Electronic Supplementary Material (ESI) for Environmental Science: Processes & Impacts.

Supporting Information

Global QSAR building procedure

Partial Least Squares (PLS)

An in-house implementation of Partial Least Squares Regression (PLS-R) (Wold et al., 2001) using Nonlinear Iterative Partial Least Squares (NIPALS) algorithm was used. A maximum of 7 Latent Variables (LV) were extracted. The model producing the best predictive quality according to Leave One Out cross-validation (LOOCV) was used to select the optimum model dimensionality. The models were further refined by using Fractional Factorial Design (FFD) variables selection [1]. ADAN [2] methodology was applied for assessing the applicability domain of predicted compounds and to provide pseudo 95% CIs.

Random Forest (RF)

Random Forest Regressor (RF-R) available in the scikit-learn library [3] was used to build the models. The number of trees (n_estimators), and maximum features (features) were optimised using a grid search algorithm to obtain optimum values according to Out Of Bag (OOB) criterion. Then, a comparable evaluation of the model predictive quality was performed by applying LOO cross-validation. The goodness-of-fit of each model was assessed using the determination coefficient (R^2). Besides, the predictive performance was assessed using LOO cross-validation and quantified using cross-validated determination coefficient (Q^2) and the Standard Deviation Error of the Prediction (SDEP). In both cases, the model was further validated by predicting the test series and computing the external Q^2 and SDEP parameters.

The descriptor importance was calculated to allow for interpretation. The random forest algorithm assigns the descriptor importance based on the effect of every variable on the error function (Gini score) drop during the algorithm progress, averaged across all decision trees. The greater the drop when the variable was chosen, the greater the importance.

The top 5 descriptors contribute 15% to the RF model (compared to 0.24% for 5 random descriptors). The top 5 descriptors characterize molecular topology/accessibility, the presence of nitrogen and other heteroatoms, and energetic features. Moreover, a descriptor relating to ionization potential contributes on average 0.65% to the model (compared to 0.05% for a random descriptor). A descriptor characterizing the presence of nitrogen contributes on average 0.11% (0.05% for a random descriptor). See also “feature-importance-paDEL-RF-BioTF.exs”.

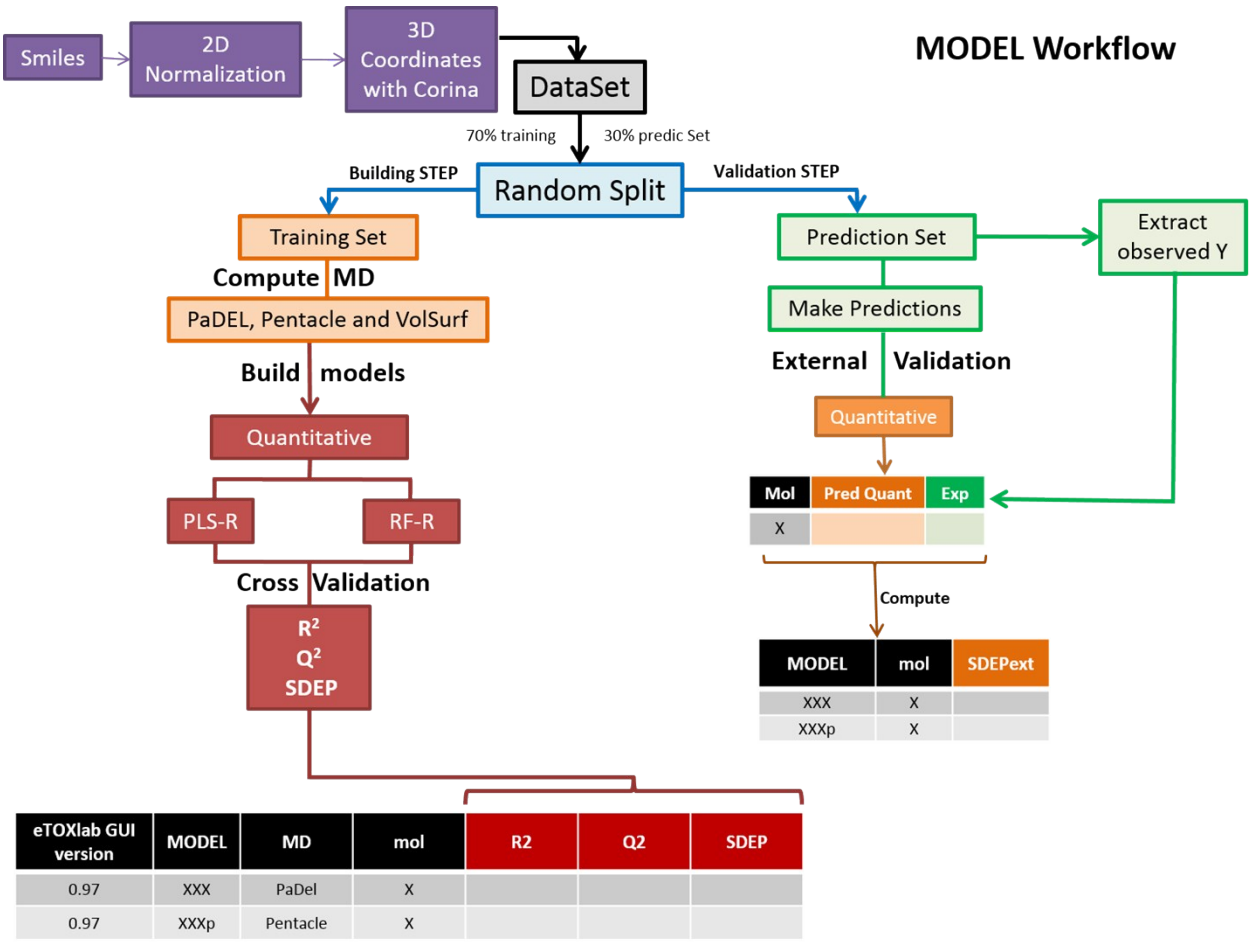

Fig SI1. Detailed modelling workflow used to develop the global biodegradation QSARs.

Table SI1. Quantitative scores for biotransformation endpoint showing both internal and external validation. Best model (highlighted in green) was selected based on the quality of the external prediction, as the best possible method to avoid overfitting.

Endpoint	Estimator	MD	Calculated		Internal Prediction, cross-validated (LOOCV)		External Prediction	
			R^2	SDEC	Q^2	SDEP	Q^2	SDEP
log k_{b2}	PLS-R	PaDEL	0.40	0.75	0.19	0.87	0.32	0.78
		GRIND2	0.47	0.70	0.26	0.83	0.35	0.76
		VolSurf	0.46	0.71	0.28	0.82	0.33	0.77
	RF-R	PaDEL	0.91	0.29	0.32	0.79	0.46	0.69
		GRIND2	0.91	0.29	0.31	0.80	0.39	0.73
		VolSurf	0.91	0.29	0.35	0.78	0.43	0.71

Table SI2. Training set for the global QSAR (RF-R + PaDEL) model which contains the 70% of the full dataset (163 molecules).

name	log k_{b2}		
	experimental	**calculated**	**Cross-validated**
1,1,2-trichloroethane	-11.60	-11.46	-11.24
1,2-dichloroethane	-11.40	-11.17	-10.73
1,2-xylene	-10.10	-10.16	-10.27
1,3-butadiene	-10.20	-10.18	-10.06
1,4-dioxane	-11.00	-10.87	-10.62
17b-estradiol	-8.90	-9.71	-10.85
1-butoxymethylimidazolium L-lactate	-11.80	-11.51	-11.04
1-butylimidazolium L-lactate	-12.10	-11.53	-10.76
1-methylimidazolium L-lactate	-12.10	-11.55	-10.54
2,3,4,6-tetrachlorophenol	-11.80	-11.55	-11.24
2,4,5-trichlorophenol	-11.50	-11.39	-11.20
2,4,6-trichlorophenol	-11.00	-11.16	-11.43
2,4,6-trinitrophenol	-11.40	-11.33	-11.15
2,4-D methylester	-9.20	-9.71	-10.62
2,4-D propylester	-8.60	-9.39	-10.82
2,4-dichlorophenoxyacetic acid (2,4-D)	-11.50	-11.06	-10.33
2,4-dimethylphenol	-9.80	-9.89	-10.09
2,4-dinitrotoluene	-11.20	-11.12	-11.05
2,6-dinitrotoluene	-11.20	-11.15	-11.11
2-amino-5-chlorobenzophenone	-11.40	-11.30	-11.18
2-chlorophenol	-10.00	-10.13	-10.52
2-methylaminoethanol	-10.60	-10.53	-10.52
2-methylpyridin	-9.80	-9.98	-10.41
2-nitrophenol	-11.10	-11.07	-11.00
2-nitropropane	-10.80	-10.80	-10.77
3,3'-dichlorobenzidine	-10.50	-10.83	-11.28
3-nitrophenol	-9.60	-10.01	-10.75
3-pentanone	-10.10	-10.08	-10.01
4,6-dinitro-o-cresol	-10.90	-11.01	-11.19
4-chloroaniline (p-chloroaniline)	-11.60	-11.35	-10.96
4-chloro-m-cresol	-10.70	-10.51	-10.29
4-chloro-o-cresol	-10.50	-10.46	-10.41
4-hydroxybenzoic acid	-11.20	-10.80	-10.00
4-nitrophenol	-10.10	-10.29	-10.51
abacavir	-10.00	-10.45	-11.25
acebutolol	-11.60	-11.55	-11.51
acesulfame	-11.20	-11.15	-11.14
acetophenone	-10.10	-10.06	-9.89
acrylonitrile	-9.50	-9.84	-10.39

acyclovir	-9.90	-10.25	-10.93
alachlor	-9.80	-10.39	-11.43
Albuterol	-11.80	-11.70	-11.46
allyl alcohol	-9.80	-9.89	-9.95
Amantadine	-11.90	-11.45	-10.72
Amphetamine	-10.40	-10.49	-10.57
aniline	-10.50	-10.35	-10.18
aspartic acid	-10.20	-10.44	-10.82
aspirin	-10.50	-10.55	-10.48
atenolol	-11.50	-11.49	-11.44
atorvastatin	-11.80	-11.62	-11.33
atrazine	-12.10	-11.80	-11.19
azinphos-methyl	-12.10	-11.87	-11.29
benzanilide	-11.50	-11.31	-10.85
bis(2-chloroethyl)ether	-11.40	-11.23	-10.89
bis(chloromethyl)ether	-10.50	-10.79	-11.22
bisphenol A	-9.70	-10.15	-11.03
caffeine	-11.80	-11.41	-10.84
candesartan	-10.80	-11.00	-11.39
carbaryl	-10.00	-10.33	-10.90
carbon tetrachloride	-11.50	-11.32	-10.96
carbonmonoxide	-9.50	-9.88	-10.41
chloroquine	-11.70	-11.62	-11.52
Ciprofloxacin	-10.70	-11.01	-11.54
citric acid	-10.20	-10.45	-10.93
crizotinib	-12.00	-11.80	-11.51
cyclohexanone	-9.40	-9.77	-10.29
cyclophosphamide	-11.00	-11.01	-11.10
desipramine	-11.80	-11.52	-11.12
diazepam	-12.00	-11.78	-11.32
dicamba	-11.60	-11.39	-10.95
dichloromethane	-10.80	-10.87	-10.94
diclofenac	-11.80	-11.70	-11.53
diethanolamine	-10.50	-10.54	-10.69
diethylene glycol dinitrate	-12.20	-11.64	-10.83
dimethyl phthalate	-8.80	-9.47	-10.49
di-n-propyl phthalate	-8.80	-9.68	-11.05
diuron	-12.50	-12.01	-11.17
emtricitabine	-10.70	-10.69	-10.74
epichlorohydrin	-10.50	-10.62	-10.79
ethinylestradiol	-10.10	-10.22	-10.41
ethyl acetate	-9.60	-9.84	-10.25
ethyl acrylate	-9.80	-9.86	-10.04

ethylbenzene	-10.70	-10.47	-10.06
ethylene glycol	-10.10	-10.19	-10.40
ethylene oxide	-10.80	-10.71	-10.53
ethylester of 4-chlorophenoxyacetic acid	-9.50	-9.71	-10.13
fenitrothion	-12.30	-12.00	-11.47
Fenoprofen	-11.80	-11.61	-11.31
Flumequine	-11.20	-11.20	-11.14
fluphenazine	-11.70	-11.61	-11.47
formaldehyde	-9.80	-10.15	-10.67
fructose	-9.10	-9.52	-10.31
fulvic acid	-12.70	-12.20	-11.46
gemfibrozil	-12.00	-11.61	-10.99
glucose	-9.10	-9.35	-9.96
Glutamic acid	-10.20	-10.38	-10.64
glycolic acid	-11.00	-10.78	-10.47
hydroquinone	-9.80	-9.88	-10.07
ibuprofen	-11.70	-11.39	-10.96
Indomethacin	-10.80	-11.11	-11.64
isophorone	-10.20	-10.31	-10.39
Isopropanol	-9.80	-10.01	-10.25
Isopropylamine	-10.60	-10.60	-10.55
isoproturon	-10.80	-10.98	-11.35
Ketoprofen	-11.80	-11.63	-11.35
lamivudine	-10.10	-10.34	-10.87
Laninamivir	-11.90	-11.82	-11.71
levamisole	-10.80	-10.94	-11.20
levofloxacin	-11.90	-11.72	-11.40
lindane	-10.40	-10.68	-11.26
lysine	-10.30	-10.43	-10.77
malathion	-9.90	-10.49	-11.42
malic acid	-11.50	-11.09	-10.36
m-chloroaniline	-11.70	-11.45	-10.99
Mefenamic acid	-11.40	-11.39	-11.33
Methamphetamine	-10.60	-10.73	-10.87
methyl acrylate	-9.80	-9.91	-10.15
methyl ethyl ketone	-9.50	-9.67	-9.97
methyl isobutyl ketone	-9.80	-9.99	-10.26
methyl methacrylate	-10.50	-10.34	-10.03
methylthiouracil	-10.50	-10.76	-11.09
metobromuron	-11.80	-11.59	-11.19
mitomycin C	-10.50	-10.77	-11.28
m-nitroaniline	-12.30	-11.67	-10.55
monochloro acetic acid	-10.50	-10.66	-11.03

Naproxen	-11.80	-11.37	-10.78
n-butanoic acid	-10.00	-10.10	-10.23
nitroquanidine (NQ)	-12.40	-11.80	-10.87
Novobiocin	-11.20	-11.46	-11.82
O-desmethylvenlafaxine	-11.70	-11.57	-11.30
oseltamivir carboxylate	-11.90	-11.71	-11.36
oxalic acid	-11.50	-11.33	-11.02
oxolinic acid	-12.10	-11.73	-11.01
paracetamol	-11.40	-10.98	-10.22
p-Cresol	-9.00	-9.40	-10.13
pentyl acetate	-10.50	-10.29	-10.08
peramivir	-11.90	-11.69	-11.31
Phenylbutazone	-10.70	-10.92	-11.31
polyacrylic acid	-12.70	-12.28	-11.51
pronamide	-12.90	-12.31	-11.16
propanil	-9.50	-10.05	-11.16
propionaldehyde	-9.80	-9.86	-10.01
quinine	-11.70	-11.56	-11.30
Salicylic acid	-9.70	-10.07	-10.66
sec-butyl alcohol	-9.80	-9.86	-9.96
sertraline	-11.90	-11.66	-11.29
simazine	-11.90	-11.53	-10.91
sucralose	-12.60	-12.14	-11.23
sulfadiazine	-11.90	-11.65	-11.28
sulfamethazine	-11.80	-11.70	-11.57
sulfamethoxazole	-11.90	-11.72	-11.50
Sulfamethoxypyridazine	-11.20	-11.38	-11.58
tert-butanol	-11.00	-10.76	-10.22
tetrachloroethylene	-11.50	-11.45	-11.30
toluene	-9.90	-10.02	-10.30
trichloroethylene	-11.60	-11.52	-11.34
Triclocarban	-11.80	-11.71	-11.58
Triclosan	-10.90	-10.99	-11.15
Trimethoprim	-11.80	-11.69	-11.50
tylosin	-12.70	-12.36	-11.74
vinyl chloride	-11.00	-10.91	-10.78
Zanamivir	-11.90	-11.77	-11.47
zidovudine	-10.10	-10.61	-11.43

Table SI3. Test set for the global QSAR model (RF-R + PaDEL) which contains the 30% of the full dataset (70 molecules)

name	Log k_{b2}	
	experimental	**predicted**
1-butyl-3-methylimidazolium hexafluorophosphate	-11.90	-11.31
2,4-D ethylester	-9.30	-9.83
2,4-dichlorophenol	-9.80	-10.71
2,4-dinitrophenol	-11.40	-11.24
2-acetamido-fluorene	-11.10	-10.92
2-phenylphenol	-9.40	-10.70
3-cresol (m-cresol)	-9.50	-10.04
4-chlorophenol	-10.30	-10.34
acetic acid	-10.20	-10.42
acetone	-9.80	-10.19
acrolein (2-propenal)	-9.80	-9.97
acrylic acid	-9.60	-10.50
allylamine	-10.60	-10.26
Amitriptiline	-11.80	-11.27
Anthranilic acid	-11.20	-10.55
benzene	-10.80	-10.39
benzenesulfonate	-11.00	-10.82
benzoic acid	-9.60	-10.47
benzylamine	-9.70	-10.39
bis(2-chloroisopropyl)ether	-10.50	-10.96
carbamazepine	-11.80	-11.03
catechol	-9.80	-10.05
cellobiose	-9.10	-11.27
chloromethyl methyl ether	-10.50	-10.74
chlorpropham (CICP)	-12.10	-11.22
clofibric acid	-12.00	-11.03
cyclohexane carboxylic acid	-11.60	-10.32
deisopropylatrazine	-12.20	-11.25
desethylatrazine	-12.20	-11.34
dichloro acetic acid	-10.00	-11.03
diethyl phthalate	-9.50	-10.12
dimethoate	-11.30	-10.83
dimethylamine	-9.60	-10.72
Diquat	-11.60	-11.08
endothall	-10.90	-10.95
estrone	-9.00	-10.12
ethanol	-9.80	-10.15
ethyl carbamate	-9.80	-10.45
Fluoxetine	-11.20	-11.56

fluridone	-10.40	-11.54
formate	-9.80	-10.45
glyphosate	-11.20	-11.07
Hydrochlorothiazide	-10.80	-11.34
irgarol 1051	-12.40	-11.44
Isobutanol	-9.90	-10.01
lactic acid	-10.20	-10.45
Methacrylic acid	-10.60	-10.28
methanol	-9.40	-10.23
metoprolol	-11.20	-11.20
Nadolol	-10.90	-11.48
n-butanol	-9.80	-10.17
n-butyraldehyde	-9.80	-9.96
nitrobenzene	-10.50	-10.85
Norfloxacin	-10.70	-11.44
o-Cresol	-9.60	-9.91
ofloxacin	-11.50	-11.72
pentachlorophenol	-11.50	-11.39
phenobarbital	-10.50	-11.05
phenol	-9.20	-9.98
propachlor	-10.30	-11.05
propranolol	-11.00	-11.09
styrene	-10.40	-10.43
styrene oxide	-9.80	-10.32
succinic acid	-11.50	-10.65
sulfadimethoxine	-12.50	-11.69
tartaric acid	-11.20	-10.65
Timolol	-11.20	-11.63
trichloro acetic acid	-11.00	-11.29
trifluoroacetic acid	-12.60	-11.20
venlafaxine	-12.00	-11.49

1. Baroni, M., et al., *Generating Optimal Linear Pls Estimations (Golpe) - an Advanced Chemometric Tool for Handling 3d-Qsar Problems.* Quantitative Structure-Activity Relationships, 1993. **12**(1): p. 9-20.
2. Carrio, P., et al., *Applicability Domain Analysis (ADAN): A Robust Method for Assessing the Reliability of Drug Property Predictions.* Journal of Chemical Information and Modeling, 2014. **54**(5): p. 1500-1511.
3. Pedregosa, F., et al., *Scikit-learn: Machine Learning in Python.* Journal of Machine Learning Research, 2011. **12**: p. 2825-2830.