An OECD compliant *in silico* model predicting ready, inherent biodegradability or potential P of organic chemicals <1000 Da

LARRAS Floriane[1] , THOMAS Paul[1] , AY-ALBRECHT Emel[1]

[1] KREATiS, 38080 L'Isle d'Abeau – France

Contact us : www.kreatis.eu / contact@kreatis.eu

INTRODUCTION

Aerobic microbial biodegradation of organic chemicals is a key environmental process mandatory for the Persistence Bioaccumulation Toxicity (PBT) triad in various regulatory frameworks. The **OECD 301/310 Guidelines** (OECD, 1992) set out the criteria to experimentally determine if a chemical is biodegradable: Readily (RB) is 60% mineralisation within 28 days meeting the 10-day window. Non-P is >60% within 60 days. However, the criteria are kinetic based, the test is rather long to conduct and the methodology suffers from a high degree of variability because of the spatiotemporal structural and functional variability of the inoculum. Implementaion of a Safe and Sustainable by Design approach, quick and robust predictions for biodegradable are important. Thus, we describe a **new Structure Activity Relationship (SAR) model** compliant with the five OECD principles for (Q)SAR method validation (OECD, 2007) using **two major curation steps** and **two machine learning approaches** while enhancing the uncertainty assessment and mechanistic understanding.

DATA MANAGEMENT

DATA AQUISITION

7326 substances

CURATION 1 — Data reliability or relevance

- ✓ Endpoint of purpose
- ✓ Experimental study
- ✓ OECD 301/310 guidelines compliant
- ✓ heterogenous results check
- ✓ Study reliability check
- ✓ Monoconstituants

CURATION 2 — Clustering approach

Review of 677 clusters

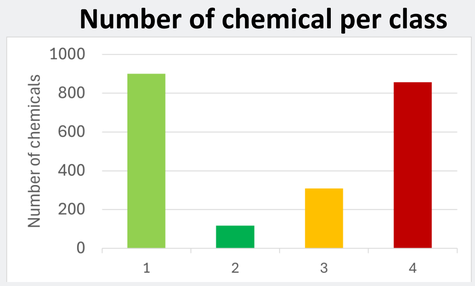
- ✓ Ca. 477 clusters — In accordance
- ✗ Ca. 200 clusters — Expert investigation

2183 substances

OECD301/310 RESULTS

- **RB class** (readily biodegradable): classes 1 and 2 merged.
- **Non-RB class** (Not readily biodegradable) : classes 3 and 4 merged.
- RB and non-RB classes represent 47% and 53% of the dataset. **The dataset is considered as balanced.**

Number of chemical per class



1: readily biodegradable (≥60% at 28-day & 10-day window);
2: readily biodegradable failing 10-day window (≥60% at 28-day);
3: inherently biodegradable (≥20% & 60%< at 28-day; ≥60% at 60-day);
4: persistent (<20% at 28-day; <60% at 60-day)

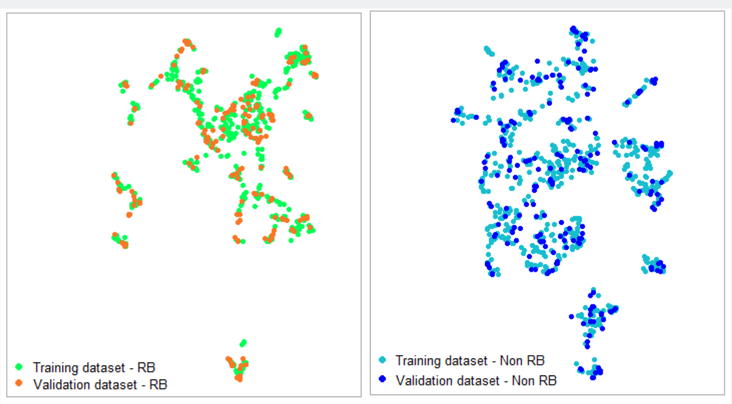
DATASET SPLITTING

Dataset splitting using genetic algorithm (**GA**) and resulting to :

- **training set** (1311 compounds; 60%),
- **internal test set** (436 compounds; 20%)
- **external validation set** (436 compounds; 20%)

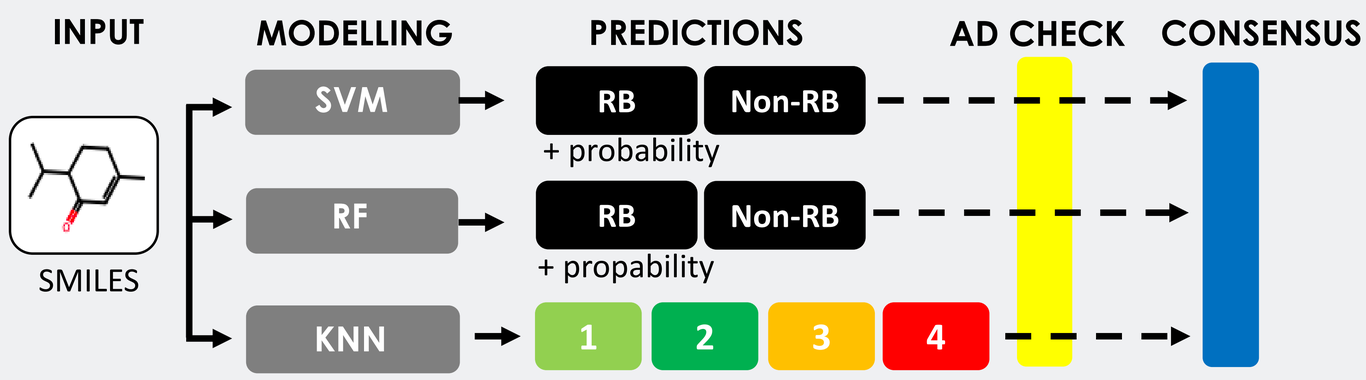
Uniform Manifold Approximation and Projection (**UMAP**) visualization confirms that the validation set is representative of the training set for molecular structure and label distribution, for RB and non-RB classes.

Validation and training UMAPs for RB and Non-RB classes



MODEL WORKFLOW

The workflow consists of 3 models: i) a **Support Vector Machine (SVM)** and ii) a **Random Forest (RF**) predicting a RB or a non-RB class, iii) a **K-nearest neighbor (KNN)** approach determining a finer level of RB class (1, 2, 3 or 4). An applicability domain **(AD)** assessment is also performed to enhance the confidence to the prediction. **SVM** : hyperparameters and best features selected using **GA**. MACCS, PubChem, Estate, and FCFP6, were evaluated to improve the model's performance. **RF**: Best fingerprint features explored using the Boruta approach. The best results for both models were obtained with **91 features of MACCS key fingerprints** identified using GA.



THE 5 OECD PRINCIPLES IN A NUTSHELL

Principle 1: a defined endpoint
Ready biodegradability (OECD301/310).

Principle 2: an unambiguous algorithm
SVM, RF, KNN (parameters and hyperparameters).

Principle 3: a defined domain of applicability (AD)
AD is defined from an in-house algorithm using accuracy and concordance criteria, among others.

Principle 4: Appropriate measures of goodness of fit, robustness and predictivity
Statistics in line with the OECD guidance (OECD, 2007) for QSAR model validation and showing very good goodness-of-fit and predictivity results for both classes and both models.

Principle 5: Mechanistic interpretation
Provided from genetic algorithm selected features and random forest selected metrics.

FOCUS ON PRINCIPLE 4

	TRAINING SET (n=1311)			VALIDATION SET (n=436)		
	Sensitivity (RB)	Specificity (Non-RB)	F1-score (RB)	Sensitivity (RB)	Specificity (Non-RB)	F1-score (RB)
SVM	0.91	0.90	0.90	0.88	0.88	0.88
RF	0.88	0.87	0.87	0.83	0.84	0.83

More than 1000 model versions using various hyperparameter values were tested. The best version was identified as the one providing the **highest statistics in the validation set in line with the statistics obtained in the training set** to avoid any overfitting. Statistics for both models demonstrate a **very good goodness-of-fit and predictivity**.

FOCUS ON PRINCIPLE 3

	Model prediction / AD prediction	TRAINING Out AD	TRAINING In AD	VALIDATION Out AD	VALIDATION In AD
SVM	✗	**123**	1	**31**	20
	✓	275	**912**	185	**200**
RF	✗	**164**	0	**71**	0
	✓	249	**898**	167	**198**

- Training set:

For both models, the AD algorithm classified **« in AD » ca. 80% of the correctly predicted compounds**. **Nearly 100% of the non-correctly predicted compounds fall « out AD ».**

- Validation set:

Around 50% of the compounds correctly predicted fall « out of AD ». The **AD algorithm is particularly effective for both models** and most of substances (even all of them in the case of the RF model). The question about **model misclassification vs. model generalization** arises.
The AD is currently considered as restrictive and will be optimized.

CONCLUSION

- This first version of the KREATiS model predicting the readily biodegradation potential of organic monoconstituent chemicals relies on a unique 2 step curation and combination of three models plus an in-house AD algorithm strongly enhancing the statistics of the model.
- To enhance mechanistic understanding, biodegradation pathways will be implemented in a future version of the model.
- This method, successfully applied to monoconstituent compounds, will be extended to multiconstituent and polymeric compounds.

REFERENCES

OECD (1992), *Test No. 301: Ready Biodegradability*, OECD Guidelines for the Testing of Chemicals, Section 3, OECD Publishing, Paris, https://doi.org/10.1787/9789264070349-en; OECD (2007), *Guidance Document on the Validation of (Quantitative) Structure-Activity Relationship [(Q)SAR] Models*, OECD Series on Testing and Assessment, No. 69, OECD Publishing, Paris, https://doi.org/10.1787/9789264085442-en.

Scan QR code to download the poster