Predicting Biodegradability

and saving our planet

Context

“90% of the trash floating in our oceans is made of plastic”

Why is plastic filling our Oceans?

- low reactivity
- high durability
- not dissolvable in water

=> Bacteria just don’t like to digest plastics

Research Question

Can we predict whether or not a compound will be biodegradable?

Yes, we can!

The Data

Research Data-Set : QSAR biodegradation Datase

Contains quantitative chemical analysis data for 1055 chemical Compounds

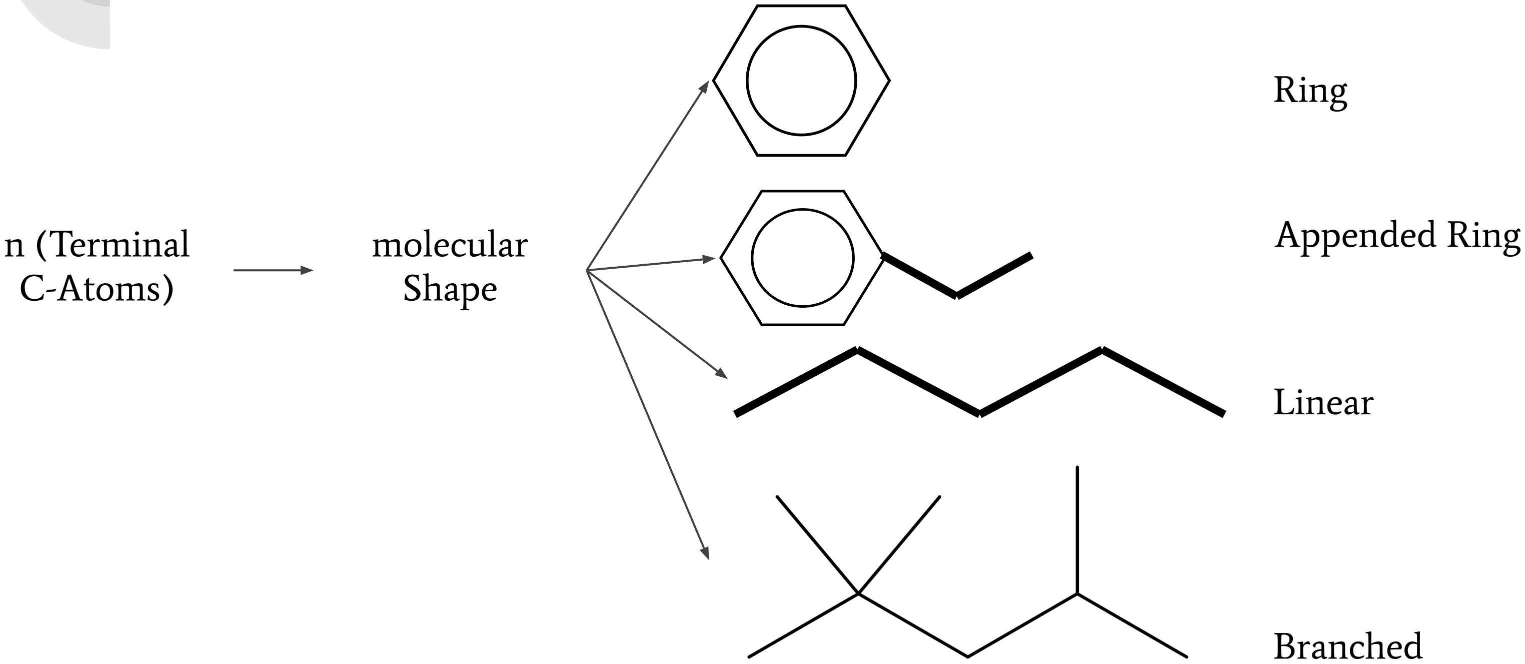
What is it measuring?

Observation contains summary statistics for a kind of “Social network” of each Atom in the Molecule.

How is it derived?

- **mathematical computation of properties**
- **lab tests**
- **data base scraping**

Engineered Features - Molecule Shape



n (Terminal C-Atoms) → molecular Shape

Ring

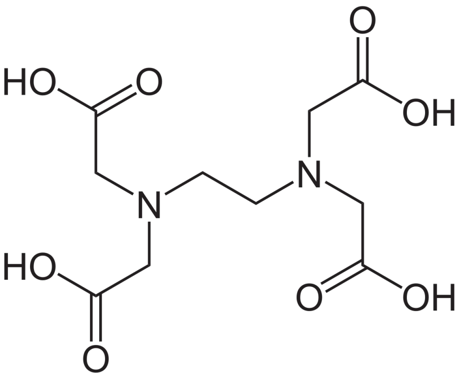
Appended Ring

Linear

Branched

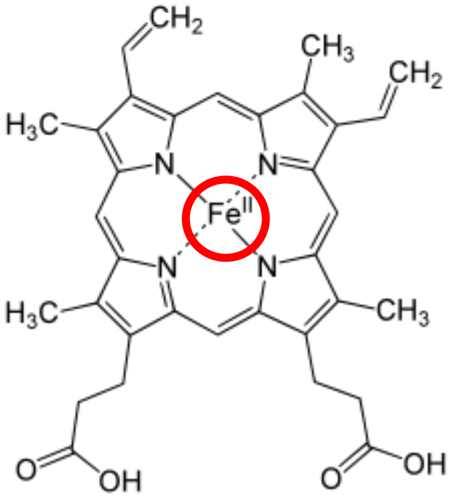
Engineered Features - Functionality

n (heavy Metals) → Funtionality

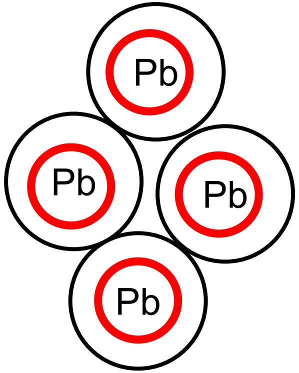


○ : heavy Metals

Non Complex
nHM = 0

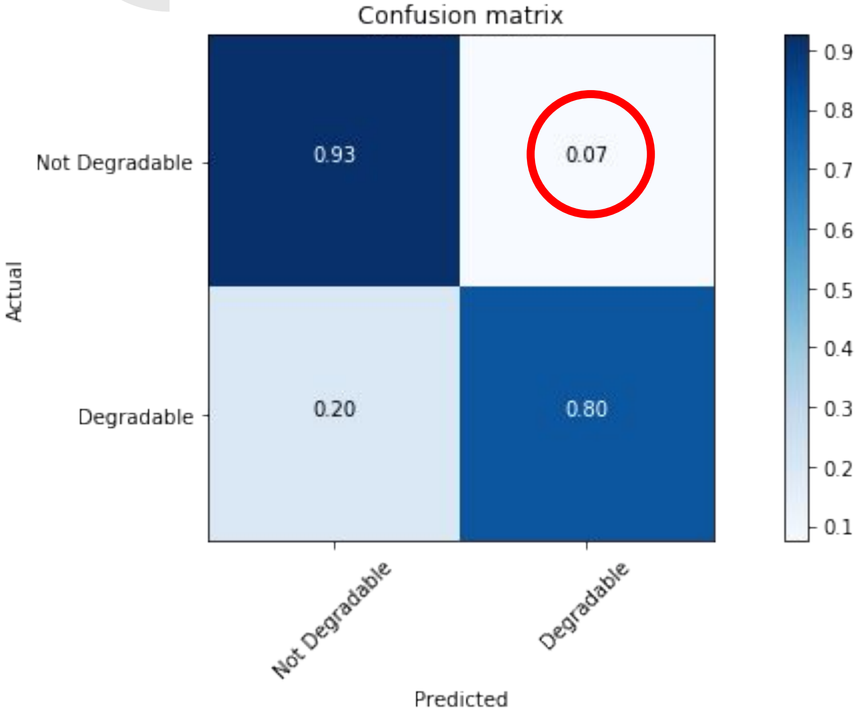


Complex
nHM = 1



Poisenous
nHM>=2

Results



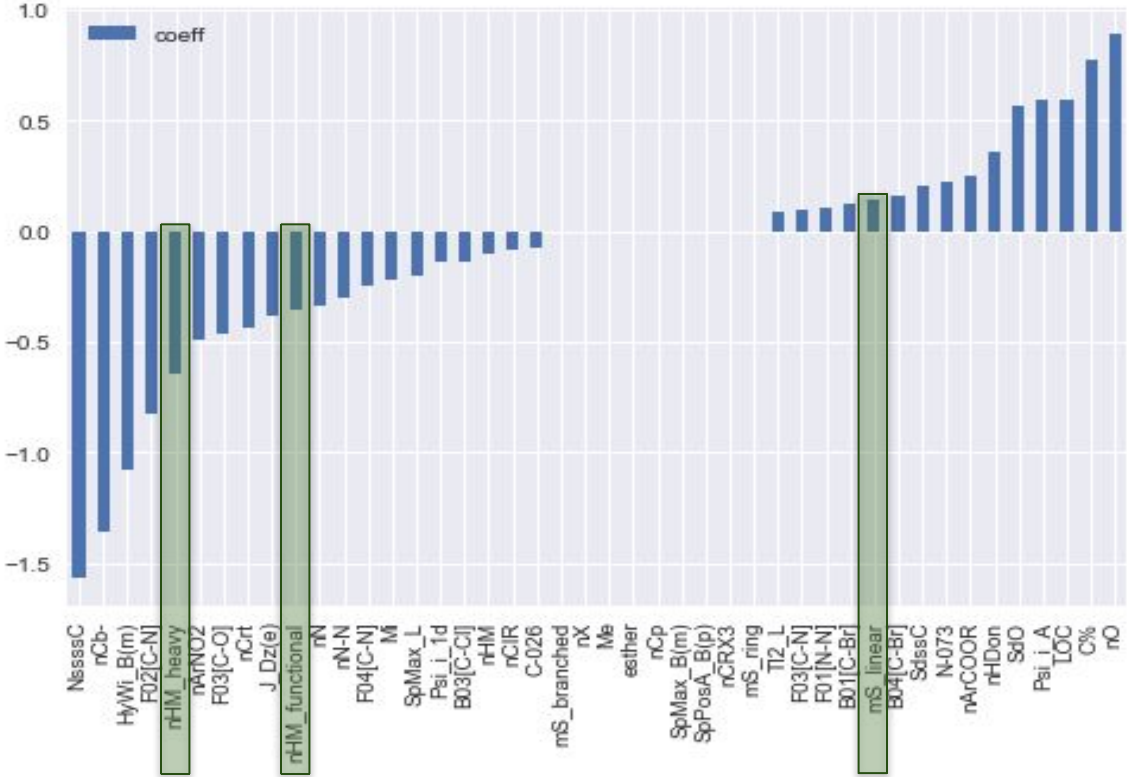
Final Model

Logistic Regression model with Lasso penalty.

Precision	0.86
Accuracy:	0.88
Specificity:	0.93
Sensitivity:	0.80

Citing the Datasets Authors results:
“*The model presented specificity and sensitivity close to 0.8* “

Feature Importance



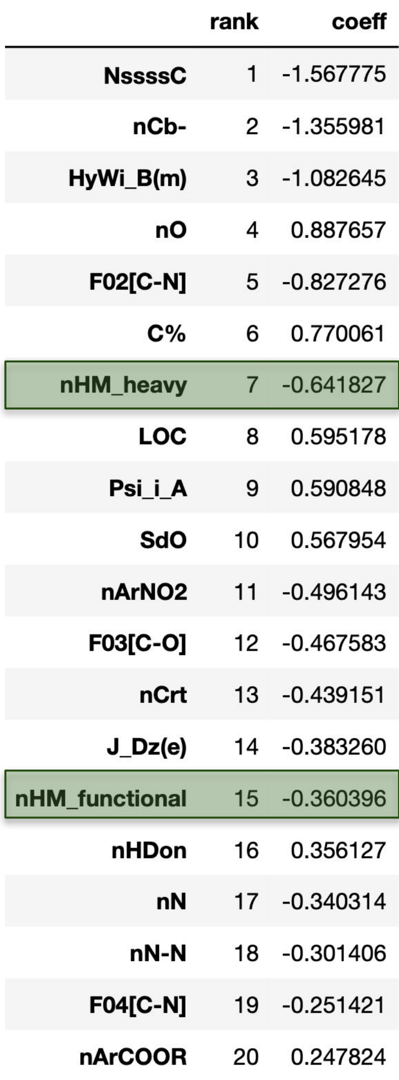
Keynotes

both engineered features:

“nHM” & “mS”

are present in the final model

Top 20 Features



	rank	coeff
NssssC	1	-1.567775
nCb-	2	-1.355981
HyWi_B(m)	3	-1.082645
nO	4	0.887657
F02[C-N]	5	-0.827276
C%	6	0.770061
nHM_heavy	7	-0.641827
LOC	8	0.595178
Psi_i_A	9	0.590848
SdO	10	0.567954
nArNO2	11	-0.496143
F03[C-O]	12	-0.467583
nCrt	13	-0.439151
J_Dz(e)	14	-0.383260
nHM_functional	15	-0.360396
nHDon	16	0.356127
nN	17	-0.340314
nN-N	18	-0.301406
F04[C-N]	19	-0.251421
nArCOOR	20	0.247824

Application

OECD provides sufficient Data on Chemical Compounds, ready to be scraped for future Analysis.

→ Create new biodegradable compounds based on Feature importance

→ Run Lab tests on waste to determine whether compostable or not

→ Save money on drug tests

→ certify biodegradable products based on compound data

Summary

- **Biodegradability** of compounds can be classified with a Logistic Regression Model at a precision of **86 %** given basic QSAR-Data
- Model **performs better*** than the datasets author's model.
- Using this model multiple **profitable** and **sustainable** applications are possible.

For further information contact me:
www.linkedin.com/in/tinopietrassyk
pietrassyk@googlemail.com

* at a 13 % higher specificity

Scources

Dataset:
https://archive.ics.uci.edu/ml/datasets/QSAR+biodegradation#

Paper:
Mansouri, K., Ringsted, T., Ballabio, D., Todeschini, R., Consonni, V. (2013). Quantitative Structure - Activity Relationship models for ready biodegradability of chemicals. Journal of Chemical Information and Modeling, 53, 867-878

Picture:
by Brian Yurasits:
https://images.unsplash.com/photo-1558640476-437a2b9438a2?ixlib=rb-1.2.1&ixid=eyJhcHBfaWQiOjEyMDd9&auto=format&fit=crop&w=2098&q=80

GitHub-Repo:
https://github.com/Pietrassyk/P_4_4_Biodegradability