# Prediction of High Energy Molecules properties using **R**ecursive **Mo**lecular **S**earch (R.Mo.S)

S.Aguero[1], G. Jacob[2], R. Terreux[1]

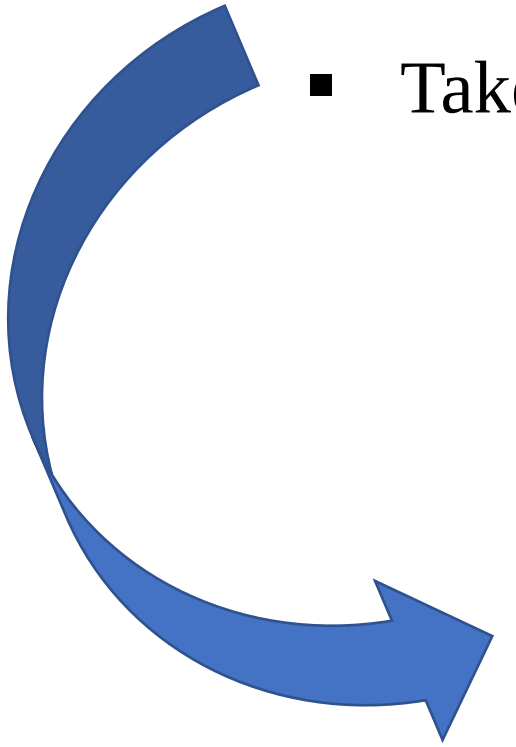
1) EC-Mo Team / Laboratoire de Biologie Tissulaire et d'Ingénierie Thérapeutique (LBTI) - UMR 5305 CNRS - Université Lyon 1. Institut de Biologie et Chimie des Protéines (IBCP), 7 passage du Vercors, 69367 Lyon – France.
2) Centre de Recherches du Bouchet (Ariane Group), 9 Rue Lavoisier, 91710 Vert le Petit – France.

**AVT-340 Research Workshop on "Preparation and Characterization of Energetic Materials"**

**11 February 2021**

# Introduction

- Focus on the improvement of explosives performance
- Take into account toxicological and environmental concerns



Reduction of environmental and health hazard

=

**Major consideration**

# Project Toolbox

In collaboration with Ariane Group (AGS) Since 2012


<u>Project Toolbox:</u> Build optimized tools for predicting the properties of High Energy Molecules (HEM)

- Toxicology
- Physico-chemical properties
- Generators of molecules under constraints
- …

# Toxicological Context

- **Regulatory requirement = REACh**
(**R**egistration, **E**valuation, **A**uthorization and Restriction of **Ch**emicals)

- Large number of *in vitro* tests commonly used

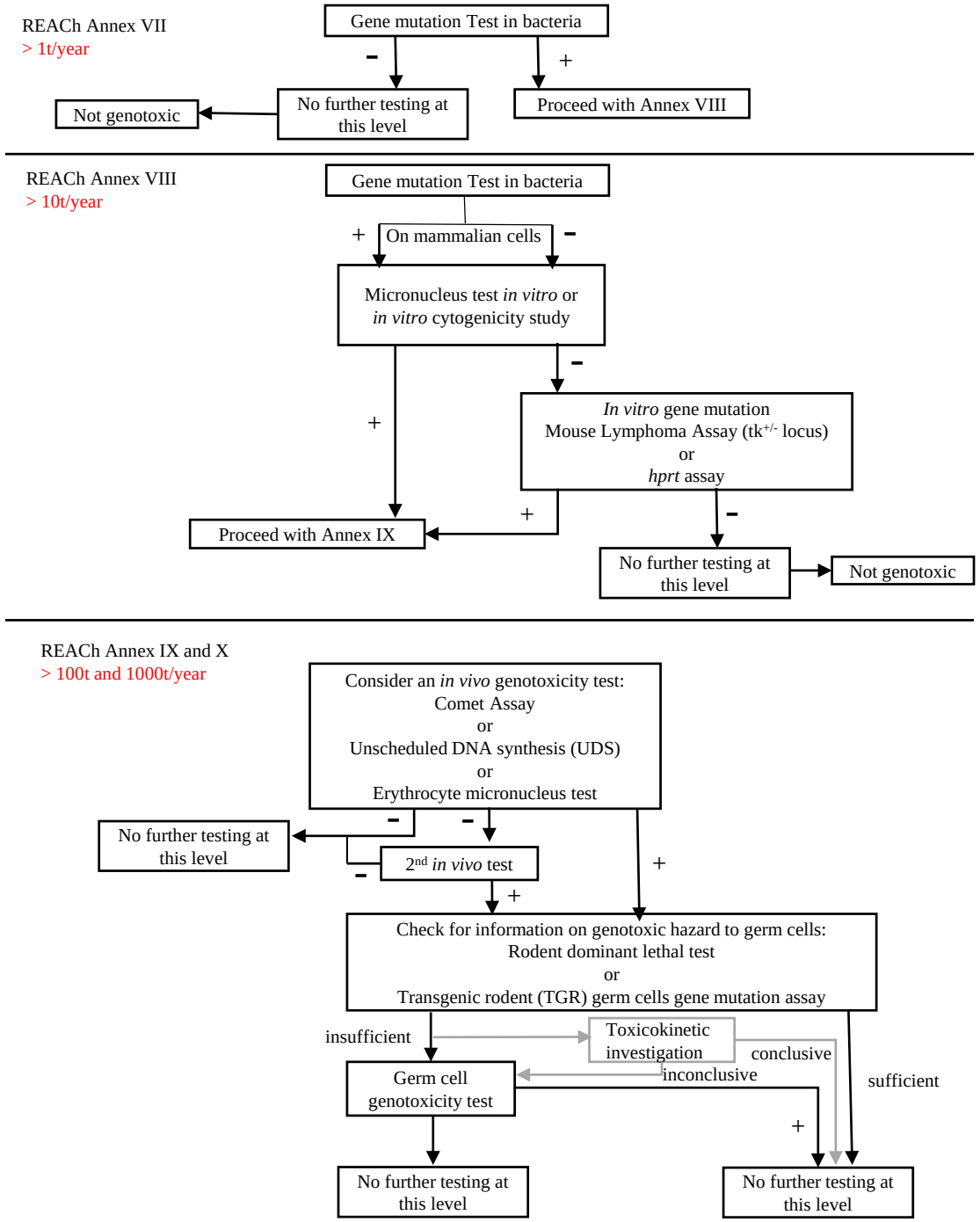
- Set up by REACh and the OECD
(Organization for the Economic Cooperation and Development)

**a variety of harmonized approaches**

- Only 10 categories of toxicity tests are routinely/regulatory used

**mutagenicity**

described by REACh annexes from VI to XI



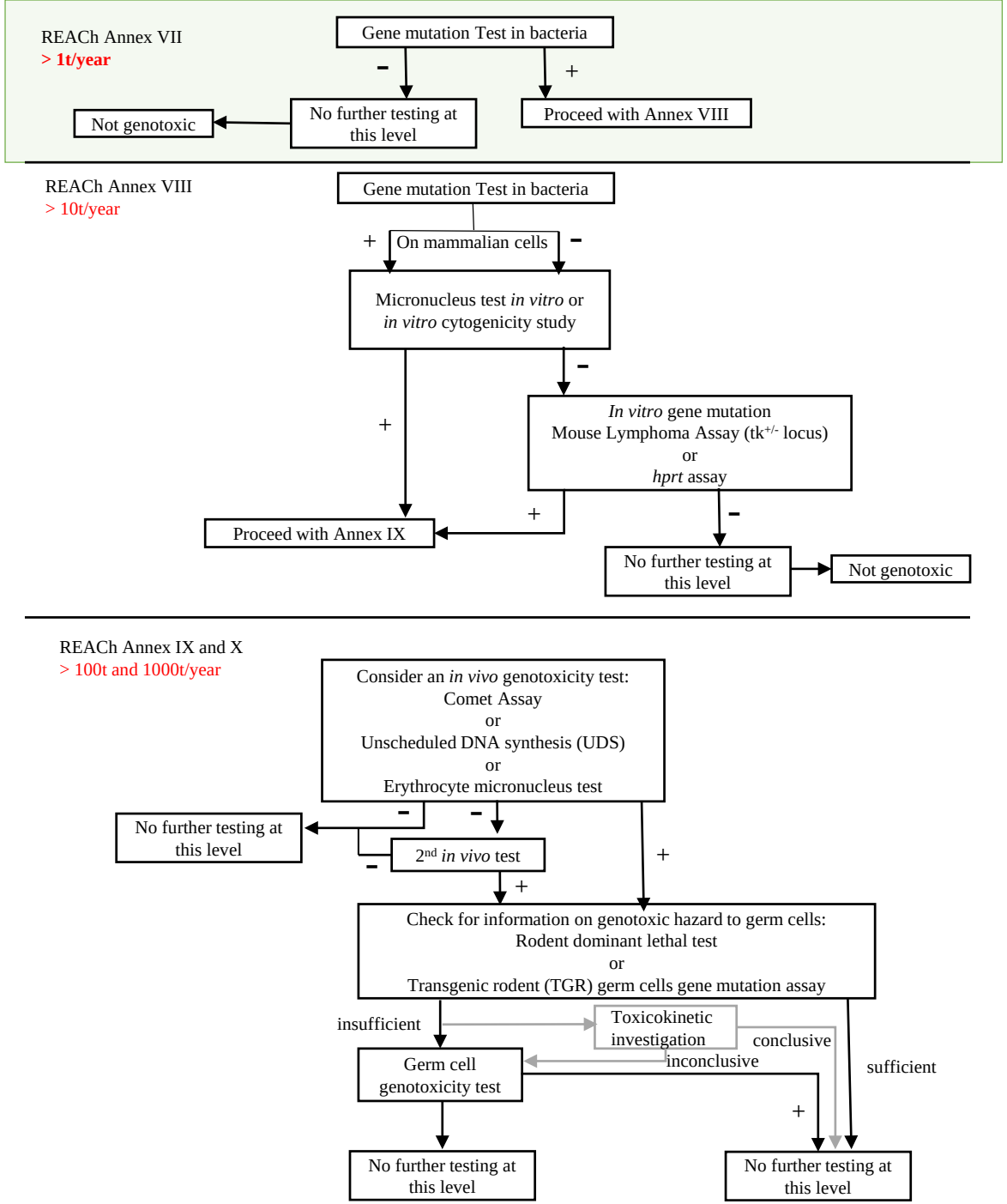

# Context

- Different processus of tests

- Specific of the quantity of production

Officially adopted guidelines for the mutagenicity testing procedure

Schema of the mutagenicity category



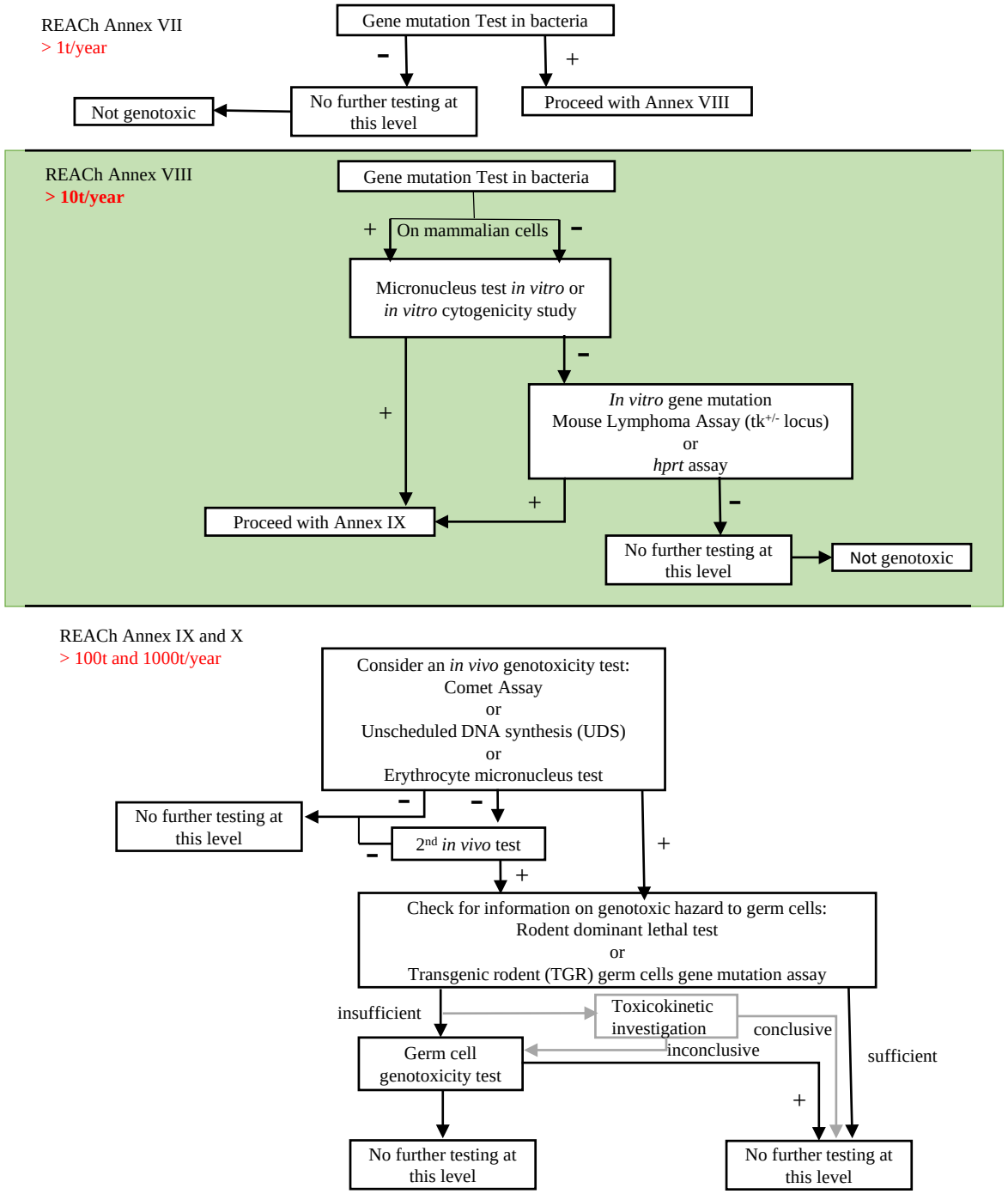

# Context

- Different processus of tests

- Specific of the quantity of production

Officially adopted guidelines for the mutagenicity testing procedure

Schema of the mutagenicity category



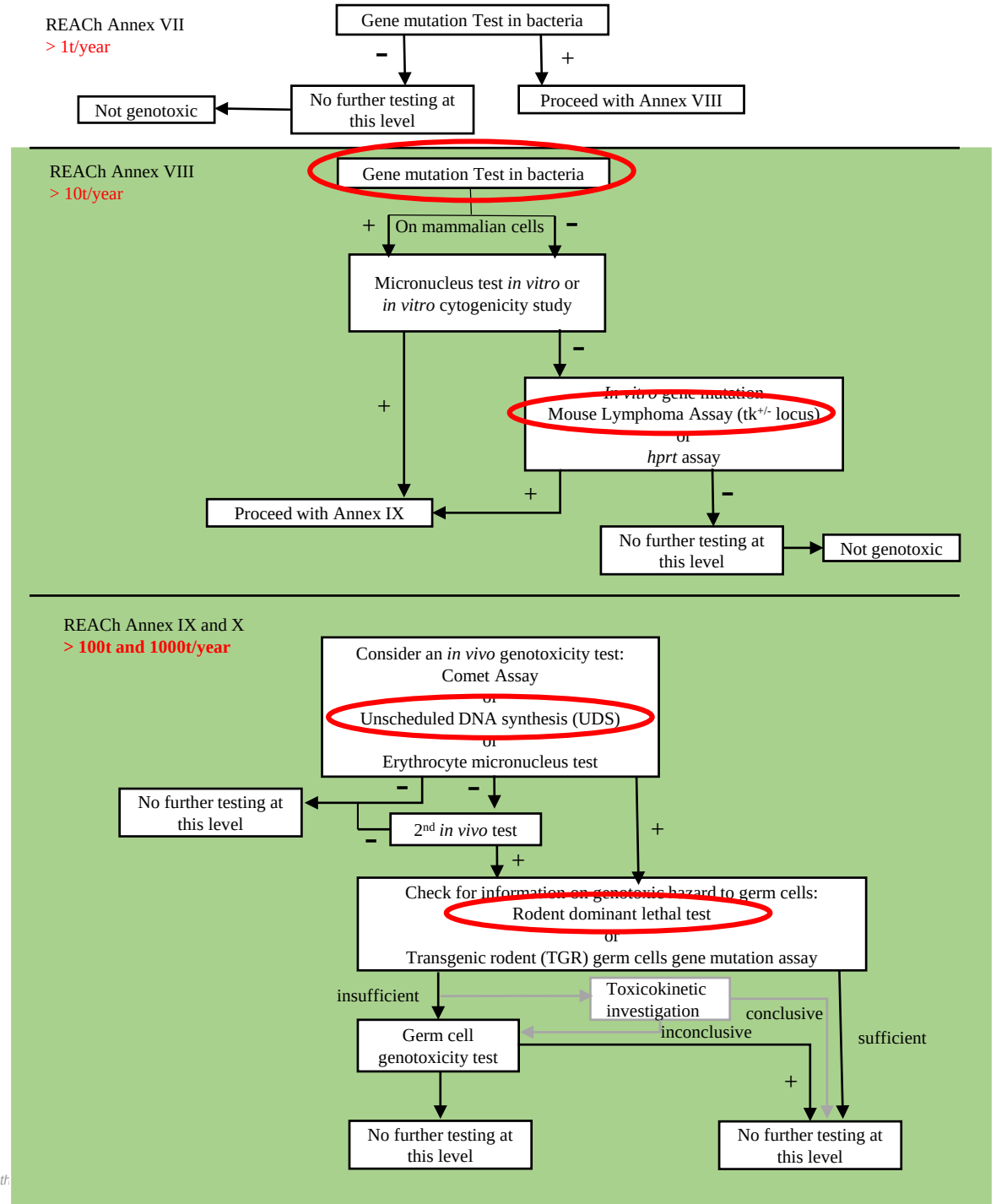

# Context

- Different processus of tests

- Specific of the quantity of production

Officially adopted guidelines for the mutagenicity testing procedure

Schema of the mutagenicity category




# Context

- Different processus of tests

- Specific of the quantity of production

Officially adopted guidelines for the mutagenicity testing procedure

Schema of the mutagenicity category

# Quantitative Structure Activity Relationship (QSAR)

- Quantitative Structure Activity Relationship setup a mathematical equation to link the structure of a set of molecules to their activity / property.
- With molecules set **which the property has been evaluated experimentally**, train your QSAR:

$$Activity = \sum_{1}^{i} k_i . Descriptor_i$$

- Descriptors : functions who answer a finite number using the molecule structure . i.e number of nitrogen Atom.

# Building database (Version 1.0)

## Problem: Lack of experimental data for HEM molecules

Selection of 5 European harmonised databases:

- EURL-ECVAM (European Union Reference Laboratory-European Commission for Alternatives to Animal Testing)
- JRC (Joint Research Center) QSAR Model database
- Carcinogenicity Genotoxicity eXperience (CGX)
- Carcinogenic Potency DataBase (CPDB)
- ISSTOX (Instituto Superiore di Sanità TOXicity databases)

One program: NTP (National Toxicity Program)

**Obtain data for a large number of molecules**

# Table of available test

| | Number of molecules in Db (R.Mo.S) (V1.0) | (V2.0) |
|---|---|---|
| **➔ Mutagenicity :** | | |
| ▪ Ames test | 7.723 | 9.856 |
| ▪ Chromosomal Aberration (CA) test | 1.250 | in progress |
| ▪ Mouse Lymphoma Assay (MLA) | 1.468 | in progress |
| ▪ Unscheduled DNA Synthesis test (UDS) | 650 | in progress |
| ▪ Dominant Lethal Test (DLT) | 70 | in progress |
| **➔ Carcinogenicity** | 1.801 | in progress |
| **➔ Reprotoxicity** | 379 | in progress |

# Quantitative Structure Activity Relationship (QSAR)

- Quantitative Structure Activity Relationship allow to determine a mathematical equation to link the structure of molecule to its activity (properties).
- On molecules set which the property has been evaluated experimentally, train your QSAR:

$$Activity = \sum_{1}^{i} k_i . Descriptor_i$$

- Descriptors : function who answer a finite number using the molecule structure . i.e number of nitrogen Atom.
- Starting this research program with commercial software → 45 % Errors !!!

- <u>QSAR assumption</u> :
  - ~~Molecules of the training set should be similar to m~~
  - ~~The predicted property must have **one** biological / physical mechanism~~

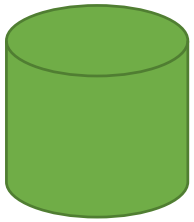
Commercial softwares are made for pharma companies. Drugs… are not similar to HEM

There are million's of toxic mechanisms for a cell.

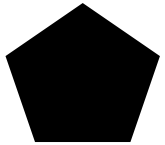
# Predictive methods used in the project

1

2013



Descriptors / Statistical method



Prediction

0/1

- QSAR : no good result

2

2015

3

2017

# Predictive methods used in the project

Large database: Molecules / Prop. X

| Molecule | Ames Test (Exp.) | Descriptor 1 | Descriptor 2 | Descriptor 3 | Descriptor i |
|---|---|---|---|---|---|
| 1 | 0 | 2 | 0,8 | 9,2 | 8 |
| 2 | 1 | 1 | 0,5 | 8,1 | 4 |
| 3 | 1 | 2 | 1,11 | 2,31 | 1 |
| 4 | 0 | 5 | 1,23 | 9,45 | 3 |
| … | | | | | |

Statistical method

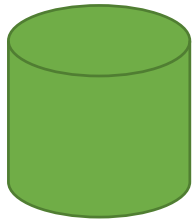
$$Activity = \sum_{1}^{i} k_i . Descriptor_i$$

⬟ Molecule to predict

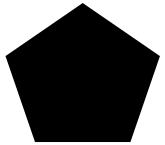
| Molecule | Ames Test PREDICTION | Descriptor 1 | Descriptor 2 | Descriptor 3 | Descriptor i |
|---|---|---|---|---|---|
| ⬟ | $Activity = \sum_{1}^{i} k_i . Descriptor_i$ | 3 | 0,9 | 10,5 | 7 |

# Predictive methods used in the project

1
2013



Descriptors / Statistical method

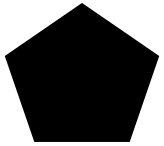


Prediction

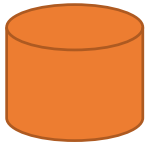
0/1

- QSAR : no good result

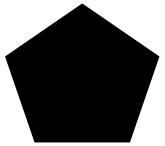
2
2015



Selection of subset by Similarity



Descriptors / Statistical method



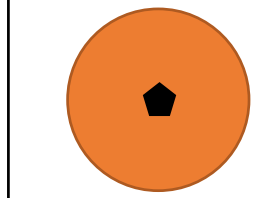
Prediction

0/1

- OTF-QSAR : Good result but lack of adaptation to unknown molecule.

3
2017

# Predictive methods used in the project



Large database: Molecules / Prop. X

| Molecule | Ames Test (Exp.) | Descriptor 1 | Descriptor 2 | Descriptor 3 | Descriptor i |
|---|---|---|---|---|---|
| S1 | 0 | 2 | 0,8 | 9,2 | 8 |
| S2 | 1 | 1 | 0,5 | 8,1 | 4 |
| S3 | 1 | 2 | 1,11 | 2,31 | 1 |
| S4 | 0 | 5 | 1,23 | 9,45 | 3 |
| … | | | | | |

Statistical method

$$Activity = \sum_{1}^{i} k_i . Descriptor_i$$

⬟ Molecule to predict

| Molecule | Ames Test PREDICTION | Descriptor 1 | Descriptor 2 | Descriptor 3 | Descriptor i |
|---|---|---|---|---|---|
| ⬟ | $Activity = \sum_{1}^{i} k_i . Descriptor_i$ | 3 | 0,9 | 10,5 | 7 |

# Predictive methods used in the project

1
2013


- QSAR : no good result

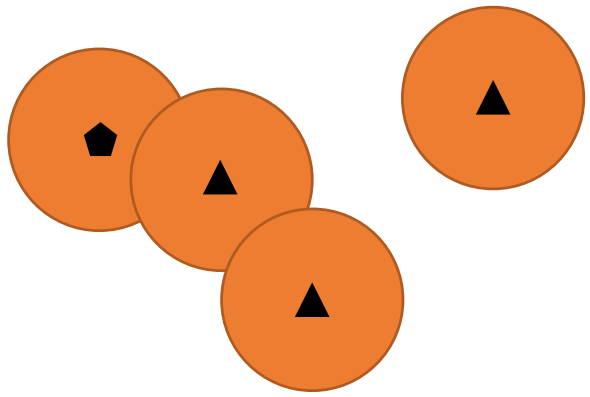
2
2015


- OTF-QSAR : Good result but lack of adaptation to unknown molecule.

3
2017


- R.Mo.S / ML : Few errors ...

# Recursive Molecular Search algorithm (R.Mo.S)

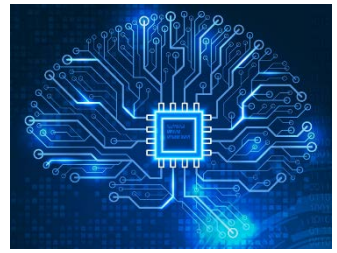


Large database: Molecules / Prop. X

| Molecule | Ames Test (Exp.) | Descriptor 1 | Descriptor 2 | Descriptor 3 | Descriptor i |
|---|---|---|---|---|---|
| S1 | 0 | 2 | 0,8 | 9,2 | 8 |
| S2 | 1 | 1 | 0,5 | 8,1 | 4 |
| S3 | 1 | 2 | 1,11 | 2,31 | 1 |
| S4 | 0 | 5 | 1,23 | 9,45 | 3 |
| … | | | | | |

Output | Input

Machine Learning

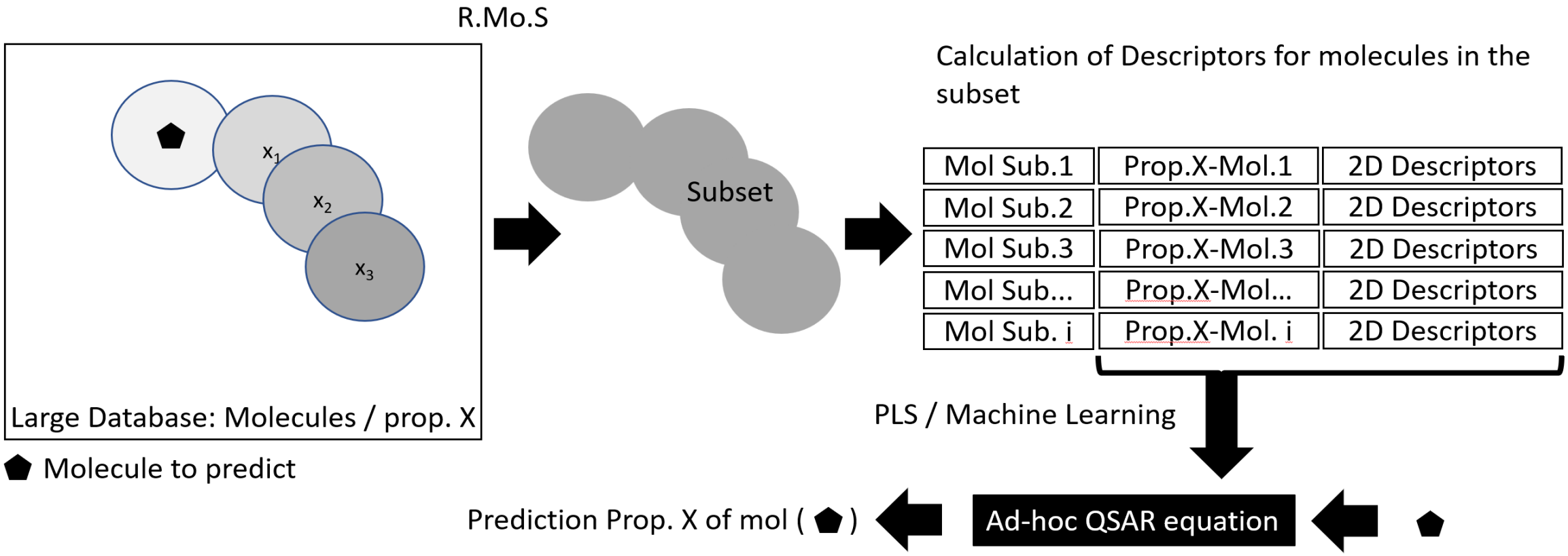


Prediction system

⬟ Molecule to predict

| Molecule | Ames Test PREDICTION | Descriptor 1 | Descriptor 2 | Descriptor 3 | Descriptor i |
|---|---|---|---|---|---|
| ⬟ |  => 0/1 | 3 | 0,9 | 10,5 | 7 |

# Recursive Molecular Search algorithm (R.Mo.S)




- The R.Mo.S algorithm will create a new training set of molecule far away from the unknown molecule.
- Set up a predictive system on unknown molecules
- R.Mo.S patent AG / CNRS / UCBL patent WO 2018 / 234718 extension 2020 : EU / USA
- Alliod et al., PEP 2017.

# Recursive Molecular Search algorithm (R.Mo.S) Development

- Optimization of 2D descriptors for machine learning Algorithm.
  - Reduction of number of 2D descriptors and incorporation of MACCS


Parameter optimization


reduction of number of descriptors

Cross-validation method

- Optimization of database and fill database with news data from literature.
  - Ames DB V1 7.723 molecule ⇨ 9.856 molecules.

# R.Mo.S Testing : Test Set

## ➔Test Set

- Molecules not in database
- Specialized Test set :
  - Fillers 20 molecules
  - Plasticizer 5 Molecules
  - Oxidizer 6 Molecules
  - Pharmacology 45 Molecules
  - Pyrotechnics 9 Molecules

**i.e pharmacology test set predicted by R.Mo.S and ACD percepta.**

**Errors are in Red**

| Pharmacology | Exp. Data | R.Mo.S / Machine Learning | ACD Percepta |
|---|---|---|---|
| 1-methylpyrène | + | - 87% | + |
| 1-methyl-4-nitrobenzene | + | - 70% | - |
| 1,2,3-benzotriazol | + | + 60% | + |
| 1,2-dichoroethane | + | + 82% | + |
| 1,4-dithiane | - | D 55% | + |
| 1,4-oxanthiane | - | - 92% | - |
| 1,8-dinitronaphalene | + | + 69% | + |
| 2-aminopyridine | - | - 88% | - |
| 2-methylbenzamide | - | - 63% | - |
| 2-nitropropane | + | - 78% | + |
| 5-nitro-o-toluidine | + | + 70% | + |
| Acetamide | - | - 68% | - |
| Aniline | - | - 96% | - |
| Aspartame | - | - 100% | - |
| Benzothiazole | - | - 88% | - |
| Benzyl chloride | + | + 95% | + |
| Butanal oxime | + | + 93% | + |
| CI Acid Orange 3 | + | - 79% | + |
| CI Allura red 17 | - | - 99% | - |
| Caffeine | - | - 100% | - |
| Caprylyl chloride | + | + 97% | + |
| DEPH | - | - 76% | - |
| Dibenz[a,h] anthracene | + | + 88% | + |
| Disobutylphtalate | - | - 84% | - |
| Diisodecylphtalate | - | - 86% | - |
| Di-n-octyl phthalate | - | - 87% | - |
| Estradiol | - | - 72% | - |
| Ethylene glycol | - | - 90% | - |
| HNF | + | | D 50% |
| Hydrazine hydrate | + | + 61% | |
| Hydrazine perchlorate | - | - 64% | |
| Isopropyl methylphosphonic acid | - | - 66% | - |
| p-chlorophenyl methyl sulfide | - | - 84% | - |
| p-chlorophenyl methyl sulfone | - | - 81% | - |
| p-chlorophenyl methyl sulfoxide | - | - 92% | - |
| Phenylacetonitrile | - | - 74% | - |
| Yellow 74 pigment | - | - 93% | - |
| Pitavastatin | - | - 100% | + |
| Propionitrile | - | - 81% | - |
| o-cresol | - | - 95% | - |
| Stearic acid | - | - 95% | - |
| L-taurine | - | - 88% | - |
| Theophylline | - | - 91% | - |
| Thiourea | - | - 96% | - |
| Vitamin E | - | - 97% | - |

# Prediction results comparison

| | R.Mo.S | ACD percepta |
|---|---|---|
| **Ames test** | 72/85 | 64/85 (2 impossible) |
| **CA Test** | 71/85 | 49/85 (1 impossible) |
| **MLA Test** | 69/85 | 45/85 |
| **UDS Test** | 67/85 | 57/85 (2 impossible) |
| **DLT Test** | 75/85 | Not Available |
| **Carcinogenicity Test** | 79/85 | 69/85 (5 impossible) |
| **Reprotoxicity Test** | 75/85 | 39/85 |
| **Number of good responses** | 508/595 | 323/595 |
| **(%) of good responses** | **85,4** | **54,3** |

- Excellent result ≈ 85 % good prediction (R.Mo.S V1)
- R.Mo.S V2 : 88 % (Ames Test)
- Best result than Gold standard (ACD Percepta, ACD Labs)
- DLT test not available for ACD.
- Real difference on reprotoxicity test.

# Conclusion

- Best prediction on prediction for “R.Mo.S / Machine learning” approach.
- Prediction could be performed on HEM, liquid, salts, energetic or not molecules, pharmaceuticals…
- Around 85% of good prediction on different sets of molecules.
- “R.Mo.S” is patented since December 2018 by AG / CNRS / UCBL.

- “R.Mo.S / ML” software development :
  - Deployment on web service and commercial distribution is in progress.
  - Databases are improved and algorithm too… R.Mo.S Version 2 soon !

# Acknowledgment

- *In silico* tox team
  - Pr. R. Terreux
  - Dr. JA Chemelle
  - S. Aguero
  - M. Fournier
- AG team (CRB)
  - J.F. Guery
  - G. Jacob
  - L. Blarasin

We are thankful to DGA for funding this project