

Supporting Information

In Silico Assessment of Chemical Biodegradability

Feixiong Cheng[†], Yutaka Ikenaga[§], Yadi Zhou[†], Yue Yu[†], Weihua Li[†], Jie Shen[†], Zheng Du[†],

Lei Chen[†], Congyin Xu[†], Guixia Liu[†], Philip W. Lee^{*,†,‡} and Yun Tang^{*,†}

*Corresponding Author: Tel: +86-21-64251052; Fax: +86-21-64253651

E-mail address: philiplee2007@gmail.com (P. L); ytang234@ecust.edu.cn (Y. T.)

Table S1. The CAS number, SMILES, ready biodegradability (RB) and not ready biodegradability (NRB) scores of 1,604 unique compounds (The training set and test set were listed). **Please see the excel file.**

Table S2. Overview of 148 physicochemistry descriptors and seven kinds of fingerprints used in this study.

Table S3. Feature subsets selected by four different features reduction methods.

Table S4. The detailed predicted results of the seven global probability models and consensus model and experimental results using OECD MITI test protocol for the external validation set of 27 novel chemicals. RB: ready biodegradability, NRB: not ready biodegradability, BOD: biological oxygen demand. If a chemical was predicted with a probability greater than 0.5, this chemical is RB, otherwise, this chemical is NRB. The detailed description about the consensus model using classic classifiers fusion techniques of Mean method and probability output can be found in our previous published work (Cheng et al. J. Chem. Inf. Model. 51, 996-1011.). **Please see the excel file.**

Table S5. The detailed results if the “Frequency of a fragment” enrichment factor of

the representative substructure fragments for ready and no ready biodegradability determined using FP4 fingerprints and Klekota-Roth fingerprint (ERFP). **Please see the excel file.**

Scheme S1. The “Frequency of a fragment” enrichment factor of 27 representative substructure fragments for ready and no ready biodegradability determined using FP4 fingerprints.

Scheme S2. The “Frequency of a fragment” enrichment factor of 15 specific substructure fragments for ready and no ready biodegradability determined by fragment analysis based on Klekota-Roth fingerprint.

Tables S

Table S1. The CAS number, SMILES, ready biodegradability (RB) and not ready biodegradability (NRB) scores of 1,604 unique compounds (The training set and test set were listed). **Please see the excel file.**

Table S2. Overview of 148 physicochemistry descriptors and seven kinds of fingerprints used in this study.

Descriptor Type	Number of Descriptors	Descriptor Type	Number of Descriptors
ALOGP	3	Longest aliphatic chain	1
APol	1	Mannhold LogP	1
Aromatic atoms count	1	McGowan volume	1
Aromatic bonds count	1	Molecular linear free energy relation	6
Atom count	8	Petitjean number	1
Autocorrelation (charge)	5	Ring count	9
Autocorrelation (mass)	5	Rotatable bonds count	1
Autocorrelation (polarizability)	5	Rule of five	1
BCUT	6	Topological polar surface area	1
Bond count	3	Vertex adjacency information	1
BPol	1	Weight	1
Carbon types	7	Weighted path	5
Chi chain	10	Wiener numbers	2
Chi cluster	6	XLogP	1
Chi path cluster	6	Zagreb index	1
Chi path	16	CDK fingerprint (FP)	1024
Eccentric connectivity index	1	CDK extended fingerprint (ExtFP)	1024
Atom type electrotopological state	24	Estate fingerprint (EStateFP)	79
Fragment complexity	1	MACCS fingerprint (MACCS)	166
Hbond acceptor count	1	Pubchem fingerprint (Pubchem)	881
Hbond donor count	1	Substructure fingerprint (FP4)	307
Kappa shape indices	3	Klekota-Roth fingerprint (KRFP)	4860
Largest chain	1	Klekota-Roth fingerprint count (KRFPC)	4860
Largest Pi system	1		

Table S3. Feature subsets selected by four different features reduction methods.

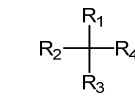
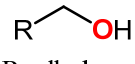
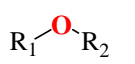
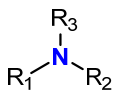
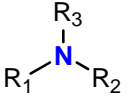

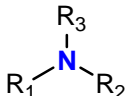

CFS	CART	CHAID	GASVM
BCUTp-1h	XLogP	VPC-5	BCUTw-1h
hmin	hmin	ATSm5	C1SP2
MLFER_E	VPC-4	ndssC	SC-4
nRing	nN	nN	VPC-4
nT6Ring	BCUTw-1h	VP-5	maxssO
n6Ring	C1SP2	C2SP2	fragC
naasC	ATSm1	nBondsS	nHBAcc
ALogP	WTPT-5	nH	MLogP
BCUTp-1l	BCUTp-1h	ALogp2	MLFER_E
naAromAtom	MLFER_E	VP-0	LipinskiFailures
---	---	---	WTPT-5
---	---	---	Zagreb

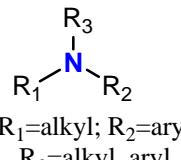
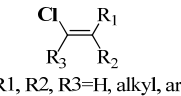
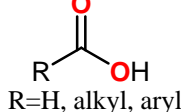
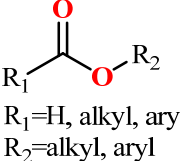
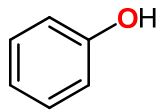

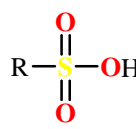
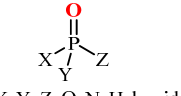
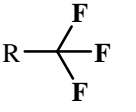
CFS: Linear correlation analysis, CART: The classification and regression tree algorithm, CHAID: chi-squared automatic interaction detector, GASVM: Genetic algorithm and Support Vector Machine.

Table S4. The detailed predicted results of the seven global probability models and consensus model and experimental results using OECD MITI test protocol for the external validation set of 27 novel chemicals. RB: ready biodegradability, NRB: not ready biodegradability, BOD: biological oxygen demand. If a chemical was predicted with a probability greater than 0.5, this chemical is RB, otherwise, this chemical is NRB. The detailed description about the consensus model using classic classifiers fusion techniques of Mean method and probability output can be found in our previous published work (Cheng et al. J. Chem. Inf. Model. 51, 996-1011.). **Please see the excel file.**

Table S5. The detailed results if the “Frequency of a fragment” enrichment factor of the representative substructure fragments for ready and no ready biodegradability determined using FP4 fingerprints and Klekota-Roth fingerprint (ERFP). **Please see the excel file.**

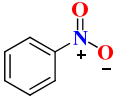

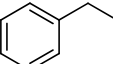
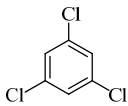
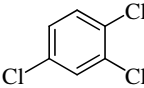
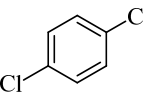
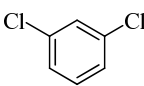
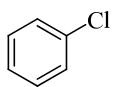
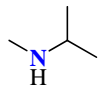
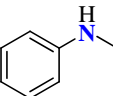
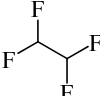
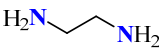
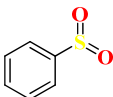
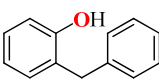
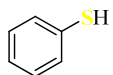
Scheme S1. The “Frequency of a fragment” enrichment factor of 27 representative substructure fragments for ready and no ready biodegradability determined using FP4 fingerprints.

NO	SMARTS	Description	General Structure	N _{RB}	N _{NRB}	F _{RB}	F _{NRB}	IG
1	[CX4]([#6])([#6])([#6])([#6])	Quaternary_carbon	 R ₁ =alkyl, aryl, R ₂ =alkyl, aryl R ₃ =alkyl, aryl, R ₄ =alkyl, aryl	7	93	0.192	1.46	0.022
2	[ClX1][CX4]	Alkylchloride	R-Cl R=alkyl	18	72	0.548	1.26	0.005
3	[FX1][CX4]	Alkylfluoride	R-F R=alkyl	0	27	0	1.57	0.011
4	[BrX1][CX4]	Alkylbromide	R-Br R=alkyl	2	28	0.183	1.47	0.007
5	[OX2H][CX4;!\$(C([OX2H])[O,S,#7,#15]))]	Alcohol	R-OH R=alkyl	119	83	1.615	0.647	0.021
6	[OX2H][CX4H2;!\$(C([OX2H])[O,S,#7,#15]))]	Primary_alcohol	 R=alkyl, aryl	84	51	1.71	0.595	0.018
7	[c][OX2][c]	Diarylether	 R ₁ =aryl, R ₂ =aryl	1	18	0.144	1.49	0.005
8	[NX3+0,NX4+;!\$(N~[#6]);!\$(N)*~[#7,#8,#15,#16])]	Amine	 R ₁ =alkyl, aryl R ₂ =alkyl, aryl R ₃ =alkyl, aryl	61	225	0.585	1.24	0.016
9	[NX3H0+0,NX4H1+;!\$(N)[!C]);!\$(N)*~[#7,#8,#15,#16])]	Tertiary_aliph_amine	 R ₁ =alkyl, R ₂ =alkyl R ₃ =alkyl	5	42	0.292	1.41	0.007
10	[NX3H2+0,NX4H3+][c]	Primary_arom_amine	R-NH ₂ R=aryl	9	104	0.218	1.45	0.023
11	[NX3H1+0,NX4H2+;!\$(N)[!c]);!\$(N)*~[#7,#8,#15,#16])]	Secondary_arom_amine	 R ₁ =aryl, R ₂ =aryl	0	21	0	1.57	0.008
12	[NX3H0+0,NX4H1+;!\$(N)[!c]);!\$(N)*~[#7,#8,#15,#16])]	Tertiary_arom_amine	 R ₁ =aryl, R ₂ =aryl R ₃ =aryl	0	18	0	1.57	0.007
13	[NX3H1+0,NX4H2+;!\$(N)([c])[C]);!\$(N)*~[#7,#8,#15,#16])]	Secondary_mixed_amine	 R ₁ =alkyl, R ₂ =aryl	1	14	0.183	1.47	0.003

14	[NX3H0+0,NX4H1+;\$([N]([c])([C])[#6]);!\$([N]*~[#7,#8,#15,#16])]	Tertiary_mixed_amine	 R ₁ =alkyl; R ₂ =aryl R ₃ =alkyl, aryl	1	30	0.088	1.52	0.009
15	[CIX1][CX3]=[CX3]	Chloroalkene	 R ₁ , R ₂ , R ₃ =H, alkyl, aryl	0	16	0	1.57	0.006
16	[CX3;\$([R0][#6]),\$([H1R0])](=[OX1])[\$([OX2H]),\$([OX1-])]	Carboxylic_acid	 R=H, alkyl, aryl	97	55	1.75	0.570	0.023
17	[CX3;\$([R0][#6]),\$([H1R0])](=[OX1])[OX2][#6;!\$(C=[O,N,S])]	Carboxylic_ester	 R ₁ =H, alkyl, aryl R ₂ =alkyl, aryl	115	45	1.97	0.443	0.041
18	[OX2H][c]	Phenol		52	152	0.699	1.17	0.006
19	[Cl][c]	Arylchloride	R-Cl R=aryl	7	168	0.109	1.51	0.052
20	[F][c]	Arylfluoride	R-F R=aryl	0	4	0	1.57	0.002
21	[Br][c]	Arylbromide	R-Br R=aryl	4	25	0.378	1.36	0.003
22	[I][c]	Aryliodide	R-I R=aryl	0	6	0	1.57	0.002
23	[sX2]	Hetero_S		0	14	0	1.57	0.006
24	[\$([NX3])(=O)=O,\$([NX3+])(=O)[O-])[!#8]	Nitro	 R=alkyl, aryl	3	84	0.094	1.52	0.026
25	[SX4;\$([H1]),\$([H0][#6])](=[OX1])(=[OX1])[\$([OX2H]),\$([OX1-])]	Sulfonic_acid	 R=alkyl, aryl	9	58	0.368	1.36	0.008
26	[PX4D4](=[!#6])([!#6])([!#6])([!#6])	Phosphoric_acid_derivative	 X, Y, Z=O, N, Hal residue	8	35	0.510	1.28	0.003
27	[FX1][CX4;!\$([H0][Cl,Br,I]);!\$([F][C]([F])([F])([F]))([FX1])]	Trifluoromethyl	 R-alkyl, aryl	0	18	0	1.57	0.007

N_{RB} is the number of ready biodegradability (RB) in ready biodegradability class with specified fragment t . N_{NRB} is the number of no ready biodegradability (NRB) in no ready biodegradability class with specified fragment t . F_{RB} is the “Frequency of a fragment” enrichment factor of a specified fragment (t) in ready biodegradability class. F_{NRB} is the “Frequency of a fragment” enrichment factor of a specified fragment (t) in no ready biodegradability class.

Scheme S2. The “Frequency of a fragment” enrichment factor of 15 specific substructure fragments for ready and no ready biodegradability determined by fragment analysis based on Klekota-Roth fingerprint.

NO	SMARTS	General Structure	N _{RB}	N _{NRB}	F _{RB}	F _{NRB}	IG
1	<chem>[O-][N+](=O)c1ccccc1</chem>		3	62	0.127	1.50	0.017
2	<chem>C1CC2CC1C=C2</chem>		0	8	0	1.57	0.003
3	<chem>CCc1ccccc1</chem>		30	154	0.447	1.32	0.018
4	<chem>Clc1cc(Cl)cc(Cl)c1</chem>		1	15	0.171	1.48	0.004
5	<chem>Clc1ccc(Cl)c(Cl)c1</chem>		0	22	0	1.57	0.009
6	<chem>Clc1ccc(Cl)cc1</chem>		0	31	0	1.57	0.013
7	<chem>Clc1cccc(Cl)c1</chem>		3	54	0.144	1.49	0.014
8	<chem>Clc1ccccc1</chem>		7	160	0.115	1.51	0.049
9	<chem>CNC(C)C</chem>		2	30	0.171	1.48	0.007
10	<chem>CNc1ccccc1</chem>		12	87	0.332	1.38	0.014
11	<chem>FC(F)C(F)F</chem>		0	13	0	1.57	0.005
12	<chem>NCCN</chem>		2	29	0.177	1.47	0.007
13	<chem>O=S(=O)c1ccccc1</chem>		7	74	0.237	1.44	0.015
14	<chem>Oc1ccccc1Cc2ccccc2</chem>		1	22	0.119	1.51	0.006
15	<chem>S(c1ccccc1)</chem>		8	83	0.241	1.44	0.017

N_{RB} is the number of ready biodegradability (RB) in ready biodegradability class with specified fragment *t*. N_{NRB} is the number of no ready biodegradability (NRB) in no ready biodegradability class with specified fragment *t*. F_{RB} is the “Frequency of a fragment” enrichment factor of a specified fragment (*t*) in ready biodegradability class. F_{NRB} is the “Frequency of a fragment” enrichment factor of a specified fragment (*t*) in no ready biodegradability class.