Electronic Supplementary Material (ESI) for Green Chemistry.
This journal is © The Royal Society of Chemistry 2020

Transition-state rate theory sheds light on 'black-box' biodegradation algorithms

T.M. Nolte, W.J.G.M. Peijnenburg, T.J.H.M. van Bergen, A.J. Hendriks

SUPPLEMENTARY INFORMATION

CONTENT

S0. Data selection

For details on the data selection and curation, we refer to our previous study [1]. Briefly, in both the current study and our previous study, the underlying data were a mix of second-order and first-order rate constants for primary aerobic biodegradation. The starting point was to include only unadapted communities [1]. However, we cannot exclude the presence of communities present due to e.g. (historic) cometabolism or from (local) background concentrations. These may be constant factors for many chemicals, moreover, for a large dataset (e.g. N=550), the effects encoded in the predicted (RF) values for k_b are 'averaged out' and only consider the differences between chemicals. Thus, the data unit was 'homogenized': when unavailable, we considered the biomasses to be constant and convert the first-order rate constants (1) to second-order rate constants (2) according to:

$k_b(1) = [\text{Biomass}] \times k_b(2)$

In total, we selected 550 compounds. Structures were drawn for their speciated form, at experiment-specific pH, where possible. We performed corrections for bioavailability via sorption to dissolved organic carbon. Fig. S1 shows the distribution of logK_{OW} and molecular volume. With exception of a few compounds, logK_{OW} was between -4 and +4. We included highly diverse molecular volumes, ≤400 Å^3:

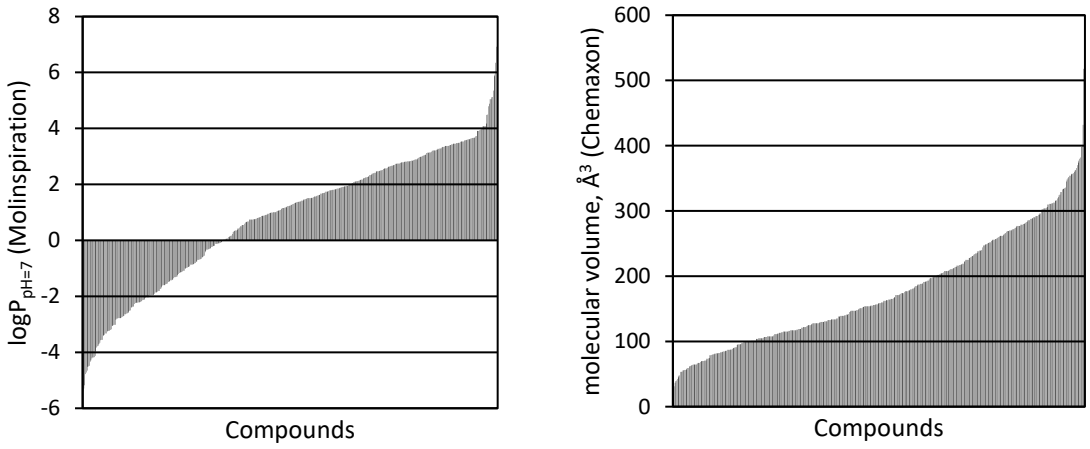

Fig. S1. The distribution of log*P* and molecular volume of the compounds (N=550).

S1. Supplemental RF-QSBR validation

The RF-QSBR has R^2_{ext} = 0.66±0.05, and root-mean-squared error ($RMSE_{ext}$) = 0.53±0.03. The RF-QSBR entailed fewer outliers (0.5<RMSE<0.6) than previously (0.7)[1] showing the RF algorithm finds more statistically significant relationships between structural aspects and k_b, i.e. 'learned' 'more' from the larger dataset (predictions are more precise). This is an intrinsic result in 'big data' science.

Fig. S2 shows the predicted k_b values by the RF-QSBR versus the biodegradation probability from CATABOL. We find a general agreement, with discrepancies for e.g. cyanobenzenes, pyridines desulfuration and beta-oxidaton not being significant. Discrepancies may arise due to: 1) either the CATABOL or current dataset carries insufficient learning data, 2) CATABOL is not parameterized to account for acclimation (implying e.g. that bacteria are 'more easily' acclimated to pyridines than to nitriles), 3) abiotic hydrolysis [2, 3], 4) naturally occurring nitrilase-like enzymatic activity may be relatively abundant [4].

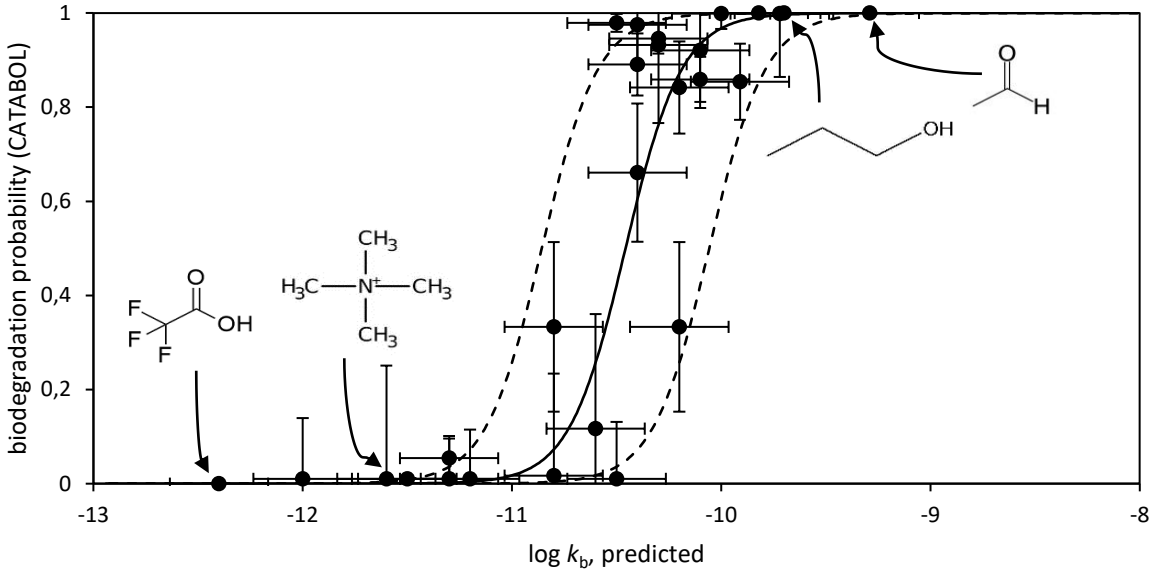

Fig. S2. Predicted logk_b versus the probability of the principle reaction (biodegradation step) as utilized by CATABOL [5, 6]. Lines denote a sigmoid fit and 1 standard error. Probability for trifluoroacetate (lower bottom-left) was taken based on structurally related compounds.

S2. Calculations

According to the collision theory, the number of molecules of product formed per unit time per unit volume is equal to the number of collisions, *A*, multiplied by a factor, which takes into account the fact that only a fraction of the collisions involve molecules that possess the excess energy, activation energy, necessary for reaction[7]. The dynamics of diffuse fronts in systems modeled with step-function kinetics and in systems modeled with the Arrhenius kinetics are qualitatively the same at time scales at which the bulk reaction ahead of the front can be ignored [8]. Based on these notions, we define:

Equation S1-1 $$k_b \propto A(i-j) \cdot P(i-j)e^{\frac{-\Delta G^{\ddagger}(i-j)}{RT}}$$

Wherein:

Equation S1-2 $$A(m_{i-j}) = \frac{\sum_{i=0}^{i} D(i-j)}{d(i-j)\cdot \Lambda}$$

The interpretation of the symbols is given in the main document. Λ is the de Broglie wavelength, $\Lambda = h/p$, with h Planck's constant and p the momentum of the particle/molecule. The latter we consider constant for all molecules.

We calculated the terms in Eq. S1 as custom descriptors via SMILES (Simplified Molecular Input Line Entry System) input. As the electronic structure of molecules and energies of their frontier orbitals can be significantly altered by (de-)protonation, we implement pH-corrected ionic speciation states for the calculations: we determined ion speciation states at experimental pH (~7.4) using pK_a/pK_b, taken from the literature or estimated using ChemAxon[9].

We refer to the spreadsheets, as supplementary information, for practical examples of application of the methods.

Calculation of $\Delta G^{\ddagger}$

From the vast number of possible bacteria, enzymes, isoforms, concentrations, geometries, co-factors, etc., we regard direct calculation of realistic activation energies $\Delta G^{\ddagger}$ using current chemo-informatic tools for the chemicals considered not realistic.

On a higher level, there is some empirical evidence that $\Delta G^{\ddagger}$ relates to delocalizability (δ) and the energy of the highest occupied molecular orbital (E_{HOMO}) of the molecule, Eq. S2. We calculate δ via atom-specific Fukui (electrophillic) delocalizability indices [10]. We take delocalizabilities as minimum values on aliphatic and maximum values among aromatic carbons in the molecule, respectively.

Equation S2 $$\Delta G^{\ddagger} = f(\delta, E_{\mathrm{HOMO}})$$

Based on previous results [11-15], we calculated δ and E_{HOMO} via MOPAC [16, 17]. Structures were pre-optimized using OpenBabel[18] and molecular orbital (MO) calculations were carried out using the semi-empirical Hamiltonian Parameterization Method 7 (PM7 Hamiltonian) within the program package MOPAC Version 2016 [16] with 92 geometrical segments (NSPA). We describe the water solvent (ε = 78.4) using the COSMO Implicit Solvation (Conductor-like Screening approximation) Model.

Semi-empirical MO theory was chosen to limit the computational effort, but we increased the criteria for terminating electronic and geometric optimizations by a factor 100 to acquire more precise results. The accuracy of MOPAC's 3D structure generation is evaluated elsewhere: relevant information, e.g. heat of formations, can be accessed here:

http://openmopac.net/PM7_accuracy/Heats_of_Formation.html

Calculation of D

Considering the complexity of biodegradation, we deemed it not realistic to discern between the potential influences of diffusion through membranes, aqueous pores or towards/within cascades of proteins/enzymes. As a more general description, we considered for diffusion limited reactions:

Equation S3 $$k_{\mathrm{b}} \propto A = (D_{\mathrm{i}} + D_{\mathrm{j}}) \cdot R_0$$

where R_0 is the minimal distance between molecule i and enzyme j active sites obtainable during the biotransformation. With virtually endless possible sizes and shapes for the enzymes active sites, it seems unlikely that we can specify D_{j}. Luckily, since the enzyme/bacteria is large, it is effectively stationary, and only D_{i} is relevant:

Equation S4 $$A \propto (D_{\mathrm{i}}) \cdot R_0$$

It is cumbersome to calculate D_{i} for 550 molecules using 'ab initio' methods. Rather, we determined the diffusion coefficient via the Stokes-Einstein relationship and volume [19-21]:

Equation S5 $$D \propto V^{-1/3}$$

With V as molecular volumes. We anticipate deviations in Eq. S5 for non-spherical molecules which we characterized by *d*. We describe deviations due to polarity influence on the diffusion of molecules by *P*. Both are detailed below.

Calculation of d

Collision theory gives good results for bimolecular gas reactions and reactions in solution involving simple ions. However, for many other reactions the predicted rates are (much) too large. The deviation appears to increase with the complexity of the reactant molecules. As a means of correcting for this deviation we need a probability or steric factor [22]. Illustratively:

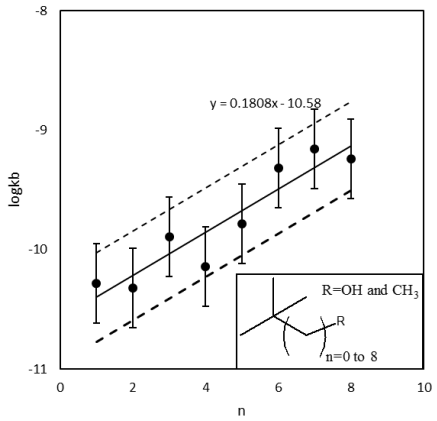

Fig. S3. Log-transformed length-normalized k_b' (alkanes+alcohols) versus a steric factor[a]. Error bars are RF prediction uncertainties

Computation of surface accessibility has importance in drug (ligand) design: most binding sites for small ligands in proteins are cavities, with specific accessibility (imposing an upper limit for a probe). Illustratively, the weight of the catalytic domain positively correlates with the catalysis[23]. We assume the minimal distance R_0 (Eq. S4) to express effective interaction (catalysis) which is proportional to the effective areas. Then, we can use the accessibility ratio, as proposed by Feldblum and Isaiah [24] to determine the characteristic distance $d(i-j)$ of the chemical and active site via:

Equation S6 $$d(i-j) \sim R_0 \sim R_{\mathrm{g}} \cdot \frac{ASA(i-j)}{vdwSA(i)}$$

in which $ASA(i-j)$ the accessible surface area (e.g. to the enzymes catalytic site), $vdwSA(i)$ is the van der Waals surface area, and R_{g} is the radius of gyration. We approximate R_{g} by substituting volumes into:

[a] Here, we show the steric factor as number of carbon bonds adjacent to t-Bu. k_b' was taken as $k_b = k_b$(n, t-Bu) / k_b(n), wherein k_b(n) is k_b for the equivalent compound (to k_b(n, t-Bu)) without the t-Bu group.

Equation S7 $$R_g \sim \left(\frac{3}{4}\frac{V}{\pi}\right)^{1/3}$$

The $ASA(i-j)$ of atom i is defined as the locus of the center of the probe j. The ASA of an atom radius r is the area on the surface of the sphere of radius $R=r+r_{probe}$ on each point of which the probe (solvent) molecule can be placed in contact with this atom without penetrating any other atoms of the molecule. Fig. S4 illustrates parameters in Eq. S6, e.g. the black circle denotes R_g, proportional to the root mean square distance of all atoms:

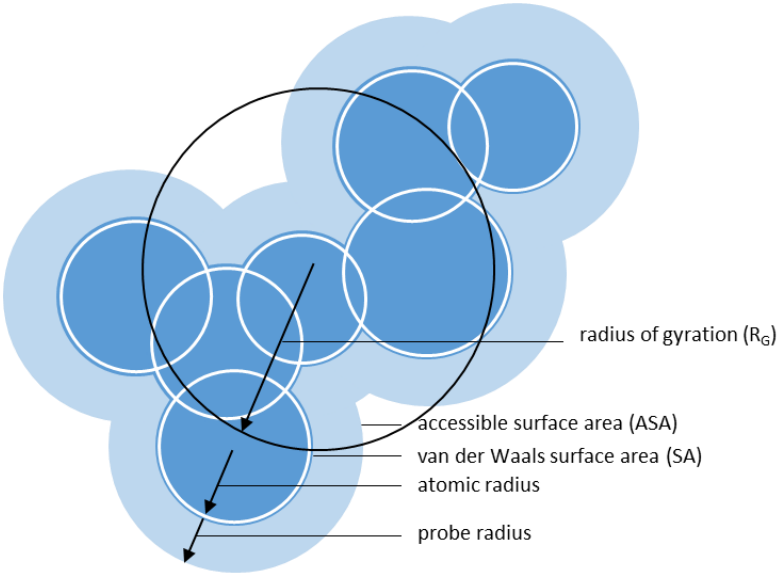

Fig. S4. Two-dimensional depiction of the accessibility ratio $\frac{ASA(i-j)}{SA(i)}$.

Probing the 550 molecules with a multitude of biochemical 3D structures is computationally intensive and laborious. As protein binding sites are accessible only to small molecules, there is a connection between cavity and solvent accessible surface area. Therefore, instead we take as the probe simply a H_2O molecule with a radius ~1.4A. Thus, we simplify $ASA(i-j)$ as determined by the solvent molecule H_2O rolling over the van der Waals (vdw) surface area of the solute molecule[25].

Since charge can affect intramolecular forces, we let Chemaxon calculate values for the areas at pH=7.4: $vdwSA(i)_{pH=7.4}$ and $ASA(i)_{pH=7.4}$.

Calculation of (Σ)

In-house preliminary analysis including the multiplicity *Σ* (number of equivalent functional groups) did not find any significant improvement of the correlations of both global and class specific sets of compounds via any known (to us) methods. Hence, the multiplicity was not taken into account in this study.

Calculation of P

The membrane and internal cellular components are main barriers for diffusion. The diffusion coefficient can be determined via the Hayduk-Laudie correlation, but applies only to uncharged molecules. For charged molecules, the solvation layer needs to be included. This is because ionic diffusion is slower when the hydration layer is thicker due to higher the ionic potential.

We calculated logK_{OW} (characterizing facilitated diffusion), for specific speciation states (pH=7.4) of the molecules, i.e. log$D_{OW,pH=7.4}$ via Chemaxon [9, 26] and validated manually via the Molinspiration webtool [27]. Then, logK_{OW} characterizes diffusion via the inclusion of the hydration layer. We consider P constant for all carboxylates on the basis that ionic binding is stronger than hydrophobic binding: k_b values for carboxylates were not corrected for K_{OW}.

To illustrate the interdependence of parameters in Eq. S1, Fig. S5 shows the dependence of K_{OW} on surface area for 'like' chemical classes e.g. ethylene glycol oligomers, alkanes, etc.:

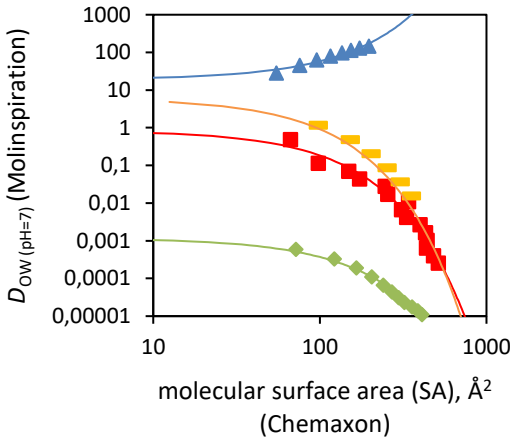

Fig. S5. The interdependence between K_{OW} and molecular surface area of oligomers. Blue are alkanes, red are ethylene glycol oligomers, green are carboxylates, yellow are alcohols.

The solid curves in Figure S6B shows expected values based on thermodynamic considerations, based on the formula:

Equation S8 $$k_b \propto \frac{1}{SA} = \pm \frac{0.13}{ln(K_{OW}) - \ln(K_{OW}')}$$

Wherein K_{OW} is cross-correlated to e.g. surface area via $K_{OW} \propto e^{\pm 0.13SA}$ (calculated via [27]). K_{OW}' is size independent and compound specific (Fig. S6). By extension, we use a relation between area-normalized k_b and K_{OW}:

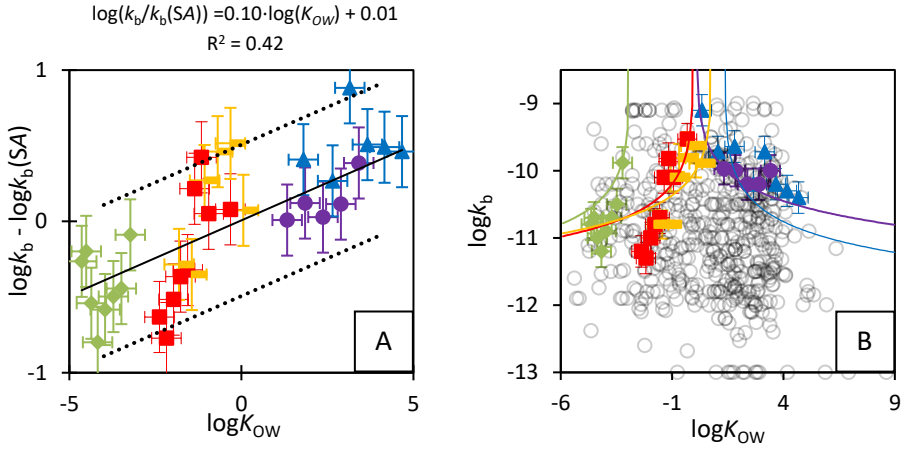

Fig. S6. A: The surface area-normalized biodegradation rate constant versus logK_{OW}. We normalized for surface area via its relationship with K_{OW} (Fig. S5). B: The biodegradation rate constant versus logK_{OW} (via log$D_{OW,\ pH=7}$ as calculated via Molinspiration). Colors indicate different families of molecules/oligomers. Green: carboxylates, red: ethylene glycol oligomers, yellow: alcohols, purple: carboxylates, blue: alkanes. Solid lines denote the expected values based on K_{OW} and K_{OW}'.

We consider k_b data for compounds with no significant variation as expected from E_{HOMO} or δ (i.e. $\Delta G^{\ddagger}$ is constant). For these data, based on Fig. S6A, the partition function relates to K_{OW}:

Equation S9 $$k_b \propto P \propto 0.10(\pm 0.02) \cdot K_{OW}$$

I.e. a factor ~10 difference in the equilibrium partitioning. In comparison the carbon density in bacteria is a factor ~3 higher than in octanol. For a better comparison, we should distinguish between the fractions of polar and non-polar carbon. Hence, we view the obtained regression with logK_{OW} to be in line with the differences in organic carbon density between octanol and active biomass in environmental matrices.

S3. Supplemental modelling results

We have used *P*, *D* and the accessibility term *d* to transform the k_b values. Via fitting all parameters (see above) to k_b data for 'similarly reactive chemicals', Eq. S1 becomes:

Equation S10 $$k_b \propto V(i-j)^{-1/3} \cdot K_{OW}(i-j)^{0.1} \cdot \left(R_g \frac{ASA(i-j)}{SA(i)}\right)^{-1.8} \cdot \sum_{i=0}^{i} e^{\frac{-\Delta G^{\ddagger}(i-j)}{RT}}$$

Wherein the apparent $\Delta G^{\ddagger}(i-j)$ is described in terms of δ and E_{HOMO} (S2).

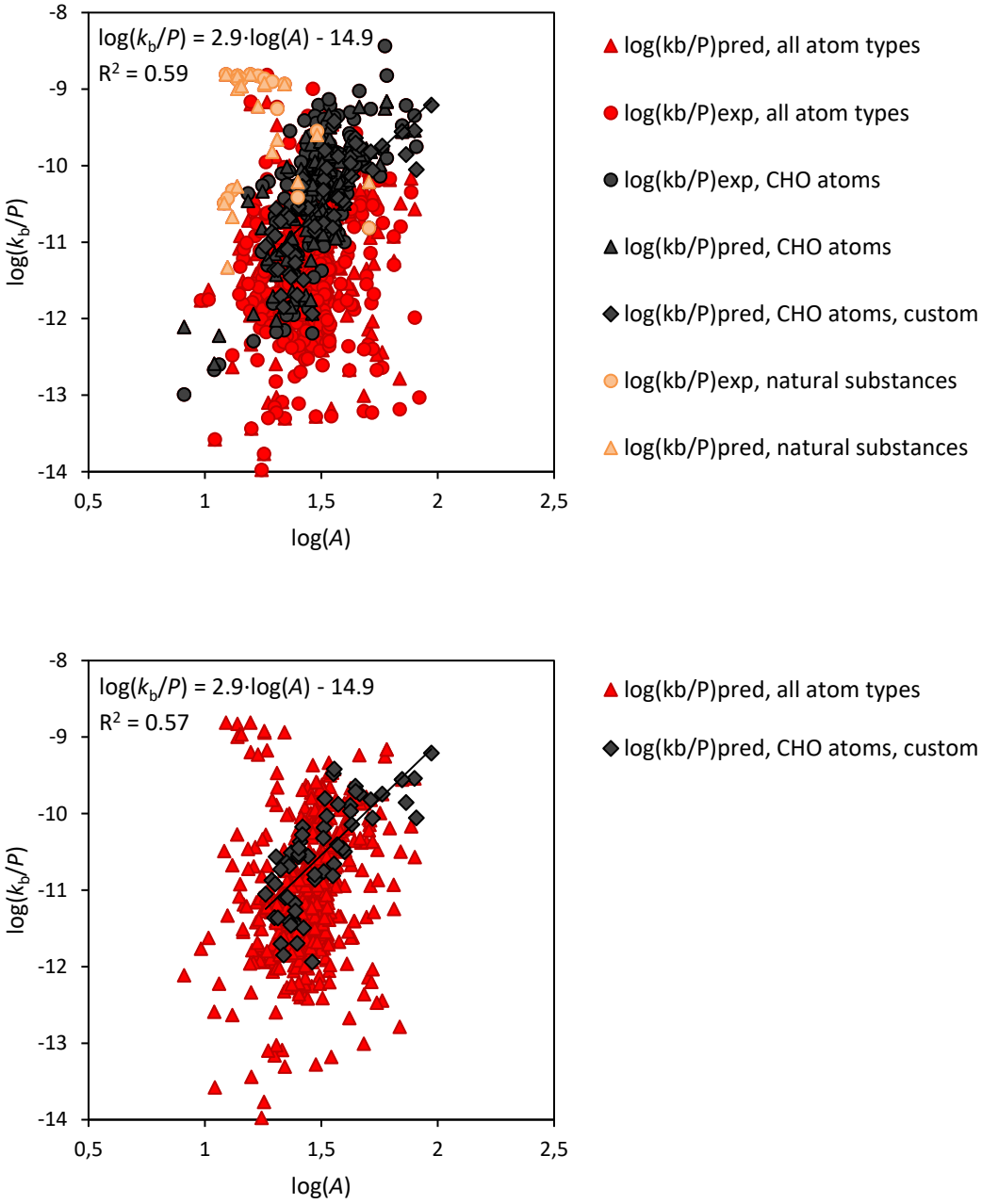

Fig. S7. The log(k_b/P), i.e. logk_b normalized for the partition function, versus the frequency factor A. Black symbols denote 'electron-rich' compounds.

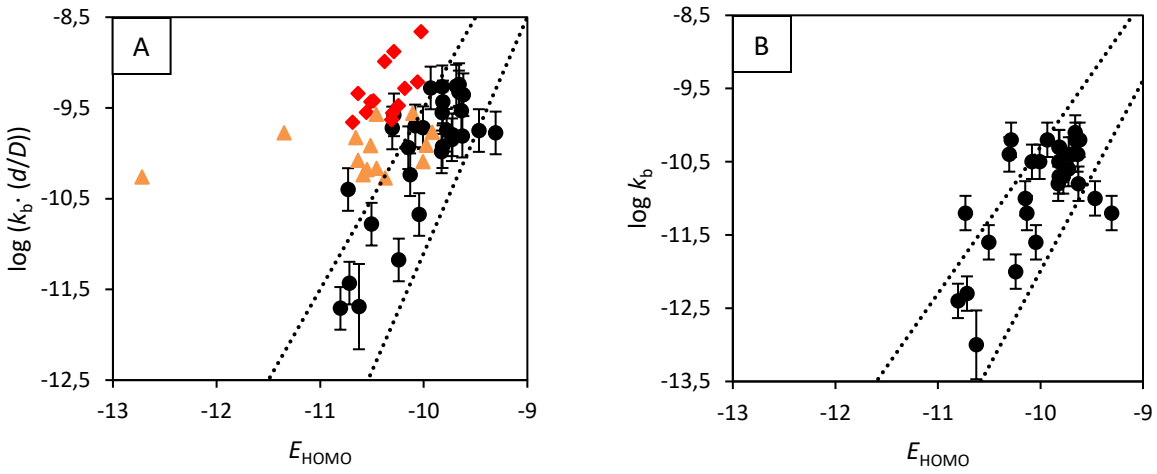

Fig. S8. A: log k_b (transformed k_b values) versus E_{HOMO}. Orange triangles denote nitrogen-containing compounds, each with more than 1 possible biotransformation pathway according to EAWAG PPS [28]; red denotes natural substances. Fig. B: log k_b (non-transformed k_b values) versus E_{HOMO}. Tricyanoacetate (orange triangle) did not adhere to the LFER in Fig. S8A.

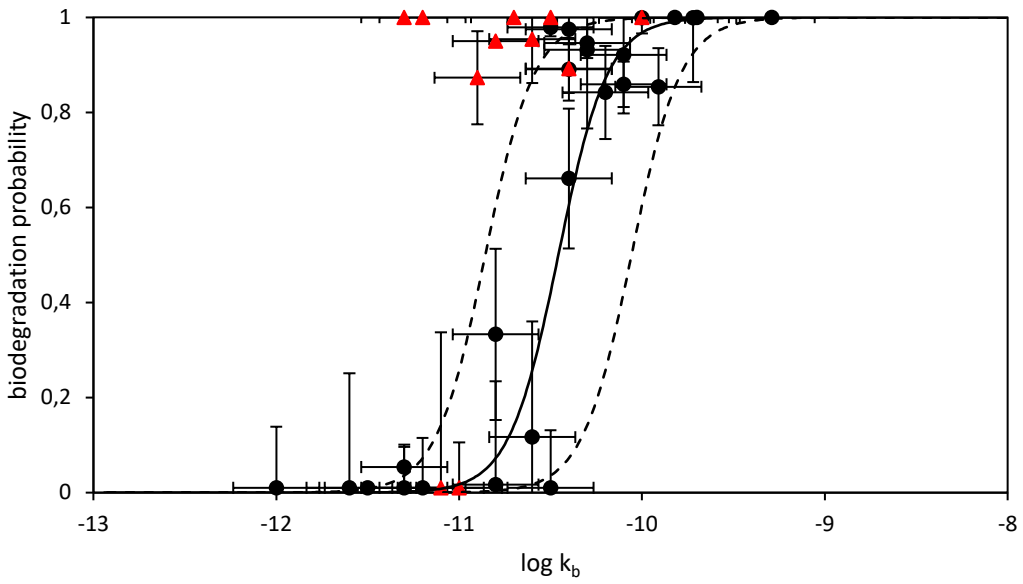

Fig. S9. CATABOL predictions (y) versus the current study (x) predictions for biodegradation. Red triangles denote compounds entailing possibly hydrolytically unstable groups, i.e. which degrade abiotically.

From the 'global' QSAR, we have found R^2 = 0.66 ± 0.05 and $RMSE_{pred}$ ~ 0.53 ± 0.03. The latter number entails both prediction uncertainty and internal variability as a result of test conditions. If we consider 'like' chemicals only, test conditions factor out (only the transformation step is considered). Then, the $RMSE_{pred}$ in k_b is a function of the combined RMSE, e.g. $0.5 \cdot RMSE_{total}$ [12]. This was used to construct error bars throughout this study.

References

1. Nolte, T. M.; Pinto-Gil, K.; Hendriks, A. J.; Ragas, A. M. J.; Pastor, M., Quantitative structure-activity relationships for primary aerobic biodegradation of organic chemicals in pristine surface waters: starting points for predicting biodegradation under acclimatization. *Environ Sci-Proc Imp* **2018,** *20* (1), 157-170.
2. Zhang, M. J.; Yao, P. Y.; Yu, S. S.; Zhang, T. C.; Wu, Q. Q.; Zhu, D. M., Efficient selective hydrolysis of terephthalonitrile to 4-cyanobenzoic acid catalyzed by a novel nitrilase from Pantoea sp. *Process Biochem* **2018,** *75*, 152-156.
3. Serra, I.; Capusoni, C.; Molinari, F.; Musso, L.; Pellegrino, L.; Compagno, C., Marine Microorganisms for Biocatalysis: Selective Hydrolysis of Nitriles with a Salt-Resistant Strain of Meyerozyma guilliermondii. *Mar Biotechnol* **2019,** *21* (2), 229-239.
4. Schnoor, J. L.; Licht, L. A.; Mccutcheon, S. C.; Wolfe, N. L.; Carreira, L. H., Phytoremediation of Organic and Nutrient Contaminants. *Environ Sci Technol* **1995,** *29* (7), A318-A323.
5. Dimitrov, S.; Pavlov, T.; Dimitrova, N.; Georgieva, D.; Nedelcheva, D.; Kesova, A.; Vasilev, R.; Mekenyan, O., Simulation of chemical metabolism for fate and hazard assessment. II CATALOGIC simulation of abiotic and microbial degradation. *Sar Qsar Environ Res* **2011,** *22* (7-8), 719-755.
6. Jaworska, J.; Dimitrov, S.; Nikolova, N.; Mekenyan, O., Probabilistic assessment of biodegradability based on metabolic pathways: Catabol system. *Sar Qsar Environ Res* **2002,** *13* (2), 307-323.
7. Hill, T. L., Diffusion Frequency Factors in Some Simple Examples of Transition-State Rate Theory. *P Natl Acad Sci USA* **1976,** *73* (3), 679-683.
8. Golovati, D., On Step-Function Reaction Kinetics Model in the Absence of Material Diffusion. *SIAM Journal on Applied Mathematics* **2007,** *67* (3), 792-809.
9. ChemAxon *Calculator Plugin for structure property prediction. MarVin Version 5.2.0*.
10. Fukui, K.; Kato, H.; Yonezawa, T., A New Quantum-Mechanical Reactivity Index for Saturated Compounds. *B Chem Soc Jpn* **1961,** *34* (8), 1111-1115.
11. Rorije, E.; Langenberg, J. H.; Richter, J.; Peijnenburg, W. J., Modeling reductive dehalogenation with quantum chemically derived descriptors. *Sar Qsar Environ Res* **1995,** *4* (4), 237-52.
12. Nolte, T. M., Chen, G. et al., Disentanglement of the chemical, physical, and biological processes aids the development of quantitative structure-biodegradation relationships for aerobic wastewater treatment. *STOTEN* **2020,** *708* (133863).
13. Nolte, T. M.; Peijnenburg, W. J. G. M., Aqueous-phase photooxygenation of enes, amines, sulfides and polycyclic aromatics by singlet (a1Δg) oxygen: prediction of rate constants using orbital energies, substituent factors and quantitative structure–property relationships. *Environmental Chemistry* **2018,** *14* (7), 442-450.
14. Nolte, T. M.; Peijnenburg, W. J. G. M., Aqueous-phase reaction between organic chemicals and superoxide/hydroperoxyl radicals (HO2·/O2·-) – Relationships between oxidation and reduction rate constants and quantum-chemical descriptors. *Free Radical Research* **2019,** *52* (10), 1118-31.
15. Nolte, T. M., Nauser, T., Gubler, L., Peijnenburg, W.J.G.M. , Thermochemical unification of molecular descriptors to predict radical hydrogen abstraction with low computational cost *PCCP* **2020,** *accepted*.
16. Stewart, J. J. P. *MOPAC*, Stewart Computational Chemistry: Colorado Springs, CO, USA, 2016.
17. Stewart, J. J. P., Optimization of parameters for semiempirical methods VI: more modifications to the NDDO approximations and re-optimization of parameters. *J Mol Model* **2013,** *19* (1), 1-32.
18. O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R., Open Babel: An open chemical toolbox. *J Cheminformatics* **2011,** *3*.

19. Costigliola, L.; Heyes, D. M.; Schroder, T. B.; Dyre, J. C., Revisiting the Stokes-Einstein relation without a hydrodynamic diameter. *J Chem Phys* **2019,** *150* (2).
20. Marrink, S. J.; Berendsen, H. J. C., Permeation process of small molecules across lipid membranes studied by molecular dynamics simulations. *J Phys Chem-Us* **1996,** *100* (41), 16729-16738.
21. Minakata, D.; Mezyk, S. P.; Jones, J. W.; Daws, B. R.; Crittenden, J. C., Development of Linear Free Energy Relationships for Aqueous Phase Radical-Involved Chemical Reactions. *Environ Sci Technol* **2014,** *48* (23), 13925-13932.
22. Stevanovic-Huffman, M. M.; Savkovic-Stevanovic, J. B., Chemical reaction and diffusion dynamics. *International Journal of Mathematical Models and Methods in Applies Sciences* **2012,** *6* (3).
23. Arcus, V. L.; Prentice, E. J.; Hobbs, J. K.; Mulholland, A. J.; Van der Kamp, M. W.; Pudney, C. R.; Parker, E. J.; Schipper, L. A., On the Temperature Dependence of Enzyme-Catalyzed Rates. *Biochemistry-Us* **2016,** *55* (12), 1681-1688.
24. Feldblum, E., S.; Arkin, I. T., Strength of a bifurcated H bond. *PNAS* **2014,** *111* (11).
25. Gromiha, M. A., S., Role of solvent accessibility in structure based drug design. *Curr. Comput. Aided Drug Des.* **2005,** *1*, 65-72.
26. Dong, J.; Cao, D. S.; Miao, H. Y.; Liu, S.; Deng, B. C.; Yun, Y. H.; Wang, N. N.; Lu, A. P.; Zeng, W. B.; Chen, A. F., ChemDes: an integrated web-based platform for molecular descriptor and fingerprint computation. *J Cheminformatics* **2015,** *7*.
27. Molinspiration Cheminformatics. http://www.molinspiration.com/.
28. Eawag Biocatalysis/Biodegradation Database and Pathway Prediction System. http://eawag-bbd.ethz.ch/predict/ (accessed December 7).