the 10th SFCi meetings

Improving SAR analysis via pharmacophoric feature reduction and feature transformation

Hajar Rehioui[1,*], Abdelkader Ouali[1], Christophe Couronne[1], Jean-Luc Lamotte[2], Alban Lepailleur[2], Jean-Luc Manguin[1], Ronan Bureau[2], Albrecht Zimmermann[1], Bertrand Cuissart[1]

[1] GREYC, Normandie Univ., UNICAEN, CNRS – UMR 6072, 14000 Caen, France

[2] Centre d’Etudes et de Recherche sur le Médicament de Normandie, Normandie Univ, UNICAEN, CERMN, 14000 Caen, France

[*] hajar.rehioui-karine@unicaen.fr

Our work is part of the European project “SCHISM”, funded by the European Union within the framework of the Operational Programme ERDF/ESF 2014-2020

01 October 2021

General goal

- Structure-activity relationship (SAR) analysis :
  - Cleaning data, searching suitable representation.
  - Clustering ligands by family, finding out activity cliffs.

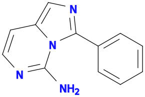
data

- 1485 ligands [1] tested on tyrosine kinase BCR-ABL, often found in patients with chronic myeloid leukemia.
- Each ligand is defined by 112048 pharmacophores (order varying from 3 to 7)
- Two classes : 711 ligands from inactive class and 774 ligands from active class.
- Activity of ligand :
  - $K_i \leq 100nM \implies$ active
  - $K_i \geq 1000nM \implies$ inactive
  - otherwise not considered

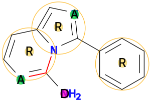
[1] GAULTON, Anna, HERSEY, Anne, NOWOTKA, Michal, et al. The ChEMBL database in 2017. Nucleic acids research, 2017, vol. 45, no D1, p. D945-D954.

Pharmacophoric features [2]

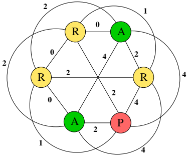
Hydrogen Bond Acceptor, Hydrogen Bond Donor, aromatic Ring, Hydrophobic area, Positively ionizable group, Negatively ionizable group



Skeletal formula



Feature occurrences



Pharmacophore graph

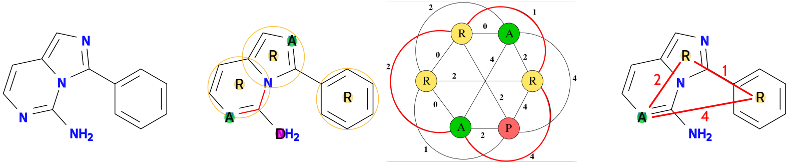
Pharmacophore graph of a molecule

- A vertex : an occurrence of a feature
- An edge : the distance between two features

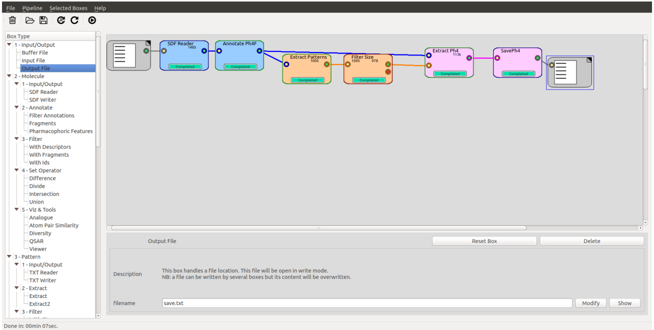
[2] OpenBabel, N. M. O'Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch, G. R. Hutchison, J. Cheminformatics 2011, 3, 33.

pharmacophore

- fragment of the overall pharmacophore graph of the ligand responsible of its biological activity (active or inactive).
- sufficiently present (e.g., appearing in at least 10 ligands)



Pharmacophoric fingerprint



Output File

Description: This box handles a file location. This file will be open in write mode.
NB: a file can be written by several boxes but its content will be overwritten.

filename: save.txt

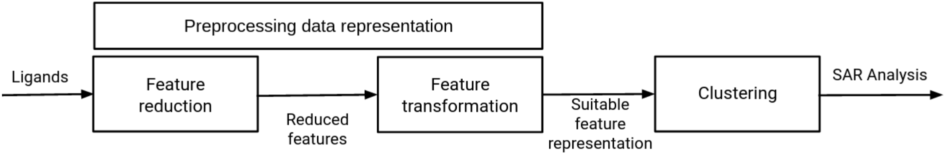
Done in: 00min 07sec.

The workflow to extract the pharmacophores by Norns [3] tools

Id_Mol	|A|A|A|-0	|R|A|R|-1	|A|A|R|-2	...	|D|D|H|H|N|N|N|-112047
CHEMBL250213	1	0	0	...	1

[3] METIVIER, Jean-Philippe, CUISSART, Bertrand, BUREAU, Ronan, et al. The pharmacophore network : a computational method for exploring structure–activity relationships from a large chemical data set. Journal of medicinal chemistry, 2018, vol. 61, no 8, p. 3551-3564.

- Work process

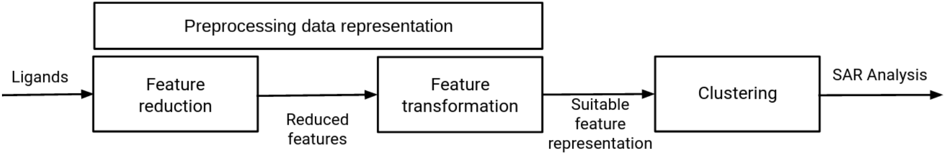


- Experimental environment

Programming language and frameworks

- language : python.
- frameworks : Keras, tensorflow, sklearn.

- Work process

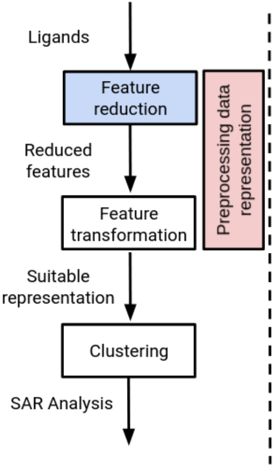


- Experimental environment

Processors	Cores	RAM	GPU	Disk space
2 Processors Intel Xeon E5-2680 v2 2.80GHz	40	512 G	2 Tesla K40M 2880 Cores 12G RAM	9.9 T

Programming language and frameworks

- language : python.
- frameworks : Keras, tensorflow, sklearn.



Ligands

Feature reduction

Reduced features

Feature transformation

Suitable representation

Clustering

SAR Analysis

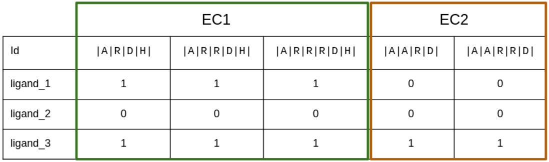
Preprocessing data representation

Objective : Simplifying data

- Eliminate redundancy.
- Keep relevant information.

Pharmacophoric equivalence class[4] (EC)

- EC : group of pharmacophores that appear in exactly the same ligands.
- Columns (pharmacophores) belonging to the same EC are reduced to one representative column.



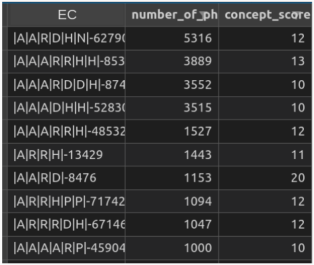
	EC1			EC2	
Id	|A|R|D|H|	|A|R|R|D|H|	|A|R|R|R|D|H|	|A|A|R|D|	|A|A|R|R|D|
ligand_1	1	1	1	0	0
ligand_2	0	0	0	0	0
ligand_3	1	1	1	1	1

[4] FOURNIER-VIGER, Philippe, GUENICHE, Ted, ZIDA, Souleymane, et al. ERMiner : sequential rule mining using equivalence classes. In : International Symposium on Intelligent Data Analysis. Springer, Cham, 2014. p. 108-119.

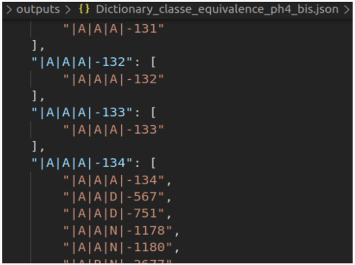
- Reduction of 86.50%

#ligands	# old features	# new features	Type of data	#classes
1485	112048	**15129**	binary (0, 1)	2

- Example of pharmacophoric equivalence classes

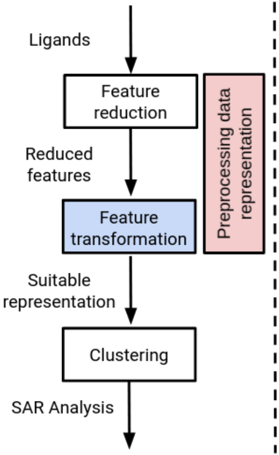


EC	number_of_ph	concept_score
|A|A|R|D|H|N|-6279	5316	12
|A|A|A|R|R|H|H|-853	3889	13
|A|A|A|R|D|D|H|-874	3552	10
|A|A|A|D|H|H|-5283	3515	10
|A|A|A|R|R|H|-48532	1527	12
|A|R|R|H|-13429	1443	11
|A|A|R|D|-8476	1153	20
|A|R|R|H|P|P|-71742	1094	12
|A|R|R|R|D|H|-67146	1047	12
|A|A|A|A|R|P|-45904	1000	10



```
        "|A|A|A|-131"
    ],
    "|A|A|A|-132": [
        "|A|A|A|-132"
    ],
    "|A|A|A|-133": [
        "|A|A|A|-133"
    ],
    "|A|A|A|-134": [
        "|A|A|A|-134",
        "|A|A|D|-567",
        "|A|A|D|-751",
        "|A|A|N|-1178",
        "|A|A|N|-1180",
        "|A|D|N|-2677"
```

- 11577 EC contain one pharmacophore, 3552 EC contain more than one



Ligands

Feature reduction

Reduced features

Feature transformation

Suitable representation

Clustering

SAR Analysis

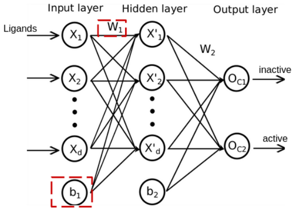
Preprocessing data representation

Objective

- Separate active ligands from inactive ones
- Find the suitable linear transformation

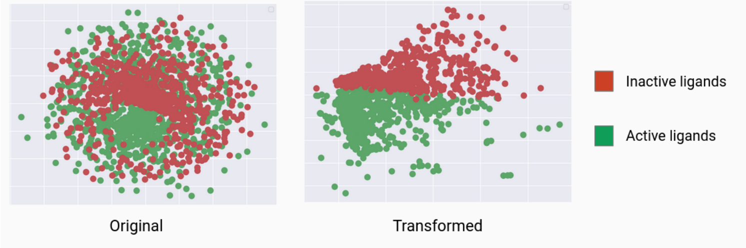
Supervised transformation

- Transform the features of ligands according to their biological activity.
- Use a neural network[5] (NN) which minimizes the optimization function "categorical_crossentropy".



$$\boxed{x'_{ij} = x_{ij} \times w_{jj} + b_j}$$

[5] BEBIS, George et GEORGIOPOULOS, Michael. Feed-forward neural networks. IEEE Potentials, 1994, vol. 13, no 4, p. 27-31.



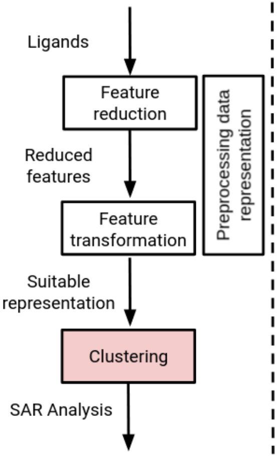
Original

Transformed

Visualization of 2D projection

The 15129 features are projected by the Multidimentional Scaling[6] (MDS) method.

[6] COX, Michael AA et COX, Trevor F. Multidimensional scaling. In : Handbook of data visualization. Springer, Berlin, Heidelberg, 2008. p. 315-347.



Ligands

Feature reduction

Reduced features

Feature transformation

Suitable representation

Clustering

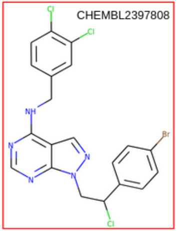
SAR Analysis

Preprocessing data representation

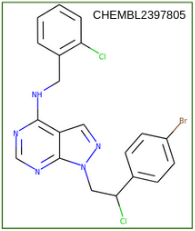
Objective : group together similar structures

Find out significant ligands

- homogeneous : families of biological activities .
- not homogeneous : potential presence of activity cliffs.



CHEMBL2397808



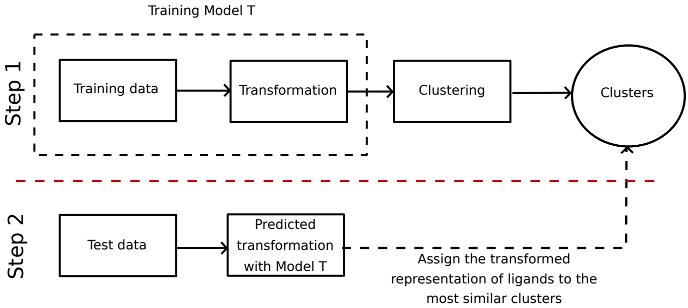
CHEMBL2397805

Objective

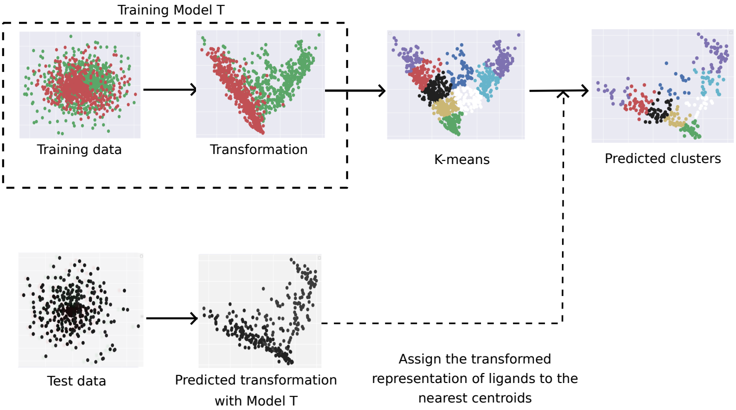
- Predict the cluster (family) of a new unlabeled ligand.

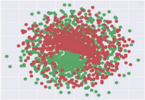
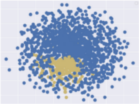
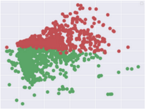
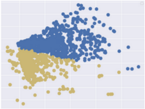
Principle

- Apply one of the clustering methods on the training set after its transformation
- Assign to each cluster built in step (1) the data that corresponds to it (the most similar)


Training Model T
Step 1
Training data
Transformation
Clustering
Clusters
Step 2
Test data
Predicted transformation with Model T
Assign the transformed representation of ligands to the most similar clusters

Example : Predictive K -means


Training Model T
Training data
Transformation
K-means
Predicted clusters
Test data
Predicted transformation with Model T
Assign the transformed representation of ligands to the nearest centroids

Data	before clustering	after clustering
Original		
Transformed		

Legend

🟥 Inactive ligands

🟩 Active ligands

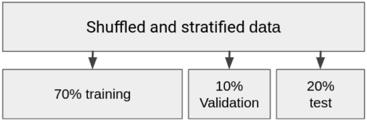
🟦 Cluster 0

🟨 Cluster 1

	Original		Transformed	
	Cluster 0	Cluster 1	Cluster 0	Cluster 1
Inactive (711)	711 (100%)	0 (0%)	710 (99.86%)	1 (0.14%)
Active (774)	450 (58.14%)	324 (41.86%)	197 (25.45%)	577 (74.55%)

Evaluation

Hold-out validation



Evaluation by normalized mutual information [7] (NMI)

- quality measure which compares the resulting clusters or classes with the ground truth.
- The results vary between 0 (no mutual information) and 1 (perfect correlation)

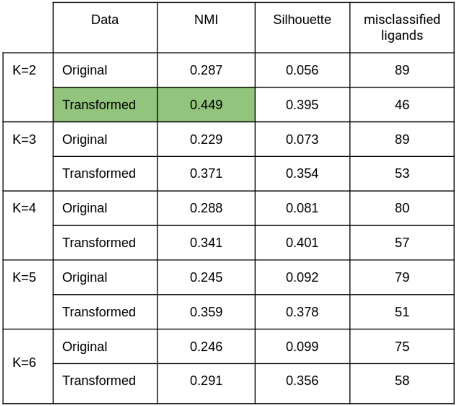
Evaluation by Silhouette [8]

- measure calculated using the mean intra-cluster distance and the mean distance to the nearest cluster.
- The best value is 1 and the worst is -1. Values close to 0 indicate that the clusters overlap. Negative values usually indicate that a sample was assigned to the wrong cluster.

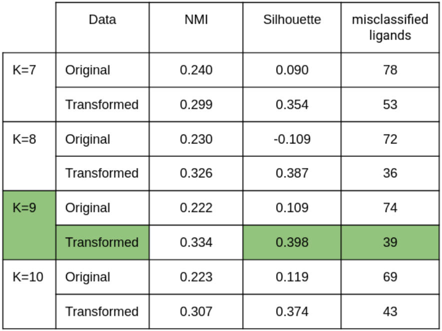
[7] ESTEVEZ, Pablo A., TESMER, Michel, PEREZ, Claudio A., et al. Normalized mutual information feature selection. IEEE Transactions on neural networks, 2009, vol. 20, no 2, p. 189-201.

[8] ROUSSEEUW, Peter J. Silhouettes : a graphical aid to the interpretation and validation of cluster analysis. Journal of computational and applied mathematics, 1987, vol. 20, p. 53-65.

Choice of the number of clusters K

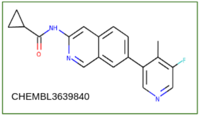


	Data	NMI	Silhouette	misclassified ligands
K=2	Original	0.287	0.056	89
	Transformed	0.449	0.395	46
K=3	Original	0.229	0.073	89
	Transformed	0.371	0.354	53
K=4	Original	0.288	0.081	80
	Transformed	0.341	0.401	57
K=5	Original	0.245	0.092	79
	Transformed	0.359	0.378	51
K=6	Original	0.246	0.099	75
	Transformed	0.291	0.356	58

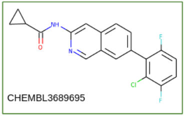


	Data	NMI	Silhouette	misclassified ligands
K=7	Original	0.240	0.090	78
	Transformed	0.299	0.354	53
K=8	Original	0.230	-0.109	72
	Transformed	0.326	0.387	36
K=9	Original	0.222	0.109	74
	Transformed	0.334	0.398	39
K=10	Original	0.223	0.119	69
	Transformed	0.307	0.374	43

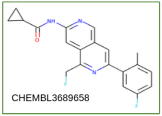
test data	Original									Transformed								
Inactive (142)	0	111	1	1	1	1	5	22	0	21	6	0	64	0	30	0	13	8
Active (155)	60	52	4	4	1	11	1	17	5	14	41	48	7	10	8	6	18	3



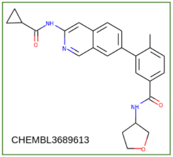
CHEMBL3639840



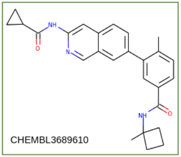
CHEMBL3689695



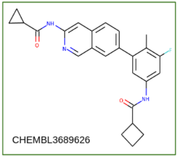
CHEMBL3689658



CHEMBL3689613

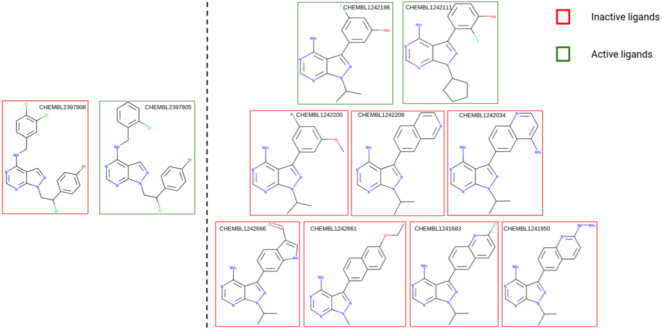


CHEMBL3689610



CHEMBL3689626

test data	Original									Transformed								
Inactive (142)	0	111	1	1	1	1	5	22	0	21	6	0	64	0	30	0	13	8
Active (155)	60	52	4	4	1	11	1	17	5	14	41	48	7	10	8	6	18	3



A cluster with two activity cliffs[9]

[9] STUMPFE, Dagmar, HU, Huabin, et BAJORATH, Jürgen. Evolving concept of activity cliffs. ACS omega, 2019, vol. 4, no 11, p. 14360-14368.

Conclusion and perspectives

Conclusion

Allowing to an expert an easy SAR analysis by :

1. preprocessing step : data cleaning.
2. clustering step : significant ligands, activity cliffs.

Perspectives

- Significant pharmacophores : analyse pharmacophores by “feature importance” study.
- Unsupervised transformation : transform unlabeled data.
- Interactivity : introduce the expert in the process.

the 10th SFCi meetings

Improving SAR analysis via pharmacophoric feature reduction and feature transformation

Hajar Rehioui[1,*], Abdelkader Ouali[1], Christophe Couronne[1], Jean-Luc Lamotte[2], Alban Lepailleur[2], Jean-Luc Manguin[1], Ronan Bureau[2], Albrecht Zimmermann[1], Bertrand Cuissart[1]

[1] GREYC, Normandie Univ., UNICAEN, CNRS – UMR 6072, 14000 Caen, France

[2] Centre d'Etudes et de Recherche sur le Médicament de Normandie, Normandie Univ, UNICAEN, CERMN, 14000 Caen, France

[*] hajar.rehioui-karine@unicaen.fr

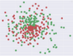
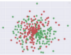
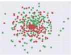
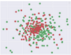
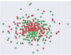
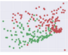
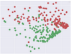
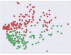
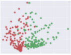
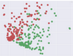
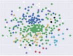
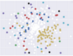
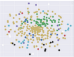
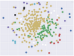
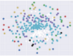
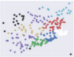
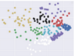
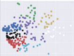
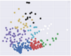
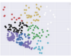
Our work is part of the European project "SCHISM", funded by the European Union within the framework of the Operational Programme ERDF/ESF 2014-2020

UNION EUROPÉENNE
Fonds européen de
développement régional

01 October 2021

appendix

Predictive clustering cross validation

	Original					Transformed				
before clustering										
K-means results										
NMI	0.277	0.260	0.264	0.287	0.236	0.353	0.346	0.376	0.297	0.327
silhouette	0.185	0.126	0.121	0.210	0.189	0.466	0.438	0.453	0.496	0.466