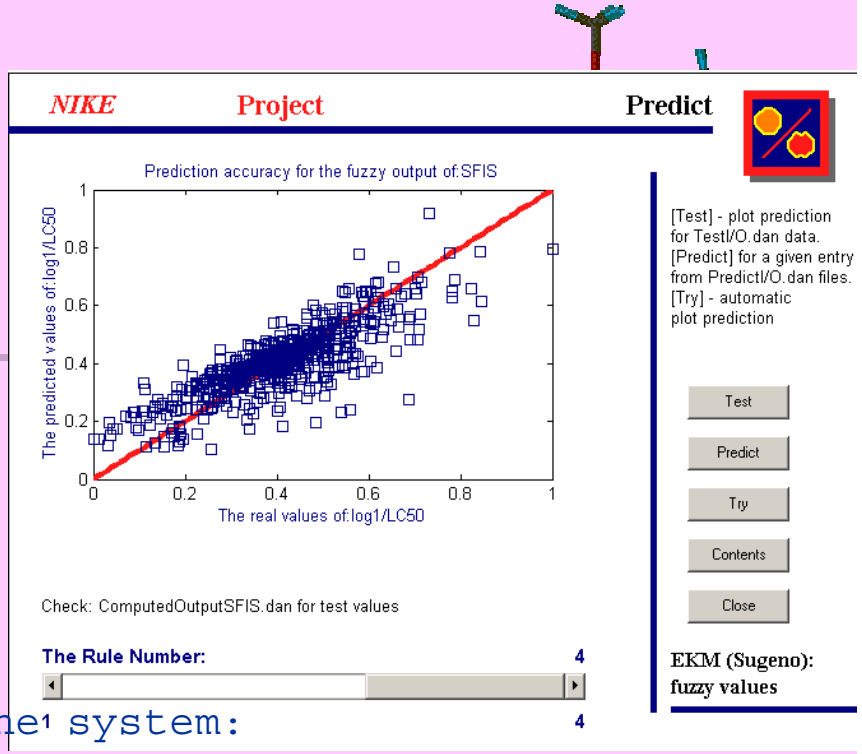




How to insert a QSAR

- Two steps
- The pre-training phase is based on a data collection generated by a selected QSAR function.
- Then the model is trained with the original data set.
- Results in some cases are better

FIS representation for QSARs



- Mamdani:
  - IF D_1 is Low AND D_2 is High THEN Tox is Medium
- zero-order Sugeno fuzzy rule:
  - IF D_1 is Low AND D_2 is High THEN Tox=k
- first order Sugeno fuzzy rule:
  - IF D_1 is Low AND D_2 is High THEN Tox=0.82+0.17**QM6*–0.79**logP*

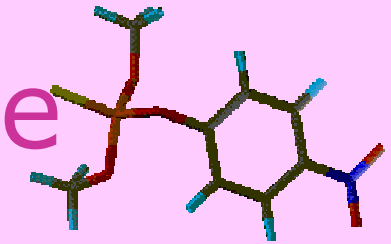
Example:

- 1. If (logP is Low) then (log1/LC50 is QSAR2) (1)
- 2. If (logP is Med) then (log1/LC50 is QSAR2) (1)
- 3. If (logP is High) then (log1/LC50 is QSAR2) (1)

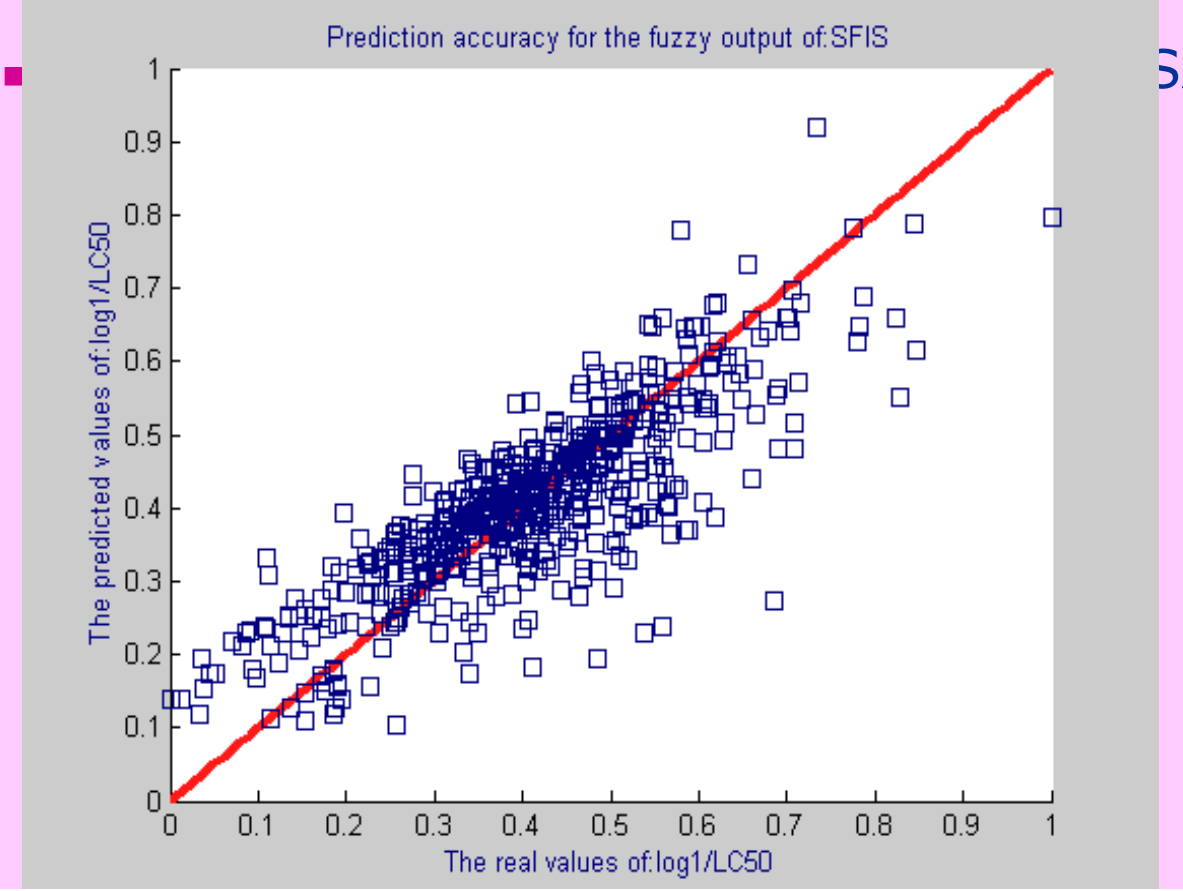
The system:

```
[Input17]
Name='logP      '
Range=[0 1]
NumMFs=3
MF1='Low':'trapmf',[0 0 0.2 0.4]
MF2='Med':'trapmf',[0.2 0.4 0.6 0.8]
MF3='High':'trapmf',[0.6 0.8 1 1]
[Output1]
Name='log1/LC50'
Range=[0 1]
NumMFs=2
MF1='QSAR2':'linear',[0 0 0.1385 0 0 -0.06703 -0.02937
     0 0 0 -0.06165 0 0 0 0 0 -0.6854 0.8779]
MF2='QSAR3':'linear',[0 0 0.1711 0 0 0 0 0 0 0 0 0 0 0
     0 0 -0.7974 0.8237]
[Rules]
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1, 1 (1) : 1
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 2, 1 (1) : 1
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 3, 1 (1) : 1
```

The regression using the fuzzy rule

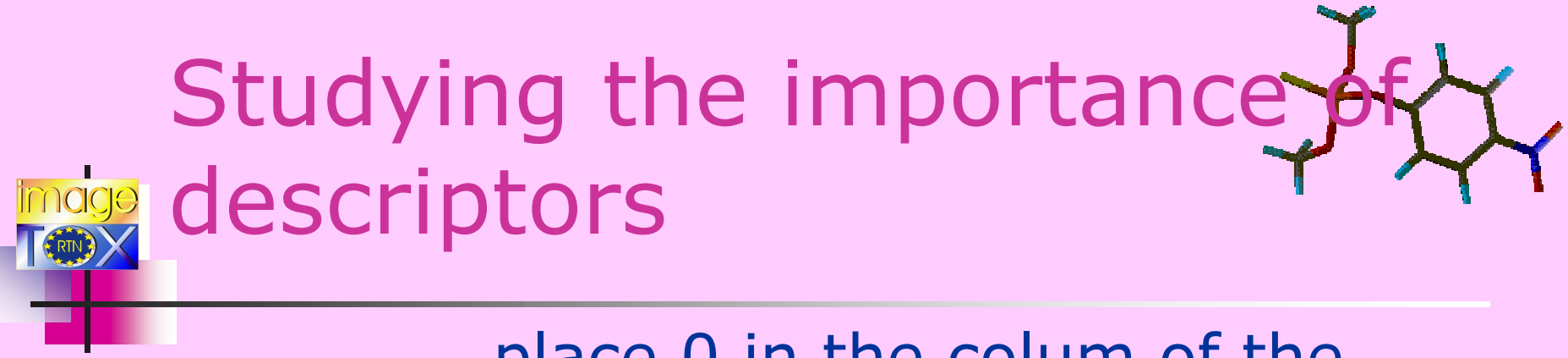


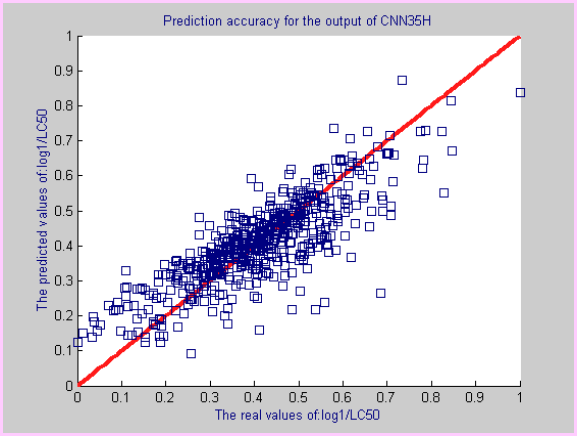
- 1. If (*logP* is Low) then (*log1/LC50* is QSAR2) (AND)
- 2. If (*logP* is Med) then (*log1/LC50* is QSAR2) (AND)
- SAR2)



Vietri 2002

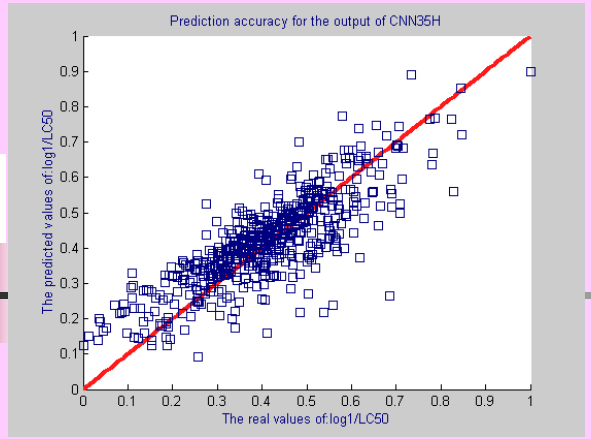
Studying the importance of descriptors

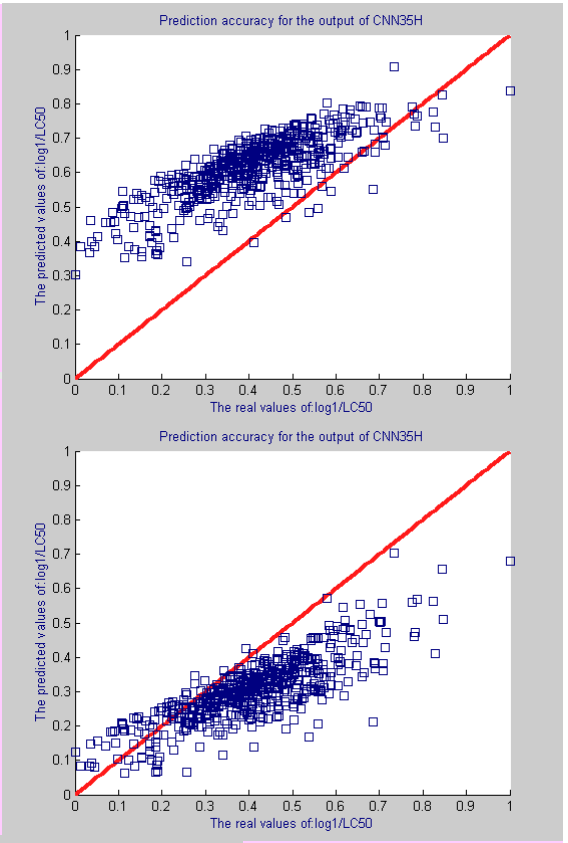


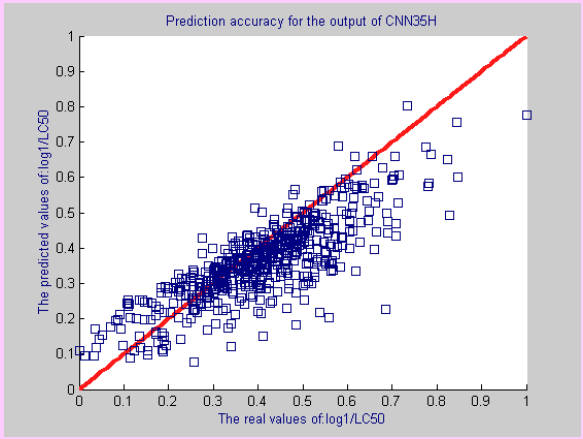


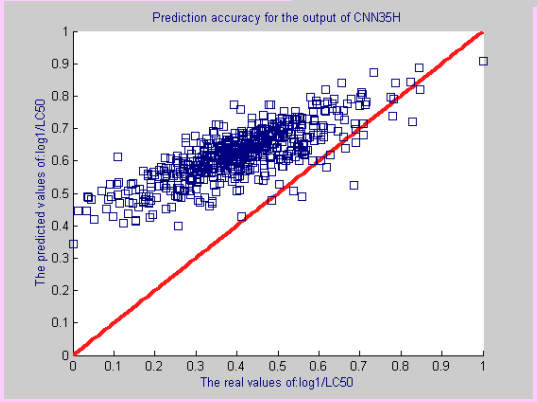
CNN performance validation (predicted data set versus real data values) for complete test data set

- place 0 in the colum of the descriptor to study, and analyze the results
- In CNN: a small increasing of absolute prediction error + predictions translation (linear dependence with the absent descriptor), or a proportional magnify of error, (rotation, a nonlinear relation between some of the current inputs)





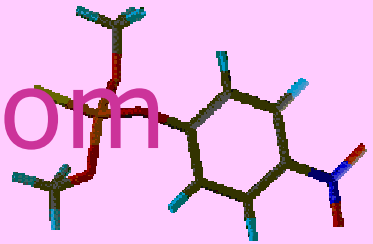




- not significant descriptor missing in test data set *QM1*, or *C9* ;

- significant descriptor missing *T6* , *G2* or the most important *logP*.

Extracting fuzzy rules from FNN



- Effect Measure Method (EMM) - combine the weights between the layers of the network.
- delete contradictory rules with small coefficient of trust:

1. if have different outputs for the same input class:

IF C9 is: Low THEN log1/LC50 is: VeryLow(42.38%)

IF C9 is: Low THEN log1/LC50 is: Medium (64.36%)

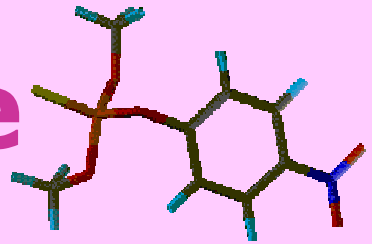
2. if big differences between the input and the output:

IF G2 is: Low THEN log1/LC50 is: Med (60.02%)

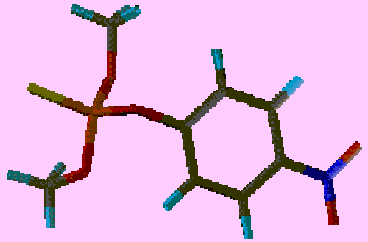
IF G2 is: Med THEN log1/LC50 is: High (33.84%)

IF G2 is: High THEN log1/LC50 is: Med (49.07%)

The integration of the experts



- three strategies
- FEM (fire each module using statistical and fuzzy integration),
- UGN (unsupervised-trained gating network for all the implied modules' fusion)
- SGN (supervised-trained gating network to integrate the expert modules).
- *Example: 5 implicit knowledge modules CNN22H, CNN35H, FNN20H, FNN25H and FNN40H + 2 explicit QSAR2, QSAR3*



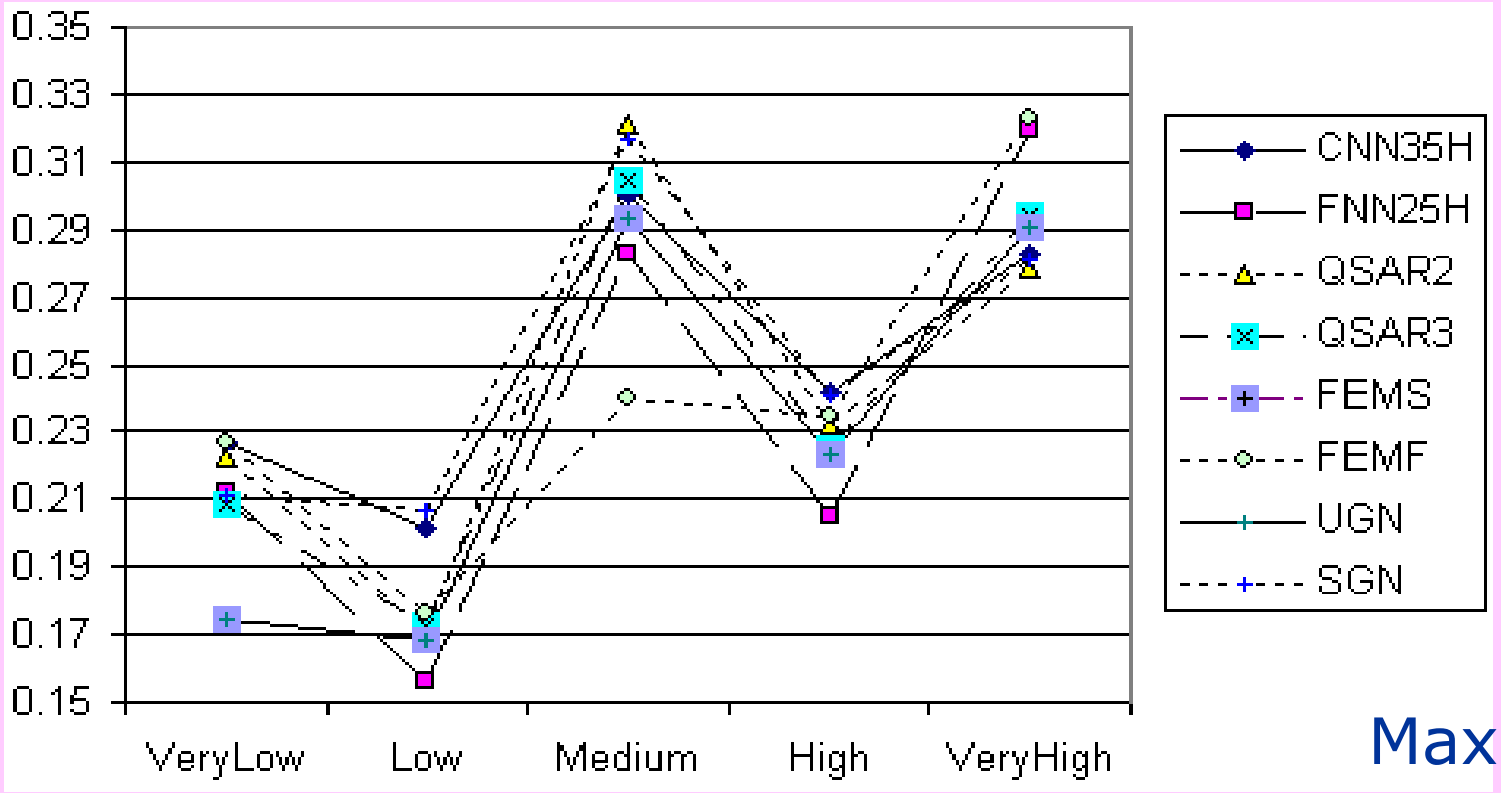
results

- The output is the *averaged* output of the modules.
- The fuzzy version uses *max T-conorm* as aggregation and *centroid* as defuzzification method
- The UGN is a 5-neurons network.
- The SGN is a CNN with 7 entries
- the number of the well predicted cases

TOXICITY	CNN35H	FNN25H	QSAR2	QSAR3	FEMS	FEMF	UGN	SGN
VeryLow (50 cases)	28	28	25	23	25	19	25	31
Low (222 cases)	197	199	199	191	201	188	201	194
Medium (245 cases)	199	211	201	209	210	217	210	197
High (46 cases)	26	30	28	28	30	25	30	26
VeryHigh (5 cases)	1	1	1	1	1	0	1	1
Total cases (568)	451	469	454	452	467	449	467	449
Percentage	79.40%	82.57%	79.93%	79.58%	82.22%	79.05%	82.22%	79.05%

The accuracy of prediction

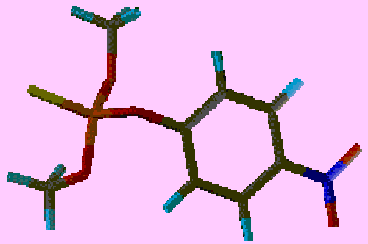
- accuracy of prediction by fuzzy classes



Max error

Average	*0.0622*	*0.0565*	*0.0629*	*0.0651*	*0.0585*	*0.0633*	*0.0585*	*0.069*

Vietri 2002

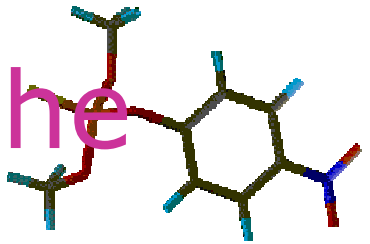


Hybrid system

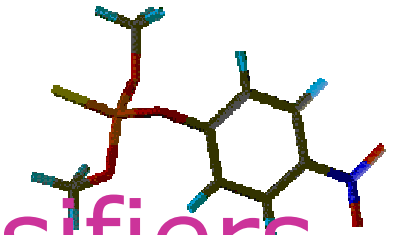
The predictions are up to 5% more accurate than those of the single approaches.

the 568 compounds used in this study do not provide a best coverage of the problem domain

Conclusions: VALUE of the predictor



- Is better than random guessing?
- ROC space analysis and the predictive toxicology challenge (Toivonen et al. 2002)



ROC for comparing classifiers

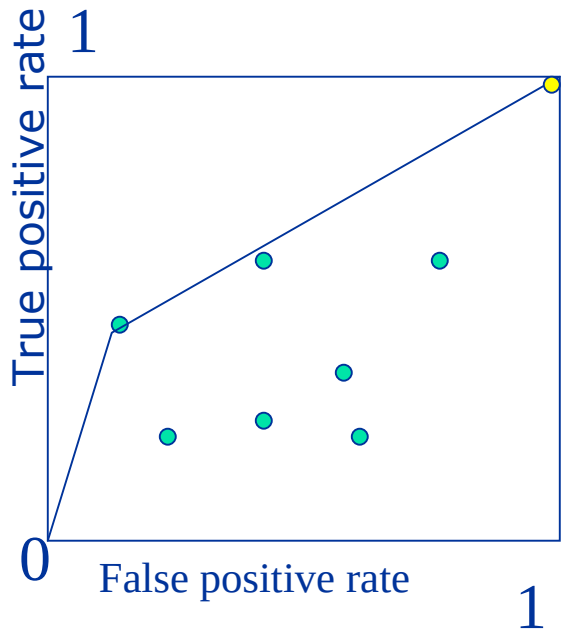
In a binary classification we can study the Receiver Operating Characteristic (ROC) space where true positive rate is plotted against false positive rate

true positive rate = sensitivity

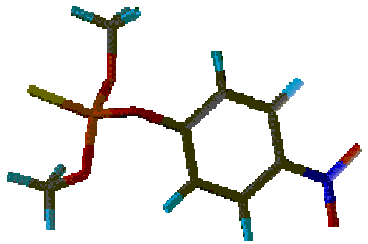
false positive rate = 1 – specificity

- Sensitivity = probability that it is predicted positive and it is positive
- Specificity = probability that it is predicted negative and it is negative
- (Bradley 95 to compare classifiers)

Graphic ROC representation

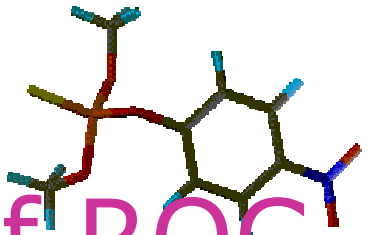


- In ROC space, the true positive rate, TP , is plotted on the Y axis and the false positive rate, FP , is plotted on the X axis. It is computed from the misclassification matrices
- ROC space is a square where N models are represented in N points.
- Convex hull from points (0,0) and (1,1): *the closer the curve to the left hand and top borders, the more accurate the predictor* (in terms of



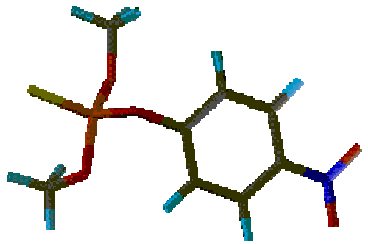
AUC

- A area under the ROC curve (AUC) (Bradley, 1997) = probability that a randomly chosen positive instance will be rated higher than a negative instance. Because random guessing produces the diagonal line between (0; 0) and (1; 1) which has an area of 0.5, no realistic classifier should have an AUC less than 0.5.
- ROC curves may be misleading: we cannot tell how much of the observed variation is due to the training#test partition. ....but
- AUC is useful in drawing conclusions across a variety of data sets for which the true misclassification costs are unknown
- **If there is not a single dominating ROC curve, multiple classifiers can be combined to form**



Statistical significance of ROC

- (Toivonen et al 2002)
- If a classifier C gives Nc predicted positive, the null hypothesis is that the selection of Nc is statistically independent of their true class.
- p value of C is the probability that random selection of Nc will give the same result as obtained by C
- METHOD: For each C compute p on all the Nc (obtained with $\chi 2$ test)
  - The smallest the value of p, the best the classifier (under the null hypothesis p values are uniformly distributed)
  - Plot in the ROC space and analyze



Conclusions

- …bad news (from the challenge – see Toivonen)
- The reason? Violation of specificity criteria
- The future? More systematic way to integrate expert knowledge in the loop.
- Mixture of experts help.