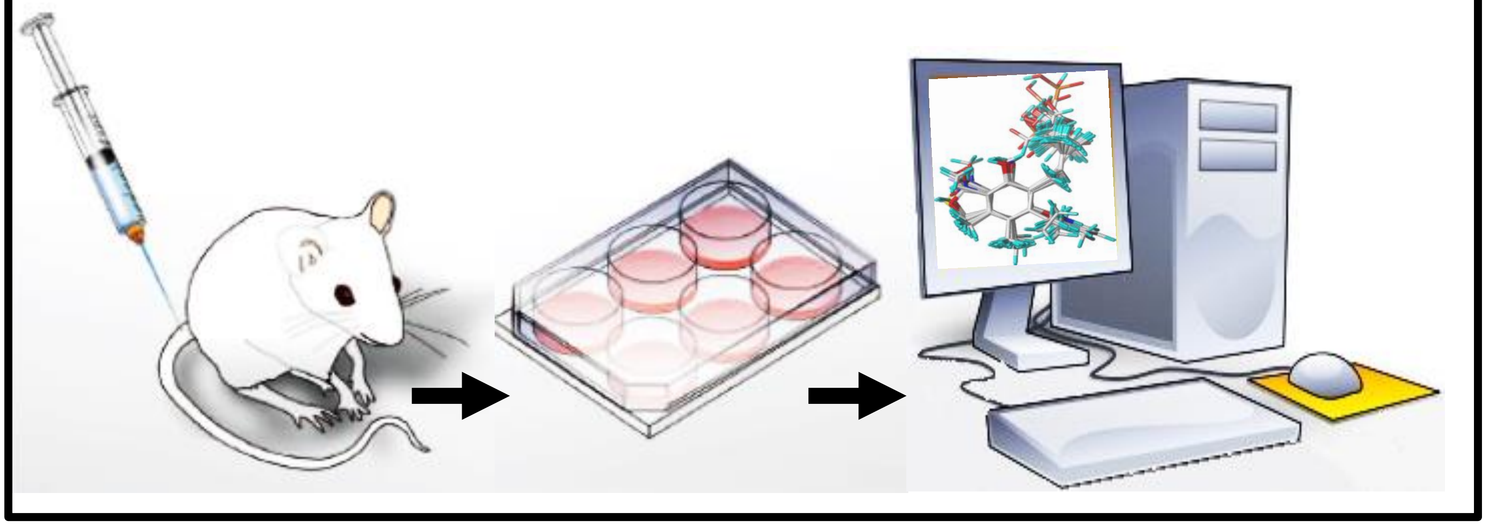
INSTITUTE OF
PARTICLE SCIENCE
& ENGINEERING

UNIVERSITY OF LEEDS

NanoSAR: Structure-Activity Relationship Model for the Toxicity of nano particles



<u>**Ceyda OKSEL**</u>

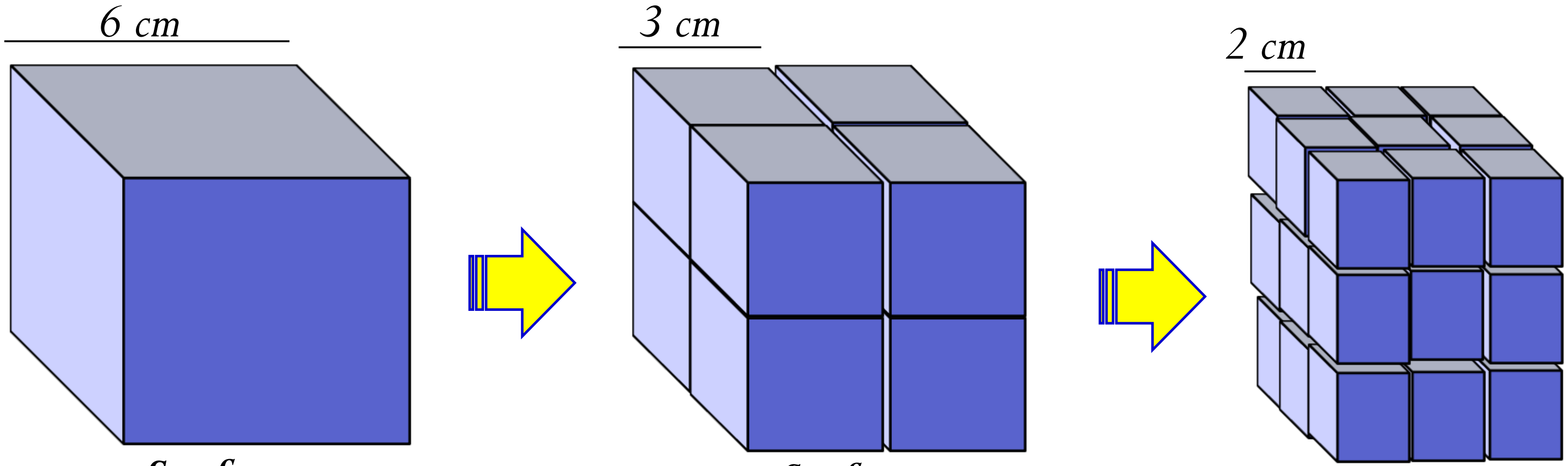
Xue Z Wang

Structure of the lecture

- BACKGROUND
  - Why are things different at nanoscale ?
  - Nanomaterial toxicity
  - Computational models for toxicity prediction
- COMPUTATIONAL MODELLING OF NANOMATERIAL TOXICITY
  - What is (nano)QSAR ?
  - 3 Case Studies
- CONCLUSIONS and FUTURE WORK

Why are things different at nanoscale?

Larger surface area



Surface area

=(6cm x 6cm x 6 faces x 1 cube)

=216cm^2

Surface area

=(3cm x 3cm x 6 faces x 8 cubes)

=432cm^2

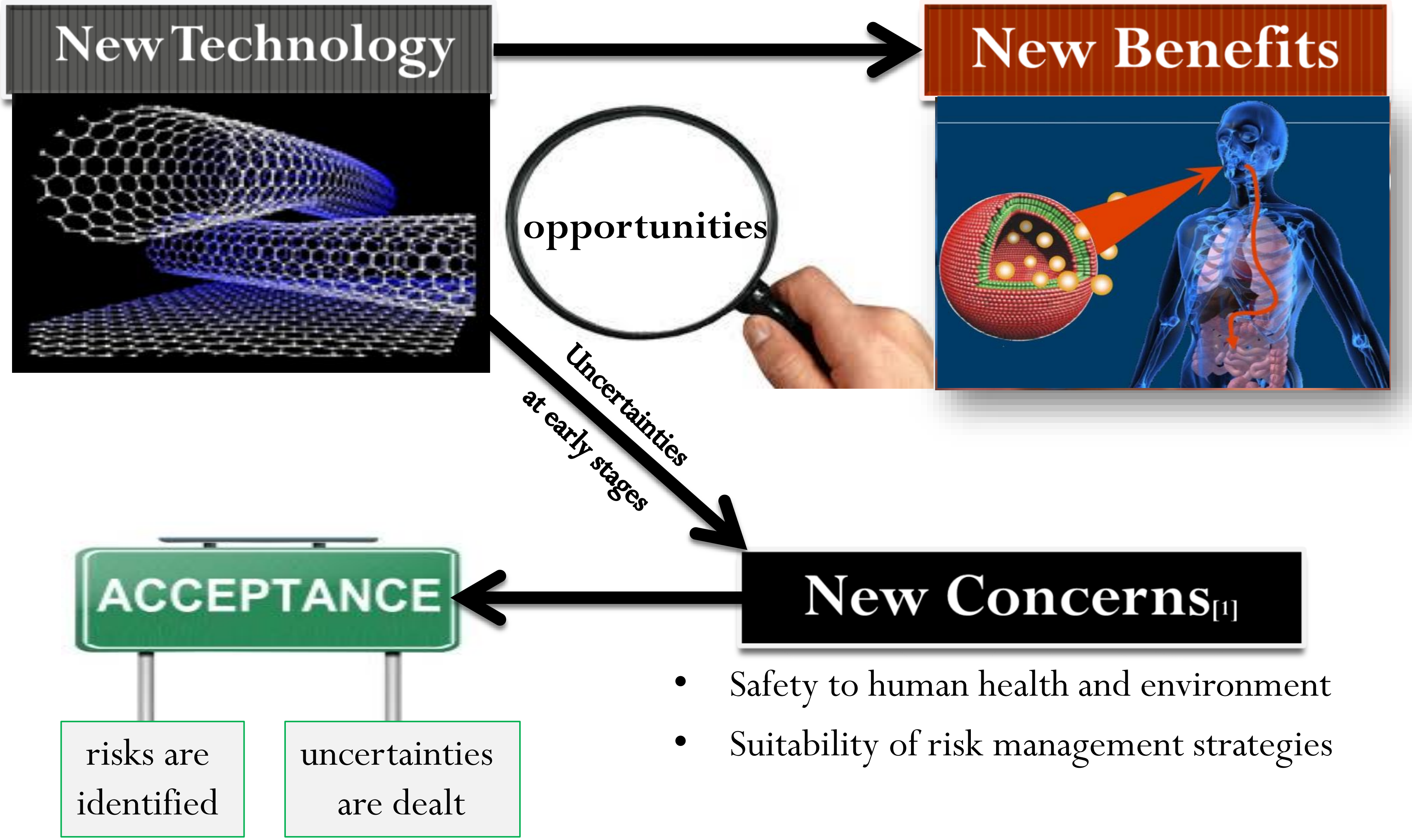
Surface area

=(2cm x 2cm x 6 faces x 27 cubes)

=648cm^2

Quantum effects

Nanomaterial Toxicity



New Technology

New Benefits

opportunities

Uncertainties at early stages

New Concerns[1]

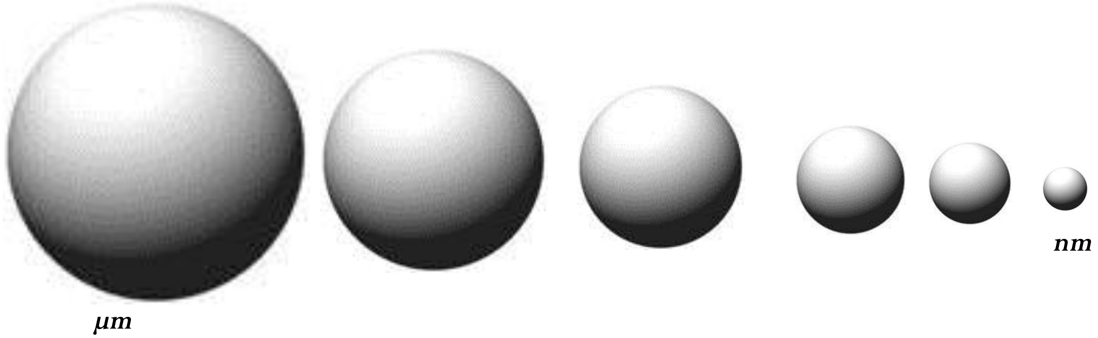
ACCEPTANCE

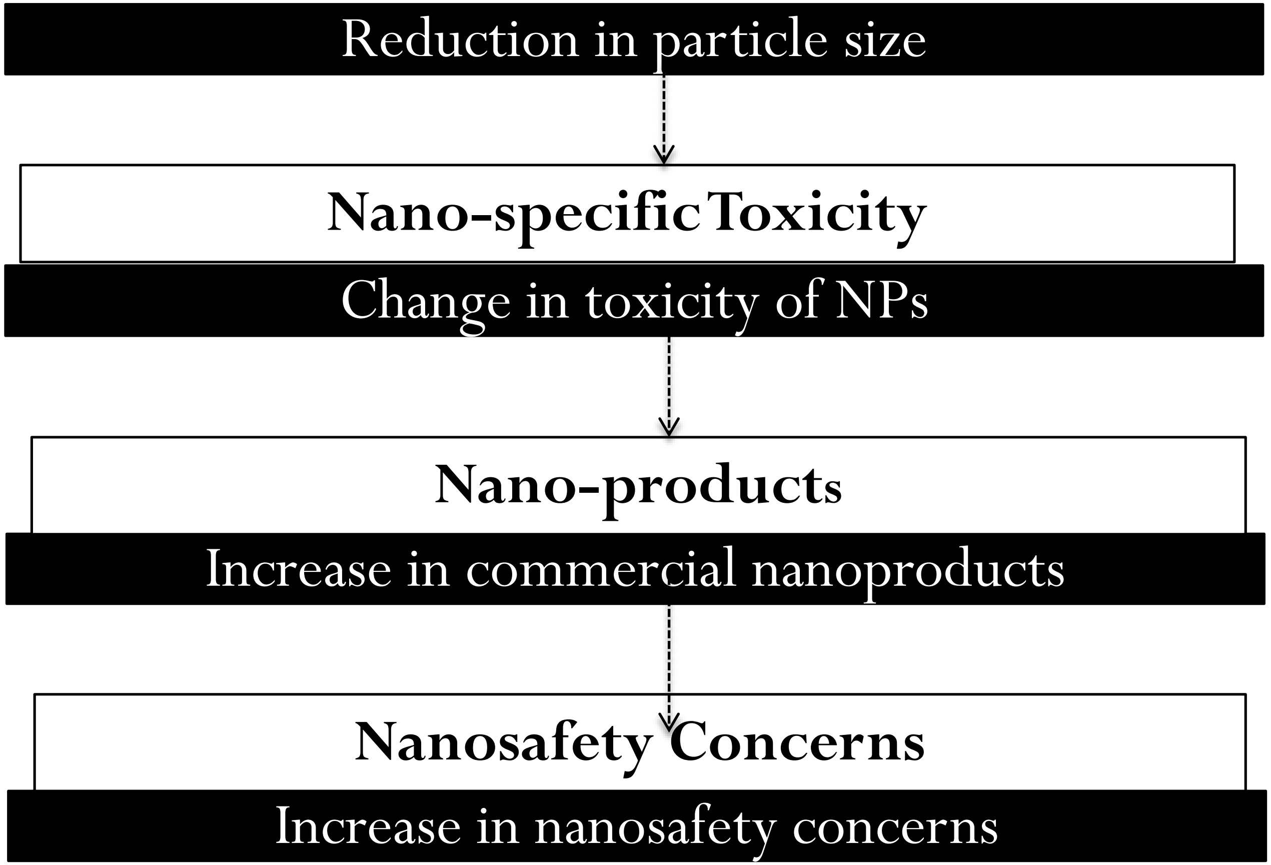
risks are identified

uncertainties are dealt

- Safety to human health and environment
- Suitability of risk management strategies

Nano Particles, Mega Problems ?





Reduction in particle size

Nano-specific Toxicity

Change in toxicity of NPs

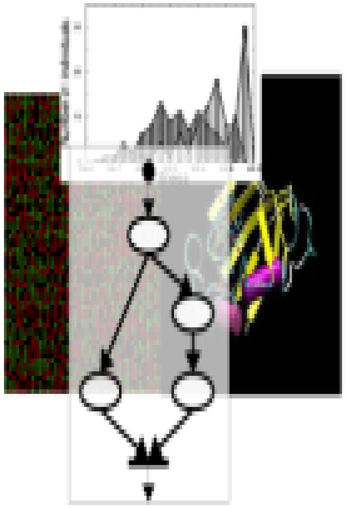
Nano-products

Increase in commercial nanoproducts

Nanosafety Concerns

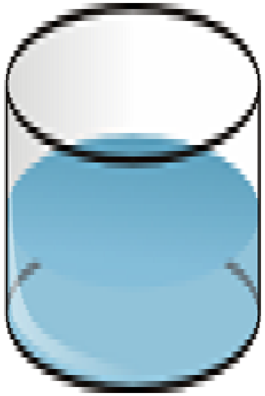
Increase in nanosafety concerns

Toxicity Testing



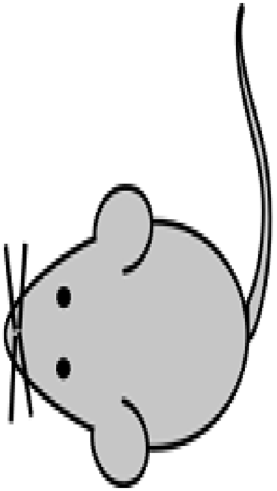
predict

validate



predict

validate



IN SILICO

(computational)

IN VITRO

(computational)

IN VIVO

(on living organism)

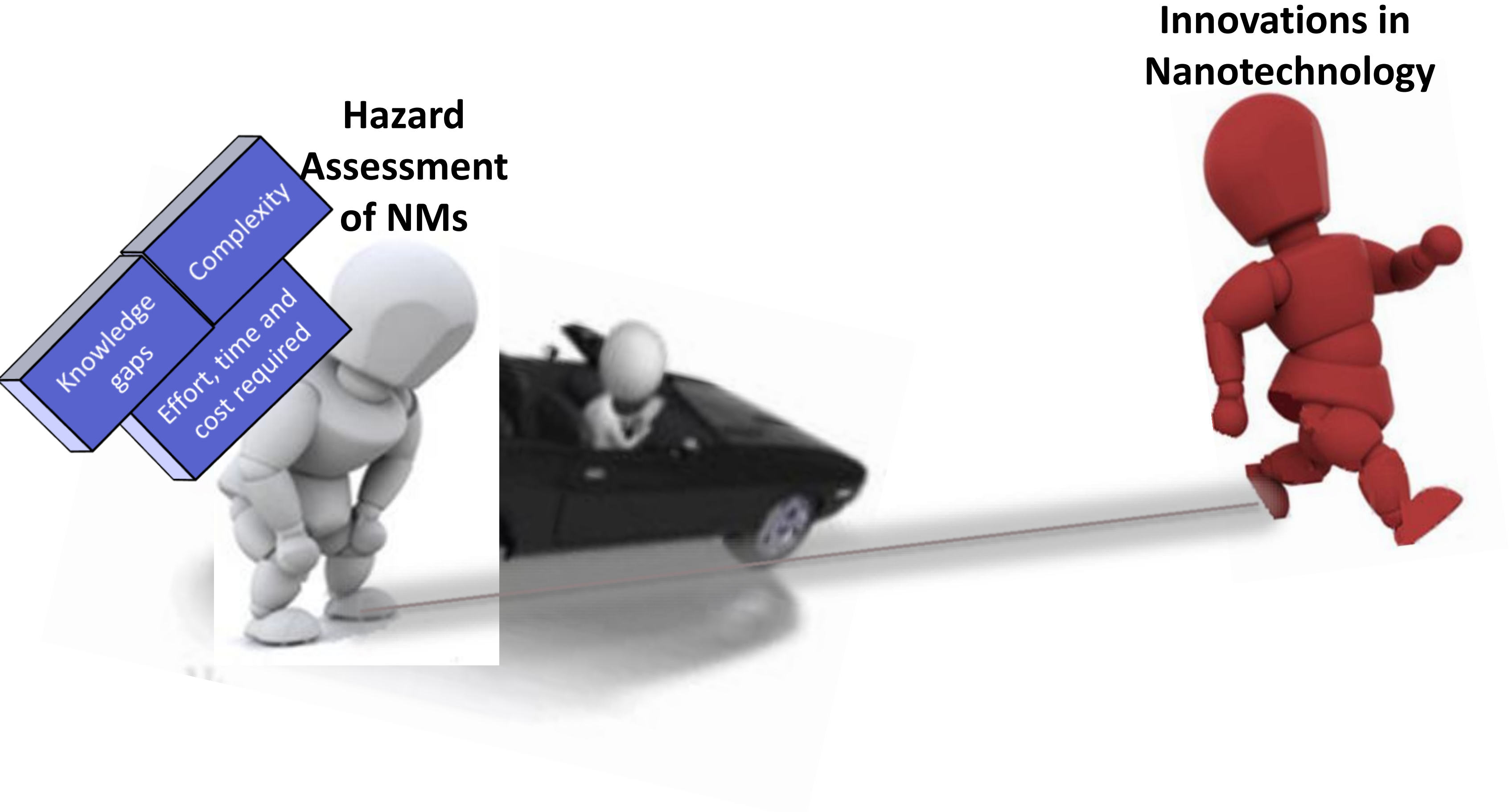
Experimental

Reduction in time,
cost and animal testing

QSAR

Quantitative
Structure-**A**ctivity
Relationship models

Why we need computational models?

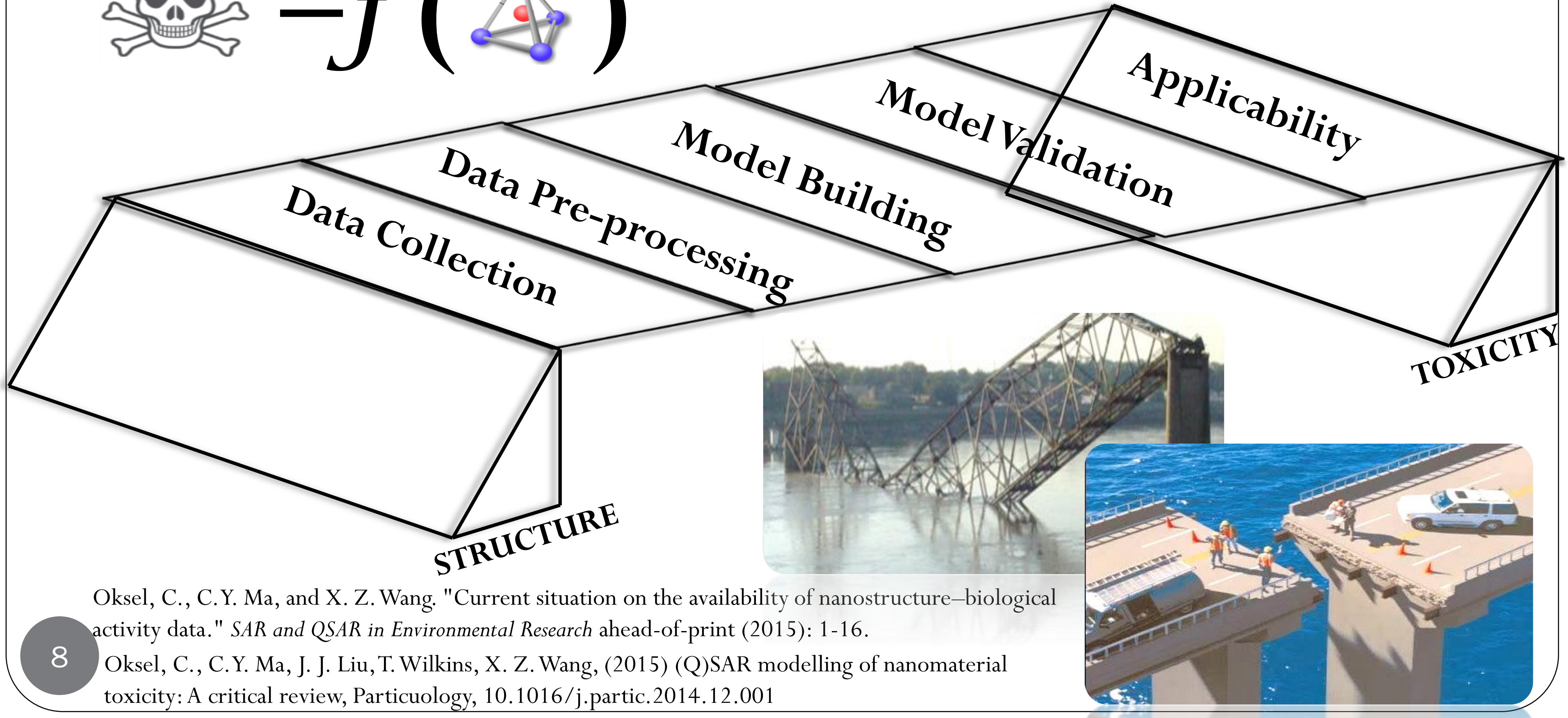


NEED: *The European REACH legislation promotes the use of non-animal testing methods*

AIM: to satisfy this need!!!

What is nano-(Q)SAR ?

A (Q)SAR is a statistical model that relates a set of physicochemical descriptors of a chemical compound to its biological activity.

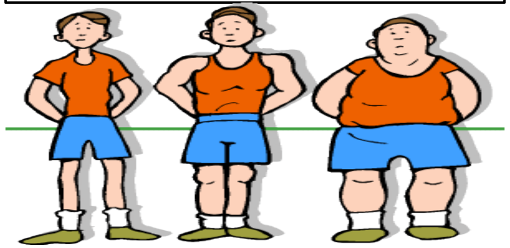


Oksel, C., C.Y. Ma, and X. Z. Wang. "Current situation on the availability of nanostructure–biological activity data." *SAR and QSAR in Environmental Research* ahead-of-print (2015): 1-16.

Oksel, C., C.Y. Ma, J. J. Liu, T. Wilkins, X. Z. Wang, (2015) (Q)SAR modelling of nanomaterial toxicity: A critical review, Particuology, 10.1016/j.partic.2014.12.001

Descriptors

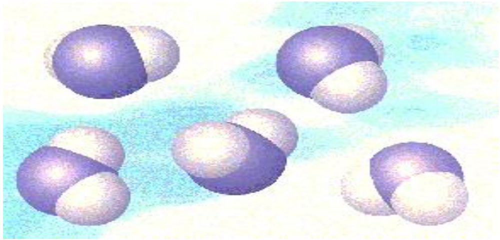
DESCRIBING A PERSON



- ✓ Height
- ✓ Weight
- ✓ Attractiveness
- ✓ …

- ✓ Eye
- ✓ Hair
- ✓ Build
- ✓ …

DESCRIBING A MOLECULE



- ✓ Molar mass
- ✓ Density
- ✓ Conductivity
- ✓ …

- ✓ Atomic prop.
- ✓ Bonds
- ✓ Chirality
- ✓ …

DESCRIBING A NANOPARTICLE

- ✓ Size
- ✓ Shape
- ✓ Composition
- ✓ …

- ✓ Coating
- ✓ Charge
- ✓ Reactivity
- ✓ …

Experimental Descriptors

Theoretical Descriptors

Descriptor Selection

Feature selection algorithms

Si ze
Shape
Charge
Compos
S.area
Structure
Coating
Solubi
Energies
HOMO
LUMO
Wiener
B-double
Count
Seric
CoMFA

Tree Induction From Genetic Programming

GPTree: **"in-house"** software

Genetic Algorithms

explore solution space

- **Starts at random points**
- **Recombining (i.e., crossover)**
- **Optionally changing (i.e., mutation)**

Genetic Algorithm

(1) Randomly generate a pre-specified number of solutions, encoded as fixed size vectors.

(2) Either form a new generation or replace individuals in the population by

2a. Selecting parents using the fitness function.

2b. Crossover the parents to form one or more offspring.

2c. Optionally mutate part of the solution.

(3) Continue with Step 2 until a pre-specified number of generations or children have been grown, or until a good solution is found.

Tree Induction From Genetic Programming

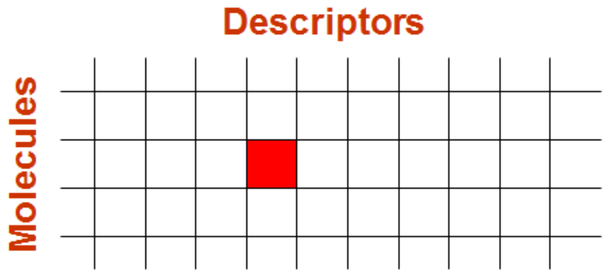
GPTree: Methodology

- **DeLisle, R. K. and Dixon, S. L.** (2004) Induction of Decision Trees via Evolutionary Programming *Journal of Chemical Information and Computer Sciences,* **44,** 862-870.- **evolutionary programming of trees**

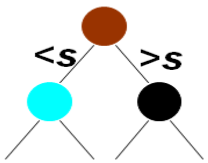
1. Divide data into training and test sets

2. Generate the 1st population of trees

- randomly choosing a row (i.e. a compound), and column (i.e. descriptor)



- Using the value of the slot, *s,* to split, left child takes those data points with selected attribute values <= *s*, whilst the right child takes those > *s*.



Tree Induction From Genetic Programming

GPTree: Methodology

- If a child will not cover enough rows (e.g. 10% of the training rows), another combination is tried.

- A child node becomes a leaf node if pure/near pure, whilst the other nodes grow children.

-When all nodes either have two children or are leaf nodes, the tree is fully grown and added to the first generation.

-A leaf node is assigned to a class label corresponding to the majority class of points partitioned there.

3. Crossover and Mutation

Tree Induction From Genetic Programming

The key parameters

y COL	Column no containing the class of the data set.
n Gen	No of generations required
n Trees	No of treesrequired in each generation
No. in tournament	No of trees in the tournament to sort out the best for crossover operation
Winn. Inc.	Winners included (The N best trees are placed directly into the next generation,This was to allow ELITISM)
L.I.I.A.T	Low increase in accuracy tolerance (It forces a mutation for every tree if no improvement in the best accuracy has been seen for this many generations.)
Mutation	% age of mutation
C in L.N	Minimum no of cases in a leaf node

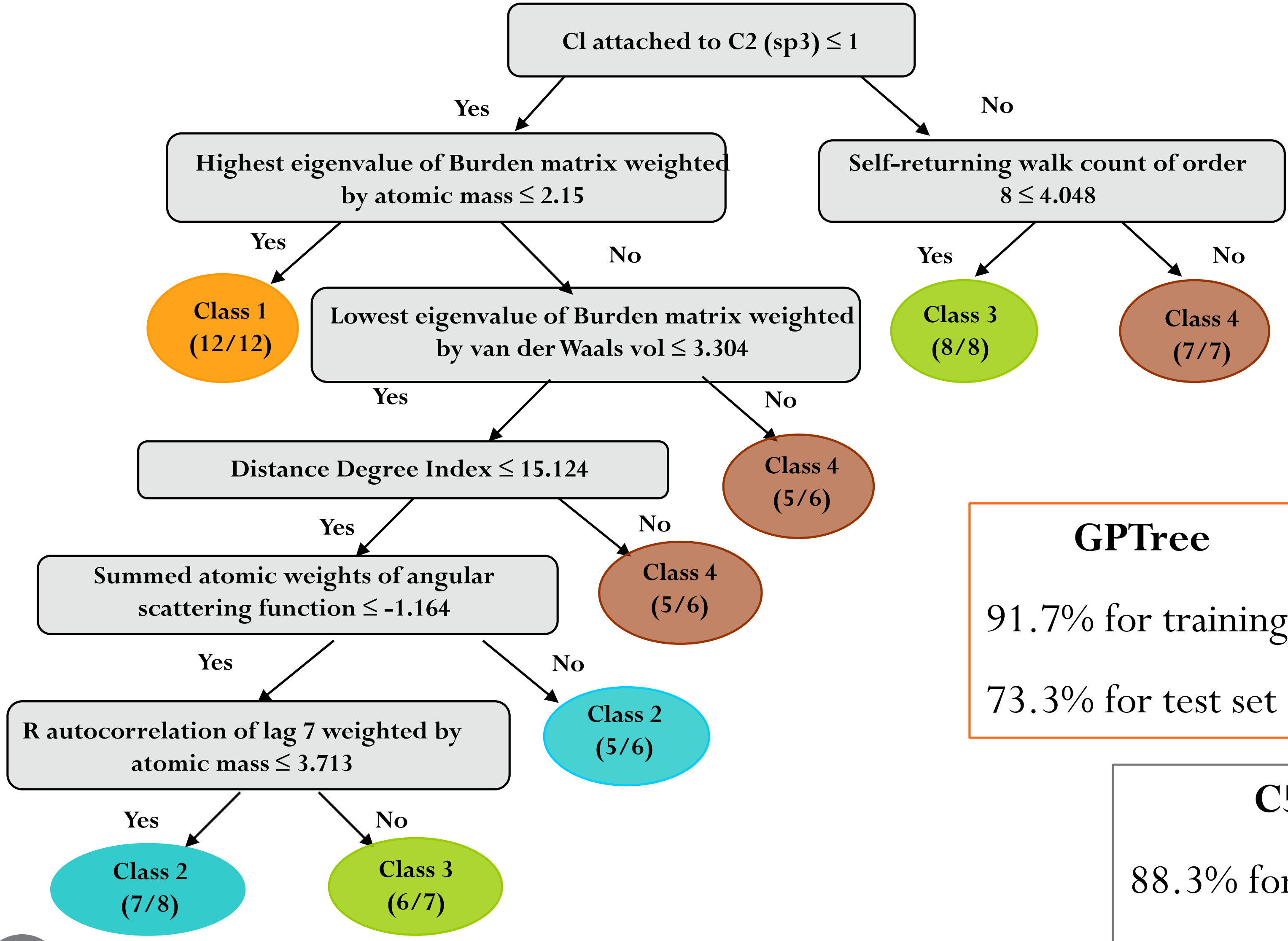
Case Study 1: Dataset

Compounds	75 Compounds
Toxicity Data (4 classes)	Concentration lethal to 50% of the population, LC50, 1/Log(LC50), of *vibrio fischeri, a biolumininescent bactorium*
Descriptors	1069 molecular descriptors calculated by DRAGON

Parameters

y COL	1070
n Gen	60
n Trees	600
No. in tournament	16
Winn. Inc.	0
L.I.I.A.T	5
Mutation	66.7%
C in L.N	2

Case Study 1: Results


Cl attached to C2 (sp3) ≤ 1
Yes
No
Highest eigenvalue of Burden matrix weighted by atomic mass ≤ 2.15
Self-returning walk count of order 8 ≤ 4.048
Yes
No
Yes
No
Class 1 (12/12)
Lowest eigenvalue of Burden matrix weighted by van der Waals vol ≤ 3.304
Class 3 (8/8)
Class 4 (7/7)
Yes
No
Distance Degree Index ≤ 15.124
Class 4 (5/6)
Yes
No
Summed atomic weights of angular scattering function ≤ -1.164
Class 4 (5/6)
Yes
No
R autocorrelation of lag 7 weighted by atomic mass ≤ 3.713
Class 2 (5/6)
Yes
No
Class 2 (7/8)
Class 3 (6/7)

GPTree

91.7% for training

73.3% for test set

C5.0

88.3% for training

60.0 % for test set

Case Study 2: Dataset

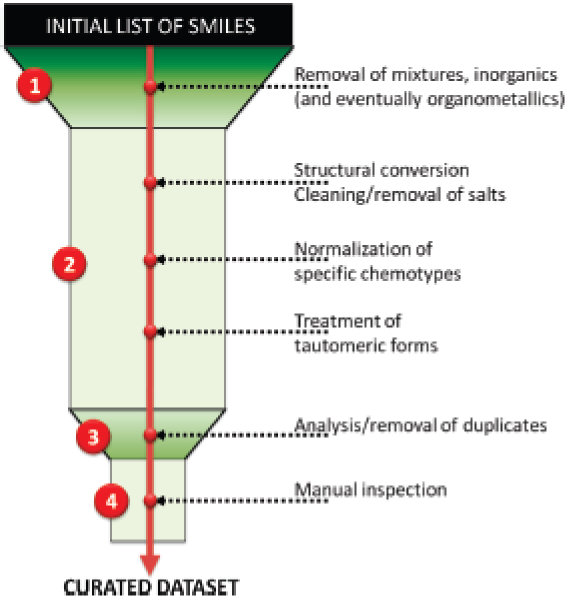
Compounds	105 nanoparticles with different surface-modifying molecules
Toxicity Data	Cellular uptake in pancreatic cancer cell lines
Threshold value	Cellular uptake values:170-27 542 nanoparticles per cell Threshold value: 10 000 nanoparticles per cell 18 nanoparticles with significant cellular uptake (CLASS 2) 87 nanoparticles with poor cellular uptake (CLASS 1)

D. Fourches, D. Pu, C. Tassa, R. Weissleder, S.Y. Shaw, R.J. Mumper, and A. Tropsha, Quantitative nanostructure–activity relationship modeling, ACS Nano 4 (2010), pp. 5703–5712.

Case Study 2: Dataset

Descriptors

Nanoparticles → Same core / Different surface-modifying molecules → Conventional descriptors

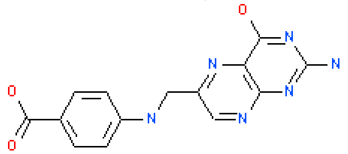

INITIAL LIST OF SMILES
1
Removal of mixtures, inorganics (and eventually organometallics)
Structural conversion
Cleaning/removal of salts
2
Normalization of specific chemotypes
Treatment of tautomeric forms
3
Analysis/removal of duplicates
4
Manual inspection
CURATED DATASET

Fourches et al. (2010)

- **Data cleaning**
- **Structural Conversion**

SMILES strings → 2D molecular graphs

(C=NC(=C(N=1)C1O)N=C(N=1)N)CNC(=CC=C(C1)C(O)=O)C



- **Manual inspection**

4 structure unmatched-excluded

- **Descriptor Calculation**

690 Dragon Descriptors

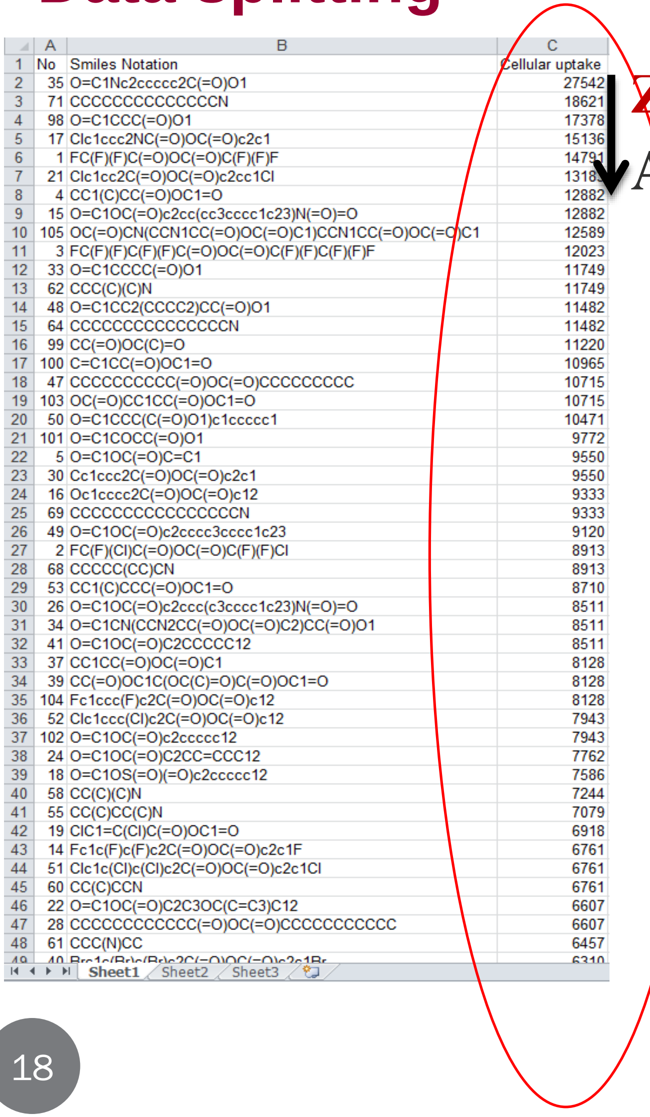
- **Descriptor Cleaning**

389 Dragon descriptors retained

Fourches, Denis, Eugene Muratov, and Alexander Tropsha. "Trust, but verify: on the importance of chemical structure curation in cheminformatics and QSAR modeling research." *Journal of chemical information and modeling* 50.7 (2010): 1189-1204.

Case Study 2: Data Pre-processing

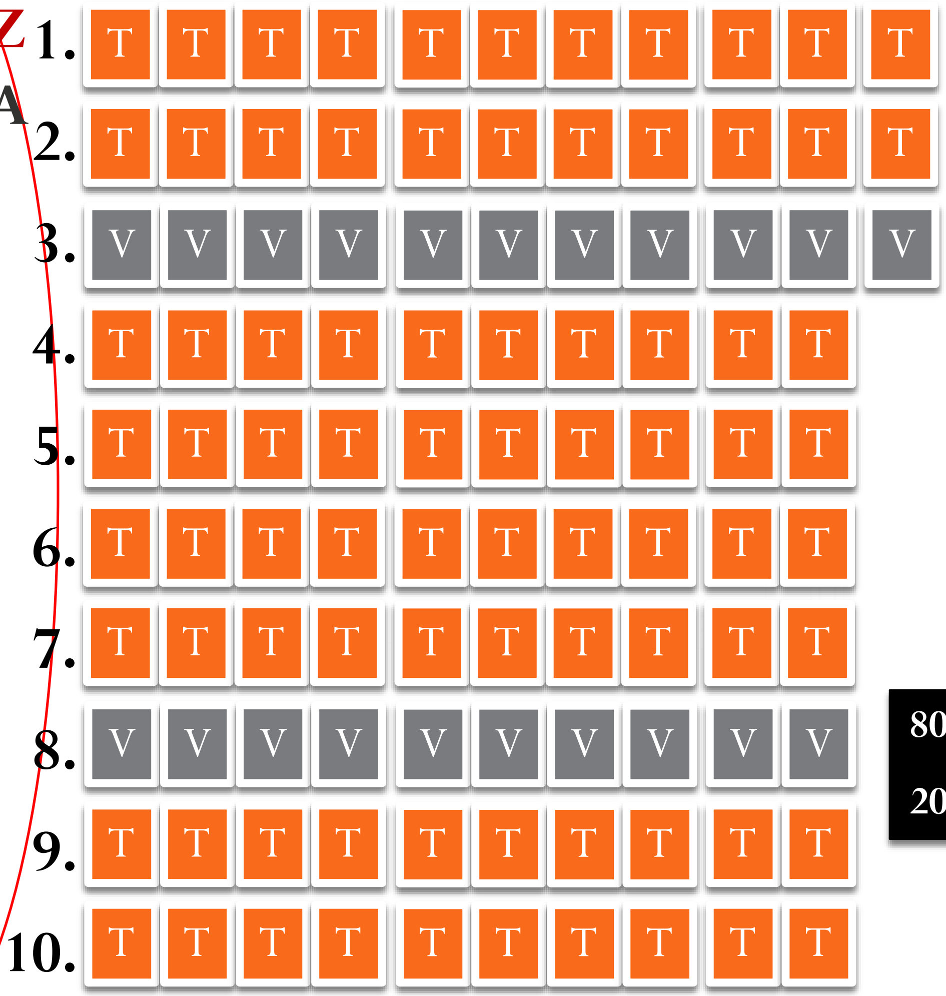
Data splitting



	A	B	C
1	No	Smiles Notation	Cellular uptake
2	35	O=C1Nc2ccccc2C(=O)O1	27542
3	71	CCCCCCCCCCCCCCN	18621
4	98	O=C1CCC(=O)O1	17378
5	17	Clc1ccc2NC(=O)OC(=O)c2c1	15136
6	1	FC(F)(F)C(=O)OC(=O)C(F)(F)F	14791
7	21	Clc1cc2C(=O)OC(=O)c2cc1Cl	1318
8	4	CC1(C)CC(=O)OC1=O	12882
9	15	O=C1OC(=O)c2cc(cc3cccc1c23)N(=O)=O	12882
10	105	OC(=O)CN(CCN1CC(=O)OC(=O)C1)CCN1CC(=O)OC(=O)C1	12589
11	3	FC(F)(F)C(F)(F)C(=O)OC(=O)C(F)(F)C(F)(F)F	12023
12	33	O=C1CCCC(=O)O1	11749
13	62	CCC(C)(C)N	11749
14	48	O=C1CC2(CCCC2)CC(=O)O1	11482
15	64	CCCCCCCCCCCCCCCN	11482
16	99	CC(=O)OC(C)=O	11220
17	100	C=C1CC(=O)OC1=O	10965
18	47	CCCCCCCCCC(=O)OC(=O)CCCCCCCCC	10715
19	103	OC(=O)CC1CC(=O)OC1=O	10715
20	50	O=C1CCC(C(=O)O1)c1ccccc1	10471
21	101	O=C1COCC(=O)O1	9772
22	5	O=C1OC(=O)C=C1	9550
23	30	Cc1ccc2C(=O)OC(=O)c2c1	9550
24	16	Oc1cccc2C(=O)OC(=O)c12	9333
25	69	CCCCCCCCCCCCCCCCN	9333
26	49	O=C1OC(=O)c2cccc3cccc1c23	9120
27	2	FC(F)(Cl)C(=O)OC(=O)C(F)(F)Cl	8913
28	68	CCCCC(CC)CN	8913
29	53	CC1(C)CCC(=O)OC1=O	8710
30	26	O=C1OC(=O)c2ccc(c3cccc1c23)N(=O)=O	8511
31	34	O=C1CN(CCN2CC(=O)OC(=O)C2)CC(=O)O1	8511
32	41	O=C1OC(=O)C2CCCCC12	8511
33	37	CC1CC(=O)OC(=O)C1	8128
34	39	CC(=O)OC1C(OC(C)=O)C(=O)OC1=O	8128
35	104	Fc1ccc(F)c2C(=O)OC(=O)c12	8128
36	52	Clc1ccc(Cl)c2C(=O)OC(=O)c12	7943
37	102	O=C1OC(=O)c2ccccc12	7943
38	24	O=C1OC(=O)C2CC=CCC12	7762
39	18	O=C1OS(=O)(=O)c2ccccc12	7586
40	58	CC(C)(C)N	7244
41	55	CC(C)CC(C)N	7079
42	19	ClC1=C(Cl)C(=O)OC1=O	6918
43	14	Fc1c(F)c(F)c2C(=O)OC(=O)c2c1F	6761
44	51	Clc1c(Cl)c(Cl)c2C(=O)OC(=O)c2c1Cl	6761
45	60	CC(C)CCN	6761
46	22	O=C1OC(=O)C2C3OC(C=C3)C12	6607
47	28	CCCCCCCCCCCC(=O)OC(=O)CCCCCCCCCCC	6607
48	61	CCC(N)CC	6457
49	40	Brc1c(Br)c(Br)c2C(=O)OC(=O)c2c1Br	6310

Sheet1 Sheet2 Sheet3

Pattern of splitting



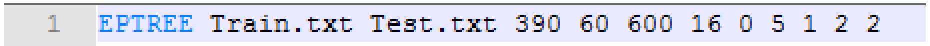
1.	T	T	T	T	T	T	T	T	T	T	T
2.	T	T	T	T	T	T	T	T	T	T	T
3.	V	V	V	V	V	V	V	V	V	V	V
4.	T	T	T	T	T	T	T	T	T	T	
5.	T	T	T	T	T	T	T	T	T	T	
6.	T	T	T	T	T	T	T	T	T	T	
7.	T	T	T	T	T	T	T	T	T	T	
8.	V	V	V	V	V	V	V	V	V	V	
9.	T	T	T	T	T	T	T	T	T	T	
10.	T	T	T	T	T	T	T	T	T	T	

80% training

20% validation

Case Study 2: GPTree settings

The key parameters

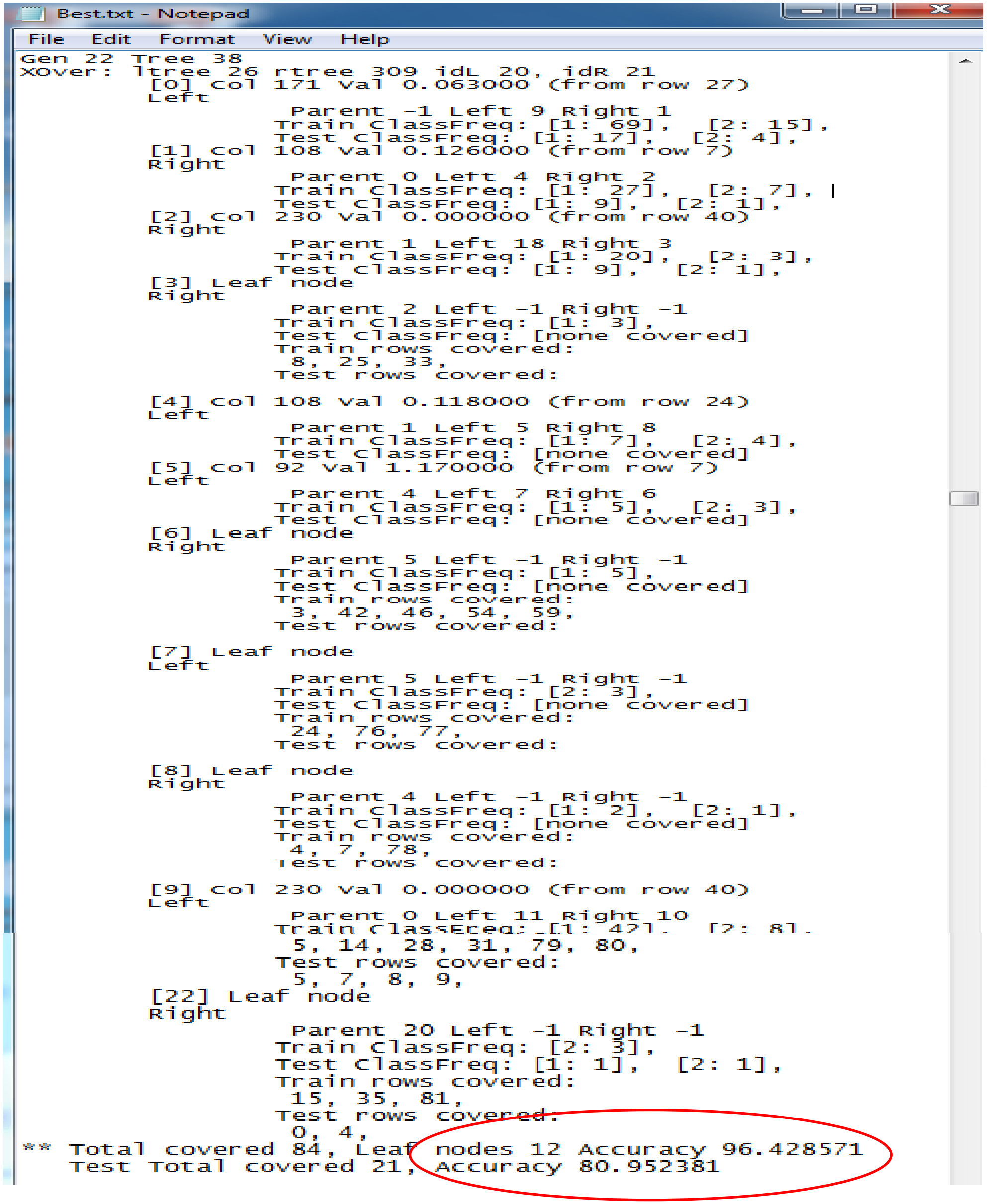


```
EPTREE Train.txt Test.txt 390 60 600 16 0 5 1 2 2
```

Column no containing the class of the data set	390
No of generations required	60
No of trees in each generation required	600
No of trees in the tournament	16
Winners included	0
Low increase in accuracy tolerance	5
% age of mutation	50%
Minimum no of cases in a leaf node	2

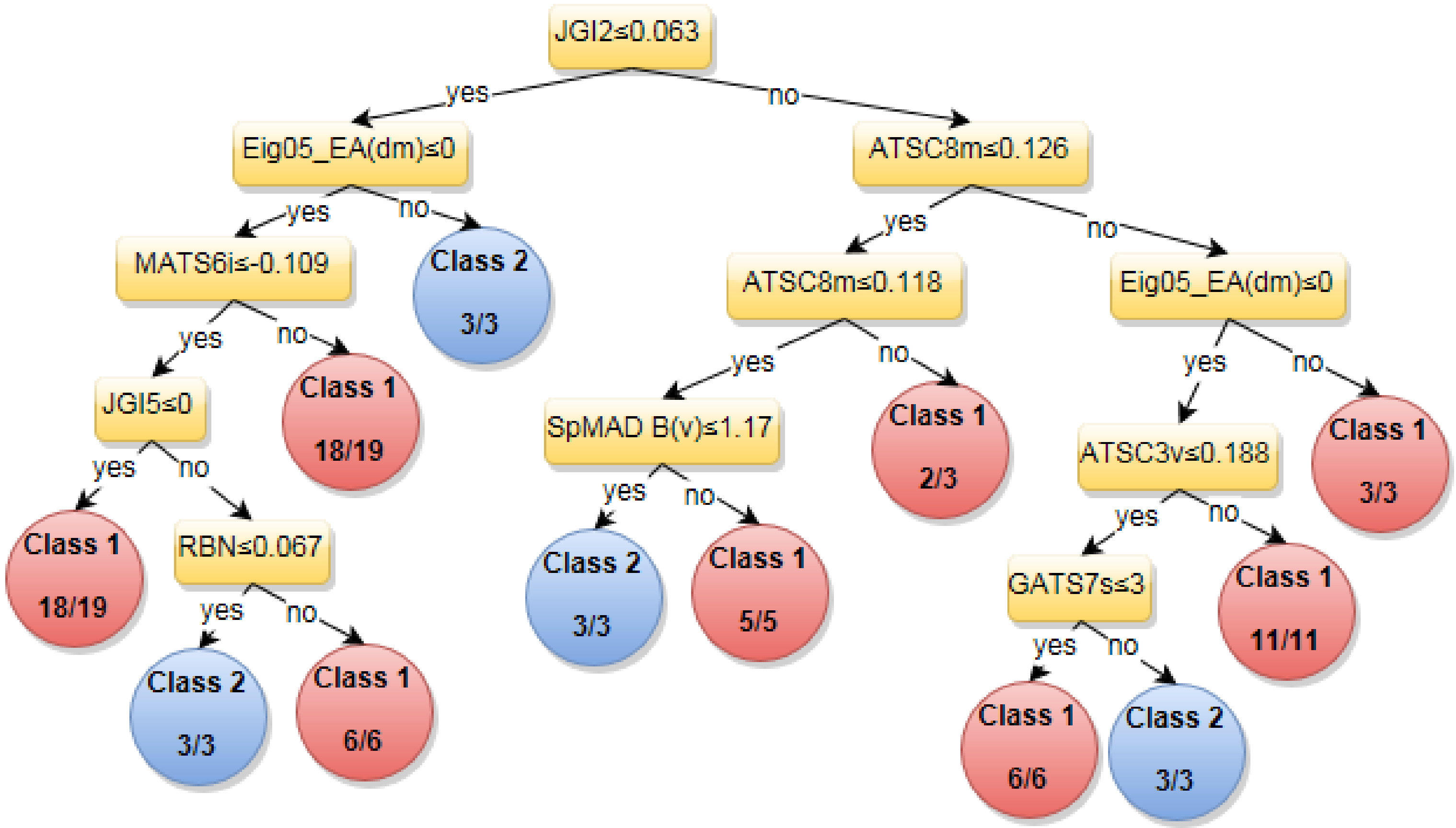
Case Study: Results

GPTree Results



```
Gen 22 Tree 38
XOver:  ltree 26 rtree 309 idL 20, idR 21
        [0] col 171 val 0.063000 (from row 27)
        Left
                 Parent -1 Left 9 Right 1
                Train ClassFreq: [1: 69],  [2: 15],
                Test ClassFreq: [1: 17],  [2: 4],
        [1] col 108 val 0.126000 (from row 7)
        Right
                 Parent 0 Left 4 Right 2
                Train ClassFreq: [1: 27],  [2: 7], |
                Test ClassFreq: [1: 9],  [2: 1],
        [2] col 230 val 0.000000 (from row 40)
        Right
                 Parent 1 Left 18 Right 3
                Train ClassFreq: [1: 20],  [2: 3],
                Test ClassFreq: [1: 9],  [2: 1],
        [3] Leaf node
        Right
                 Parent 2 Left -1 Right -1
                Train ClassFreq: [1: 3],
                Test ClassFreq: [none covered]
                Train rows covered:
                 8, 25, 33,
                Test rows covered:

        [4] col 108 val 0.118000 (from row 24)
        Left
                 Parent 1 Left 5 Right 8
                Train ClassFreq: [1: 7],  [2: 4],
                Test ClassFreq: [none covered]
        [5] col 92 val 1.170000 (from row 7)
        Left
                 Parent 4 Left 7 Right 6
                Train ClassFreq: [1: 5],  [2: 3],
                Test ClassFreq: [none covered]
        [6] Leaf node
        Right
                 Parent 5 Left -1 Right -1
                Train ClassFreq: [1: 5],
                Test ClassFreq: [none covered]
                Train rows covered:
                 3, 42, 46, 54, 59,
                Test rows covered:

        [7] Leaf node
        Left
                 Parent 5 Left -1 Right -1
                Train ClassFreq: [2: 3],
                Test ClassFreq: [none covered]
                Train rows covered:
                 24, 76, 77,
                Test rows covered:

        [8] Leaf node
        Right
                 Parent 4 Left -1 Right -1
                Train ClassFreq: [1: 2],  [2: 1],
                Test ClassFreq: [none covered]
                Train rows covered:
                 4, 7, 78,
                Test rows covered:

        [9] col 230 val 0.000000 (from row 40)
        Left
                 Parent 0 Left 11 Right 10
                Train ClassFreq: [1: 42],  [2: 8],
                 5, 14, 28, 31, 79, 80,
                Test rows covered:
                 5, 7, 8, 9,
        [22] Leaf node
        Right
                 Parent 20 Left -1 Right -1
                Train ClassFreq: [2: 3],
                Test ClassFreq: [1: 1],  [2: 1],
                Train rows covered:
                 15, 35, 81,
                Test rows covered:
                 0, 4,
** Total covered 84, Leaf nodes 12 Accuracy 96.428571
   Test Total covered 21, Accuracy 80.952381
```

Case Study: Results



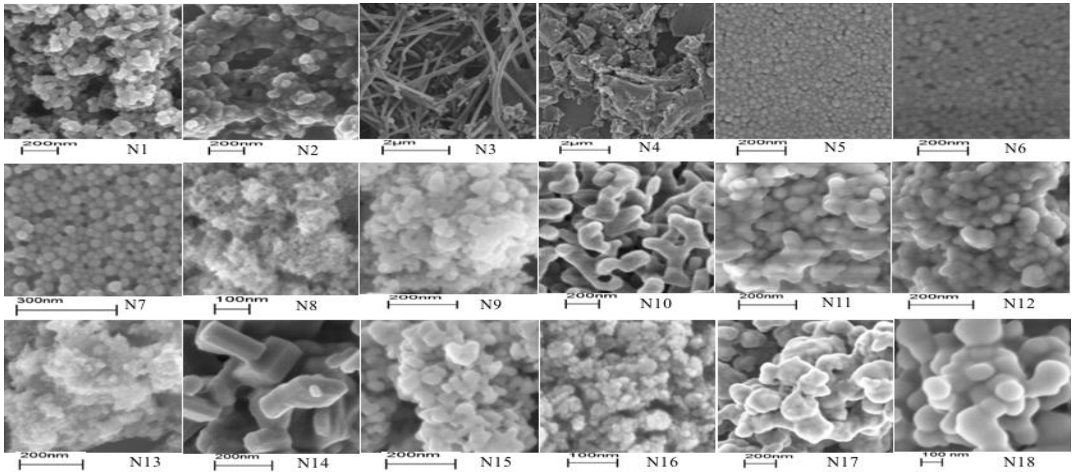
Training accuracy: 96%
Test accuracy: 81%

9 descriptors out of 389

Case Study: Results

DRAGON descriptor	Description	Block
JGI2	mean topological charge index of order 2	2D autocorrelations
JGI5	mean topological charge index of order 5	2D autocorrelations
ATSC8m	Centred Broto-Moreau autocorrelation of lag 8 weighted by mass	2D autocorrelations
ATSC3v	Centred Broto-Moreau autocorrelation of lag 3 weighted by van der Waals volume	2D autocorrelations
MATs6i	Moran autocorrelation of lag 6 weighted by ionization potential	2D autocorrelations
GATS7s	Geary autocorrelation of lag 7 weighted by I-state	2D autocorrelations
Eig05_EA(dm)	eigenvalue n. 5 from edge adjacency mat. weighted by dipole moment	Edge adjacency indices
SpMAD B(v)	spectral mean absolute deviation from Burden matrix weighted by van der Waals volume	2D matrix-based descriptors
RBN	number of rotatable bonds	Constitutional indices

Case Study 3: Data Collection



Carbon Black **N1**

Diesel Exhaust **N2**

Japanese Nanotubes **N3**

Fullerene **N4**

Polystyrene Latex Beads **N5**

Polystyrene Latex Beads **N6**

Polystyrene Latex Beads **N7**

Aluminuim Oxide **N8**

Aluminuim Oxide **N9**

Aluminuim Oxide **N10**

Cerium Oxide **N11**

Nickel Oxide **N12**

Silicon Oxide **N13**

Zinc Oxide **N14**

Titanium Dioxide Rutile **N15**

Titanium Dioxide Anatase **N16**

Silver **N17**

Silver **N18**

Characterization

•***Particle size and size distribution*** were analysed using a Malvern MasterSizer 2000

•***Particle shape*** was analysed using LEO 1530 Scanning Electron Microscope (SEM) or Philips CM20 Transmission Electron Microscope (TEM)

•***Surface area and porosity were measured using*** TriStar 3000 BET

•***The free radical activities*** were measured by EPR

•***Particle reactivity in solution****,* the dithiothreitol (DTT) consumption

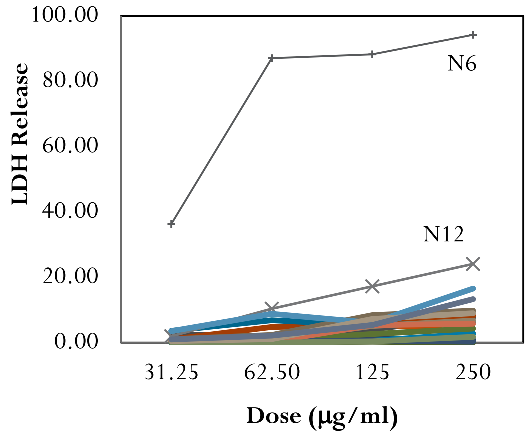
•***Metal Content*** was measured

•***Charge:*** z potential was measured using Malvern Instrument's Zetasizer Nano instrument

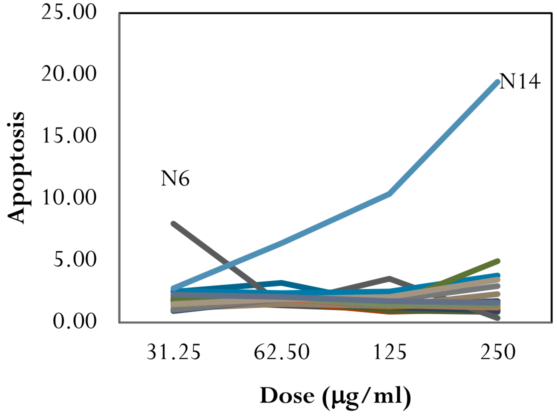
Case Study 3: Data Collection

Toxicological Evaluation

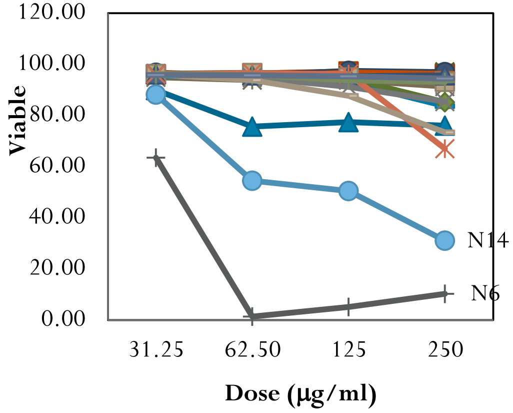
LDH Release



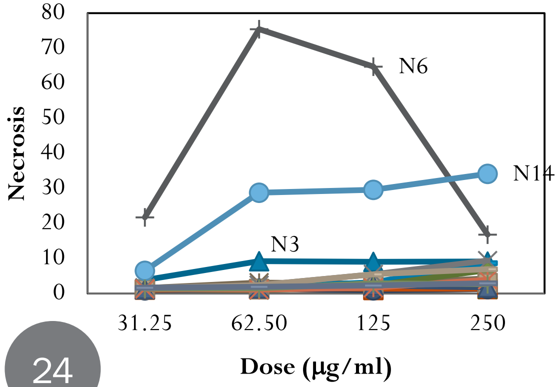
Apoptosis



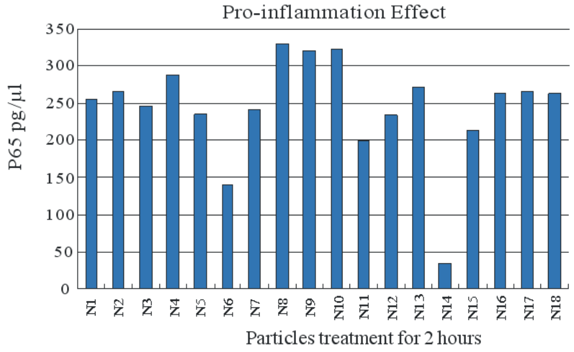
Viability



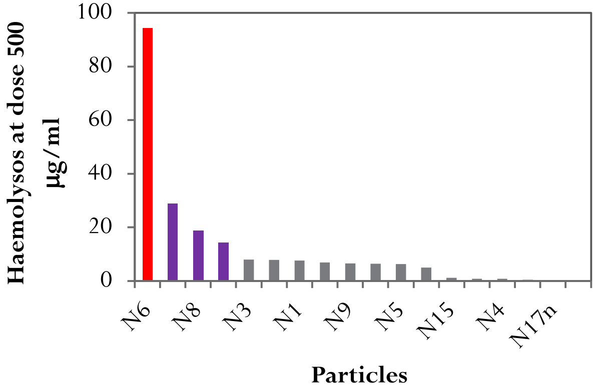
Necrosis



Pro-inflammation effects

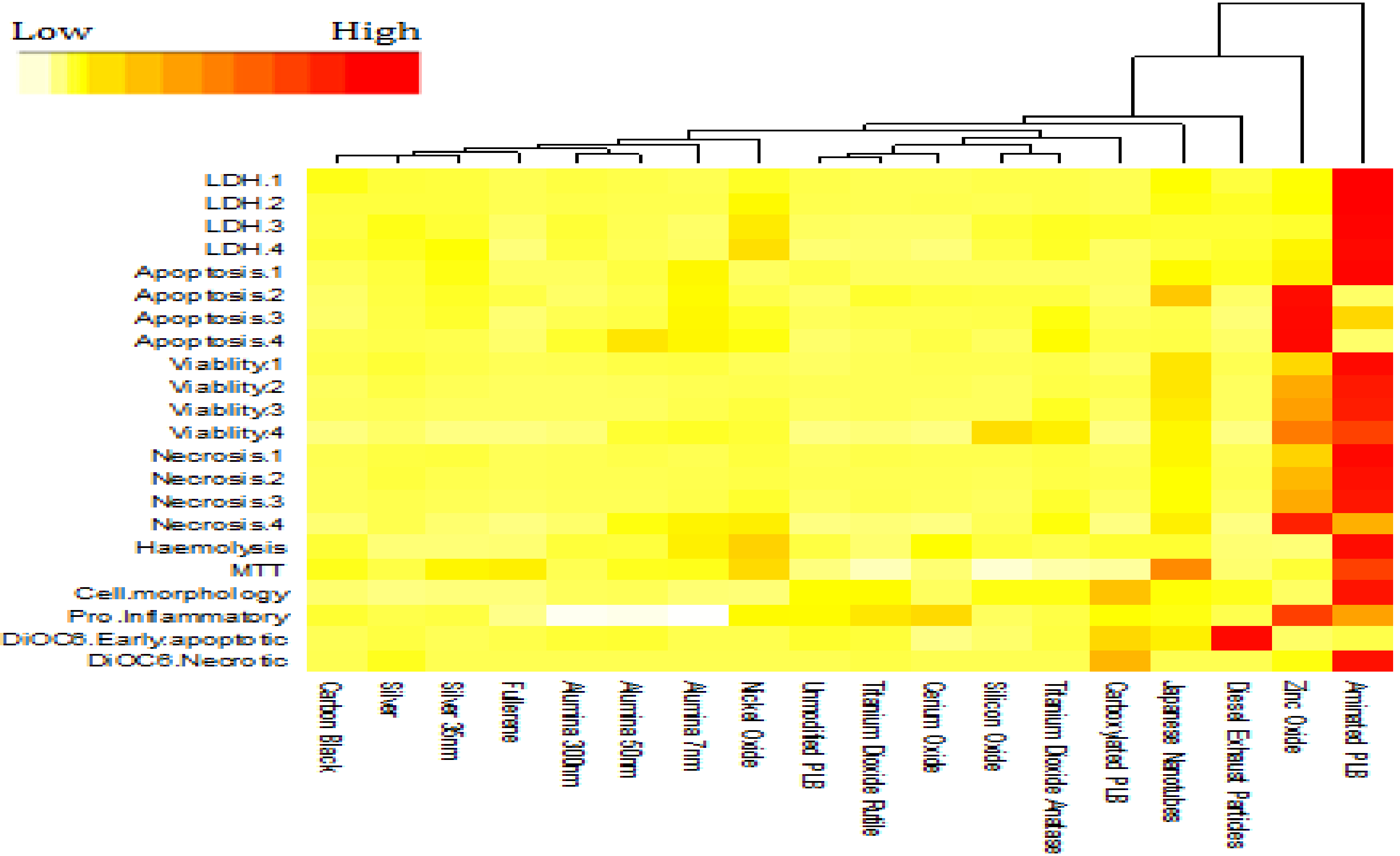


Haemolysis



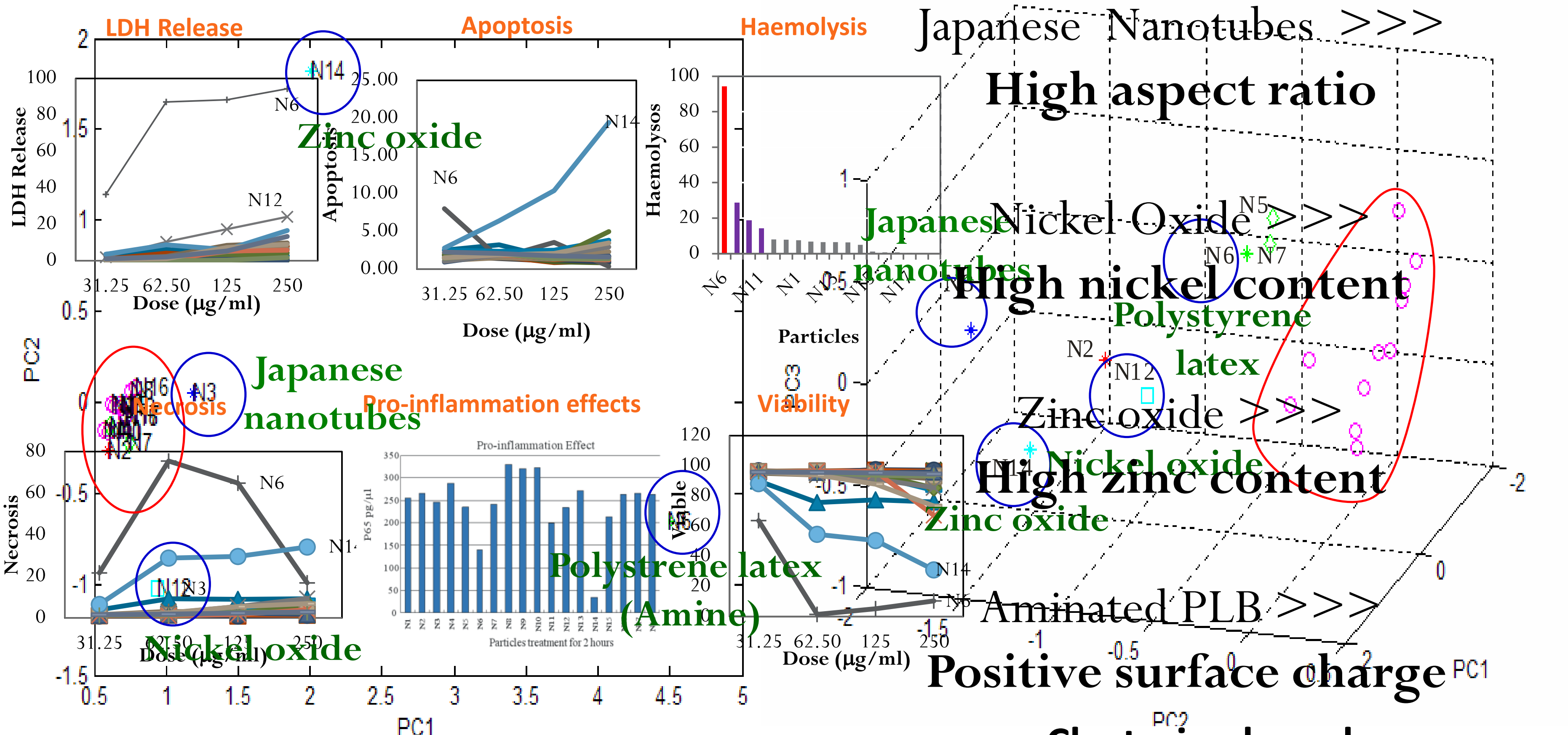
Case Study 3: Data Visualization

Multidimensional data visualization:
Heat maps with hierarchical clustering



Case Study3: Model Development

Clustering/Grouping based on Principal Component Analysis



Clustering based on toxicity data

Clustering based on Characterization data

Wang, Xue Z., et al. "Principal component and causal analysis of structural and acute in vitro toxicity data for nanoparticles." Nanotoxicology 8.5 (2014): 465-476.

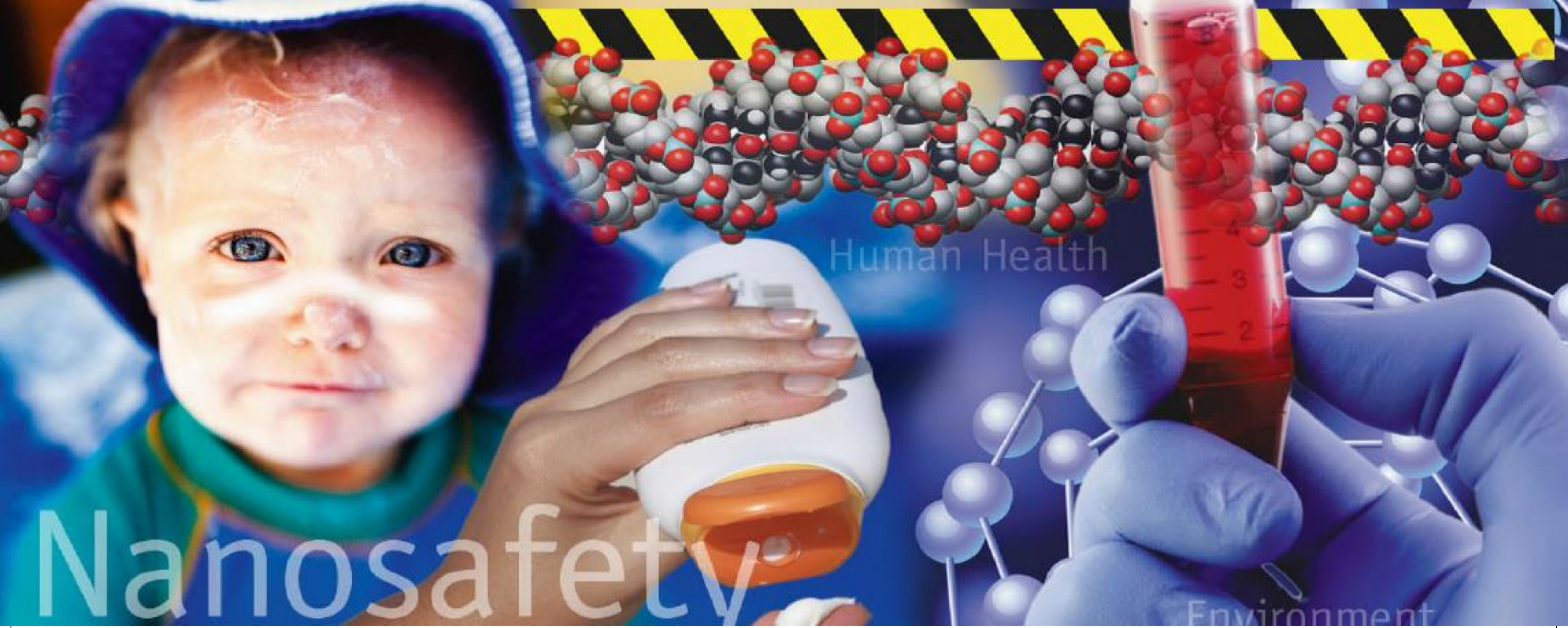
Conclusions

- In LEEDS, we have developed a decision tree software which can be successfully employed for nano-(Q)SAR investigations
- (Q)SAR tools are useful for identifying the properties that influence the toxicity

- Many potential profits:
  - An alternative, fast and cheap way of hazard assessment
  - Risk Reduction
  - Safety-by-design

Future Work

No	Dataset	Nanomaterials	Toxicity Endpoint	Characterization
1	Wang et al. (2014)	18 NMs (carbon-based and metal oxides)	LDH release, apoptosis, pro-inflammatory effects, haemolysis, MTT, DiOC6, cell morphology assay	size, surface area, morphology, metal content, reactivity, free radical generation and zeta potential
2	Shaw et al. (2008)	50 NMs with diverse core structures	ATP content, reducing equivalents, apoptosis, mitochondrial membrane potential	core composition, coating type, surface modification, size, relaxivities and zeta potential
3	NANOMMUNE project	18 NMs	In vitro assays	core, coating, 2 sizes and zeta potential
4	Puzyn et al. (2011)	17 metal oxide NMs	Cytotoxicity (EC50)	12 different quantum-mechanical descriptors
5	MARINA project	9 NMs	In vitro assays	experimental descriptors
6	Weissleder et al. (2005)	109 NMs with the same core but different surface modifiers	Cellular uptake	theoretical descriptors
7	B. Yan (private communication)	80 surface-modified MWCNTs	Protein binding activities, cell viability, nitrogen oxide generation	theoretical descriptors
8	Liu et al. (2011)	9 metal oxide NMs	Cytotoxicity (PI uptake)	a set of 10 descriptors
9	Sayes and Ivanov (2010)	42 NMs with two cores (differing in concentrations)	Cellular membrane damage (LDH release)	primary particle size, size in water and buffered solutions, concentration and zeta potential
10	ENPRA project	10 NMs	In vitro/in vivo assays	size, dustiness, surface area and impurities
11	Gajewicz et al. (2014)	18NMs	Cellular viability (LC50)	18 quantum mechanical descriptors, 11 image descriptors, 3 experimental descriptors



~~NANO-FEAR~~

SUSTAINABILITY of NANOTECHNOLOGY

Thank you !