New, Transparent, Statistical Approaches to Toxicity Prediction

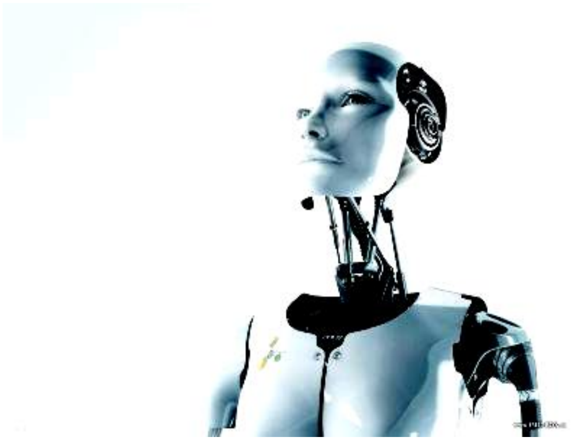
Tokyo / Osaka
March 2014

Thierry Hanser

thierry.hanser@lhasalimited.org

Toxicity prediction
The context of decision support

An ideal world

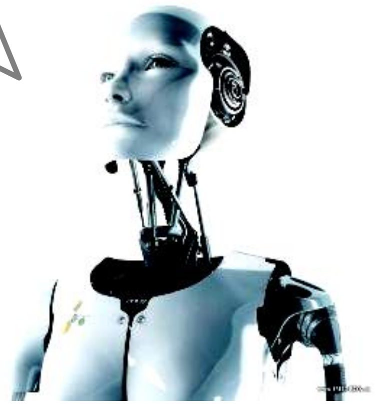


Hello Sarah !

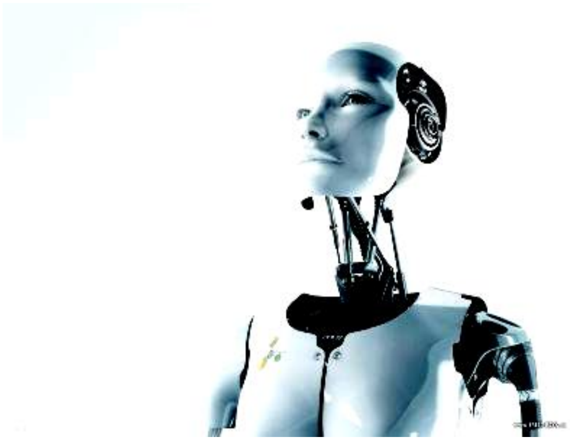
An ideal world

Hello Thierry!

How may I help you?



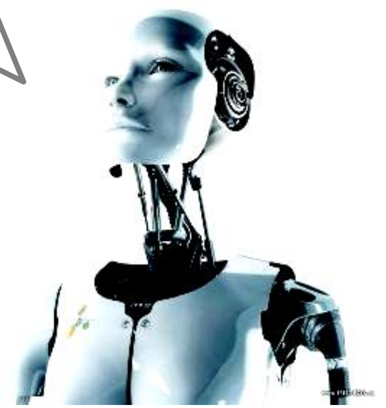
An ideal world



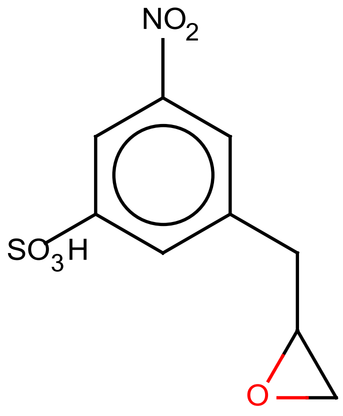
How is your knowledge about mutagenic compounds ?

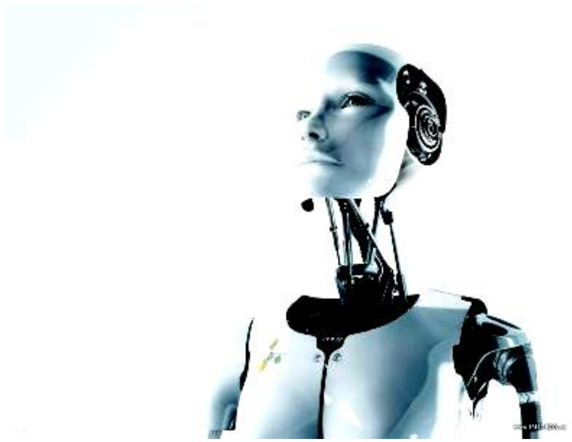
An ideal world

I have passed several tests with satisfactory results.



An ideal world

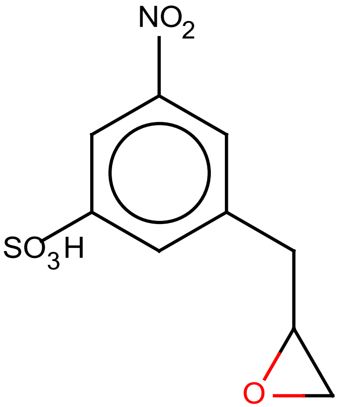




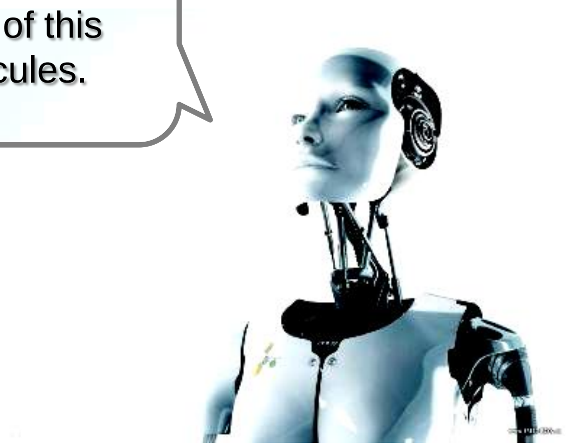
That's reassuring.

Do you know about this type of compound in the context of mutagenicity ?

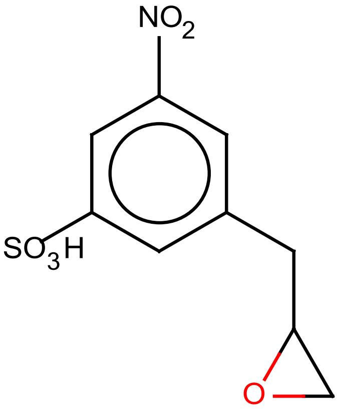
An ideal world

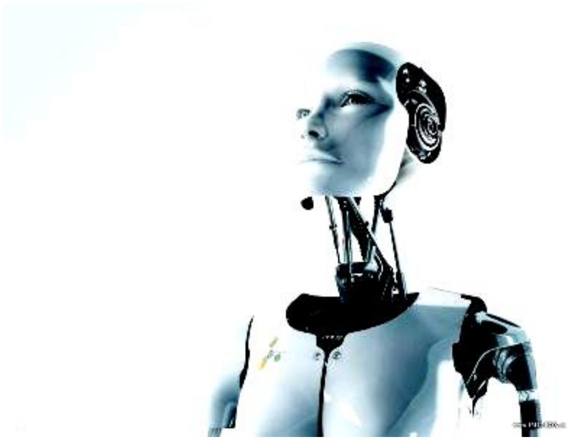


Yes, I have some understanding of this class of molecules.



An ideal world

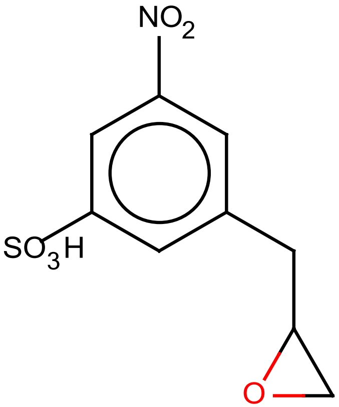




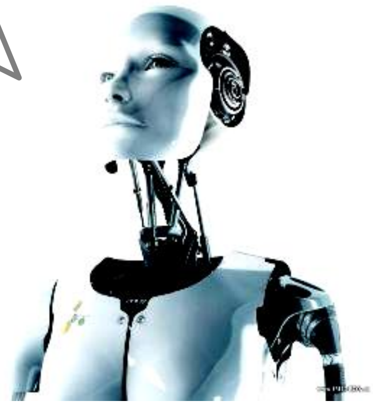
Great.

Could you please tell me if this compound is a mutagen?

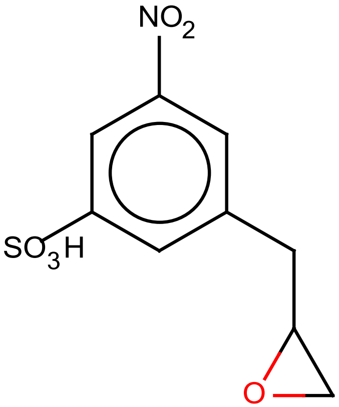
An ideal world

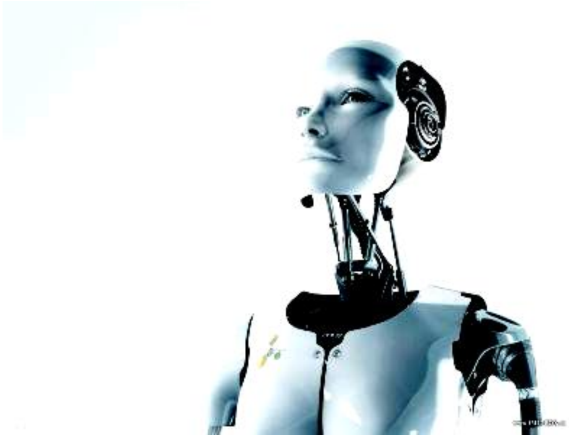


According to my knowledge it should be a mutagen; I am afraid so.



An ideal world

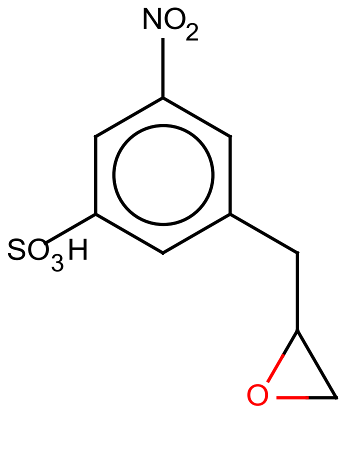




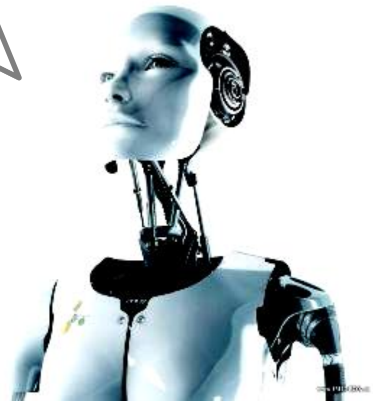
Oh, that's too bad!

Are you sure?

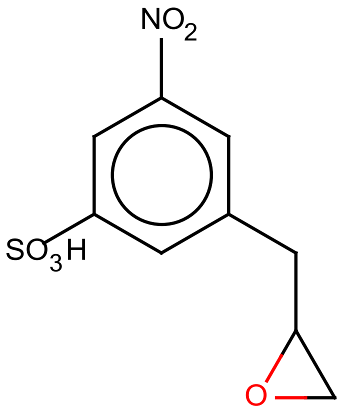
An ideal world

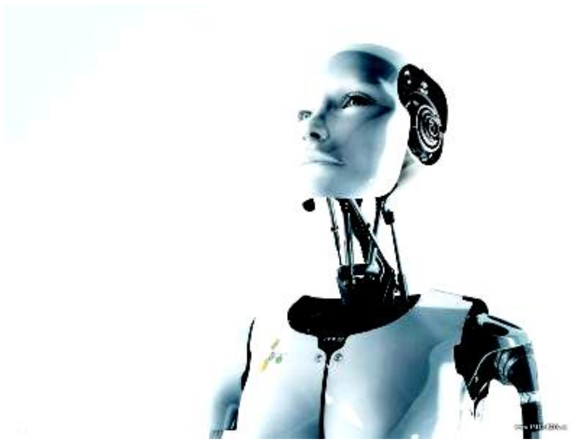


Yes I am very confident that this molecule is a mutagen!



An ideal world

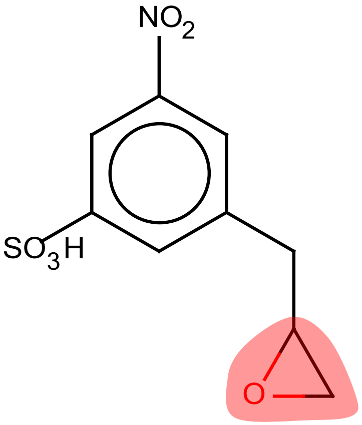




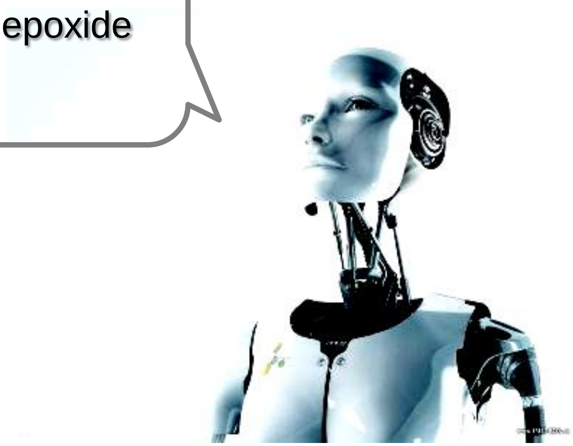
I see

Do you know the cause?

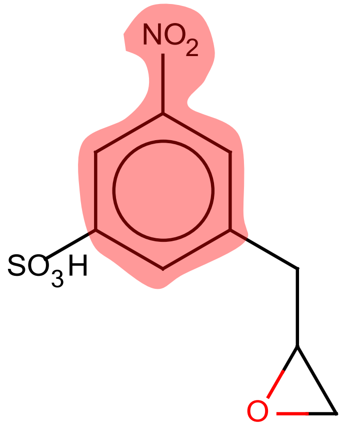
An ideal world



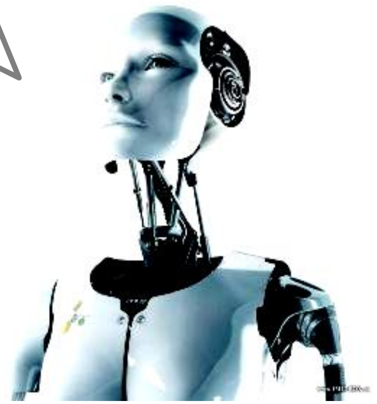
The main reason is the presence of an epoxide group



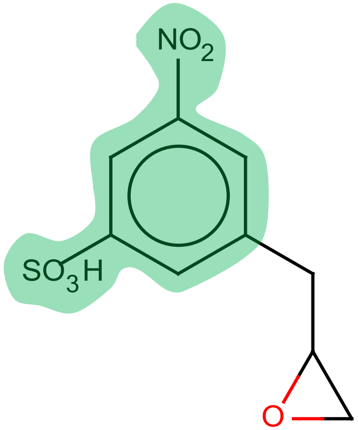
An ideal world



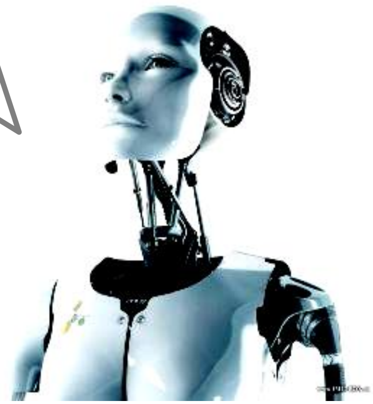
Additionally there is also a risk of mutagenicity due to the aromatic nitro group!



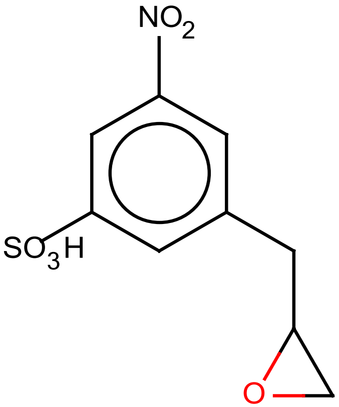
An ideal world

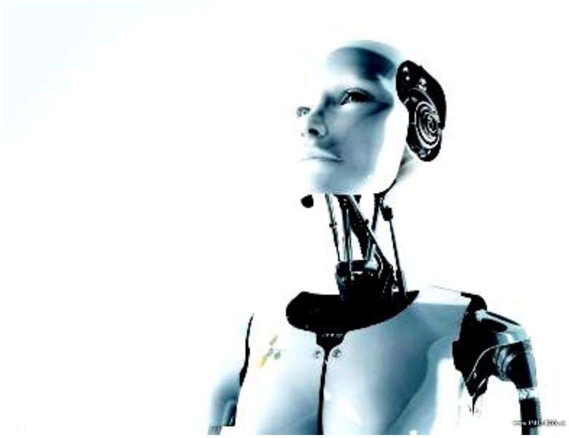


However, it seems that the sulfonic acid substituent deactivates the mutagenic effect of the nitro group



An ideal world

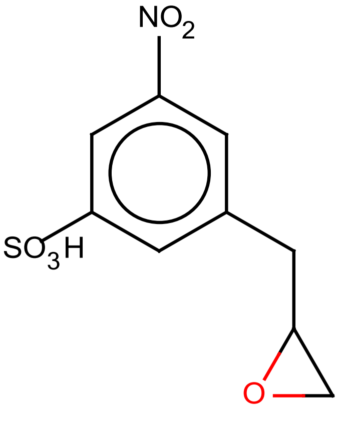




That's interesting indeed.

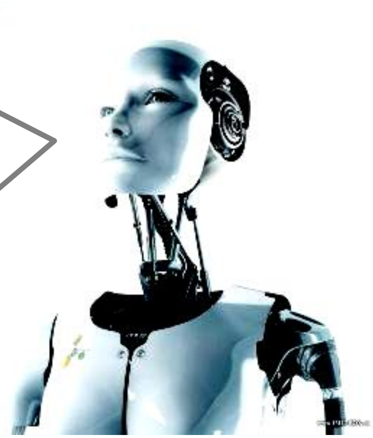
Have you got any evidence?

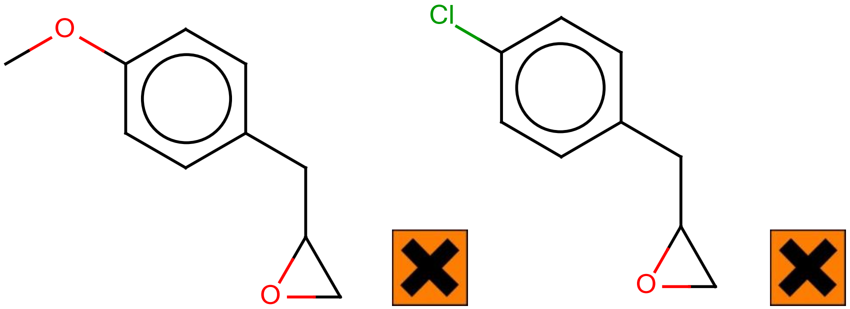
An ideal world



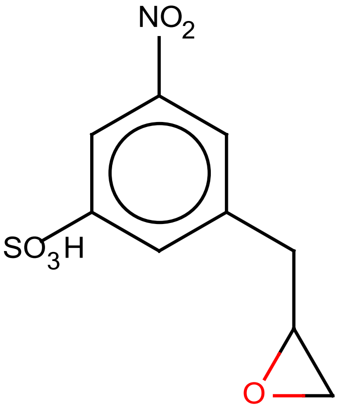
Yes.

Ames tests have shown two similar molecules containing epoxide groups as being mutagens

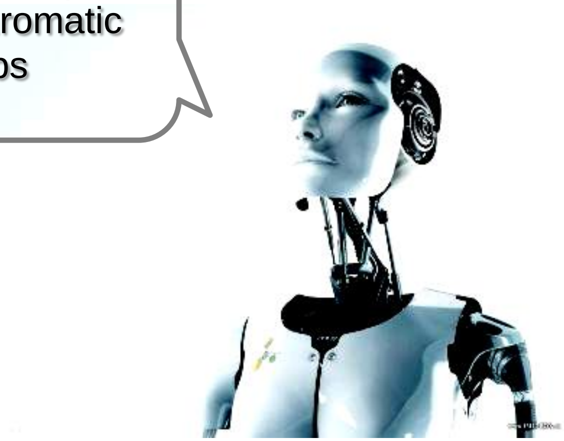


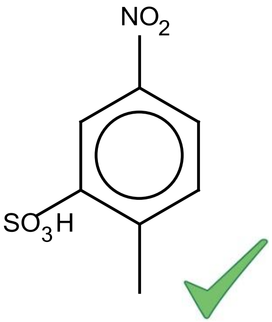


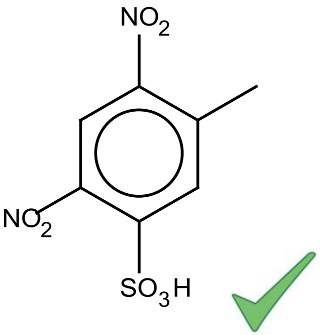
An ideal world



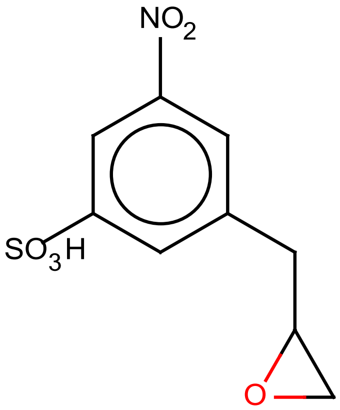
And these are examples of deactivated aromatic nitro groups



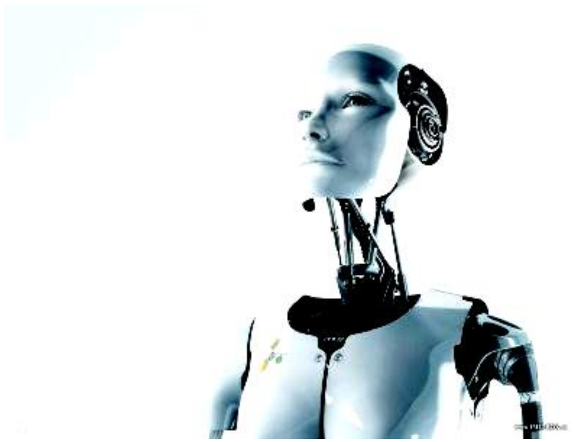




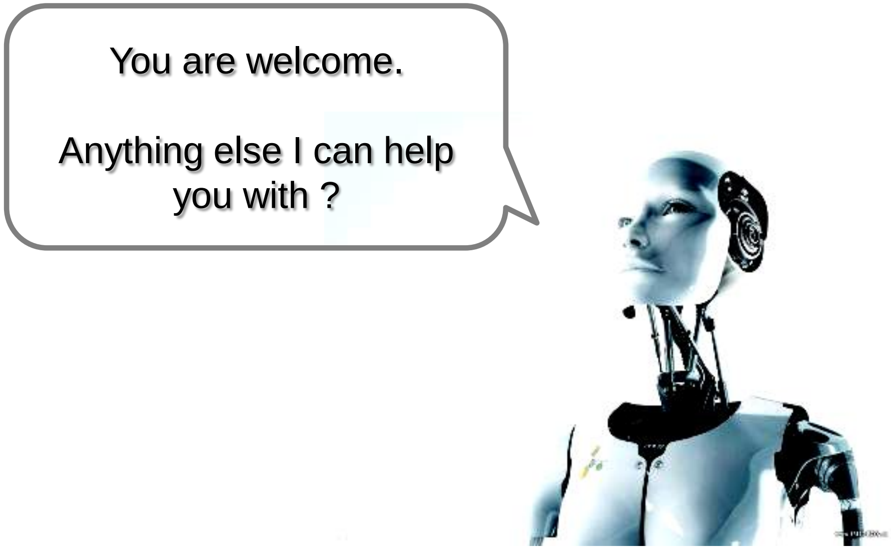
An ideal world



This will really help me take a decision about this compound.
Thank you very much!



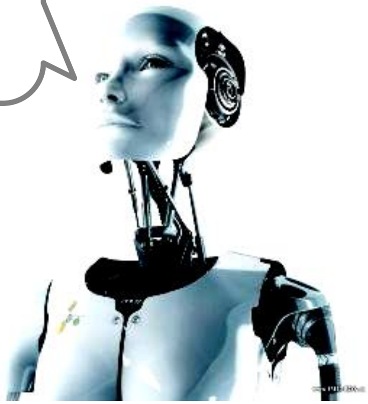
An ideal world



Model accuracy estimate

Model accuracy

I have passed several tests with satisfactory results.



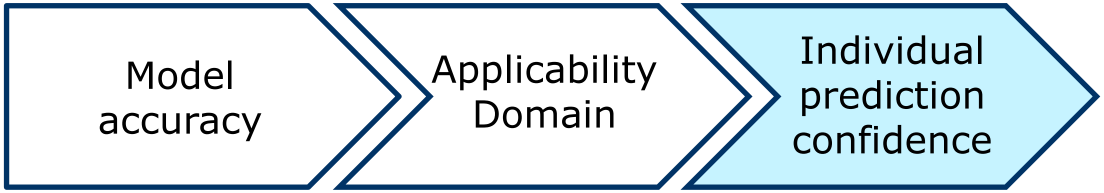
Applicability domain

Model accuracy

Applicability Domain

Yes, I have some understanding of this class of molecules.

Individual prediction confidence

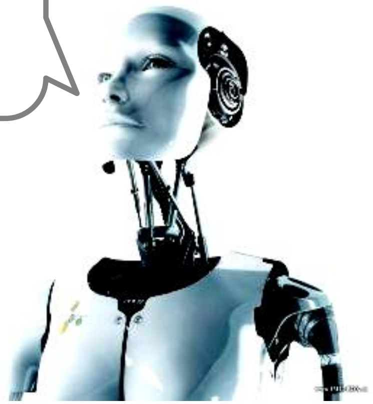


Model accuracy

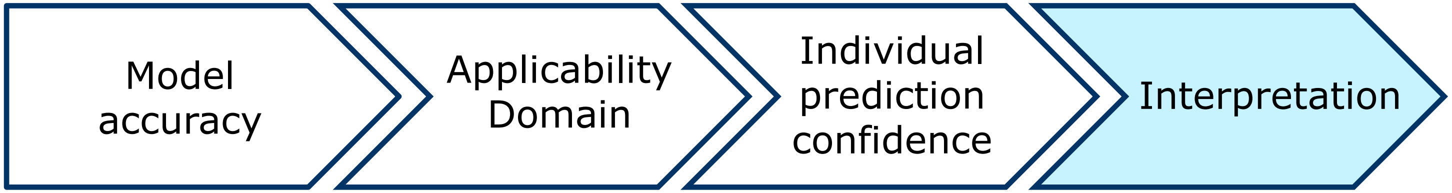
Applicability Domain

Individual prediction confidence

Yes I am very confident that this molecule is a mutagen!



Explanation

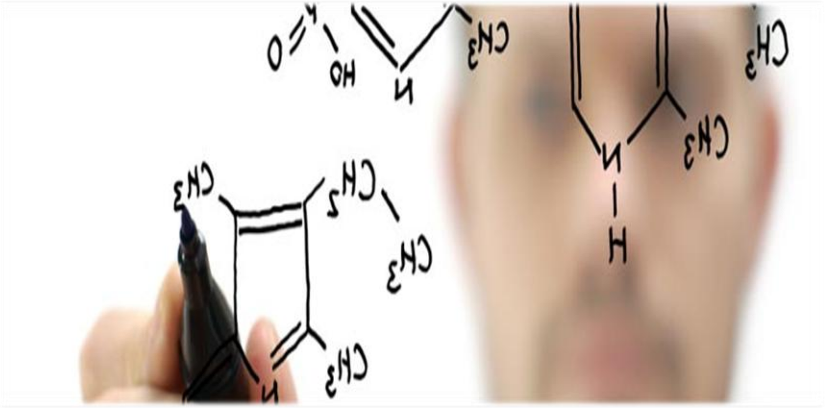


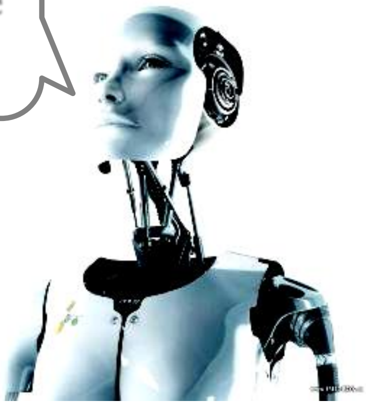
Model accuracy

Applicability Domain

Individual prediction confidence

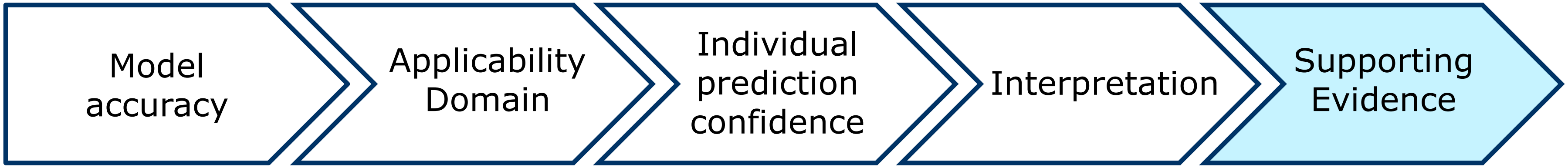
Interpretation





The main reason is probably the presence of an epoxide group

Supporting evidences



Model accuracy

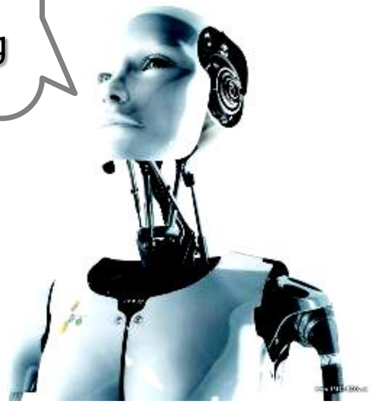
Applicability Domain

Individual prediction confidence

Interpretation

Supporting Evidence

Ames tests have shown two similar molecules containing epoxide groups as being mutagens



Interpretability accuracy trade-off

Learned
SAR

Less Accurate
(More interpretable)

Learned
SAR

More Accurate
(less interpretable)

Interpretability accuracy trade-off

Learned SAR

More intepretable
(lower accuracy)

Learned SAR

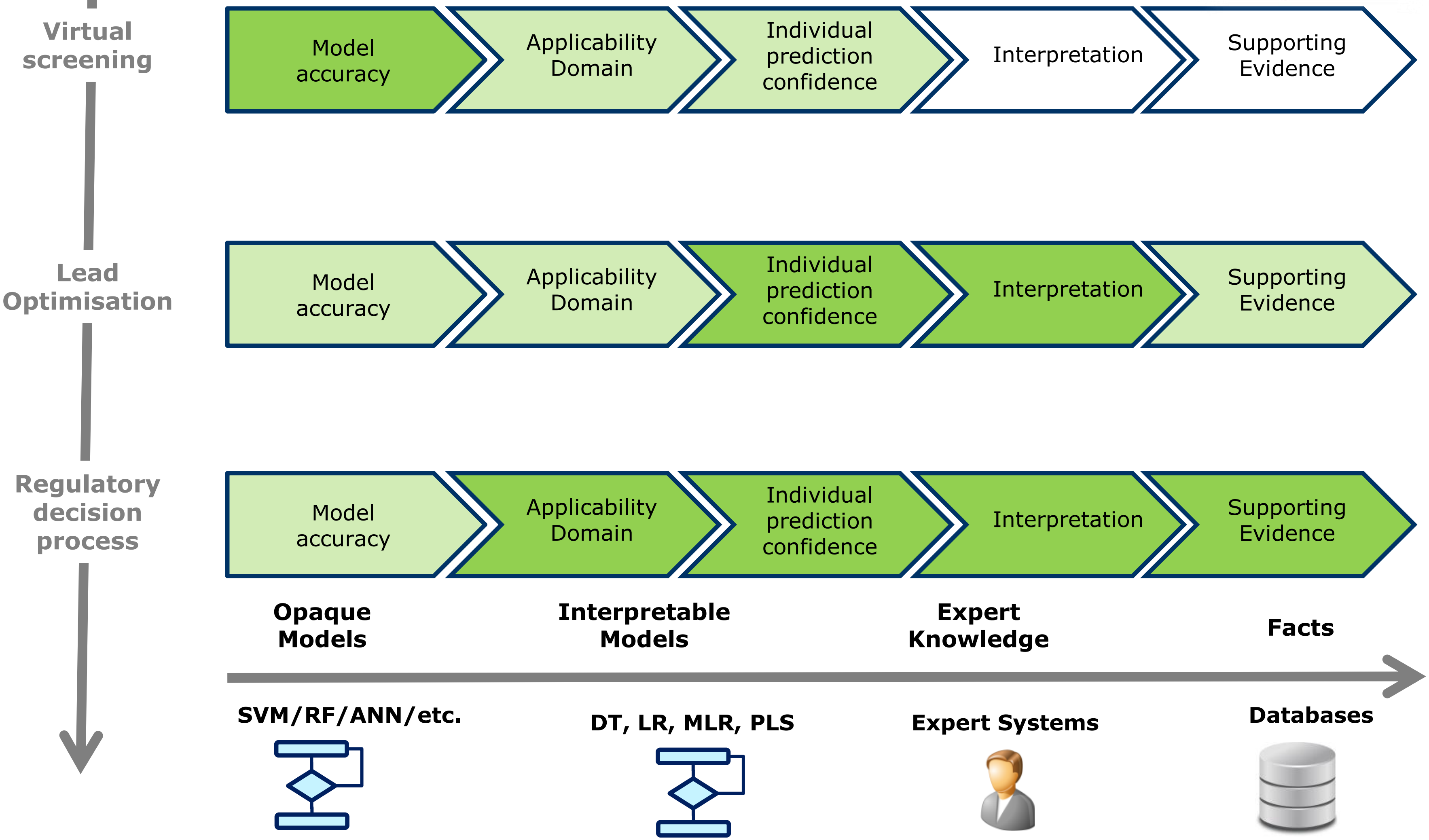
Less interpretable
(higher accuracy)

Interpretability accuracy trade-off

Learned SAR

Useful trade-off depending on the use-case

Model vs Knowledge vs Facts



Virtual screening

Model accuracy → Applicability Domain → Individual prediction confidence → Interpretation → Supporting Evidence

Lead Optimisation

Model accuracy → Applicability Domain → Individual prediction confidence → Interpretation → Supporting Evidence

Regulatory decision process

Model accuracy → Applicability Domain → Individual prediction confidence → Interpretation → Supporting Evidence

Opaque Models — **Interpretable Models** — **Expert Knowledge** — **Facts**

SVM/RF/ANN/etc.

DT, LR, MLR, PLS

Expert Systems

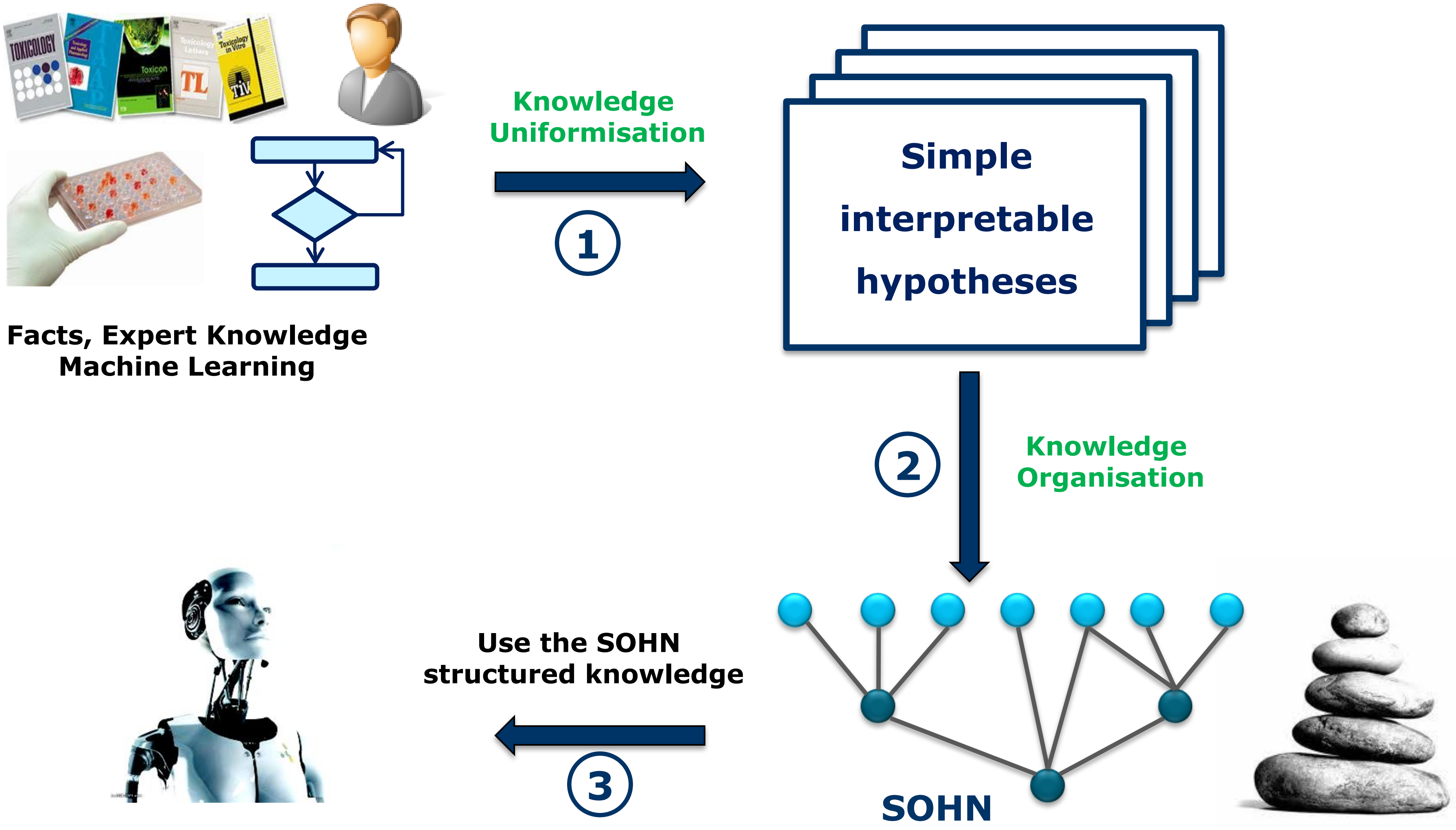
Databases

A priori knowledge unification

(**S**elf **O**rganizing **H**ypothesis **N**etworks)

Combining knowledge before prediction

Towards a unified framework



Facts, Expert Knowledge
Machine Learning

Knowledge Uniformisation

1

Simple interpretable hypotheses

2

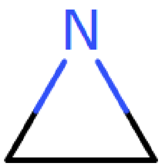
Knowledge Organisation

SOHN

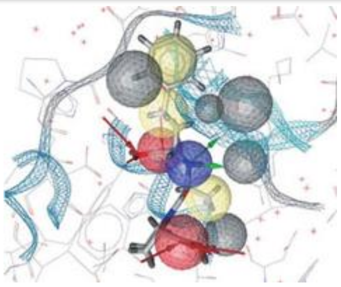
Use the SOHN structured knowledge

3

Hypothesis types



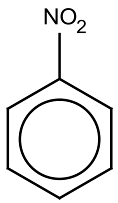
Structural Hypothesis



Pharmacophore Hypothesis

$2 < \mathrm{logP} < 3$

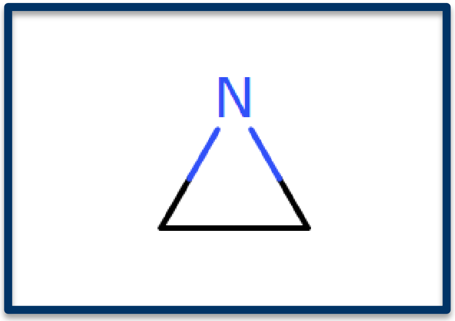
Physico-Chemical Hypothesis



$\tau > 0.8$

Similarity Hypothesis

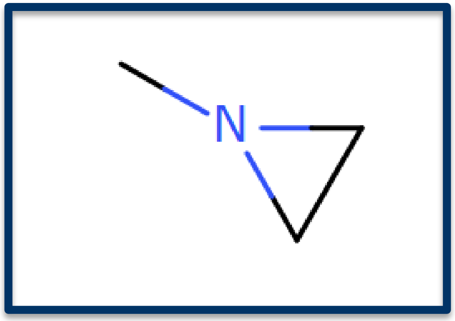
Hypotheses hierarchy



A

>

more general than



B

1 < logP < 4

C

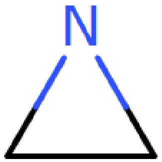
>

more general than

2 < logP < 3

D

Hypotheses hierarchy



A

$\overset{?}{>}$

more general than

2 < logP < 3

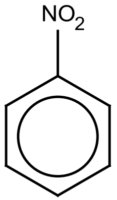
B

1 < logP < 4

C

$\overset{?}{>}$

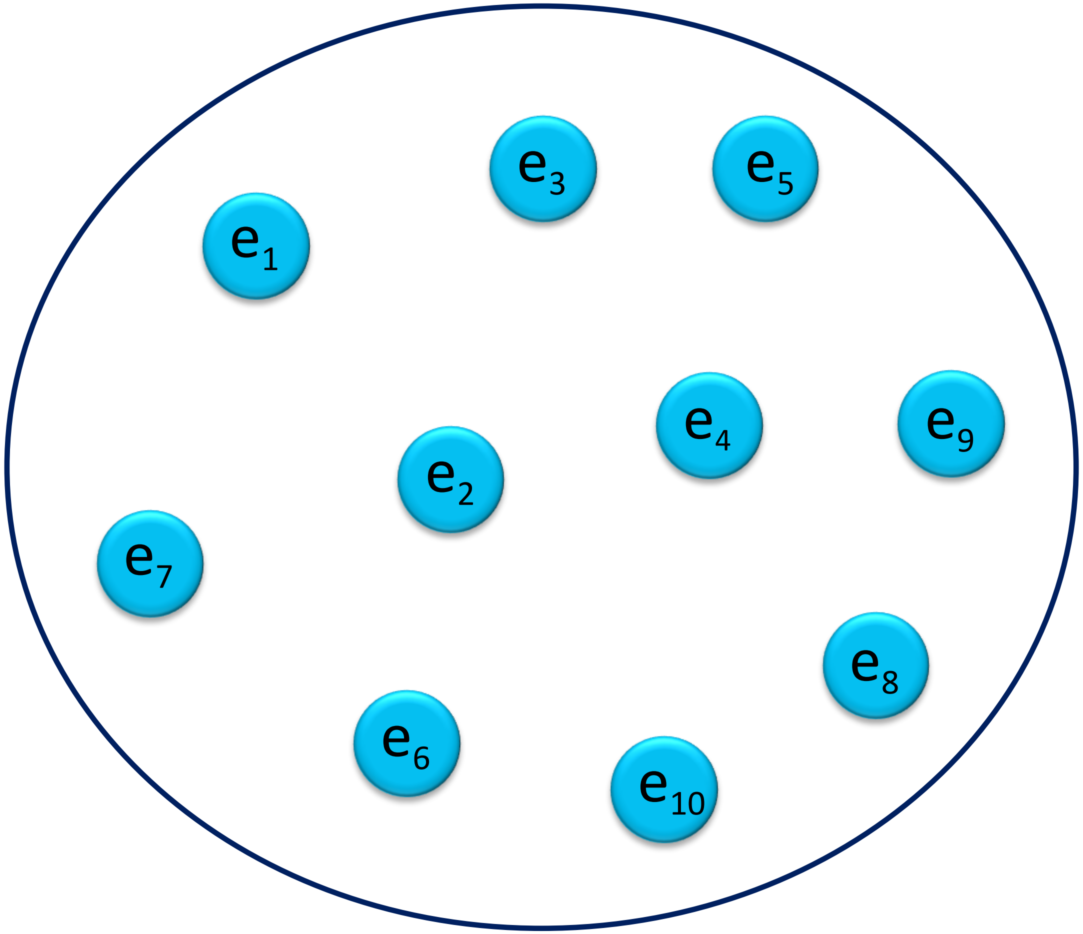
more general than



$\tau > 0.8$

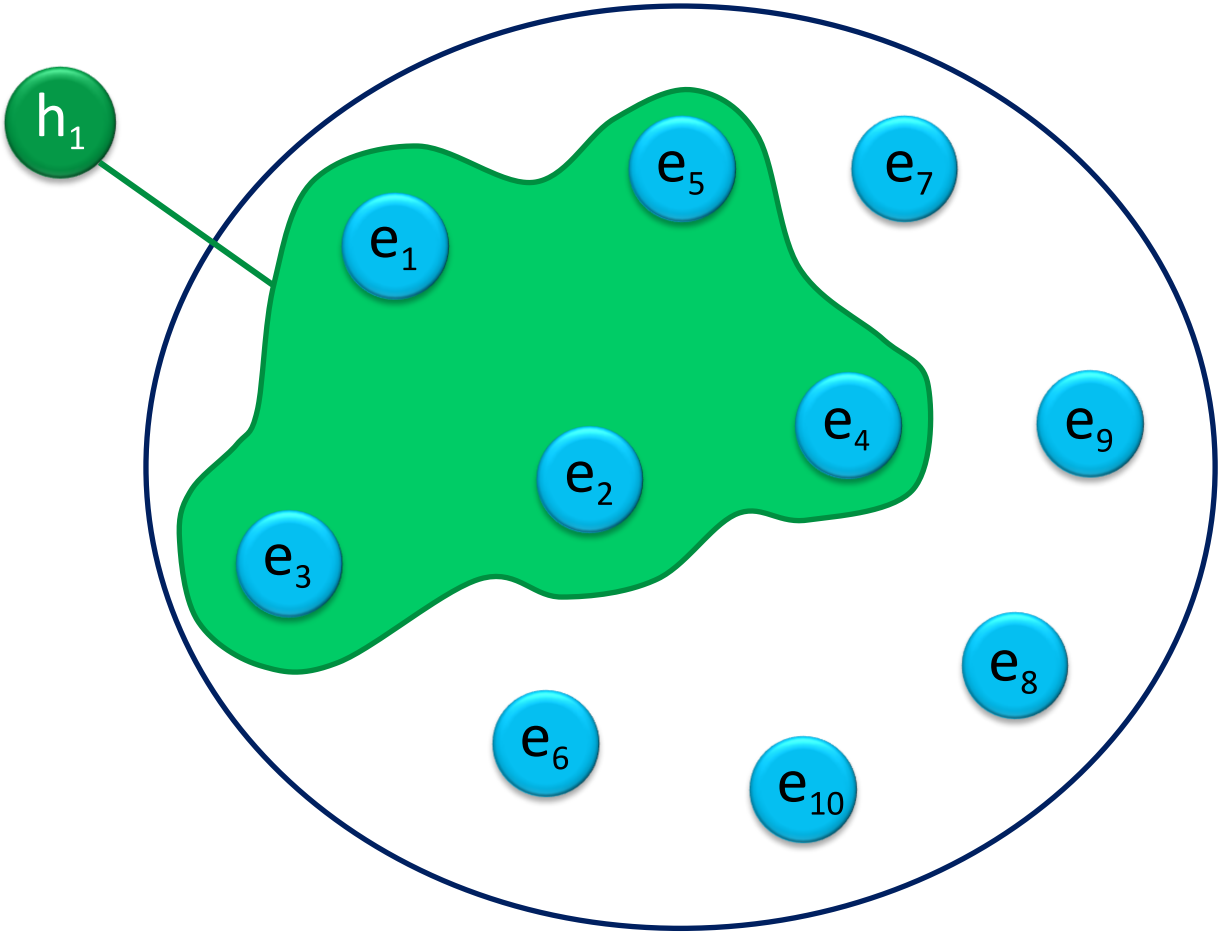
D

Hypotheses hierarchy



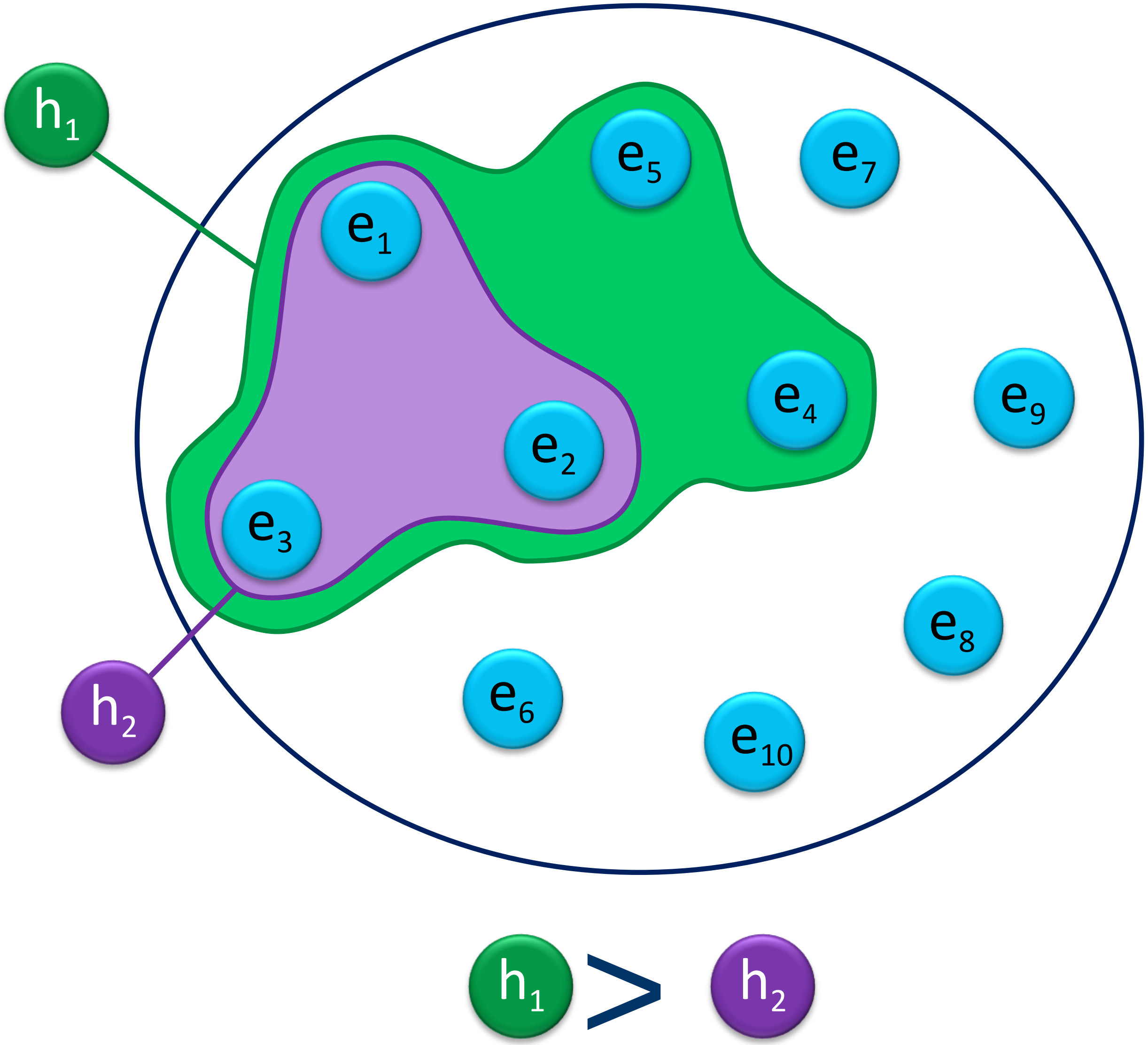
Reference dataset
(factual knowledge)

Hypotheses hierarchy



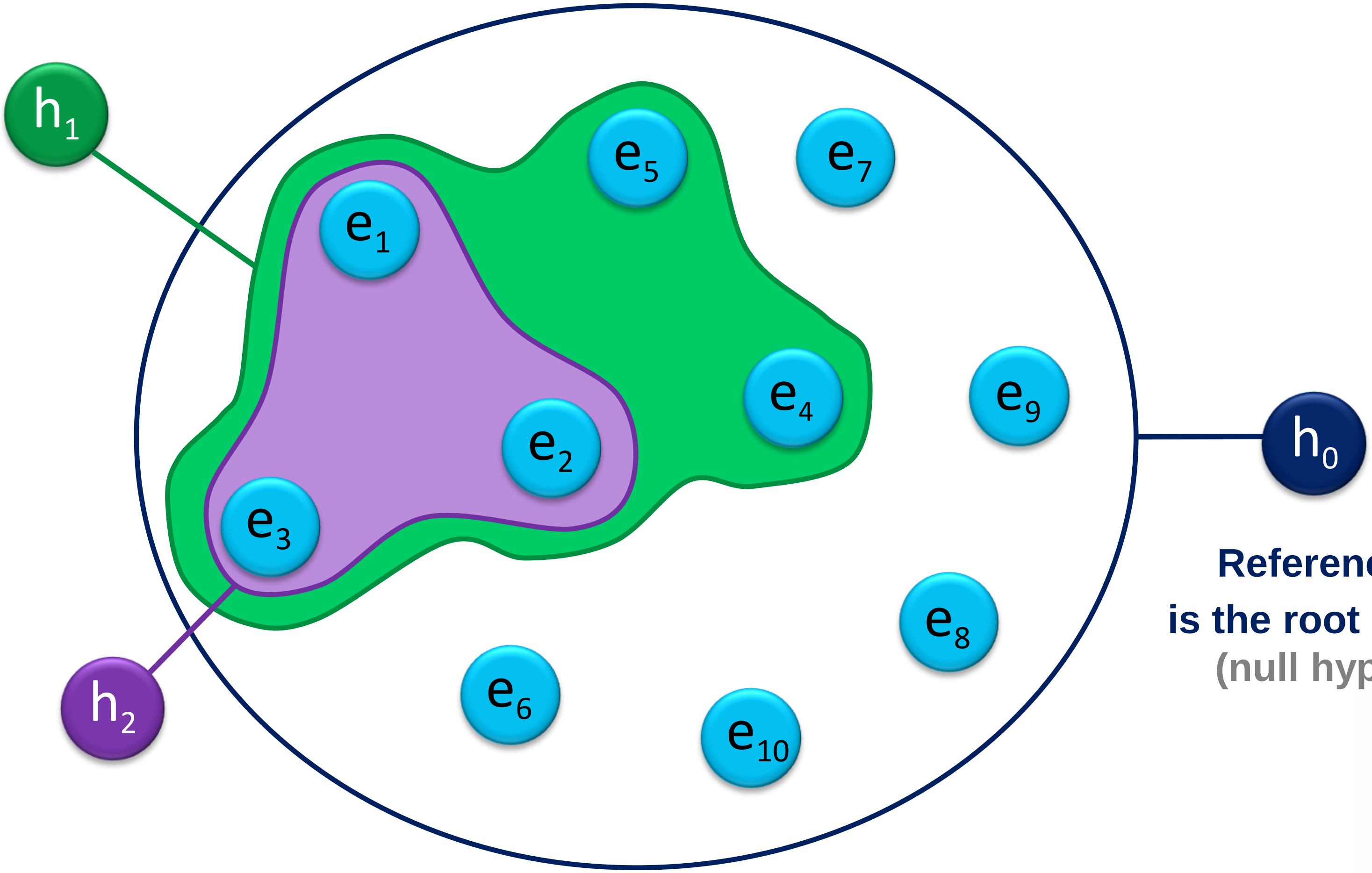
Reference dataset

Hypotheses hierarchy



Reference dataset

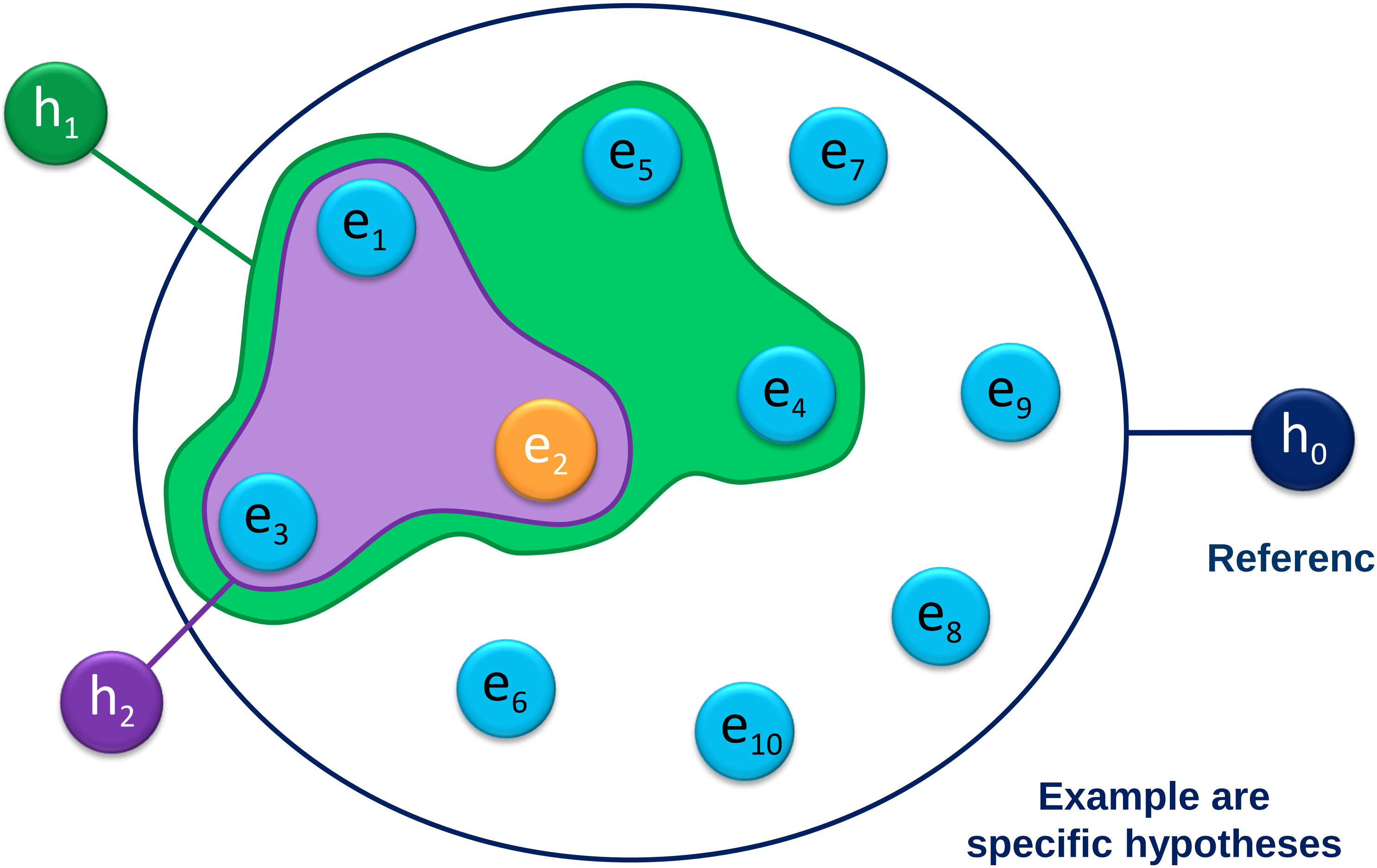
Hypotheses hierarchy



Reference dataset
is the root hypothesis
(null hypothesis)

$h_0 > h_1 > h_2 >$

Hypotheses hierarchy



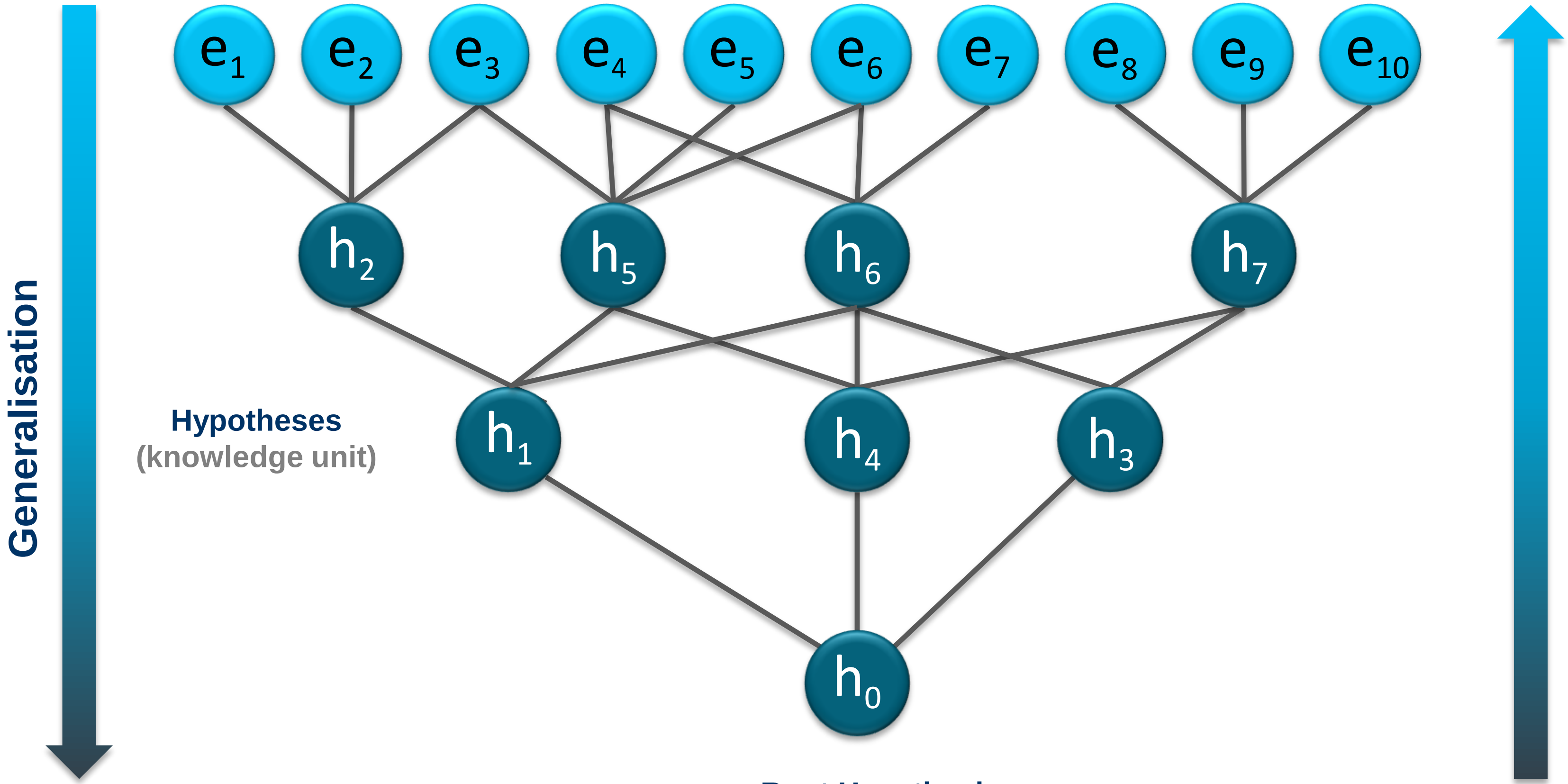
Reference dataset

Example are specific hypotheses

$h_0 > h_1 > h_2 > e_2$

Hypotheses hierarchy

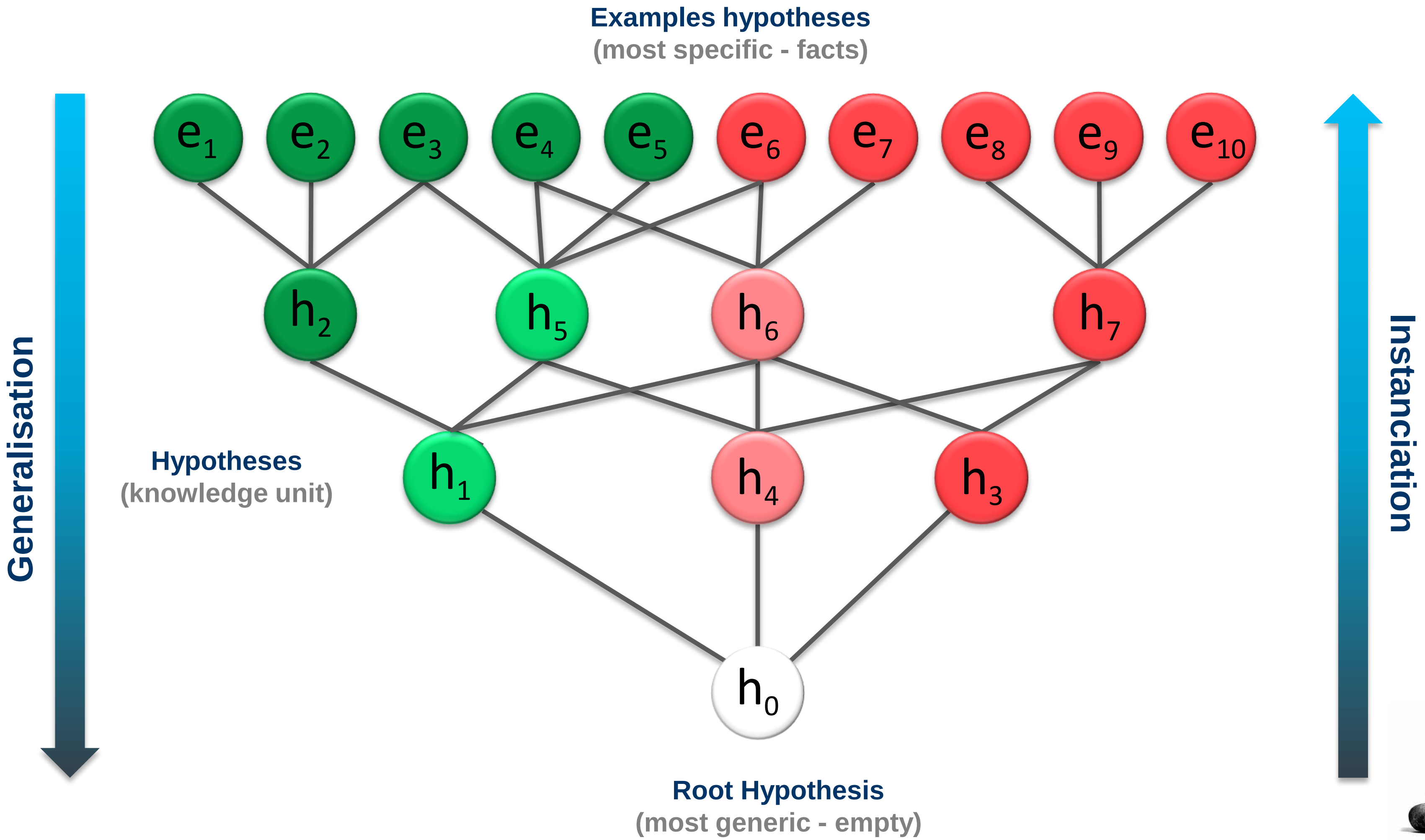
Example hypotheses
(most specific - facts)



Hypotheses
(knowledge unit)

Root Hypothesis
(most generic - null hypothesis)

Hypotheses hierarchy



Examples hypotheses
(most specific - facts)

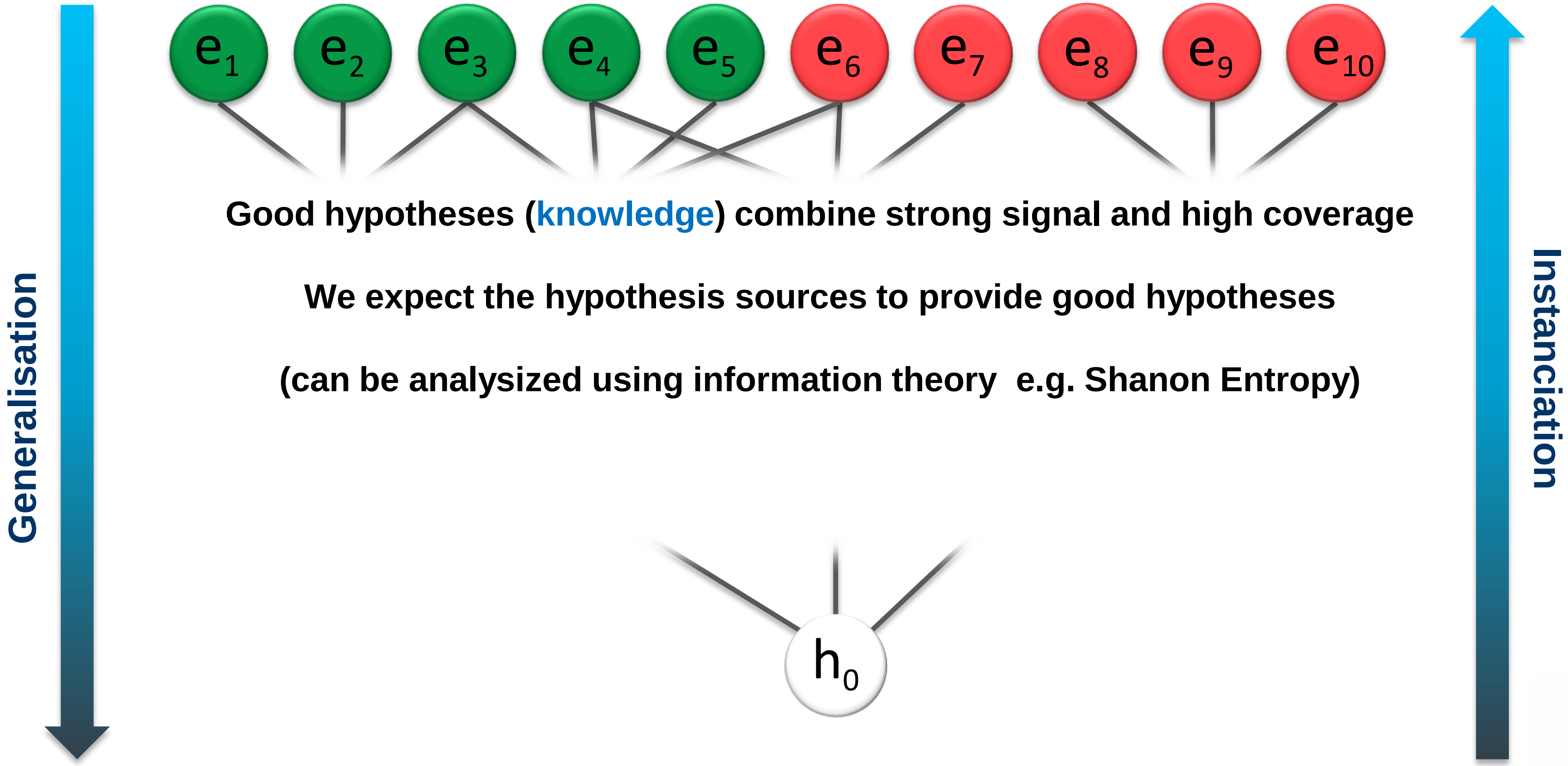
Hypotheses
(knowledge unit)

Root Hypothesis
(most generic - empty)

Generalisation

Instanciation

Hypotheses hierarchy



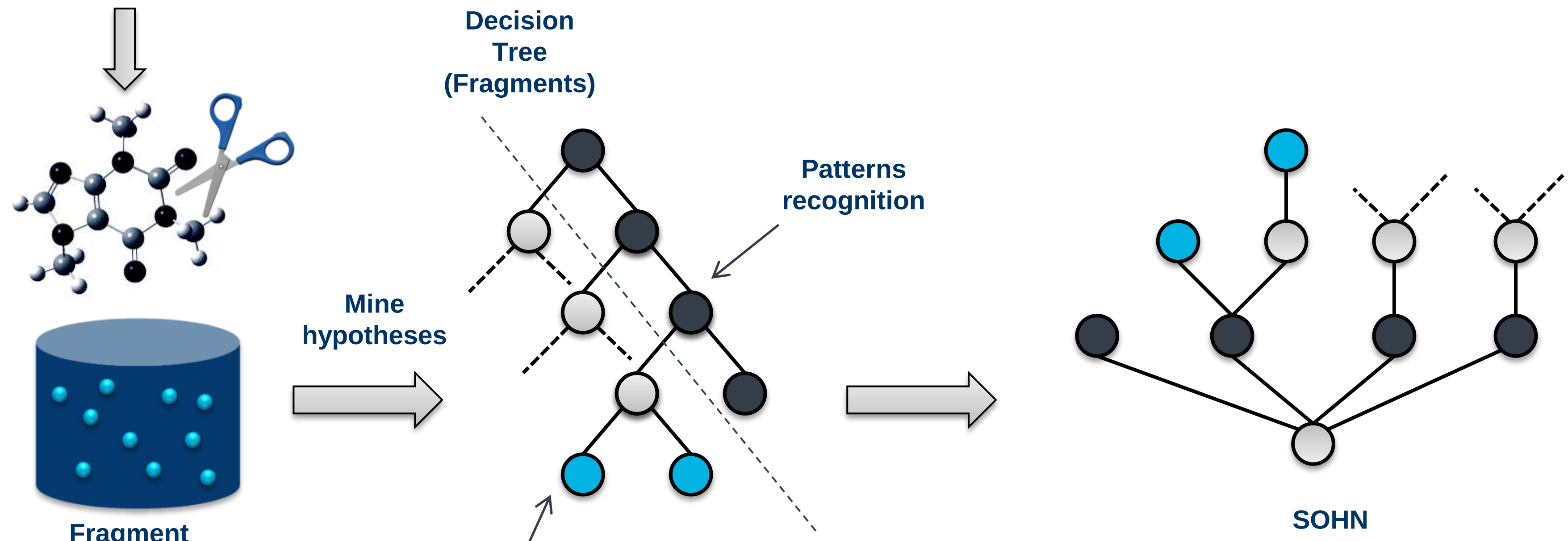
Good hypotheses (knowledge) combine strong signal and high coverage

We expect the hypothesis sources to provide good hypotheses

(can be analysized using information theory e.g. Shanon Entropy)

Example

Experimental data



Decision Tree (Fragments)

Patterns recognition

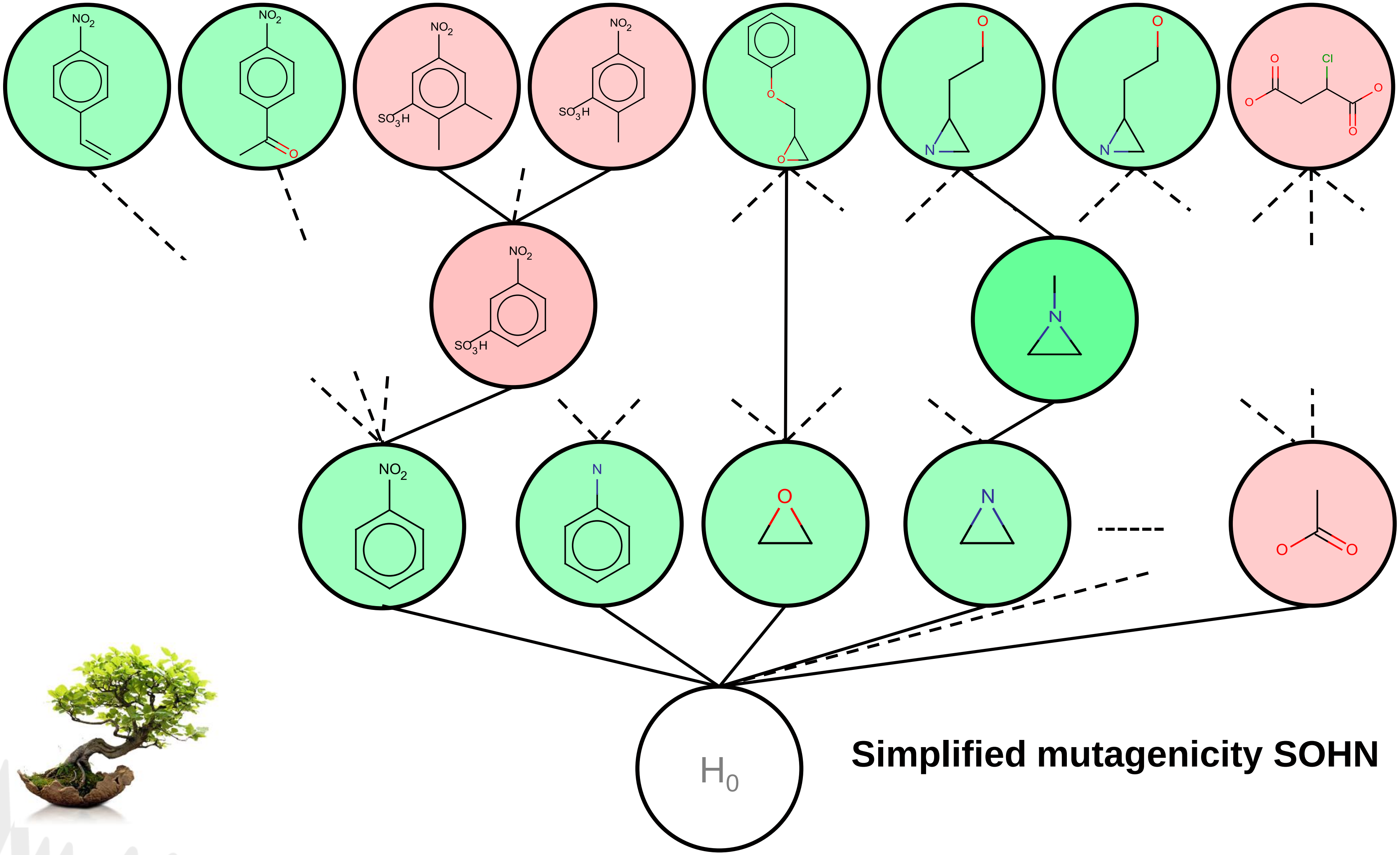
Mine hypotheses

Fragment dictionary

Patterns refinement

SOHN (Hypotheses)

Example



Simplified mutagenicity SOHN

Knowledge Mining

Paths contain valuable information

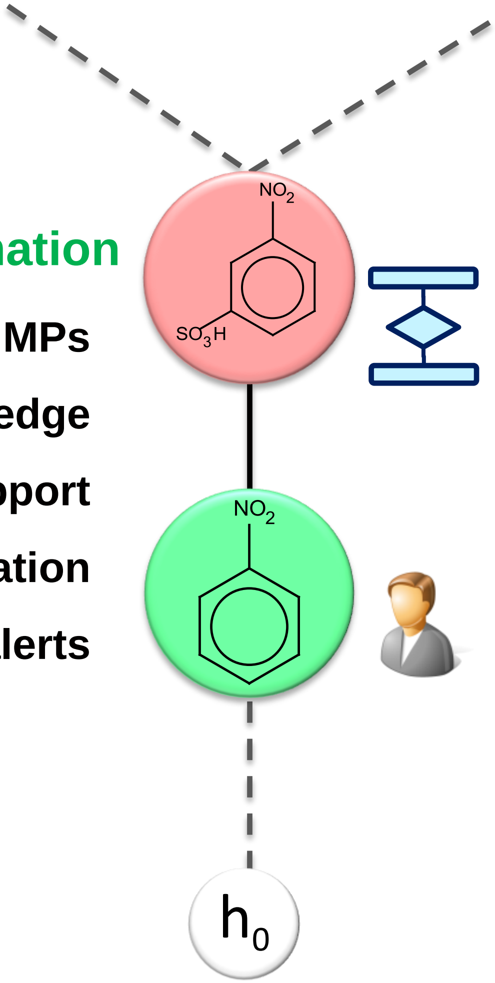
Identify Activity Cliffs / MMPs

Refine expert knowledge

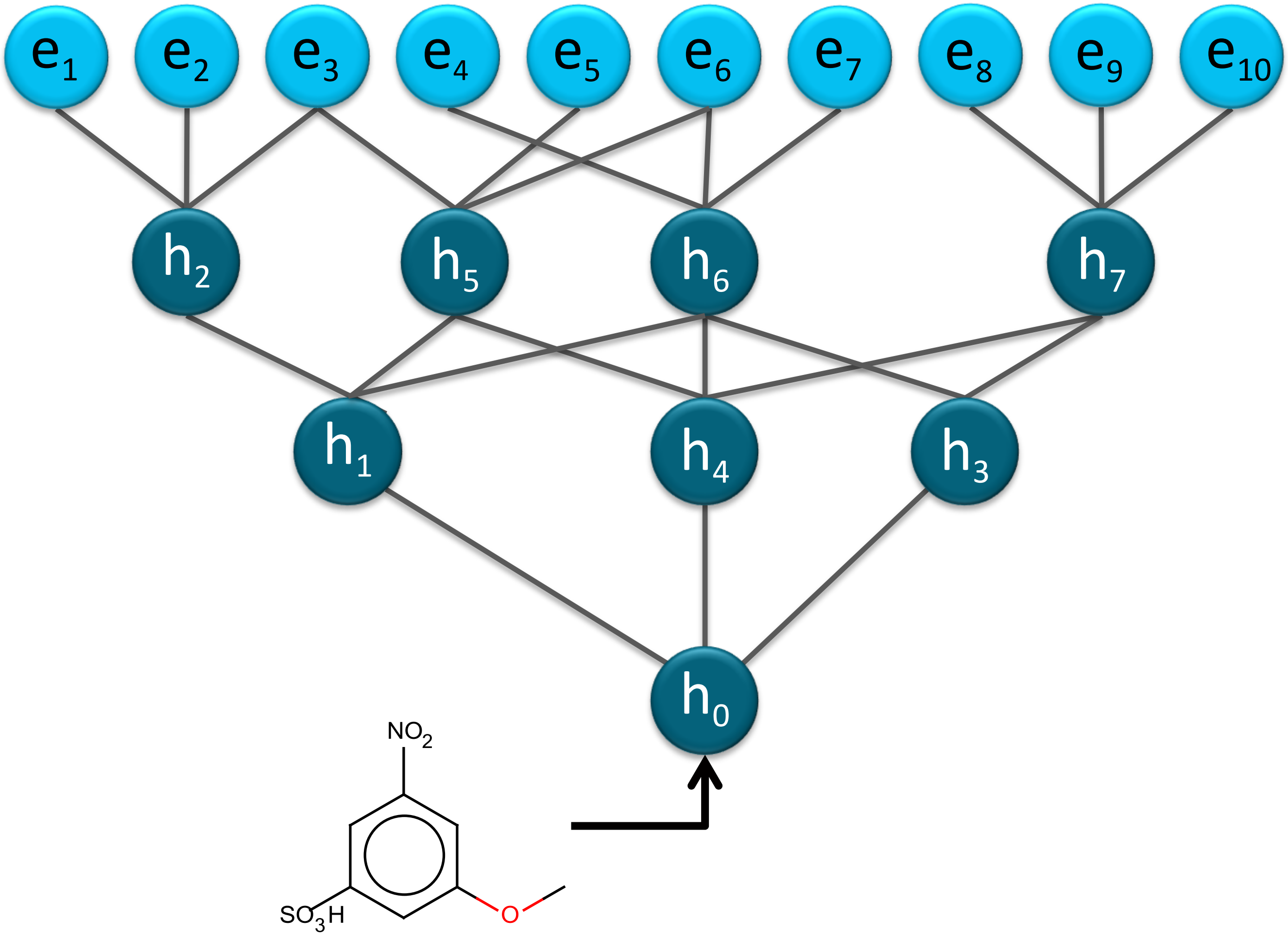
Lead optimisation support

Extended intepretation

New alerts

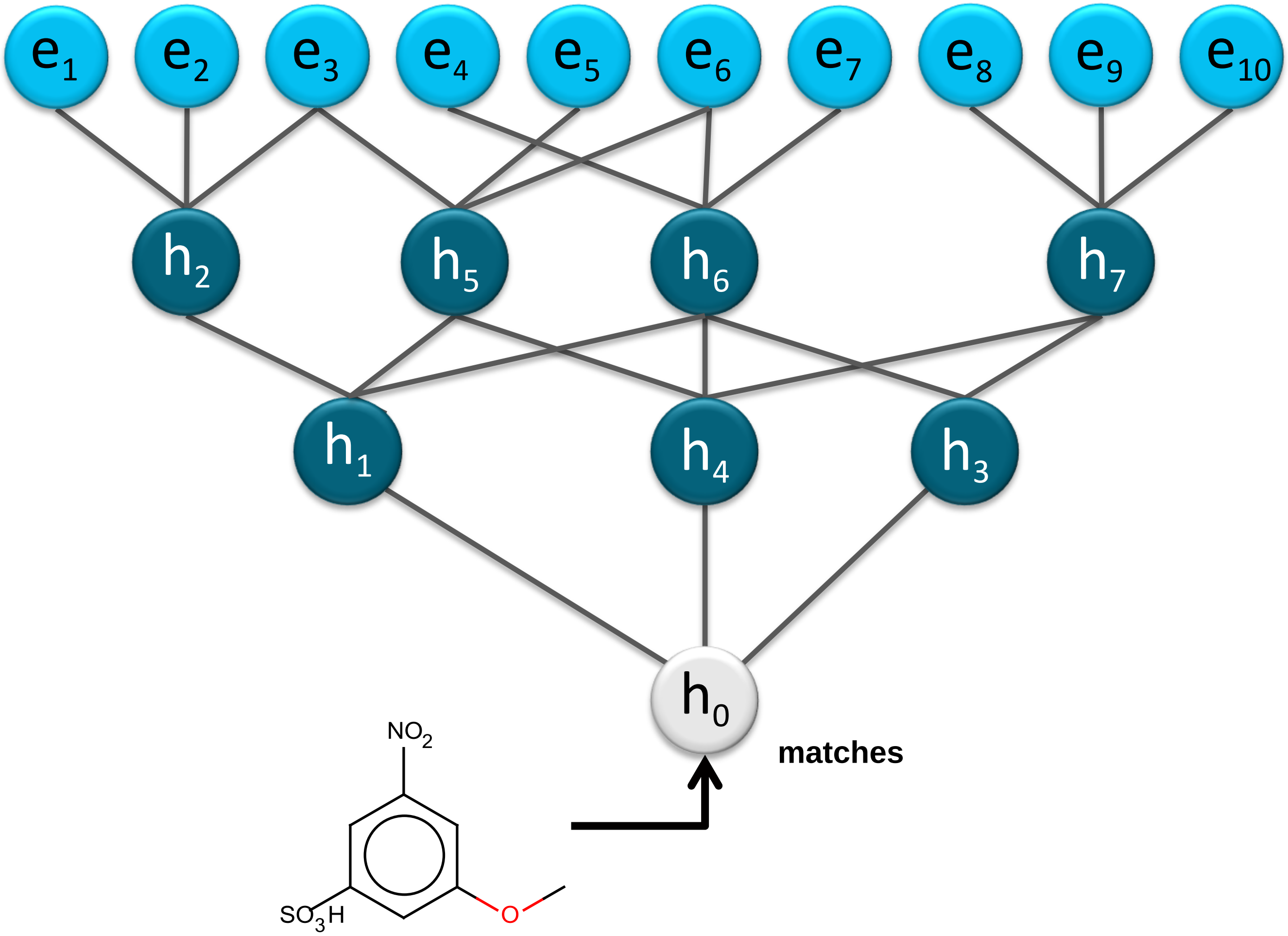


Prediction



Unseen instance

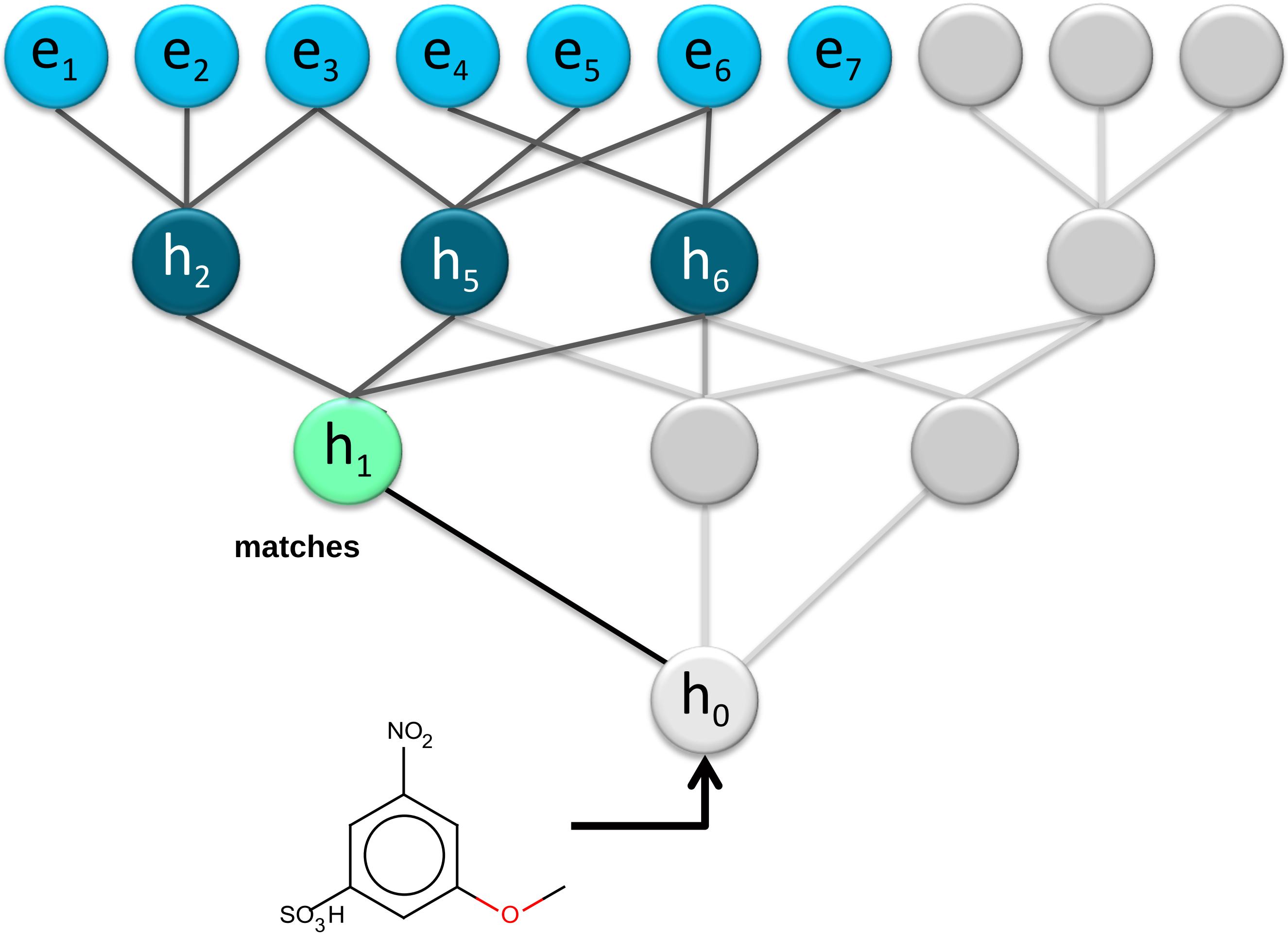
Prediction



matches

Unseen instance

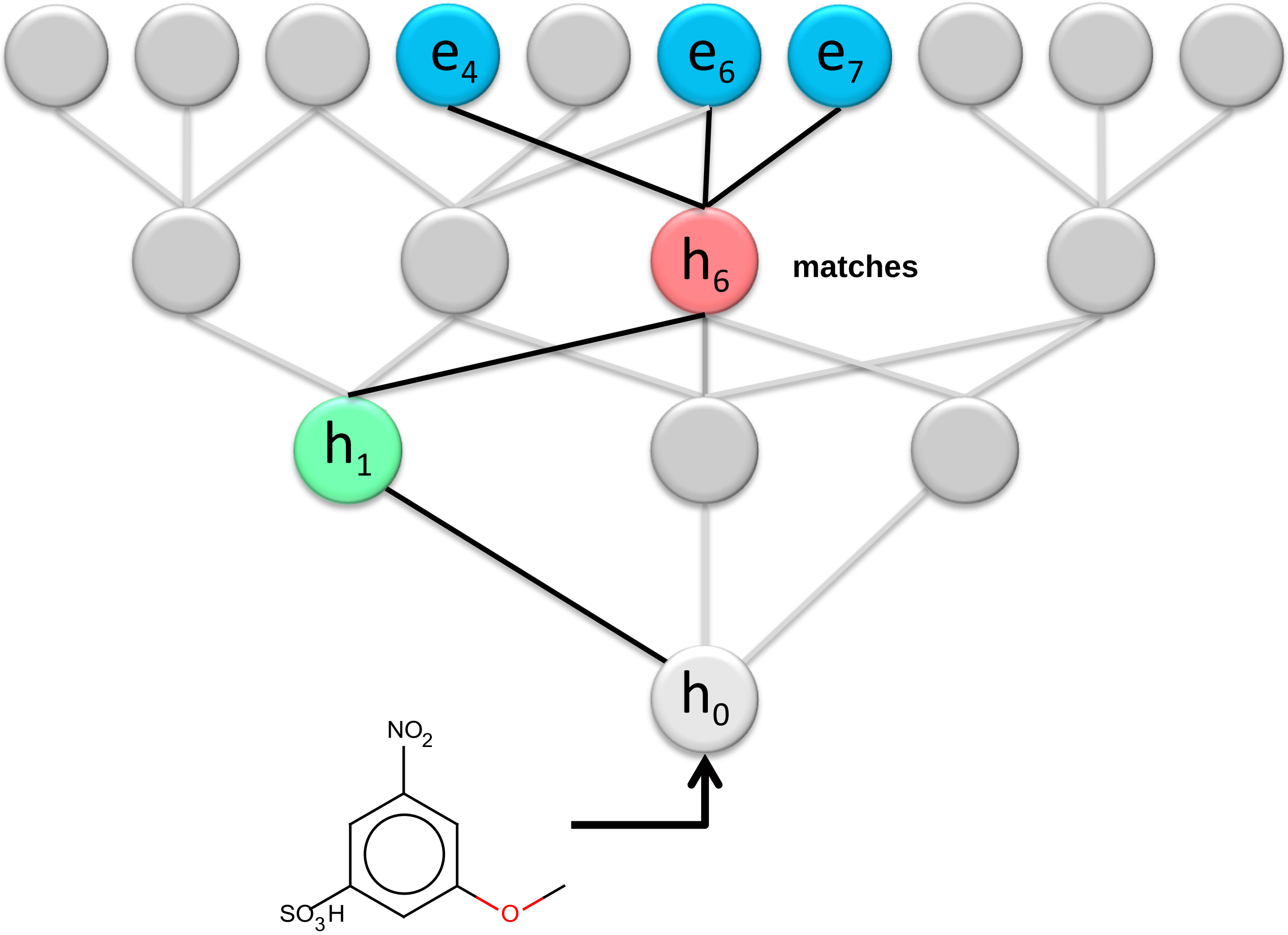
Prediction



matches

Unseen instance

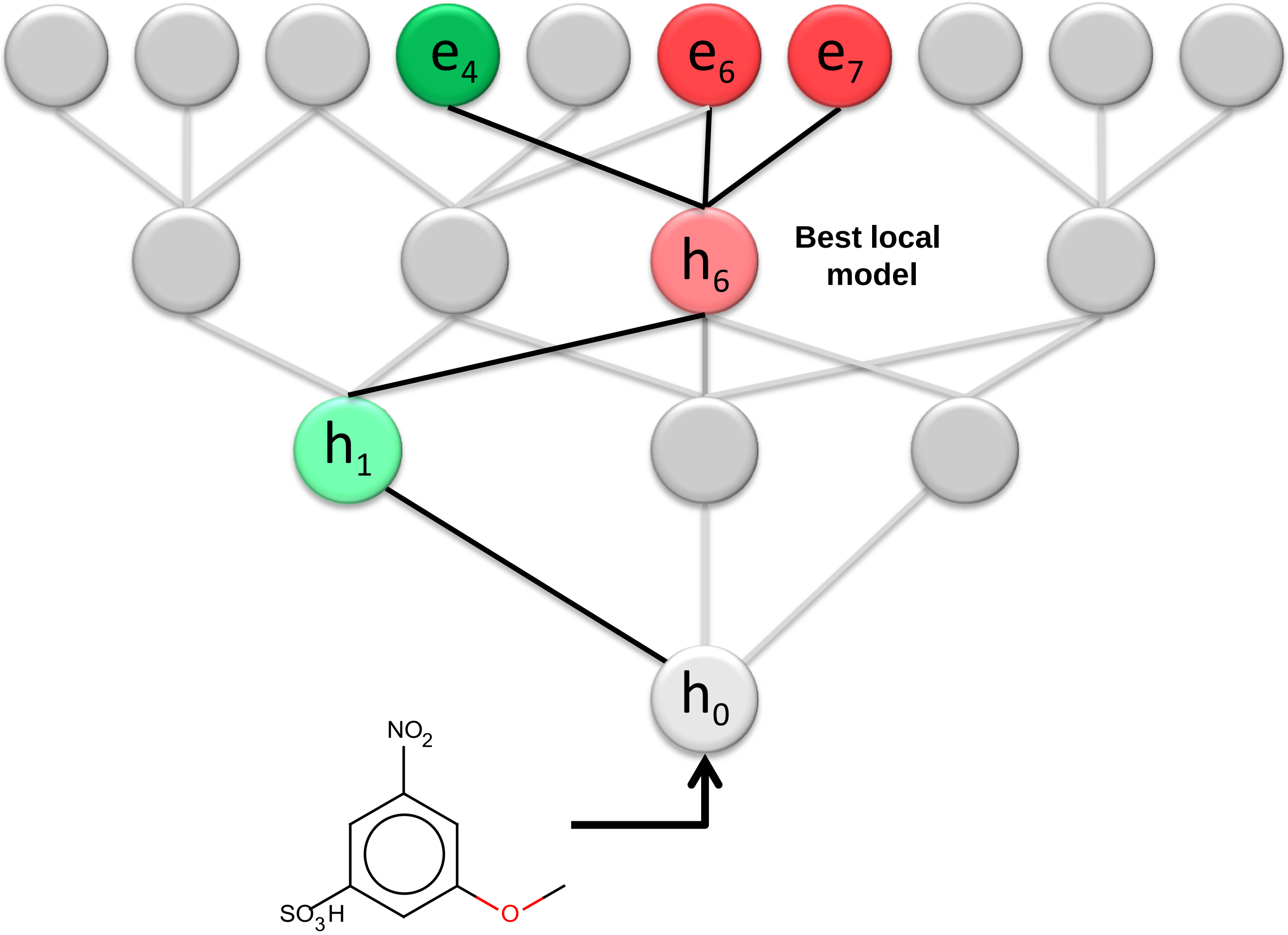
Prediction



matches

Unseen instance

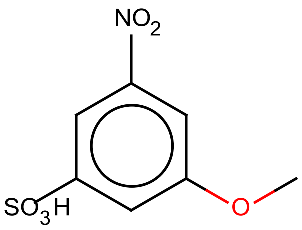
Prediction



Best local model

Unseen instance

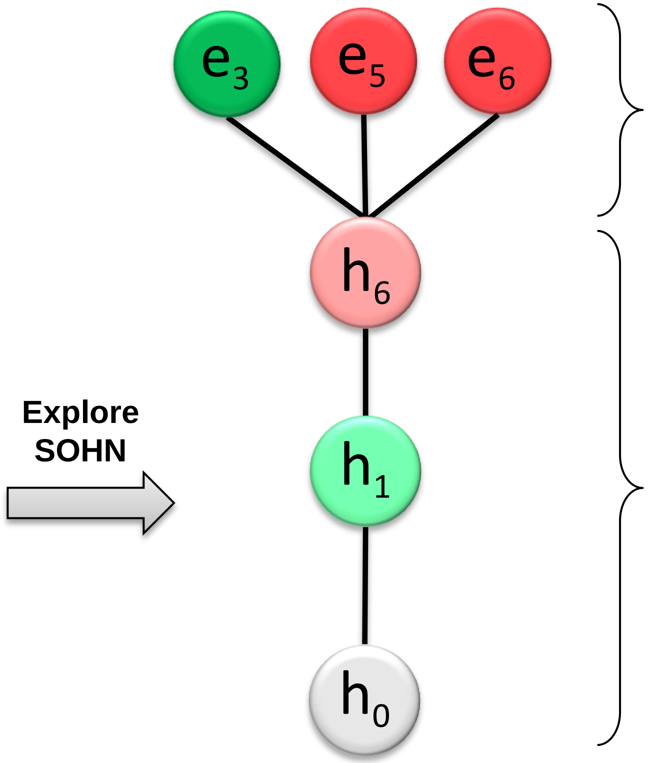
Prediction



Unseen instance

Explore SOHN

Supporting examples

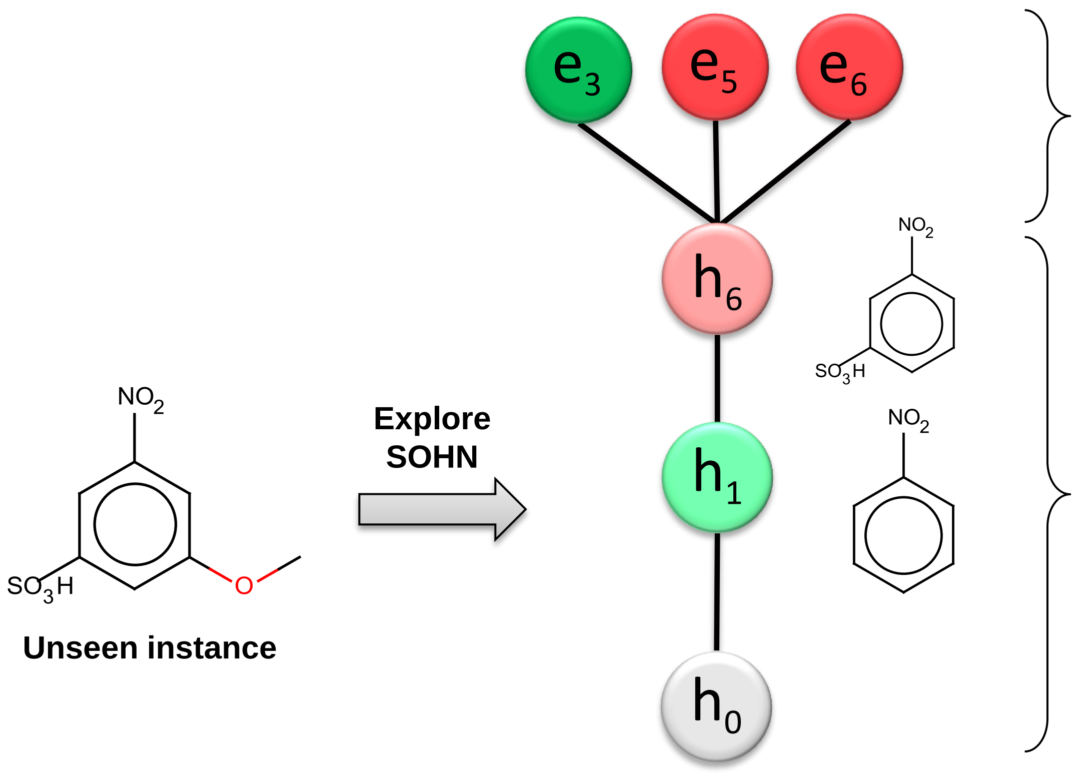


Most relevant part of the knowledge

Prediction (instance based)

Interpretation (induction based)

Prediction



Supporting examples

Explore SOHN

Unseen instance

Most relevant part of the knowledge

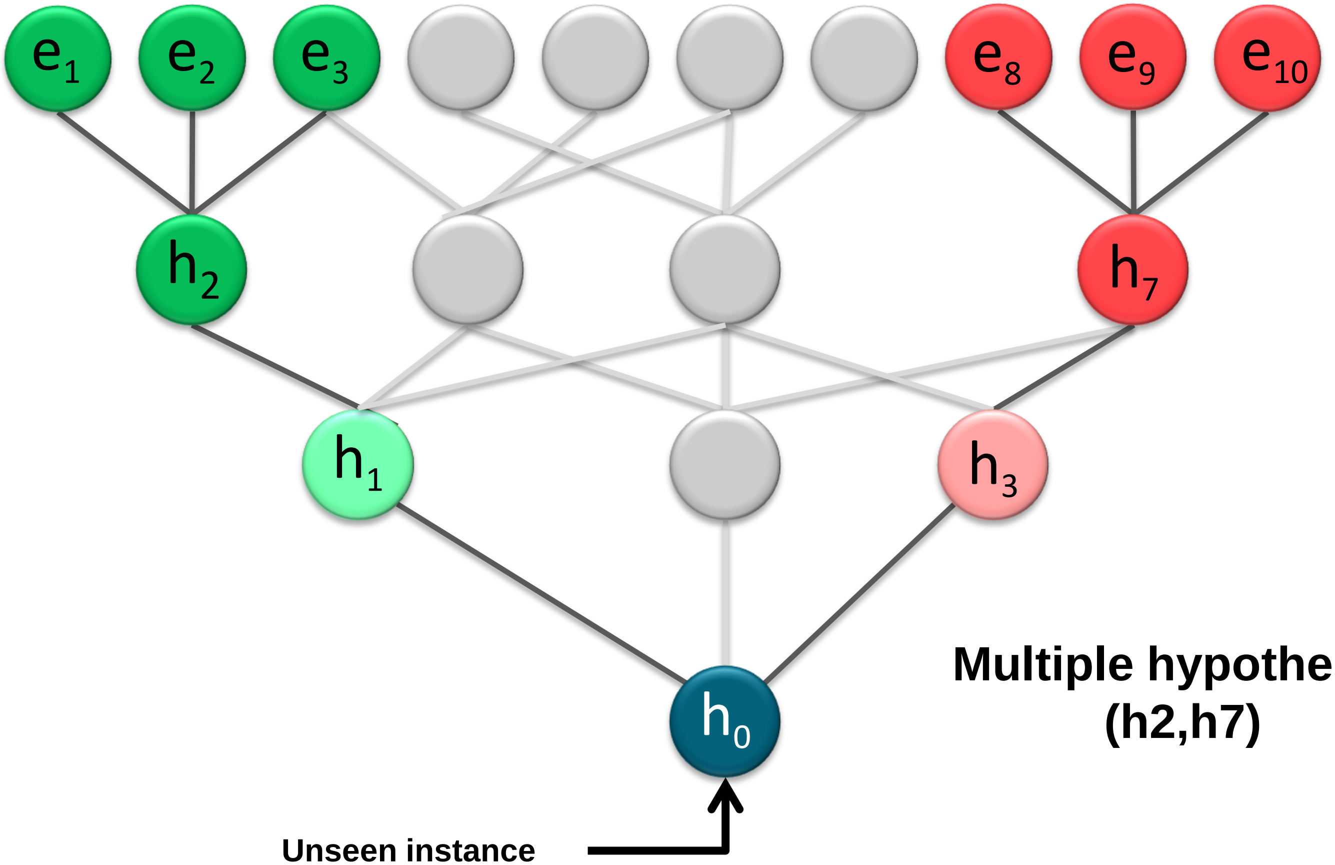
Local KNN model
CLASS

Example variance & similarity
CONFIDENCE

Path : INTERPRETATION

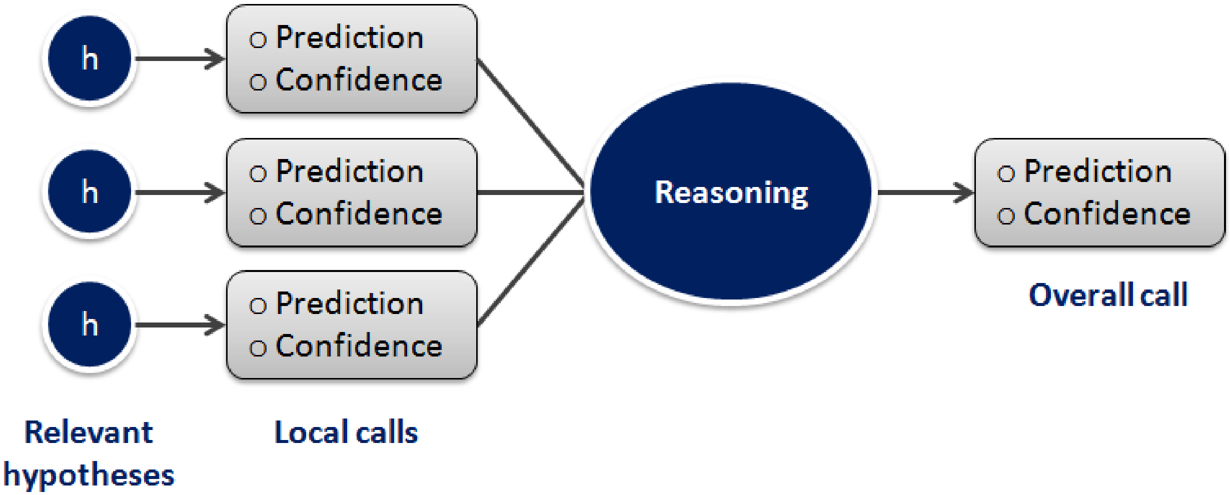
Aromatic nitro deactivated by the sulfonatic acid group in meta position

Prediction



Multiple hypotheses (h2,h7)

Prediction



Reasoning

- ➢ **Weighted / Confidence**
- ➢ **Most confident**
- ➢ **Conservative** **(1+ve enough)**
- ➢ **Average**

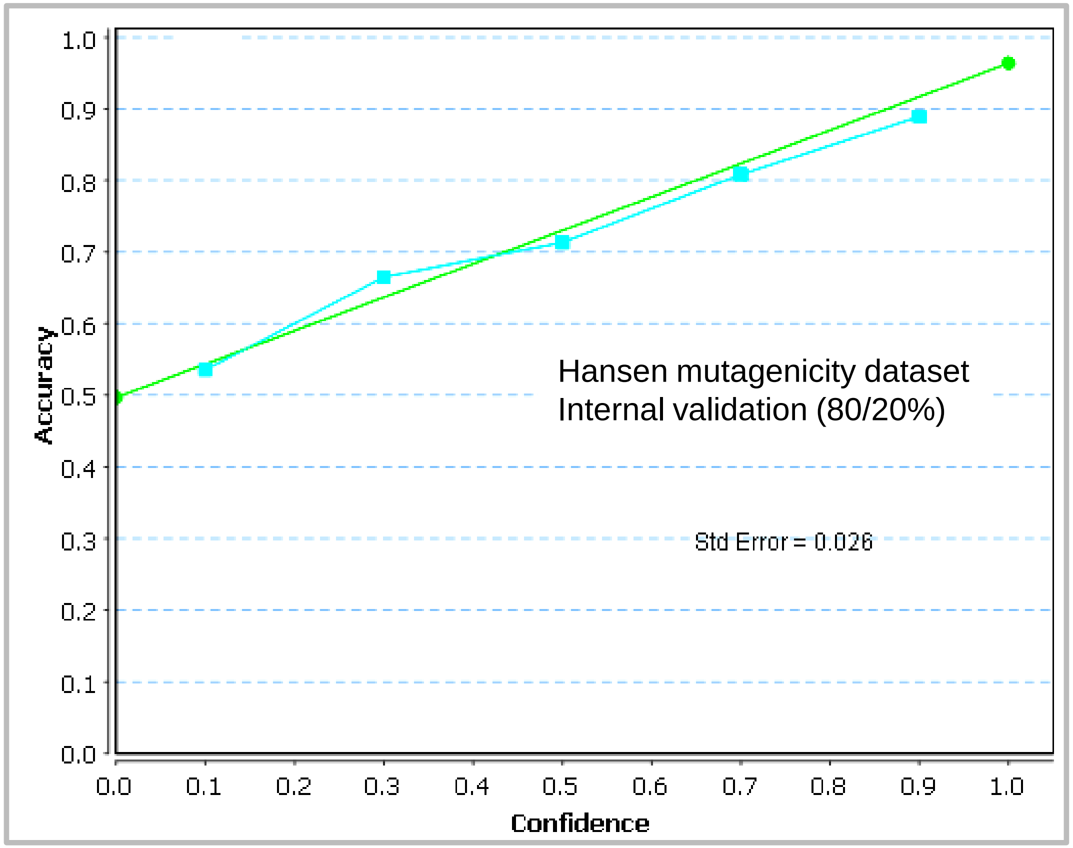
Flexibility / Use case

Weighted / Confidence

$$s_x = \frac{\sum_{h=1}^{m} s_{h,x} \times confidence_{\ h,x}}{\sum_{h=1}^{m} confidence_{\ h,x}}$$

$$confidence_x = |s_x|$$

Individual prediction confidence



Hansen mutagenicity dataset
Internal validation (80/20%)

Confidence / Accuracy correlation

Example
Mutagenicity prediction

Results: Mutagenicity

SOHN details

Dataset	Training	Test	Sensitivity	Specificity	Balanced Accuracy
A (internal 20%)	6560 (80%)	1640 (20%)	78.5	81.4	80
B (external)	8200 (100%)	800 (100%)	53	82	67

Comparison with other ML methods

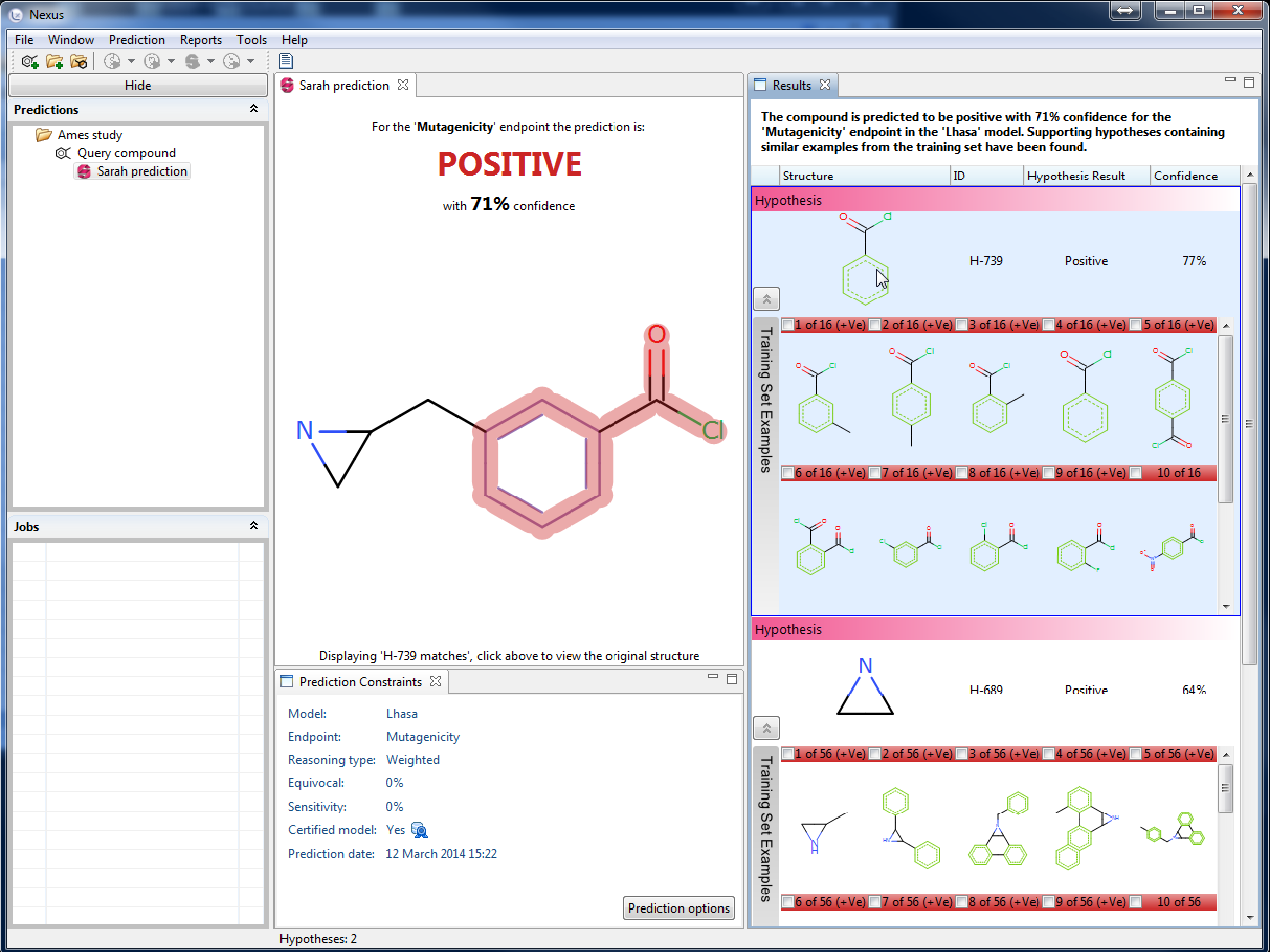
Balanced Accuracy	SVM	RF	KNN	DT	SOHN
A (internal 20%)	81	80	76	77	80
B (external)	64	63	61	60	67

Public dataset A : 8200 public structures (balanced : 50% +ve)
Proprietary dataset B : 800 proprietary structures (biased : 29% +ve)

SVM : best results using PubChem fingerprints (optimised parameters)
RF : best results using MACCS keys / 100 trees

Application

Integration into Lhasa Limited Nexus Suite



For the '**Mutagenicity**' endpoint the prediction is:

POSITIVE

with **71%** confidence

Displaying 'H-739 matches', click above to view the original structure

The compound is predicted to be positive with 71% confidence for the 'Mutagenicity' endpoint in the 'Lhasa' model. Supporting hypotheses containing similar examples from the training set have been found.

Structure	ID	Hypothesis Result	Confidence
Hypothesis			
	H-739	Positive	77%
Hypothesis			
	H-689	Positive	64%

Model: Lhasa
Endpoint: Mutagenicity
Reasoning type: Weighted
Equivocal: 0%
Sensitivity: 0%
Certified model: Yes
Prediction date: 12 March 2014 15:22

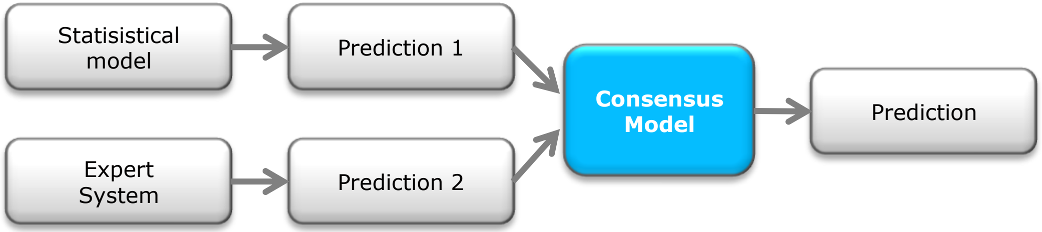
Hypotheses: 2

Combining Statistical Models with Expert Systems

(ICHM7)

Combining Statistical and Expert system

Consensus model approach

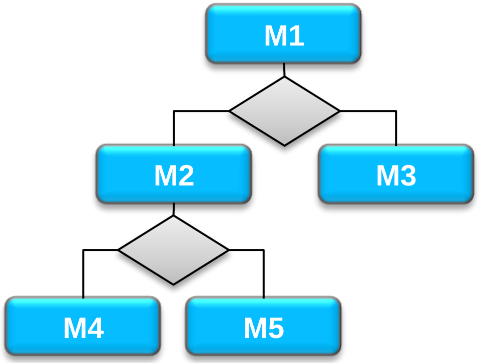


Importance of individual prediction accuracy estimate (confidence)

Model selection

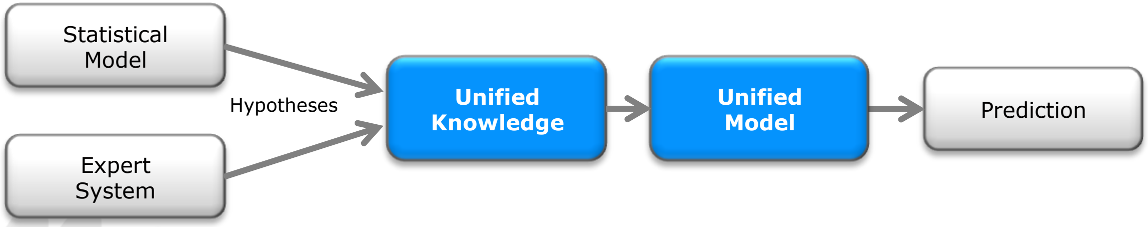
M1	M2	M3	M1	M4
M1	M1	M3	M1	M2
M3	M1	M1	M4	M1
M1	M2	M5	M1	M5
M4	M1	M1	M3	M1

Chemical space performance partition

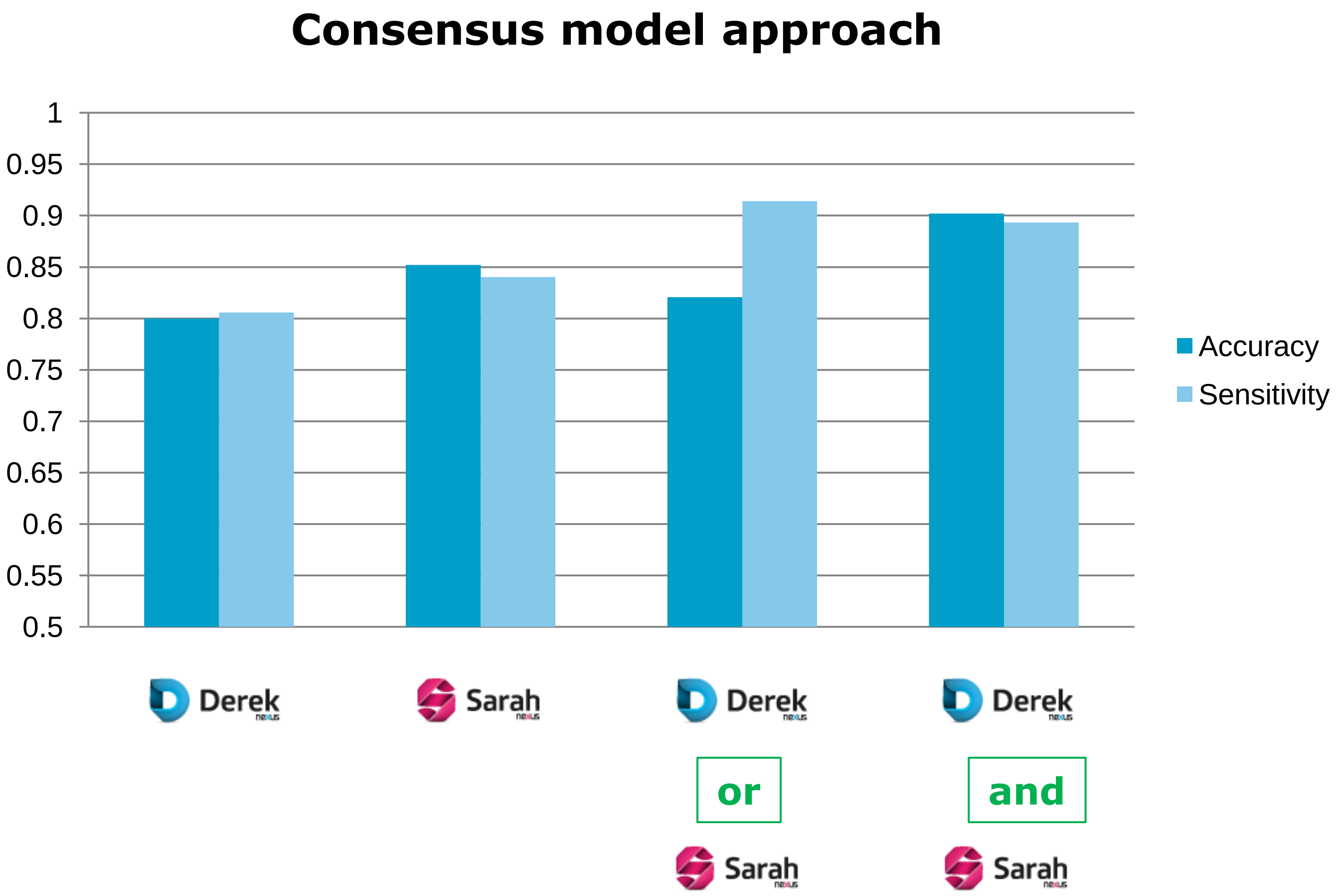


Decision trees (Meta Model)

Unified Knowledge approach

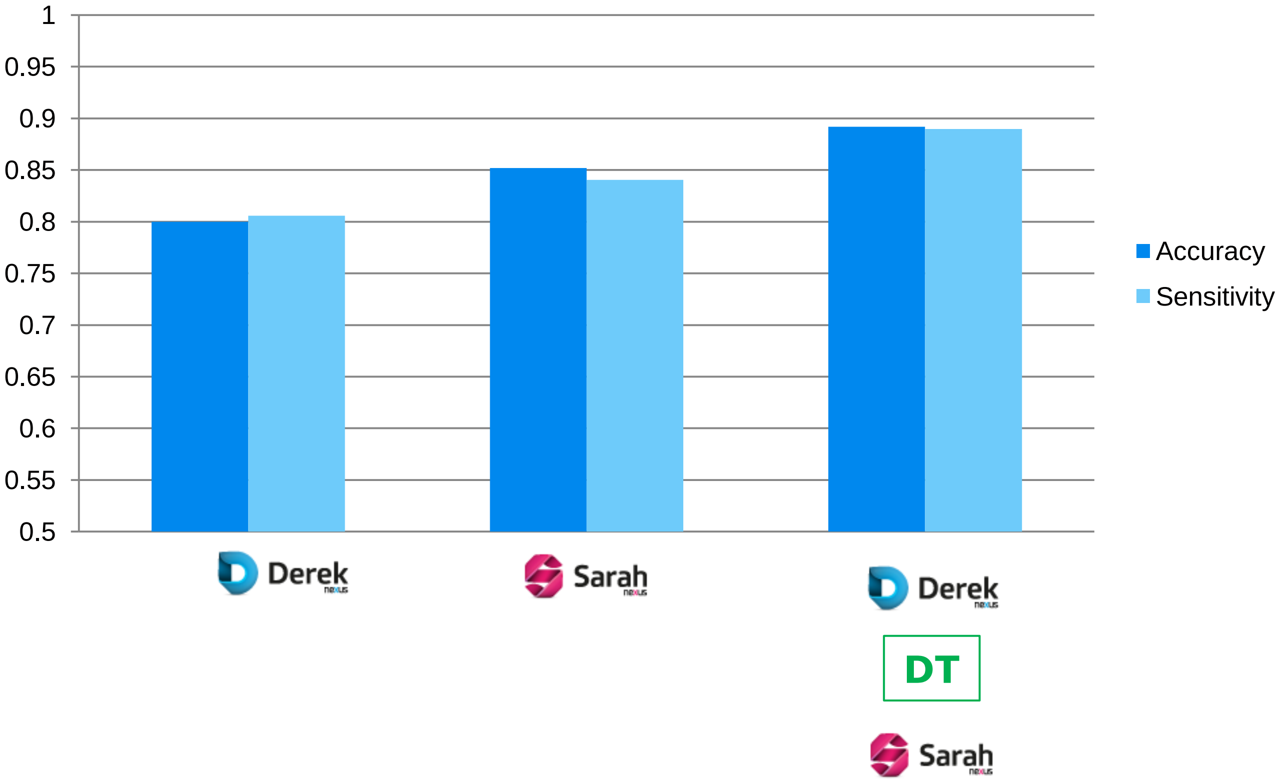


Combining Statistical and Expert system

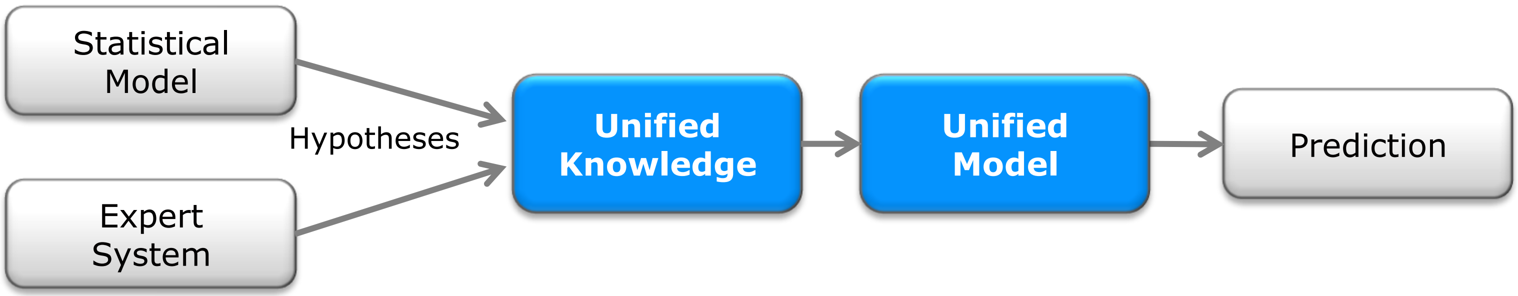


Combining Statistical and Expert system

Model selection using a Decision Tree



Unified Knowledge approach



Advantages

- Combining different source of knowledge (Machine learning, Expert Knowledge)
- Automatic knowledge organisation within a local model hiearachy
- Optimised knowledge selection
- Single prediction algorithm
- Transparent predictions with indication of the origin of the knowledge used
- Harmonised confidence level for individual prediction

Conclusion

- The expert plays the key role in accessing the toxicity of a compound and needs transparent and accurate tools to help him in this task (OECD guidelines)

- Finding the right trade-off between transparency and accuracy is challenging

- One approach is to combine the knowledge from different sources including expert systems and statistical models (ICHM7)

- These different sources can be integrated into a single framework to provide transparent and accurate predictions (SOHN approach)

Arigatou gozaimasu
有難う 御座います

Thank you