New spherical harmonic based descriptors to efficiently fuel QSAR methodology : Endocrine disruptor case study

Aurélien Stab[1], Guillaume Ollitrault[2], Arnaud Sinan Karaboga[1]

[1]Harmonic Pharma, Campus Artem, 92 rue du Sergent Blandan, 54000 Nancy, France ; [2]LORIA, Campus Scientifique, BP 239, 54506 Vandoeuvre-lès-Nancy, 54600, France

Abstract

Two-dimension quantitative structure-activity relationship (2D QSAR) has been a standard methodology for the last decade whereas multiple three-dimension (3D) descriptors have been tested with mitigated successes. In the present study, we propose a new set of highly informative and compact 3D descriptors from spherical harmonic (SH) based representations covering both the geometrical shape and the pharmacophoric features of a molecule. The process consists in placing a molecule on three different axes – each one is captured by its own set of spherical harmonics. SH related expansions are used to create compact and rotation independent descriptors - e.g. 32 floating coefficients - to describe a conformer of the molecule. These descriptors were then applied to a QSAR model of toxicity which was built from the reference dataset of the CERAPP project - a collaborative project that developed a consensus model of toxicity for the endocrine disruption [1]. The QSAR model was trained with SH based descriptors and binding activity to the estrogen receptor was considered in this study. The resulting model yielded a balance accuracy of 0.87 on the evaluation dataset. Furthermore, by combining SH and 2D descriptors from the RDKit suite [2], the subsequent QSAR model gave rise to a balance accuracy of 0.91 on the evaluation dataset, positioning its performance at the high level of the consensus model obtained by the CERAPP project.

Dataset - CERAPP Binding

The CERAPP project objective was to compute a consensus statistical model to predict molecule activity against the estrogen receptor. They provide two datasets for training and evaluation of their models that we used to implement and evaluate our own method:

- **Training set** : large dataset from the ToxCast™ [3] and Tox21 [4] programs
- **Evaluation set** : extracted from the literature to evaluate the performances of the consensus model

Table of the number of active and inactive compounds used in the training and evaluation dataset, respectively – with he binding endpoint.

	Training set	Evaluation set (All)	Evaluation set (ref > 6)
Active	237	1,982	166
Inactive	1,439	5,301	1,043

Statistical method - Random forest

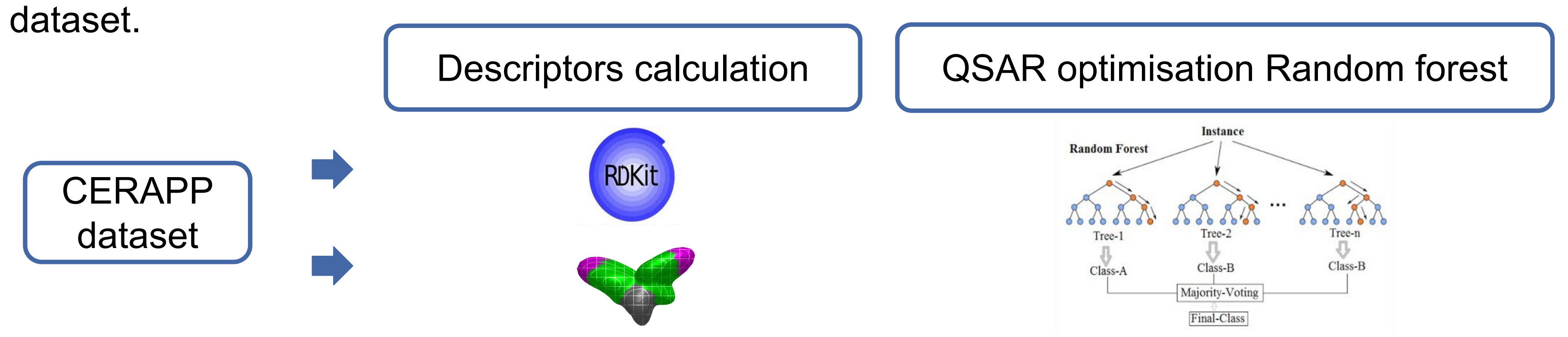
Random forest : This method creates a number of decision trees and reports the probability of belonging to a class. Each tree is trained on a sample of the data, and predictions are made by majority voting of the trees.

The performance of the model is measured with the following metrics:

- Sensitivity (Se): ability to correctly identify an active molecule
- Specificity (Sp): ability to correctly identify an inactive molecule
- Balance Accuracy (Ba) = (Se+Sp)/2
- Area Under Curve (AUC): global performance of the model

Processing 2D and 3D SH Descriptors

The RDKit suite allows the generation of a set of diverse descriptors characterizing molecules [2]. We choose to compute all the 2D descriptors available. After filtering, 108 2D descriptors were considered statistically significant and used to build the QSAR model on the CERAPP dataset.



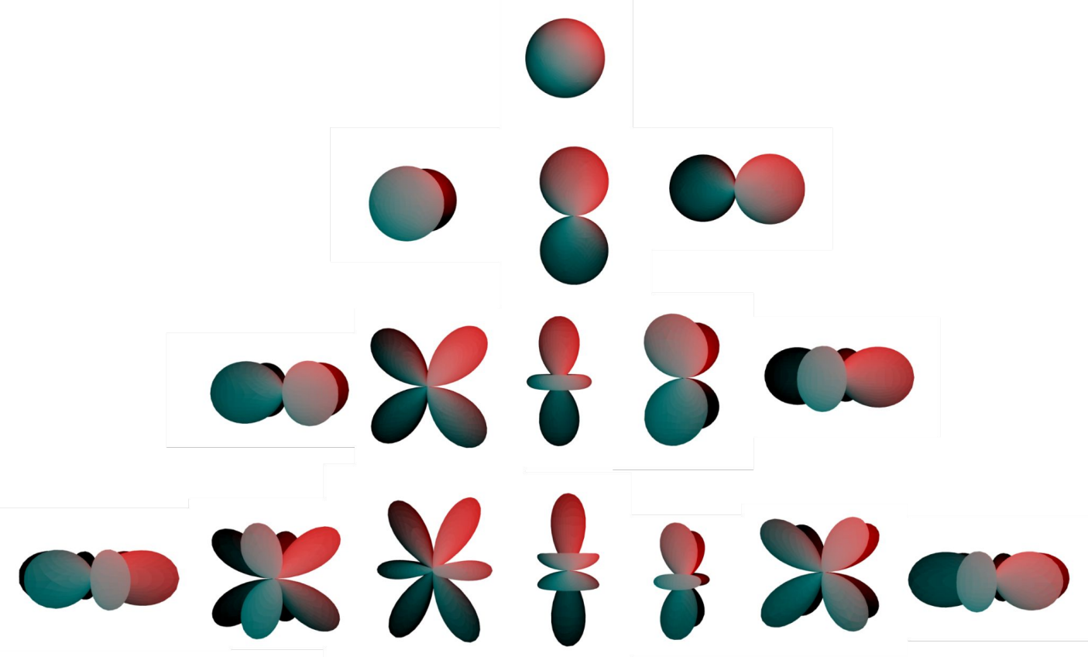
Descriptors calculation

QSAR optimisation Random forest

CERAPP dataset

Diverse models were generated with 2D RDKit descriptors alone, 3D SH descriptors alone, and by combining both categories in order to evaluate their relative contribution.

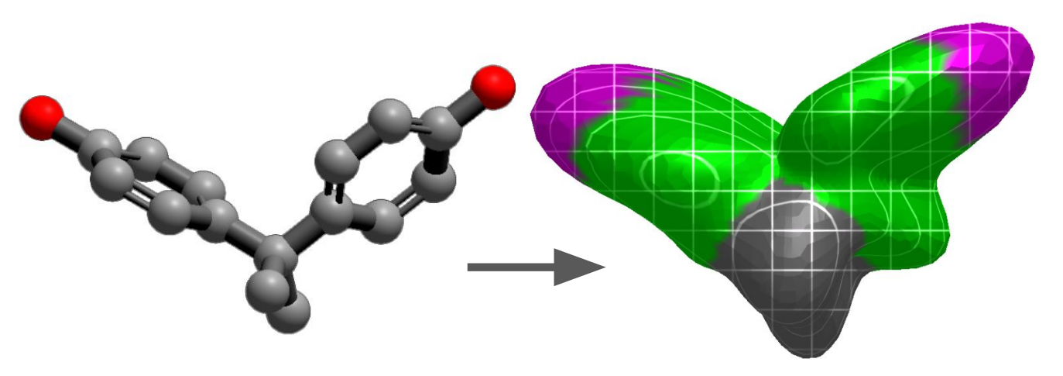
New Spherical Harmonic (SH) Method



Spherical Harmonics order 0 to 3

Spherical harmonics are a mathematical frequency based representations of 3D shapes.

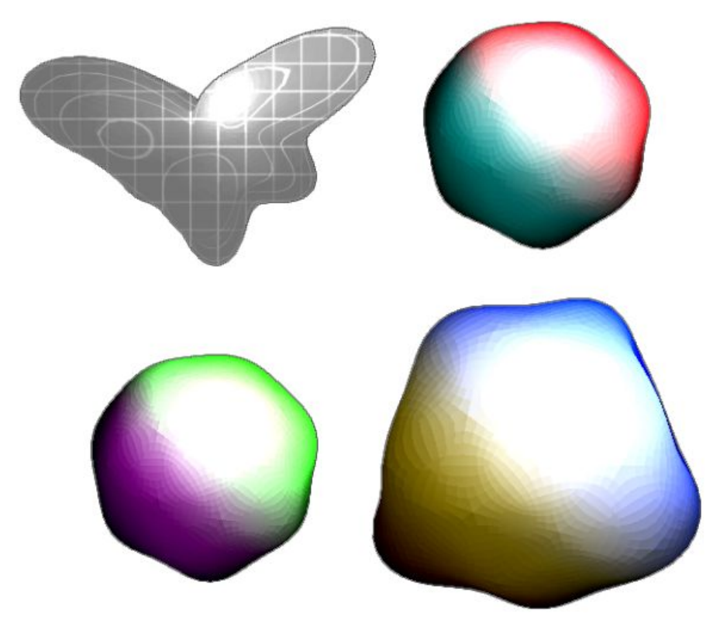
SH has already been used to describe molecular shapes and to assess the geometrical similarity between two molecules. In addition chemical features were added as complementary descriptors using the vertices of the SH to assess the chemical similarity [5].



Bisphenol A's SH representation

Here we have developed a new method to represent both the geometry and the chemical shapes of a molecule using SH expansions. The chemistry of the molecule is projected on three different axes ; each one captures by its own set of SHs.

Advantages of the SH Descriptors



The four SH based representations of Bisphenol A : the geometry displayed in grey and the chemical properties displayed in colors

SH expansions show an efficient representation of molecular features which has the advantage of being 3D and rotation independent.

These descriptors are extremely compact : a molecule is represented by 32 floating coefficients.

Each type of chemical properties is described by different descriptors and can be used separately.

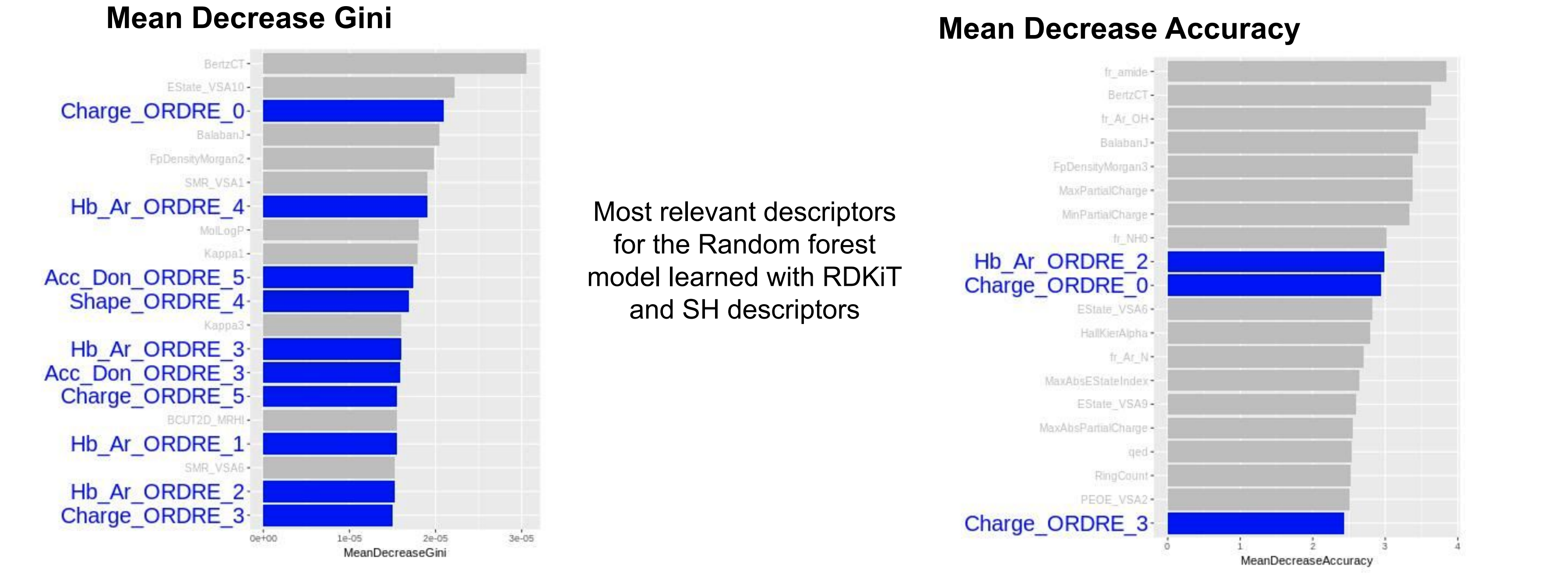
In this work seven order of SH were used to describe molecules.

Results

Table of the performances on the evaluation sets with a number of reference > 6

Random forest descriptors (N tree = 200)	CERAPP Littérature Data (Ref > 6 : 1,209 molecules)			
	Se	Sp	Ba	AUC
RDKiT (Threshold 0.72)	0.849	0.954	0.901	0.955
SH (Threshold 0.5)	0.891	0.841	0.866	0.925
RDKit + SH (Threshold 0.59)	0.892	0.924	**0.908**	0.963

- Random forest model learned with SH descriptors alone is able to efficiently predict estrogen binding of molecules (Ba: 0.87)
- SH descriptors combined with RDKiT descriptors led to the best performances (Ba: 0.91)
- The combined model showed similar performances with the CERAPP consensus model (Ba: 0.91) [1]



Mean Decrease Gini

Mean Decrease Accuracy

Most relevant descriptors for the Random forest model learned with RDKiT and SH descriptors

Barplot showing the importance of variables as measured by a Random forest method. The SH descriptors are displayed in blue

- SH descriptors are among the most relevant descriptors in our final model
- SH of low order represent the most informative descriptors of a molecule
- SH representing the charge and it's ability to interact with other molecule (Hb_Ar) are the most relevant ones in this model

Conclusion

- The present work uses new 3D descriptors that capture the geometry and the chemistry of molecular entities
- The potential endocrine disruption of new molecules is predicted with a good reliability via a Random forest method on a reference dataset using these 3D descriptors: the set of 32 SH 3D descriptors gave robust results alone or in combination with 2D RDKiT descriptors
- This new method produces highly informative and compact 3D descriptors which can be applied to diverse fields of predictive toxicology

References

[1] Kamel MANSOURI *et al.* « CERAPP : Collaborative Estrogen Receptor Activity Prediction Project ». In :Environmental Health Perspectives 124.7 (2016), p. 1023-1033.
[2] RDKit: Open-source cheminformatics; http://www.rdkit.org.
[3] David J. DIX *et al.* « The ToxCast Program for Prioritizing Toxicity Testing of Environmental Chemicals ». In :Toxicological Sciences 95.1 (2006), p. 5-12.
[4] Ruili HUANG *et al.* « Profiling of the Tox21 10K compound library for agonists and antagonists of the estrogen receptor alpha signaling pathway ». In :Scientific Reports 4.1 (2014).
[5] 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.

Support: This work has been supported by the « *Agence Innovation Défense* » (AID) through a « *Régime d'Appui Pour l'Innovation Duale* » (RAPID) program