Machine learning to improve the impact assessment of micropollutants release from WWTP

R. Servien
joint work with C. Leenknecht, E. Latrille, L. Mamy, P. Benoit, A. Helias and D. Patureau

UR 0050 LBE

INRAE

Environmental contamination

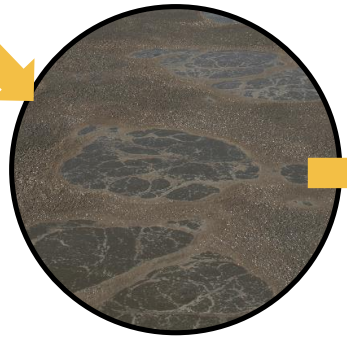
Micropollutants

Wastewater treatment plant

Atmosphere

Aquatic environment

Sludge



Soil

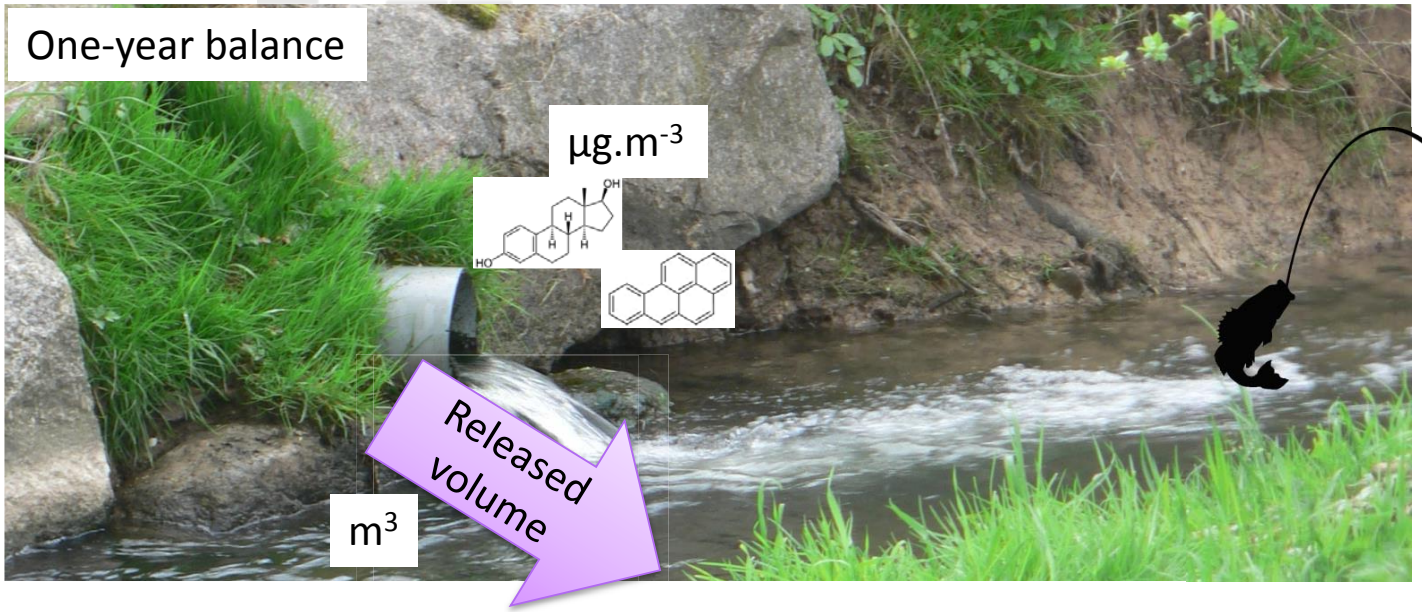
Transfer of contamination:

- Sorption to sludge
- Volatilization

Elimination from water:

- Physicochemical or biological transformation
- Mineralization

Calculation



Concentration
x
Volume
=
Mass released to aquatic environment

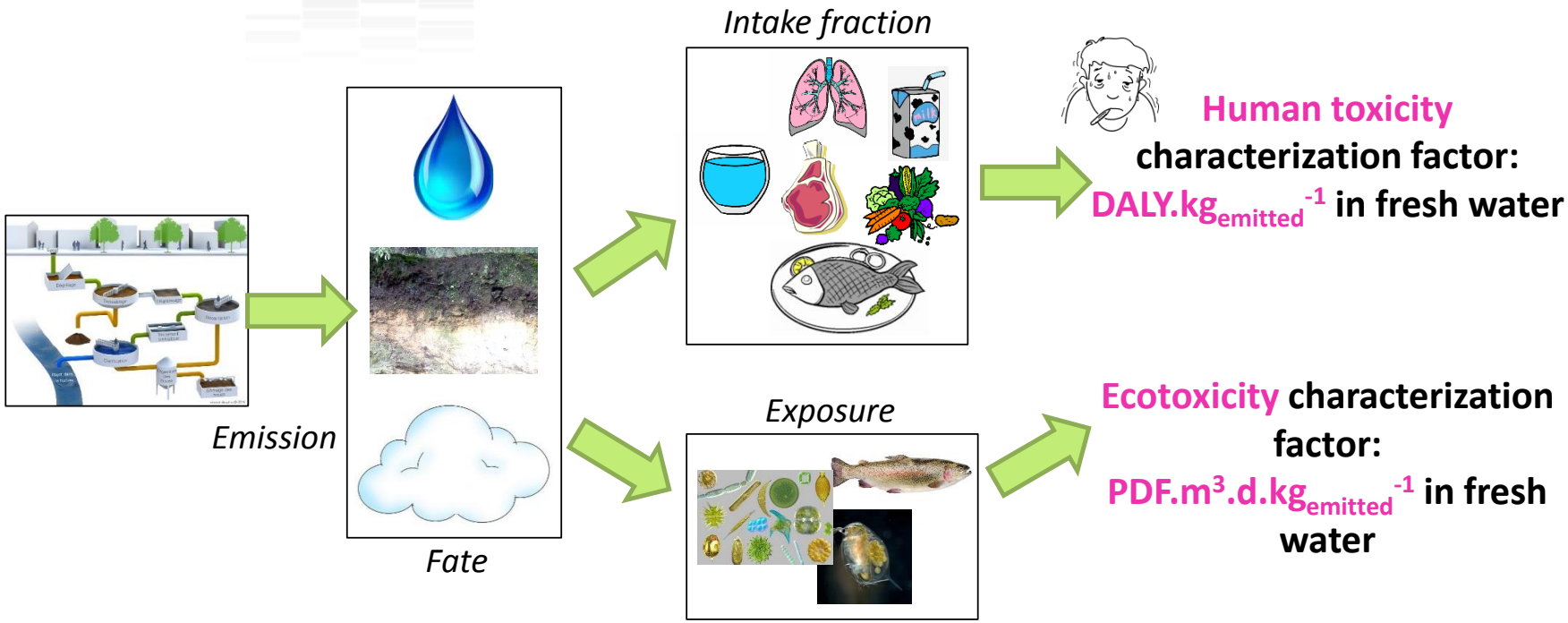
Potential impacts on Human health and aquatic environment

x characterization factor (USEtox 2.1 ®)

Characterization factor

USEtox®

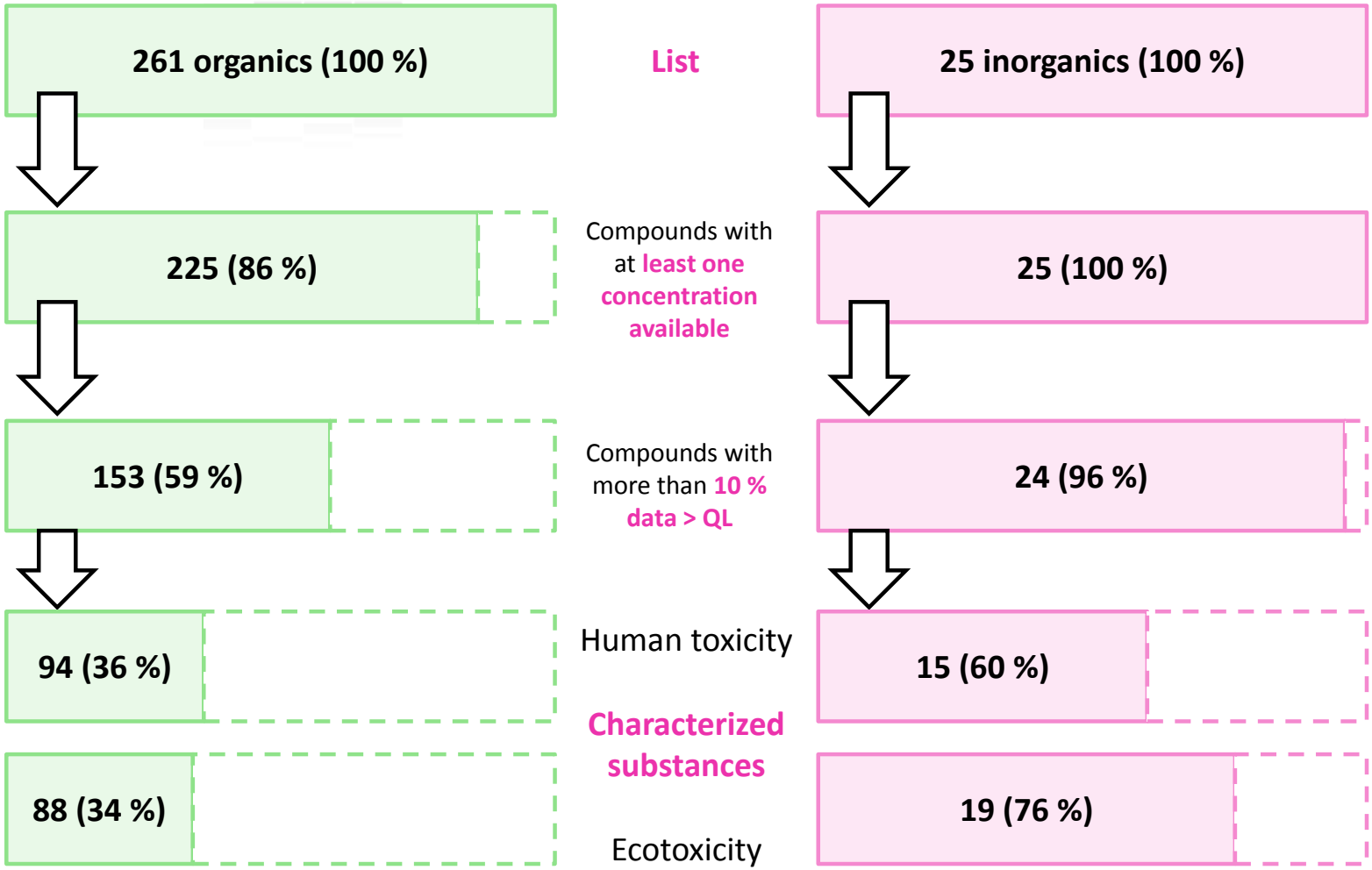
Reference method in Life Cycle Assessment for assessing human toxicity and freshwater ecotoxicity



DALY = Disability Adjusted Life Years = number of life years « lost » because of illness, handicap or death

PDF.m^3.j = Potentially Disappeared Fraction x cubic meter x day = fraction of species potentially disappeared integrated to volume and time

Available data and selecting



Lack of characterization factors (especially for pharmaceuticals)

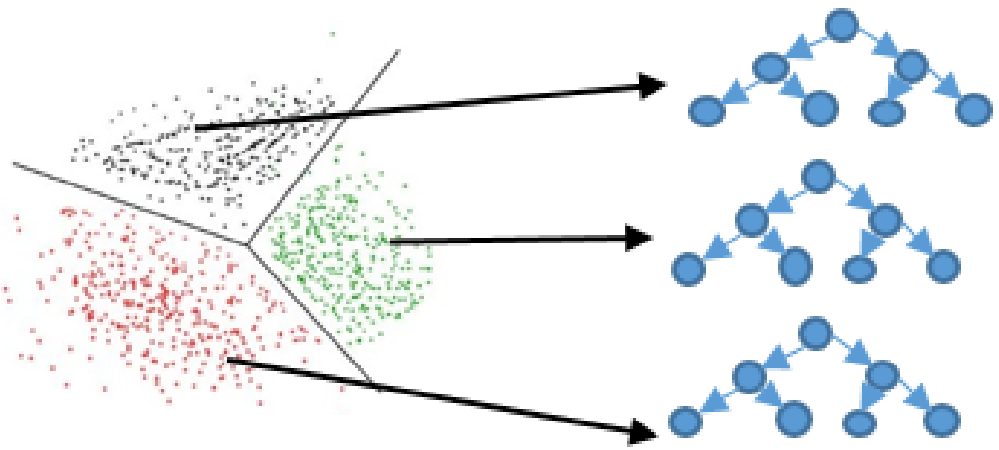
Conclusion

- Conclusions
  - Toxic for the aquatic environment $\neq$ toxic for human health
  - High mass $\neq$ high impact
  - Low mass + high CF $\rightarrow$ high impact
- Limitations
  - Lack of data (in particular for pharmaceuticals)
  - Incomplete study of impacts (only 1/3 of selected micropollutants)
  - Making experiments to obtain these CF is time-consuming and expensive
- Ideas
  - DB TyPol contains data (molecular descriptors) on this kind of compounds
  - Predict missing CF using modelisation and molecular descriptors (that are easily obtainable for a new compound)

Bibliography

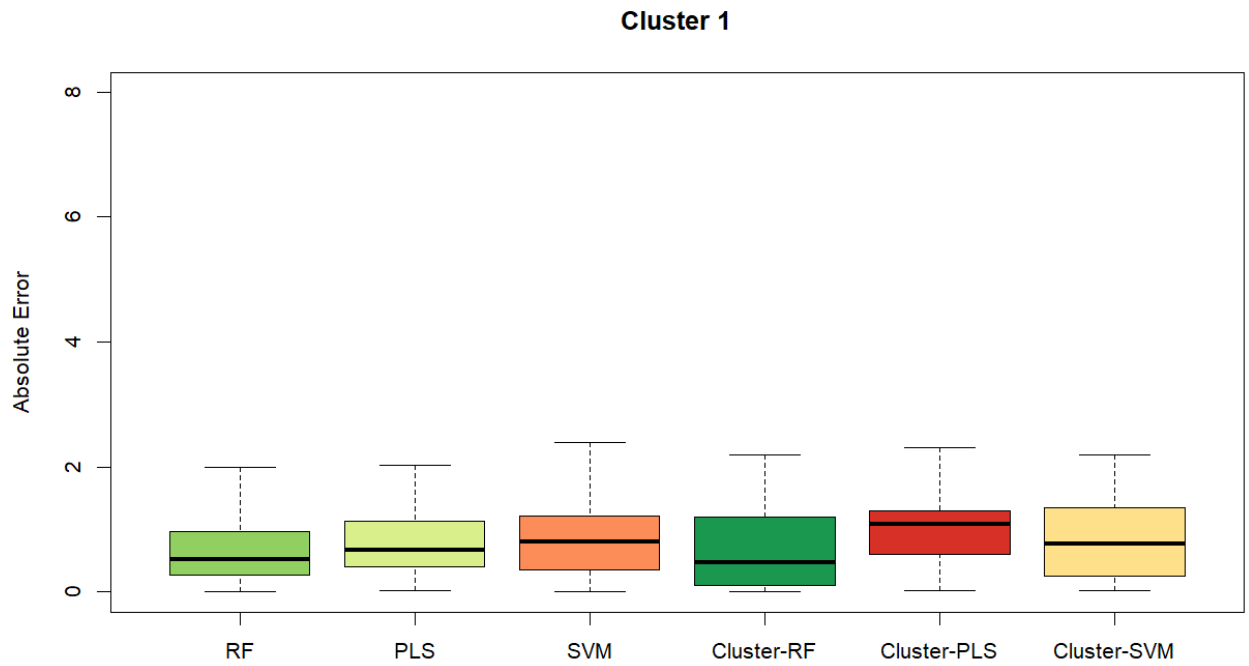
- QSAR approach
  - Danish QSAR database (DTU, 2015), ECOSAR (Mayo-Bean et al., 2011), VEGA 30 (Benfenati et al., 2013)
  - Used to obtain LC50 data
  - Remains mostly linear models
- Machine learning approaches
  - Hou, ..., Jolliet, Xu (2020a, 2020b)
  - Based on more complex data not only molecular descriptors to predict mid-points (not CF)
- Goal : Test different modeling methods based on molecular descriptors to predict CFs (one for CF_{ET} and one for CF_{HT}). Error of 1 log accepted.

- Linear model : Partial least squares (PLS)
- Machine learning (SVM and RF)
- Cluster-then-predict approaches
  - Clustering used on molecular descriptors
  - PLS, SVM and RF applied on each cluster separately



Comparison procedure

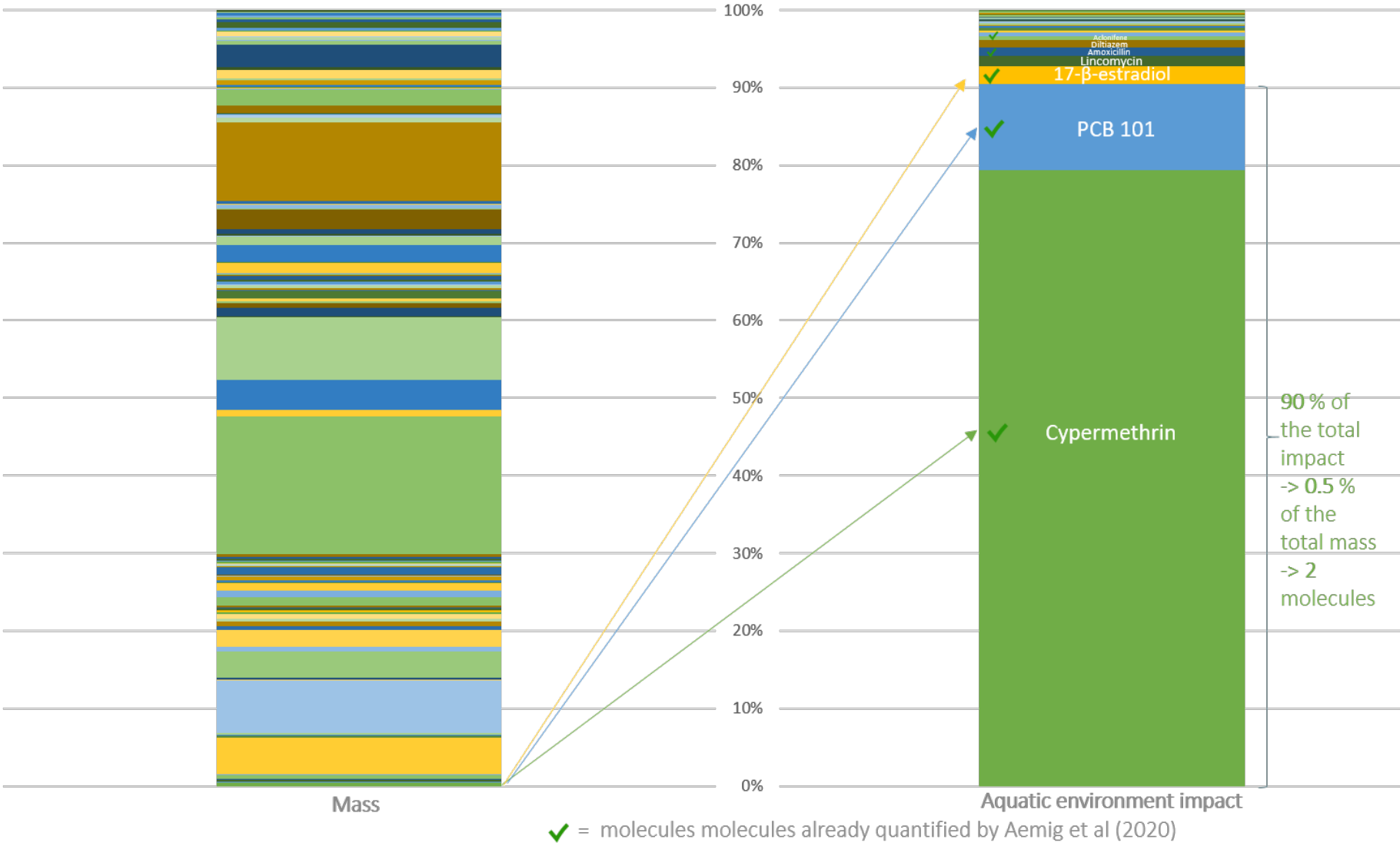
- TyPol : 529 molecules, Usetox $\sim$ 3000, 269 in common (high impact).
- Train and test
- Results for CF_{HT} for cluster 1



- Validation of the prediction models (<1 log).

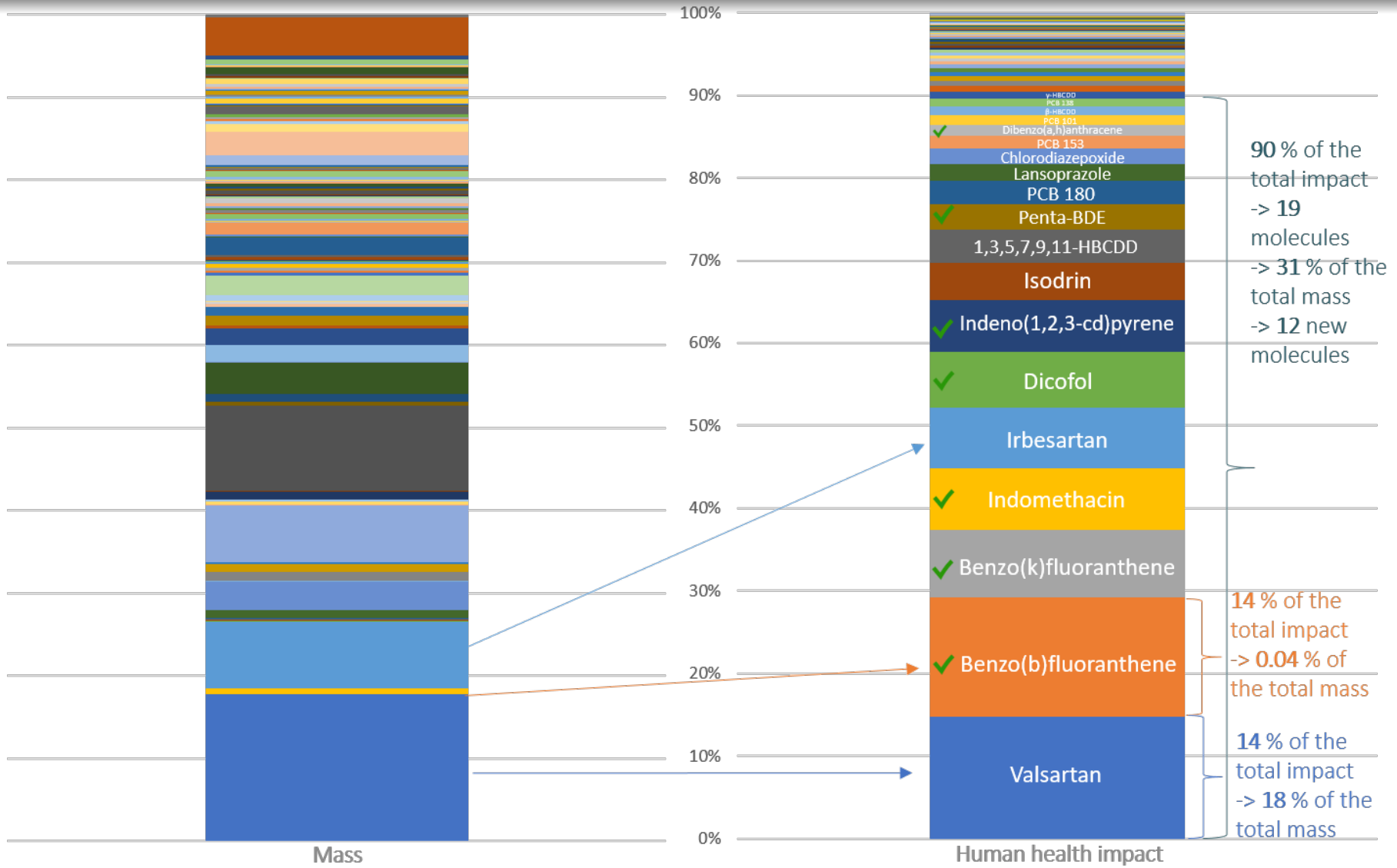
Computation of missing CFs

- Calculation of molecular descriptors for molecules without CF
- Prediction of new CFs (HT and ET) using previous models (<1 sec)
- Number of micropollutants taken into account nearly doubled
- Now 153 micropollutants on 261
- Update of the analysis of the potential impacts on human health and aquatic environment in continental freshwater


100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Aclonifene
Diltiazem
Amoxicillin
Lincomycin
17-β-estradiol
PCB 101
Cypermethrin
90 % of the total impact
-> 0.5 % of the total mass
-> 2 molecules
Mass
Aquatic environment impact
= molecules molecules already quantified by Aemig et al (2020)

Impact of compounds of the new study

- Conclusions of Aemig et al. still holds
  - 90 % of the overall impact still driven by 2 molecules with high CF
  - In 10 most impacting molecules, three are new
  - the total impact increased of only 4%


100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
γ-HBCDD
PCB 138
β-HBCDD
PCB 101
Dibenzo(a,h)anthracene
PCB 153
Chlorodiazepoxide
Lansoprazole
PCB 180
Penta-BDE
1,3,5,7,9,11-HBCDD
Isodrin
Indeno(1,2,3-cd)pyrene
Dicofol
Irbesartan
Indomethacin
Benzo(k)fluoranthene
Benzo(b)fluoranthene
Valsartan
90 % of the total impact -> 19 molecules -> 31 % of the total mass -> 12 new molecules
14 % of the total impact -> 0.04 % of the total mass
14 % of the total impact -> 18 % of the total mass
Mass
Human health impact
✓ = molecules molecules already quantified by Aemig et al (2020)

Impact of compounds of the new study

- Modified conclusions
  - 94 % of the total impact is induced by 26 (prev. 8) molecules and 41 % (prev. 4%) of the total mass
  - Total impact nearly doubled
  - Benzo(b)fluoranthene and Valsartan = same impact (14%) but not same explanation (CF/mass)

Conclusion

- CF (for human health and ecotoxicological impacts) were predicted using molecular descriptors
- Predictions using machine learning were good
- This methodology was then used to derive missing CFs
- It was used to complement study of impact assessment of WWTP
  - No change for aquatic environment impact
  - Important changes for human health impact
- Could be used for priorisation of compounds for example
  - Cypermethrin for HT
  - More complex for ET

Perspectives

- Study still incomplete ! (153/261)
- Impact assesment only limited by the mass estimation
  - Estimation of the missing masses using available data (and bootstrap)
  - Completion of the study for all the micropollutants
  - With confidence intervals
- Submitted preprints
  - Models `https://www.biorxiv.org/content/10.1101/2021.07.20.453034`
  - Applications `https://hal.archives-ouvertes.fr/hal-03346134`

Thank you for your attention !