Combination of QSAR models and chemical clustering as a NAM for computational toxicology.

Florent Petronin*, Aurélien Stab°, Michel Souchet°, Sinan A. Karaboga°

* LORIA, Campus Scientifique, 615 Rue du Jardin-Botanique, 54506 Vandœuvre-lès-Nancy ° HARMONIC PHARMA ,Campus Artem, 92 rue du Sergent Blandan, 54000 Nancy

HARMONIC PHARMA

SAFETY BY DESIGN

Loria

Laboratoire lorrain de recherche en informatique et ses applications

Abstract

In the course of setting up consistent alternatives to animal testing in agreement with regulatory guidelines, QSAR based methodologies have been thoroughly assessed. The present retrospective study exemplifies the robustness of five Quantitative Structure-Activity Relationship (QSAR) models – i.e. carcinogenic, mutagenic, reprotoxic, persistent/bio-accumulative/toxic (PBT), and endocrine disruption. Those models which are related to the founding endpoints of the REACH SVHC list were built from respective datasets via selections of chemical descriptors and a Random Forest type algorithm. The performance of the models was measured through the sensitivity factor with a dataset of 256 compounds deriving from the REACH SVHC list. Results showed an overall agreement rate of 75.4% with regards to the actual classification of the SVHC dataset with an applicability score of 100% (only organic compounds are considered). Moreover, each of the models was able to correctly classify the compounds with a significant sensitivity for carcinogenic, mutagenic, reprotoxic, PBT, and endocrine disruption – i.e. 93.3%, 85.7%, 68.0%, 62.1%, and 100.0%, respectively. It is noteworthy that carcinogenic, mutagenic ,and endocrine disruption QSARs showed the best prediction rates with a sensitivity greater than 85%.

In order to investigate further the selected SVHC dataset, a complementary clustering analysis was carried out: for each toxicity endpoint, the subset molecules were grouped by chemical similarity using the proprietary spherical harmonic (SH) descriptors taking into account the shape and the physico-chemical properties of the molecules. Interestingly, the resulting clustering obtained for each endpoint pointed out “representative compounds” that could be further used as molecular templates to screen novel compounds with the aim of designing safer ones.

Altogether, consistency and robustness of the QSARs combined with a clustering analysis support the suitability of this new approach methodology (NAM) to a prospective screening to characterize potential toxicity of novel substances from synthetic or natural sources. This NAM is a part of SAFETY BY DESIGN®, the new solution of services and software for toxicity prediction and characterization of chemical substances.

Dataset – REACH Substances of Very High Concern

Reach provides a SVHC list [2] of 548 compounds which are progressively replaced by less dangerous substances or technologies where technically and economically feasible alternatives are available. Thus, this list is a prime choice to constitute a dataset to assess the respective sensitivity of QSAR models presented in this study.

A selection process where compounds with no structure/CAS number or being inorganic were discarded, provided a dataset of 256 compounds.

Toxicity	Unique CAS	Number of Structures
Carcinogenic	49	45
Mutagenic	11	7
Reprotoxic	77	75
PBT	38	29
Endocrine disruption	102	87

Methodologies

QSAR Statistical Method - Random Forest

The QSAR models were built via the Random Forest statistical method using RDKit toolkit [3], which creates decision trees and reports the probability of belonging to a class. Each tree is trained on a sample of the data, and predictions are made by majority voting of the trees [4].

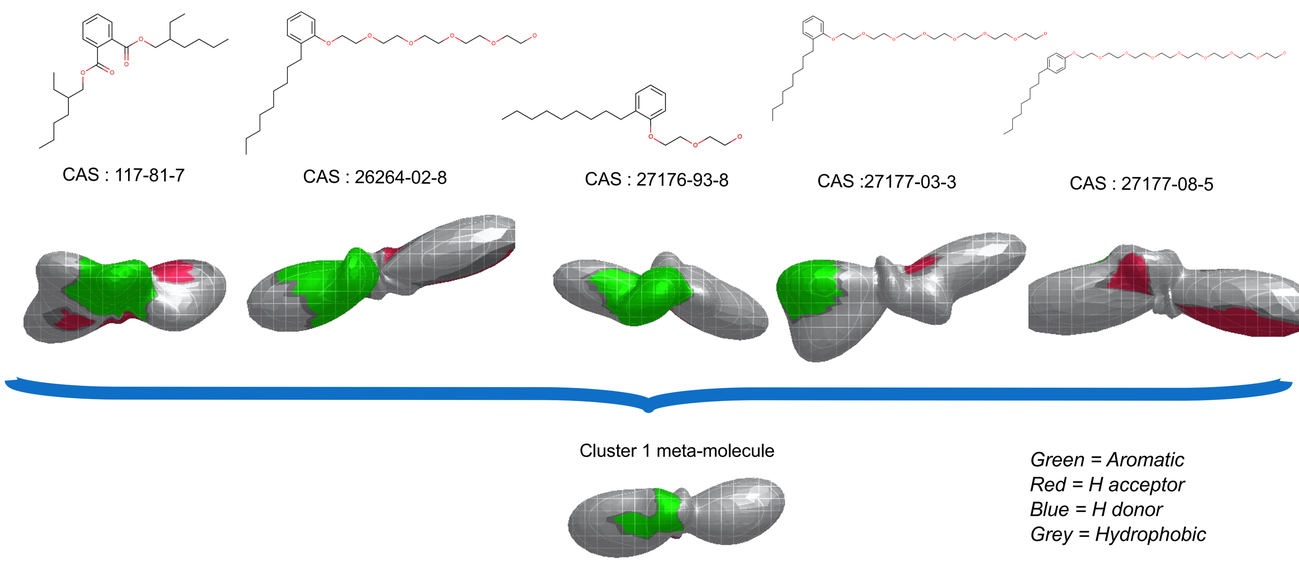
The Sensitivity (Se): [TP] / [TP]+ [FN] measures the ability to correctly identify a molecule as toxic. Since the cross-toxicity of molecules was not tested, TN and FP are unknown and Specificity and Accuracy can not be computed.

TP = True Positives, TN = True Negatives, FP = False Positives, FN = False Negatives

Clustering based on Spherical Harmonics (SH) and Meta-Molecules

SH descriptors were computed for every molecules which were then submitted to a hierarchical clustering. Each resulting cluster gave rise to a representative « Meta-Molecule » - i.e. a 3D molecular object incorporating the average chemical and geometrical properties of molecules composing the respective cluster.

Meta-molecules example : cluster 1 (see below)



QSAR Results

QSAR models correctly predicted the toxicity for the vast majority of compounds belonging to the carcinogenic and mutagenic datasets – i.e. 93.33% and 85.71%, respectively – , and for the majority of compounds belonging to the reprotoxic and PBT datasets – i.e. 68.00% and 62.07%, respectively.

Endocrine disruption (ED) « one is enough » is a consensus model deriving from four specific QSAR models namely : estrogen receptor (ER), androgen receptor (AR), thyroid perturbation (ThP) and steroidogenesis (StG). The respective sensitivity of this model is of 100 % meaning that all molecules of the dataset possess a potential ED character ; this high value results from a positive call from at least one of the four specific ED QSAR models i.e. ER, AR, ThP, StG.

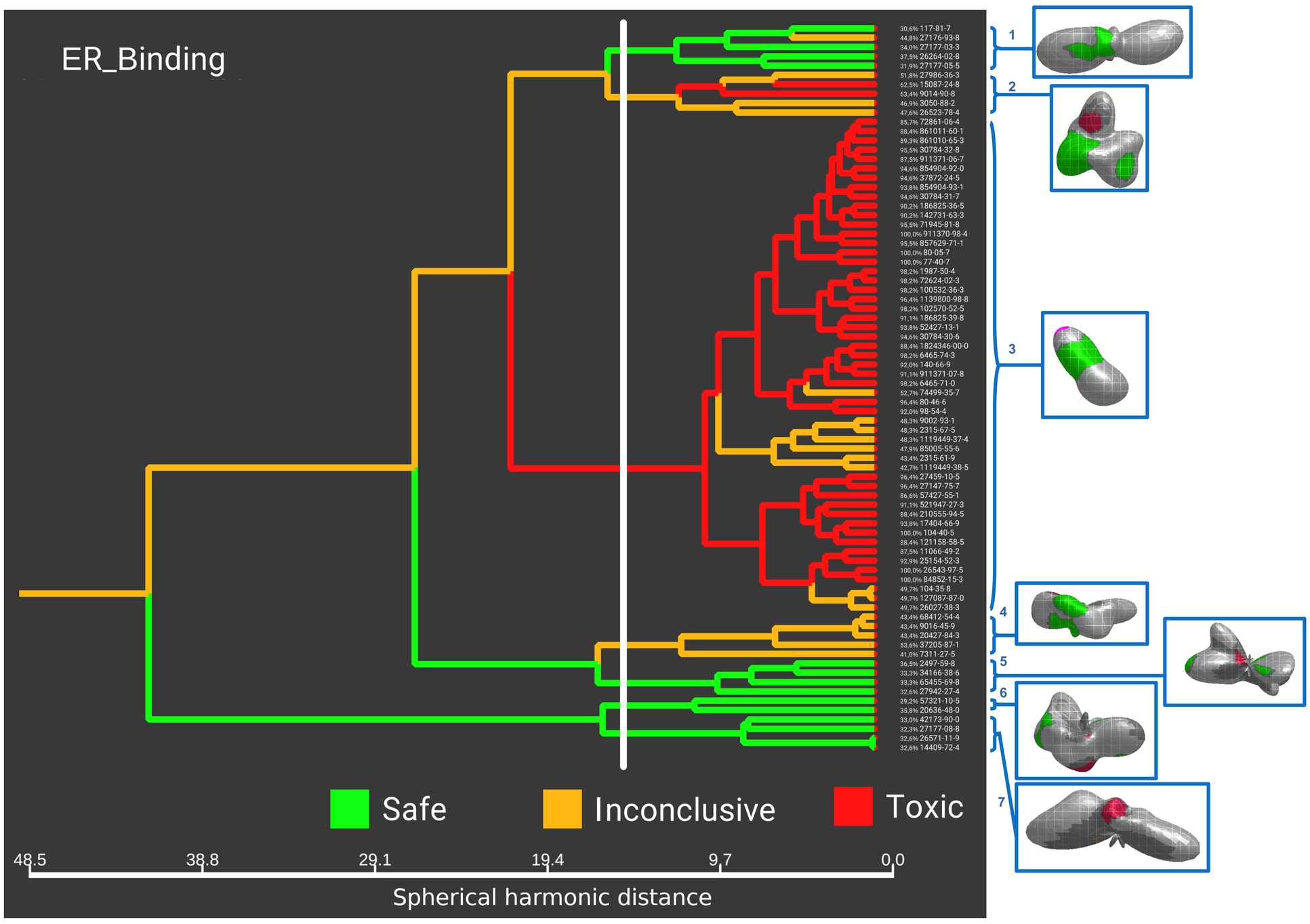
Toxicity	Sensitivity	%of molecules within the applicability domain of the model	Number of TP	Number of FN	Number of Inconclusives
Carcinogenic	93.33 %	100 %	42	1	2
Mutagenic	85.71 %	100 %	6	1	0
Reprotoxic	68.00 %	100 %	51	8	16
PBT	62.07 %	100 %	18	5	6
Endocrine disruption « One is enough »	100 %	Not Applicable	57	0	0

Clustering based on Spherical Harmonics

The hierarchical clustering approach was carried out on 78 molecules identified as estrogen receptor binders. The graph opposite illustrates diverse clusters of molecules based on their respective SH based similarity. By applying a similarity cut-off at position 14 of the scale (white line), seven different clusters clearly stood out.

It is noteworthy that those clusters are associated with different trends of predicted ER property, namely : subgroups in green, orange, and red may be representative of low, medium, and high binder to ER, respectively.

In addition, each of those seven clusters led to a representative 3D meta-molecule which represents a powerful tool to decipher molecular features linked to a specific toxicity.



Conclusions

- The QSAR models correctly predicted the toxicity of the molecules deriving from the REACH SVHC list with an overall agreement rate of 75.4 %.
- Hierarchical clustering based on Spherical Harmonics (SH) was used to group molecules of the dataset. Each cluster was characterized by a meta-molecule recapitulating the main features of underlying compounds. Thus, a meta-molecule might be used as a molecular probe/template to screen already known or novel compounds with the aim of selecting or designing safer chemical entities.
- The combination of QSAR methods with a SH based clustering allowed identifying clusters of molecules linked to specific toxicities which constitutes a new approach methodology (NAM).
- This NAM is a part of SAFETY BY DESIGN®, the new solution of services and software for toxicity prediction and characterization of chemical substances. This innovative solution is part of an alternative method to animal experimentation and is suitable for regulation purposes in the pharmaceutical, cosmetics and specialty chemical fields.

References

[1] Arnaud S. KARABOGA et al. « Benchmarking of HPCC : A novel 3D molecular representation combining shape and pharmacophoric descriptors for efficient molecular similarity assessments ». In :Journal of Molecular Graphics and Modelling 41 (2013), p. 20-30.

[2] REACH SVHC List : https://echa.europa.eu/fr/candidate-list-table

[3] RDKit: Open-source cheminformatics; http://www.rdkit.org.

[4] Breiman L. 2001. Random forests. Machine Learning 45(1):5–32]