Development of a QSAR model for predicting aqueous reaction rate constants of organic chemicals with hydroxyl radicals

Xiang Luo[a], Xianhai Yang[a], Xianliang Qiao[a], Ya Wang[a], Jingwen Chen[a,*], Xiaoxuan Wei[a], Willie J. G. M. Peijnenburg[b,c]

[a] Key Laboratory of Industrial Ecology and Environmental Engineering (MOE), School of Environmental Science and Technology, Dalian University of Technology, Dalian 116024, China

[b] Institute of Environmental Sciences (CML), Leiden University, Leiden 2300 RA, The Netherlands

[c] National Institute of Public Health and the Environment (RIVM), Center for Safety of Substances and Products, Bilthoven 3720 BA, The Netherlands

(25 Pages, 4 Figure, 4 Tables)

Information for some Dragon descriptors of the QSAR model.

(1) GETAWAY descriptors

The GETAWAY descriptors are chemical structure descriptors derived from the Molecular Influence Matrix, denoted by H and defined as the following:

$$H = M(M^{T}M)^{-1}M^{T}$$

where M is the molecular matrix consisting of the centred Cartesian coordinates x, y, z of the molecule atoms (hydrogens included) in a chosen conformation, and the superscript T refers to the transposed matrix. Atomic coordinates are assumed to be calculated with respect to the geometrical centre of the molecule in order to obtain translation invariance. The molecular information matrix is a symmetric matrix and shows rotational invariance with respect to the molecule coordinates, thus resulting independent of molecule alignment.

(2) 3D-MoRSE descriptor

3D-MoRSE (3D-Molecule Representation of Structures based on Electron diffraction) descriptors are based on the idea of obtaining information from the 3D atomic coordinates by the transform used in electron diffraction studies for preparing theoretical scattering curves. The following expression is used for 3D-MoRSE descriptor calculation:

* Correspondence to: Jingwen Chen, e-mail: jwchen@dlut.edu.cn; Phone: +86-411-8470 6269

$$Morsw = \sum_{i=1}^{nAT-1} \sum_{j=i+1}^{nAT} w_i w_j \frac{\sin(s \cdot r_{ij})}{s \cdot r_{ij}}$$

where *Morsw* is the scattered electron intensity, *w* is an atomic property, r_{ij} indicates interatomic distances and *nAT* is the number of atoms. The term *s* represents the scattering in various directions by a collection of *nAT* atoms. In order to obtain uniform length descriptors, the intensity distribution is made discrete, calculating its value at a sequence of evenly distributed values; in particular, in Dragon, it is assumed that *s* takes integer values in the range 0 ~ 31 (for $s = 0$ the scattering ratio is assumed equal to one).

In addition, 3D-MoRSE descriptors are calculated for the unweighted case (u) and for each selected atomic weighting scheme.

(3) Moriguchi octanol-water partition coefficient

Moriguchi octanol-water partition coefficient (*MLogP*) is calculated from Moriguchi logP model consisting of a regression equation based on 13 structural parameters (I. Moriguchi, S. Hirono, Q. Liu, I. Nakagome, and Y. Matsushita, Chem. Pharm. Bull., 1992, 40, 127-130; I. Moriguchi, S. Hirono, I. Nakagome, H. Hirano, Chem. Pharm. Bull., 1994, 42, 976-978). The regression coefficients were evaluated by a training set of 1230 organic molecules, including general aliphatic, aromatic, and heterocyclic compounds, containing the following atoms: C, H, N, O, S, P, F, Cl, Br, I. The statistical parameters of the model are: $r^2 = 0.906$; $s = 0.422$.

MlogP = -1.041 + 1.244(CX)0.6 - 1.017(NO)0.9 + 0.406(PRX) - 0.145(UB)0.8 + 0.511(HB) + 0.268(POL) - 2.215(AMP) + 0.912(ALK) -0.392(RNG) -3.684(QN) + 0.474(NO2) + 1.582(NCS) + 0.773(BLM)

The model variables are frequencies (denoted by N) or presence/absence (denoted by D) of some molecular features.

(4) Geary autocorrelations

Geary autocorrelations (*GATSkw*), *w* being the atomic property used to weight the molecular graph and *k* the lag, is calculated by applying Geary coefficient to the H-filled molecular graph:

$$GATSkw = \frac{\frac{1}{2\Delta?} \sum_{i=1}^{nAT-1} \sum_{j=i+1}^{nAT} \delta_{ij} (w_i - w_j)^2}{\frac{1}{(nAT-1)} \sum_{i=1}^{nAT} (w_i - w_{\text{mean}})^2}$$

where w_i is any atomic property, w_{mean} means its average value on the molecule, *nAT* is the number of atoms, δ_{ij} is the topological distance between two considered atoms). Δ is the sum of the Kronecker deltas, that is, the number of atom pairs at distance equal to *k*. Geary

coefficient is a distance-type function varying from zero to infinite. Strong spatial autocorrelation produces small values of this index; moreover, positive autocorrelation translates in values between 0 and 1 whereas negative autocorrelation produces values larger than 1; therefore, the reference "no correlation" is coefficient value equal to 1.

(5) Edge adjacency indices

These are topological indices calculated from the edge adjacency matrix of a molecule. The edge adjacency matrix (*EA*) is derived from the H-depleted molecular graph and encodes information about connectivity between graph edges. It is a square symmetric matrix of dimension nBO x nBO, where nBO is the number of bonds between non-hydrogen atom pairs. The entries of the matrix equal one if the considered bonds are adjacent and zero otherwise.

By replacing the zero diagonal elements of the edge adjacency matrix with specific bond properties, augmented edge adjacency matrices (*AEA*) are obtained, each encoding specific chemical information.

Moreover, weighted edge adjacency matrices (*EA(w)*) are unsymmetrical edge matrices derived from an edge-weighted molecular graph obtained by applying an edge weighting scheme w, as follows:

$$[EA(w)]_{ij} = w_j \text{ (if } i, j \text{ are adjacent bonds)}$$

$$[EA(w)]_{ij} = 0 \text{ (otherwise)}$$

where the off-diagonal elements of the matrix for a weighted graph are zero for non-adjacent edges, while, if two edges *i* and *j* are adjacent, the entry *i*-*j* is defined by the weight w_j of the *j*th edge and the symmetric entry *j*-*i* is defined by the weight w_i of the *i*th edge.

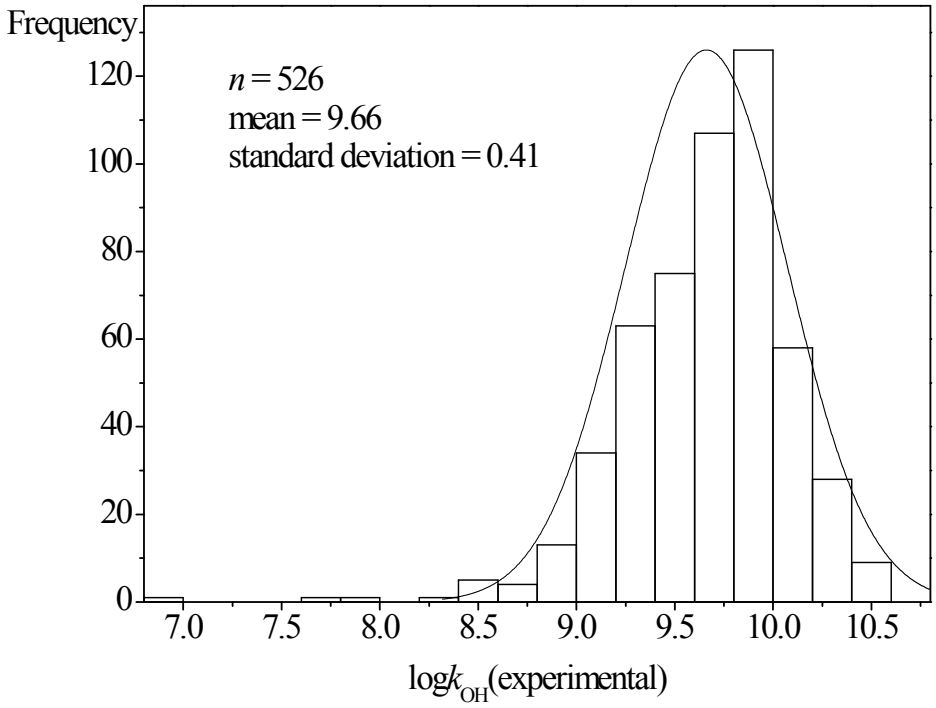

Fig. S1 Distribution of experimental logk_{OH} values across the total dataset

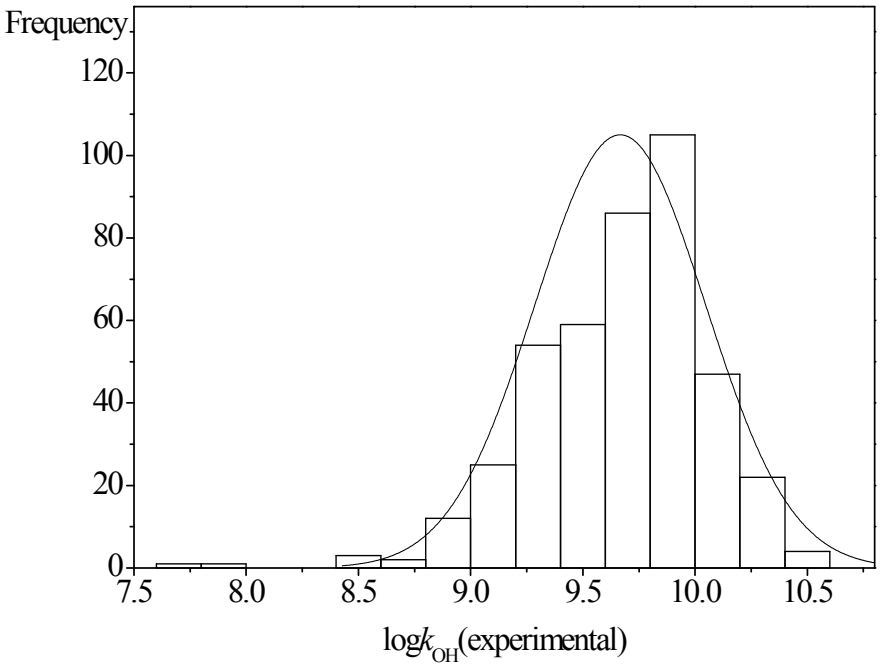

Fig. S2 Distribution of experimental logk_{OH} values of the training set

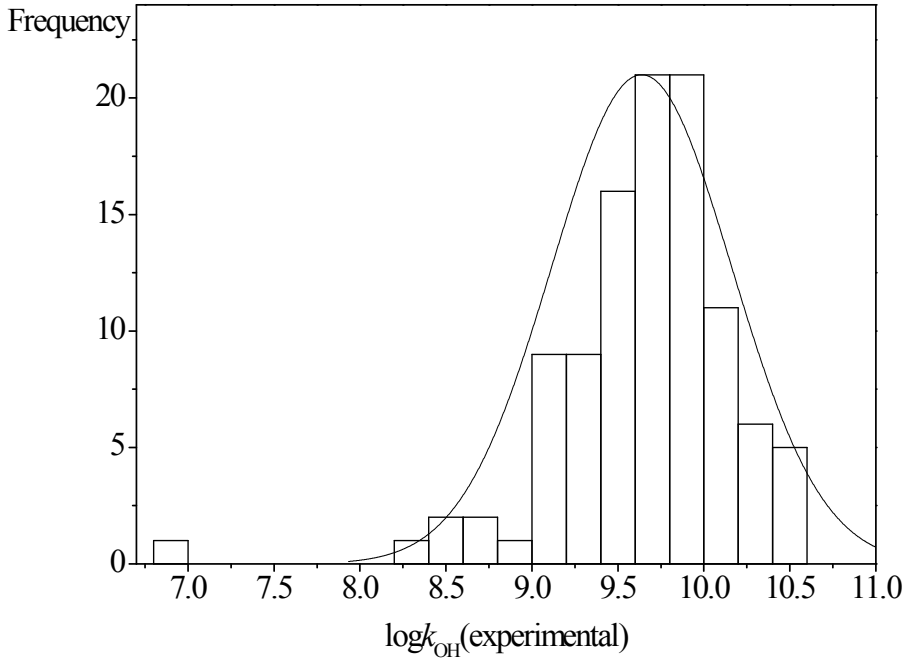

Fig. S3 Distribution of experimental logk_{OH} values of the validation set

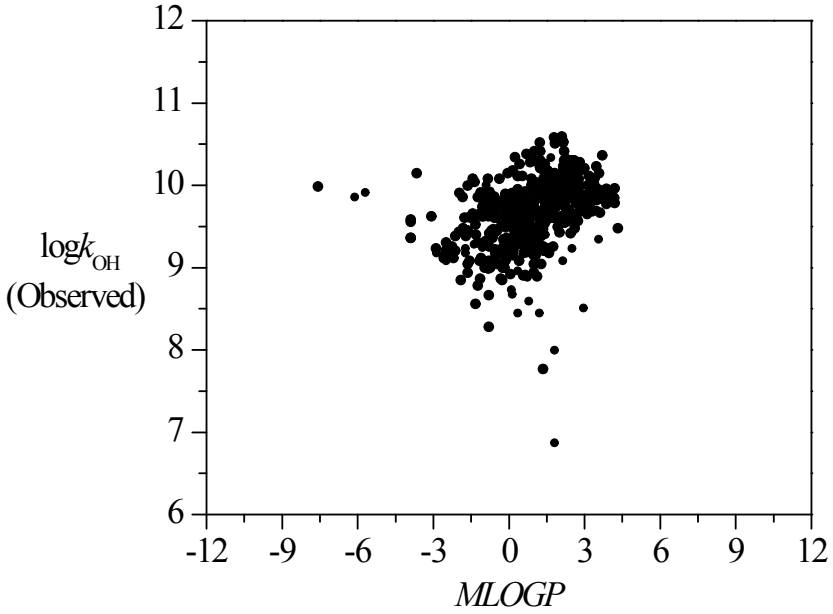

Fig. S4 Plot of the observed logk_{OH} values versus *MLOGP*

Table S1 Experimental and predicted logk_{OH} values of the organic compounds included in this study

ID	CAS No.	Names	Set[a]	Observed	Predicted	Ref[b]
1	000067-63-0	2-Propanol	T	9.28	9.34	1
2	000071-36-3	1-Butanol	T	9.62	9.53	1
3	000078-93-3	methyl ethyl ketone	T	8.91	9.17	1

ID	CAS No.	Names	Set[a]	Observed	Predicted	Ref[b]
4	000078-98-8	Methylglyoxal	R	8.66	9.04	1
5	000123-86-4	n-butylacetate	T	9.26	9.21	1
6	000637-92-3	Ethyl-ter-butylether	R	9.18	9.45	1
7	036133-35-4	2-methoxy-2-methyl propionaldehyde	T	9.48	9.37	2
8	000072-14-0	sulfathiazole	T	9.85	9.88	3
9	000127-69-5	sulfisoxazole	T	9.82	9.66	3
10	000144-82-1	sulfamethizole	T	9.69	9.69	3
11	000723-46-6	sulfamethoxazole	T	9.76	9.69	3
12	000057-68-1	sulfamethazine	T	9.70	9.76	4
13	000068-35-9	sulfadiazine	T	9.57	9.77	4
14	000080-32-0	sulfachloropyridazine	T	9.64	9.79	4
15	000122-11-2	sulfadimethoxine	T	9.79	9.84	4
16	000127-79-7	sulfamerazine	T	9.58	9.77	4
17	000298-12-4	glyoxylic acid	T	8.56	8.47	5
18	000598-42-5	Glycolamide	R	9.04	8.81	6
19	002043-43-8	2-Hydroxypropionamide	R	9.11	9.04	6
20	000067-63-0	Isopropyl alcohol	T	9.28	9.34	7
21	000616-45-5	2-Pyrrolidinone	T	9.34	9.45	8
22	000625-69-4	2,4-Pentanediol	T	9.36	9.37	9
23	026171-83-5	1,2-Butanediol	T	9.36	9.47	9
24	000094-26-8	n-butylparaben	T	9.70	9.93	10
25	000084-66-2	Diethyl phthalate	T	9.60	9.52	11
26	000084-74-2	Dibutyl phthalate	T	9.67	9.73	11
27	000131-11-3	Dimethyl phthalate	T	9.43	9.52	11
28	000131-16-8	Dipropyl phthalate	T	9.65	9.60	11
29	000060-35-5	Acetamide	R	8.28	8.74	12
30	000075-07-0	Acetaldehyde	T	8.86	9.03	12
31	000103-84-4	Acetanilide	T	9.72	9.91	12
32	000106-48-9	4-Chlorophenol	T	9.88	10.05	12
33	000108-43-0	3-Chlorophenol	T	9.86	9.99	12
34	000109-09-1	2-Chloropyridine	T	9.26	9.47	12
35	000626-61-9	4-Chloropyridine	T	9.49	9.41	12
36	001003-67-4	4-Methylpyridine-*N*-oxide	R	9.45	9.63	12
37	000050-00-0	Formaldehyde	T	9.00	8.82	12
38	000051-20-7	5-Bromouracil	T	9.64	9.37	12
39	000051-43-4	Adrenaline	T	10.34	10.03	12

ID	CAS No.	Names	Set[a]	Observed	Predicted	Ref[b]
40	000054-25-1	6-Azauridine	T	9.86	9.47	12
41	000055-21-0	Benzamide	T	9.53	9.58	12
42	000056-65-5	Adenosine triphosphate	T	9.95	9.73	12
43	000056-73-5	glucose-6-phosphate	T	9.15	9.31	12
44	000057-48-7	D-Fructose	T	9.20	9.40	12
45	000057-50-1	Sucrose	T	9.36	9.40	12
46	000058-27-5	2-methyl-1,4-naphthoquinone	T	9.74	9.54	12
47	000058-55-9	Theophylline	T	9.80	9.71	12
48	000058-61-7	Adenosine	R	9.76	9.60	12
49	000058-96-8	Uridine	T	9.61	9.63	12
50	000058-97-9	Uridine 5’-monophosphate	T	9.65	9.59	12
51	000059-23-4	D-galactose	T	9.30	9.35	12
52	000059-56-3	glucose-1-phosphate	R	9.15	9.22	12
53	000060-12-8	Phenethyl alcohol	T	9.81	9.99	12
54	000060-24-2	2-Mercaptoethanol	T	9.83	9.77	12
55	000060-29-7	Diethyl ether	T	9.56	9.49	12
56	000060-56-0	2-mercapto-1-methylimidazole	T	10.15	10.08	12
57	000061-19-8	Adenosine 5'-monophosphate	T	9.61	9.49	12
58	000062-53-3	Aniline	T	10.15	10.20	12
59	000062-56-6	Thiourea	T	9.59	9.71	12
60	000063-74-1	Sulfanilamide	R	9.28	9.44	12
61	000063-91-2	Phenylalanine	R	9.81	9.71	12
62	000065-23-6	Pyridoxine	T	9.80	9.59	12
63	000066-22-8	Uracil	T	9.76	9.49	12
64	000066-71-7	1,10-Phenanthroline	R	9.92	9.74	12
65	000067-47-0	5-Hydroxymethylfurfural	T	9.76	9.57	12
66	000067-56-1	Methanol	T	8.99	9.06	12
67	000068-12-2	*N*,*N*-Dimethylformamide	T	9.23	9.41	12
68	000069-79-4	maltose	T	9.36	9.40	12
69	000069-89-6	Xanthine	T	9.72	9.61	12
70	000069-93-2	Uric acid	T	9.86	9.80	12
71	000070-18-8	glutathione	T	10.15	9.82	12
72	000070-29-1	Diethyl sulfoxide	T	9.81	9.79	12
73	000070-51-9	desferrioxamine B	R	10.11	10.27	12
74	000071-41-0	1-Pentanol	T	9.59	9.64	12
75	000071-43-2	Benzene	T	9.89	9.89	12

ID	CAS No.	Names	Set[a]	Observed	Predicted	Ref[b]
76	000073-24-5	Adenine	R	9.79	9.63	12
77	000074-84-0	Ethane	T	9.26	9.30	12
78	000074-85-1	Ethylene	T	9.64	9.79	12
79	000074-86-2	Acetylene	R	9.67	9.31	12
80	000074-98-6	Propane	T	9.56	9.49	12
81	000075-04-7	Ethylamine	T	9.81	9.69	12
82	000075-09-2	Dichloromethane	T	7.76	7.95	12
83	000075-18-3	Dimethyl sulfide	T	10.28	9.85	12
84	000075-28-5	Isobutane	T	9.66	9.49	12
85	000075-31-0	Isopropylamine	T	10.11	9.75	12
86	000075-50-3	Trimethylamine	R	10.11	9.79	12
87	000075-84-3	2,2-Dimethyl-1-propanol	T	9.60	9.47	12
88	000075-85-4	*tert*-Amyl alcohol	T	9.28	9.38	12
89	000077-86-1	2-amino-2-(hydroxymethyl)propane-1,3-diol	R	9.18	9.58	12
90	000078-40-0	Triethyl phosphate	T	9.46	9.17	12
91	000078-78-4	2-Methylbutane	R	9.72	9.59	12
92	000078-83-1	2-Methyl-1-propanol	T	9.52	9.47	12
93	000078-92-2	2-Butanol	R	9.49	9.47	12
94	000079-01-6	Trichloroethylene	T	9.62	9.57	12
95	000079-05-0	Propionamide	T	8.85	8.92	12
96	000079-14-1	Glycolic acid	T	8.78	8.81	12
97	000079-16-3	*n*-Methylacetamide	T	9.20	9.34	12
98	000083-34-1	3-methylindole	R	10.52	10.20	12
99	000083-67-0	Theobromine	T	9.76	9.73	12
100	000083-88-5	Riboflavin	T	10.08	9.78	12
101	000084-53-7	Uridine 3’-monophosphate	T	9.83	9.55	12
102	000085-61-0	Coenzyme A	R	9.49	9.84	12
103	000087-69-4	Tartaric acid	T	8.85	9.00	12
104	000087-85-4	Hexamethylbenzene	T	9.86	9.83	12
105	000087-89-8	Inositol	T	9.23	9.36	12
106	000088-12-0	N-vinyl-2-pyrrolidinone	T	9.86	9.80	12
107	000090-05-1	2-Methoxyphenol	R	10.30	10.10	12
108	000090-76-6	2-amino-2-deoxy-D-galactose	T	9.08	9.48	12
109	000091-10-1	2,6-Dimethoxyphenol	T	10.41	10.09	12
110	000091-16-7	1,2-Dimethoxybenzene	T	9.72	10.07	12

ID	CAS No.	Names	Set[a]	Observed	Predicted	Ref[b]
111	000091-55-4	2,3-dimethylindole	T	10.11	10.19	12
112	000092-52-4	Biphenyl	T	9.95	10.17	12
113	000092-62-6	3,6-diaminoacridine	T	10.00	10.22	12
114	000093-54-9	1-Phenyl-1-propanol	T	10.00	9.98	12
115	000095-20-5	2-methylindole	T	10.53	10.19	12
116	000095-47-6	*o*-Xylene	T	9.83	9.91	12
117	000095-48-7	*o*-Cresol	T	10.04	10.04	12
118	000095-57-8	2-Chlorophenol	T	10.08	10.06	12
119	000095-63-6	1,2,4-Trimethylbenzene	T	9.79	9.91	12
120	000095-93-2	1,2,4,5-Tetramethylbenzene	T	9.85	9.88	12
121	000096-22-0	3-Pentanone	T	9.15	9.32	12
122	000096-31-1	1,3-Dimethylurea	T	9.41	9.50	12
123	000096-37-7	Methylcyclopentane	T	9.85	9.69	12
124	000097-30-3	α-methylglucoside	T	9.38	9.35	12
125	000097-54-1	2-Methoxy-4-propenylphenol	R	10.59	10.19	12
126	000098-01-1	Furaldehyde	R	9.89	9.53	12
127	000098-10-2	Benzenesulfonamide	R	9.45	9.12	12
128	000098-54-4	4-tert-Butylphenol	T	10.28	10.05	12
129	000098-82-8	Cumene	T	9.88	9.96	12
130	000098-83-9	α-methylstyrene	T	9.99	10.09	12
131	000098-86-2	Acetophenone	T	9.77	9.57	12
132	000098-92-0	Nicotinamide	T	9.18	9.23	12
133	000098-95-3	Nitrobenzene	T	9.59	9.35	12
134	000099-45-6	Adrenalone	T	10.26	9.94	12
135	000100-02-7	4-Nitrophenol	T	9.58	9.62	12
136	000100-19-6	*p*-Nitroacetophenone	T	9.49	9.24	12
137	000100-22-1	N,N,N’,N’-tetramethyl-p-phenylenediamine	T	10.00	10.26	12
138	000100-41-4	Ethylbenzene	T	9.88	9.99	12
139	000100-42-5	Styrene	T	9.78	10.14	12
140	000100-47-0	Benzonitrile	R	9.64	9.65	12
141	000100-51-6	Benzyl alcohol	T	9.92	9.84	12
142	000100-52-7	Benzaldehyde	T	9.64	9.65	12
143	000100-65-2	*n*-Phenylhydroxylamine	R	10.18	9.96	12
144	000100-66-3	Anisole	T	9.73	9.96	12
145	000100-86-7	2-methyl-1-phenyl-2-propanol	T	10.23	9.93	12

ID	CAS No.	Names	Set[a]	Observed	Predicted	Ref[b]
146	000102-71-6	Nitrilotriethanol	T	9.90	9.75	12
147	000102-82-9	tributylamine	R	10.23	10.29	12
148	000103-05-9	2-Methyl-4-phenyl-2-butanol	T	9.77	10.07	12
149	000103-90-2	Acetaminophen	T	9.99	10.00	12
150	000105-37-3	Ethyl propionate	T	8.94	9.17	12
151	000105-54-4	Ethyl butyrate	T	9.20	9.26	12
152	000106-42-3	*p*-Xylene	T	9.85	9.94	12
153	000106-44-5	*p*-Cresol	T	10.08	10.05	12
154	000106-51-4	1,4-Benzoquinone	T	9.08	9.28	12
155	000106-57-0	glycine anhydride	T	9.08	9.34	12
156	000106-97-8	(n-)Butane	T	9.66	9.62	12
157	000106-98-9	1-Butene	T	9.85	9.85	12
158	000107-10-8	Propylamine	T	9.82	9.79	12
159	000107-15-3	Ethylenediamine	T	9.74	9.72	12
160	000107-18-6	Allyl alcohol	T	9.78	9.54	12
161	000107-87-9	2-Pentanone	T	9.28	9.35	12
162	000107-88-0	1,3-Butanediol	T	9.34	9.40	12
163	000107-93-7	Crotonic acid	T	9.46	9.21	12
164	000108-38-3	*m*-Xylene	R	9.88	9.94	12
165	000108-46-3	Resorcinol	T	10.08	10.02	12
166	000108-47-4	2,4-Dimethylpyridine	T	9.49	9.46	12
167	000108-48-5	2,6-Dimethylpyridine	T	9.48	9.53	12
168	000108-67-8	Mesitylene	T	9.81	9.89	12
169	000108-68-9	3,5-Dimethylphenol	T	10.30	9.95	12
170	000108-87-2	Methylcyclohexane	T	9.85	9.88	12
171	000108-88-3	Toluene	T	9.48	9.92	12
172	000108-95-2	Phenol	T	9.82	10.07	12
173	000108-96-3	4-Pyridone	T	9.72	9.71	12
174	000108-99-6	3-Methylpyridine	T	9.38	9.51	12
175	000109-00-2	3-Pyridinol	T	9.79	9.73	12
176	000109-04-6	2-Bromopyridine	T	9.38	9.25	12
177	000109-06-8	2-Methylpyridine	T	9.40	9.50	12
178	000109-60-4	*n*-Propyl acetate	T	9.15	9.13	12
179	000109-66-0	Pentane	T	9.73	9.72	12
180	000109-86-4	2-Methoxyethanol	T	9.11	9.51	12
181	000109-87-5	Dimethoxymethane	T	9.08	9.19	12

ID	CAS No.	Names	Set[a]	Observed	Predicted	Ref[b]
182	000109-97-7	Pyrrole	T	10.18	10.00	12
183	000109-99-9	Tetrahydrofuran	R	9.60	9.48	12
184	000110-00-9	Furan	R	9.59	9.66	12
185	000110-01-0	Tetrahydrothiophene	R	10.15	9.99	12
186	000110-02-1	Thiophene	T	9.91	9.92	12
187	000110-54-3	Hexane	T	9.82	9.82	12
188	000110-58-7	*n*-Amylamine	T	9.92	9.97	12
189	000110-71-4	Ethylene glycol dimethyl ether	T	9.20	9.43	12
190	000110-80-5	2-Ethoxyethanol	T	9.23	9.52	12
191	000110-81-6	Diethyl disulfide	T	10.15	9.89	12
192	000110-81-6	Diethyl disulfide	T	10.15	9.89	12
193	000110-82-7	Cyclohexane	T	9.79	9.82	12
194	000110-83-8	Cyclohexene	T	9.94	9.92	12
195	000110-86-1	Pyridine	T	9.49	9.49	12
196	000110-88-3	1,3,5-Trioxane	T	9.18	9.22	12
197	000110-94-1	glutaric acid	T	8.92	9.12	12
198	000111-16-0	Pimelic acid	T	9.54	9.32	12
199	000111-20-6	Sebacic acid	T	9.81	9.71	12
200	000111-26-2	Hexylamine	T	10.11	10.05	12
201	000111-27-3	1-Hexanol	R	9.85	9.74	12
202	000111-28-4	2,4-Hexadien-1-ol	T	9.99	9.91	12
203	000111-46-6	Diethylene glycol	R	9.32	9.49	12
204	000111-65-9	Octane	T	9.96	9.99	12
205	000111-70-6	1-Heptanol	R	9.87	9.83	12
206	000111-86-4	*n*-Octylamine	T	10.11	10.25	12
207	000111-87-5	1-Octanol	T	9.89	9.99	12
208	000111-92-2	Dibutylamine	T	10.26	10.14	12
209	000112-36-7	Diethyleneglycol diethyl ether	T	9.51	9.46	12
210	000115-07-1	Propylene	T	9.85	9.76	12
211	000115-10-6	Dimethyl ether	R	9.00	9.34	12
212	000115-11-7	Isobutylene	T	9.73	9.73	12
213	000115-77-5	Pentaerythritol	T	9.52	9.40	12
214	000119-61-9	Benzophenone	T	9.94	9.84	12
215	000120-72-9	indole	T	10.51	10.26	12
216	000121-69-7	*N*,*N*-Dimethylaniline	T	10.15	10.22	12
217	000121-79-9	Propyl 3,4,5-trihydroxybenzoate	T	10.04	9.93	12

ID	CAS No.	Names	Set[a]	Observed	Predicted	Ref[b]
218	000122-39-4	Diphenylamine	R	10.08	10.54	12
219	000123-25-1	Diethyl succinate	R	8.89	9.18	12
220	000123-31-9	Hydroquinone	T	10.00	10.11	12
221	000123-33-1	maleic hydrazide	T	9.51	9.63	12
222	000123-39-7	*n*-Methylformamide	R	9.08	9.24	12
223	000123-51-3	3-Methyl-1-butanol	T	9.58	9.51	12
224	000123-63-7	2,4,6-Trimethyl-1,3,5-trioxane	T	9.34	9.31	12
225	000123-91-1	1,4-dioxane	T	9.45	9.47	12
226	000123-93-3	Thiodiacetic acid	T	9.78	9.76	12
227	000123-99-9	Azelaic acid	T	9.73	9.58	12
228	000124-04-9	Adipic acid	R	9.30	9.24	12
229	000127-18-4	Tetrachloroethylene	T	9.41	9.47	12
230	000127-19-5	*N*,*N*-Dimethylacetamide	T	9.54	9.48	12
231	000135-77-3	1,2,4-Trimethoxybenzene	T	9.91	10.23	12
232	000138-89-6	N,N-Dimethyl-4-nitrosoaniline	T	10.10	9.95	12
233	000139-85-5	3,4-Dihydroxybenzaldehyde	T	9.92	9.97	12
234	000142-08-5	2-Pyridone	T	9.81	9.76	12
235	000142-29-0	Cyclopentene	T	9.85	9.81	12
236	000142-30-3	2,5-dimethyl-3-hexyne-2,5-diol	T	9.52	9.32	12
237	000142-82-5	Heptane	T	9.89	9.90	12
238	000144-80-9	Sulfacetamide	T	9.74	9.66	12
239	000147-81-9	Arabinose	T	9.26	9.37	12
240	000150-39-0	2-hydroxyethylethylenediaminetriacetic acid	T	10.04	9.85	12
241	000150-76-5	4-Methoxyphenol	T	10.41	10.12	12
242	000150-78-7	1,4-Dimethoxybenzene	T	9.85	10.08	12
243	000151-10-0	1,3-Dimethoxybenzene	T	9.86	9.99	12
244	000154-17-6	2-deoxy-D-glucose	T	9.45	9.46	12
245	000156-60-5	*trans*-1,2-Dichloroethylene	R	9.79	9.72	12
246	000287-92-3	Cyclopentane	R	9.57	9.66	12
247	000288-32-4	imidazole	R	10.08	9.67	12
248	000291-64-5	Cycloheptane	T	9.89	9.82	12
249	000298-81-7	8-methoxypsoralen	T	10.04	9.98	12
250	000314-50-1	Orotidine	T	9.60	9.49	12
251	000363-72-4	Pentafluorobenzene	T	9.85	9.66	12

ID	CAS No.	Names	Set[a]	Observed	Predicted	Ref[b]
252	000366-18-7	2,2'-Bipyridine	T	9.79	9.58	12
253	000367-11-3	*o*-Difluorobenzene	T	9.88	9.72	12
254	000392-56-3	hexafluorobenzene	T	9.48	9.49	12
255	000443-72-1	N6-methyladenine	T	9.73	9.72	12
256	000452-86-8	3,4-Dihydroxytoluene	T	10.20	10.05	12
257	000461-89-2	6-Azauracil	R	9.65	9.31	12
258	000462-06-6	Fluorobenzene	T	10.00	9.75	12
259	000462-95-3	Diethoxymethane	T	9.20	9.36	12
260	000488-23-3	1,2,3,4-Tetramethylbenzene	T	9.86	9.89	12
261	000490-78-8	2,5-Dihydroxyacetophenone	T	9.90	9.91	12
262	000492-97-7	2,2'-Bithiophene	T	10.20	10.15	12
263	000496-15-1	indoline	R	10.58	10.18	12
264	000497-26-7	2-Methyl-1,3-dioxolane	T	9.54	9.24	12
265	000500-99-2	3,5-Dimethoxyphenol	T	10.30	9.96	12
266	000502-41-0	cycloheptanol	R	9.23	9.68	12
267	000503-74-2	3-Methylbutyric acid	T	9.04	9.19	12
268	000504-07-4	5,6-dihydrouracil	T	9.08	9.34	12
269	000504-24-5	4-Aminopyridine	T	9.70	9.77	12
270	000504-29-0	2-Aminopyridine	T	9.92	9.85	12
271	000505-29-3	1,4-Dithiane	R	10.26	10.01	12
272	000505-48-6	Suberic acid	T	9.68	9.46	12
273	000513-85-9	2,3-Butanediol	T	9.11	9.36	12
274	000521-31-3	luminol	R	9.94	9.88	12
275	000526-73-8	1,2,3-Trimethylbenzene	T	9.85	9.91	12
276	000526-75-0	2,3-Dimethylphenol	T	10.30	9.98	12
277	000527-53-7	1,2,3,5-Tetramethylbenzene	T	9.85	9.88	12
278	000528-50-7	D-cellobiose	R	9.56	9.30	12
279	000529-33-9	1,2,3,4-Tetrahydro-1-naphthol	T	9.85	10.09	12
280	000533-67-5	2-Deoxy-D-ribose	T	9.40	9.43	12
281	000534-15-6	1,1-Dimethoxyethane	T	9.34	9.38	12
282	000538-86-3	Benzyl methyl ether	T	10.00	9.86	12
283	000540-36-3	*p*-Difluorobenzene	T	10.00	9.72	12
284	000540-51-2	2-Bromoethanol	T	8.90	9.09	12
285	000540-84-1	2,2,4-Trimethylpentane	T	9.79	9.63	12
286	000544-25-2	Cycloheptatriene	T	10.00	10.01	12
287	000551-62-2	1,2,3,4-Tetrafluorobenzene	T	9.90	9.68	12

ID	CAS No.	Names	Set[a]	Observed	Predicted	Ref[b]
288	000553-26-4	4,4'-Bipyridine	T	9.72	9.45	12
289	000563-83-7	Isobutyramide	T	9.20	9.03	12
290	000565-70-8	2-Hydroxybutyric acid	T	9.00	9.19	12
291	000576-26-1	2,6-Dimethylphenol	R	10.41	10.00	12
292	000583-08-4	Nicotinylglycine	R	9.04	9.35	12
293	000584-02-1	3-Pentanol	T	9.49	9.55	12
294	000590-90-9	3-Hydroxy-2-butanone	T	9.00	9.29	12
295	000591-50-4	Iodobenzene	R	9.70	9.95	12
296	000592-57-4	1,3-Cyclohexadiene	T	10.00	10.00	12
297	000598-41-4	Glycinamide	R	9.45	9.01	12
298	000598-50-5	Methylurea	T	9.30	9.29	12
299	000603-76-9	1-methylindole	T	10.18	10.16	12
300	000609-38-1	2-furancarboxamide	T	9.74	9.58	12
301	000611-69-8	2-methyl-1-phenyl-1-propanol	T	9.98	9.94	12
302	000614-96-0	5-methylindole	T	10.23	10.17	12
303	000617-78-7	3-Ethylpentane	T	9.77	9.67	12
304	000617-94-7	2-Phenyl-2-propanol	T	9.66	9.87	12
305	000620-02-0	5-Methylfurfural	T	9.86	9.53	12
306	000621-23-8	1,3,5-Trimethoxybenzene	T	9.91	10.05	12
307	000623-42-7	Methyl butyrate	R	9.23	9.23	12
308	000626-55-1	3-Bromopyridine	T	9.04	9.26	12
309	000628-41-1	1,4-Cyclohexadiene	T	9.89	9.90	12
310	000629-11-8	1,6-Hexanediol	T	9.67	9.71	12
311	000630-10-4	Selenourea	T	10.08	9.84	12
312	000632-22-4	1,1,3,3-Tetramethylurea	T	9.90	9.70	12
313	000634-36-6	1,2,3-Trimethoxybenzene	R	9.90	10.10	12
314	000636-38-4	2,3-dihydroxy-2-propenal	T	10.00	9.83	12
315	000638-02-8	2,5-Dimethylthiophene	T	9.86	9.85	12
316	000646-06-0	1,3-dioxolane	R	9.60	9.23	12
317	000653-63-4	2'-deoxyadenosine-5'-monophosphate	R	9.54	9.66	12
318	000694-59-7	Pyridine-*N*-oxide	T	9.48	9.64	12
319	000696-04-8	5,6-dihydrothymine	T	9.00	9.34	12
320	000700-00-5	9-methyladenine	T	9.76	9.66	12
321	000700-12-9	Pentamethylbenzene	T	9.88	9.87	12
322	000754-10-9	Trimethylacetamide	R	9.18	9.05	12
323	000875-30-9	1,3-dimethylindole	T	10.04	10.16	12

ID	CAS No.	Names	Set[a]	Observed	Predicted	Ref[b]
324	000875-79-6	1,2-dimethylindole	T	10.00	10.13	12
325	000922-65-6	1,4-Pentadien-3-ol	T	10.00	9.88	12
326	000931-85-1	6-Azacytosine	T	9.65	9.48	12
327	000931-86-2	5-Azacytosine	T	9.32	9.42	12
328	000932-53-6	6-Azathymine	T	9.45	9.42	12
329	000938-55-6	*N*,*N*-Dimethyladenine	T	9.85	9.78	12
330	000945-51-7	Diphenyl sulfoxide	R	10.00	10.23	12
331	000958-09-8	2'-deoxyadenosine	R	9.66	9.55	12
332	001003-73-2	3-Methylpyridine-*N*-oxide	T	9.62	9.58	12
333	001006-94-6	5-methoxyindole	T	10.18	10.16	12
334	001022-31-7	N-formylkynurenine	T	9.90	9.71	12
335	001074-89-1	6-Methoxypurine	T	9.30	9.48	12
336	001123-65-5	2,4,6-Trimethyl-3-hydroxypyridine	T	9.40	9.70	12
337	001192-62-7	2-Acetylfuran	T	9.65	9.48	12
338	001193-24-4	4,6-Dihydroxypyrimidine	T	9.76	9.45	12
339	001193-82-4	Methyl phenyl sulfoxide	T	9.99	10.02	12
340	001194-02-1	p-fluorobenzonitrile	R	9.54	9.63	12
341	001197-09-7	3,4-Dihydroxyacetophenone	T	10.00	9.85	12
342	001464-44-4	Phenyl-β-D-glucopyranoside	T	9.71	9.80	12
343	001517-69-7	1-Phenylethanol	R	10.04	9.86	12
344	001518-72-5	Di-*tert*-butyl disulfide	T	9.81	10.03	12
345	001600-44-8	Tetramethylene sulfoxide	T	9.85	9.86	12
346	001611-35-4	Xylenol orange	T	10.38	10.20	12
347	001820-81-1	5-chlorouracil	R	9.72	9.64	12
348	001824-94-8	Methylgalactoside	R	9.20	9.38	12
349	001867-73-8	N6-methyladenosine/6-methylaminopurine 9-ribofuranoside	R	9.78	9.70	12
350	001899-24-7	5-bromofurfural	T	9.59	9.34	12
351	001953-54-4	5-hydroxyindole	T	10.23	10.23	12
352	001986-81-8	Nicotinamide-N-oxide	T	9.32	9.24	12
353	002168-93-6	Dibutyl sulfoxide	T	9.90	10.12	12
354	002211-89-4	Diisopropyl sulfoxide	R	9.83	9.77	12
355	002211-92-9	Di(*tert*-butyl) sulfoxide	T	9.72	9.74	12
356	002224-15-9	Ethylene glycol diglycidyl ether	R	9.36	9.45	12
357	002308-57-8	2-hydroxypurine	R	9.70	9.62	12
358	002344-70-9	1-Phenyl-3-butanol	R	10.30	10.07	12

ID	CAS No.	Names	Set[a]	Observed	Predicted	Ref[b]
359	002434-49-3	dihydro-6-methyluracil	T	9.04	9.35	12
360	002492-87-7	p-nitrophenyl-β-D-glucopyranoside	T	9.46	9.66	12
361	002568-30-1	2-(Chloromethyl)-1,3-dioxolane	T	9.54	9.27	12
362	002620-62-4	N6,N6-dimethyladenosine	T	9.81	9.77	12
363	002675-88-9	*n*-Methylisobutyramide	T	9.28	9.49	12
364	002687-91-4	1-ethyl-2-pyrrolidinone	R	9.43	9.55	12
365	002816-24-2	o-nitrophenyl-β-D-glucopyranoside	T	9.62	9.68	12
366	002916-31-6	2,2-dimethyl-1,3-dioxolane	T	9.32	9.25	12
367	003085-45-8	2,2'-Sulfinyldiethanol	T	9.72	9.87	12
368	003268-49-3	Methional	R	9.91	9.76	12
369	003317-61-1	5,5-dimethyl-1-pyrroline-1-oxyl	T	9.63	9.53	12
370	003376-24-7	Phenyl-*N*-tert-butylnitrone	T	9.86	9.85	12
371	003585-88-4	4-nitrophenyl-N-tert-butylnitrone	T	9.87	9.63	12
372	003835-64-1	2,2-Dimethyl-1-phenyl-1-propanol	T	10.00	9.94	12
373	003867-83-2	Methyl α-D-arabinopyranoside	T	9.38	9.40	12
374	003914-34-9	isouramil	T	9.70	9.77	12
375	004151-19-3	Ribose-5-phosphate	T	9.11	8.95	12
376	004170-69-8	Isopropyl phenyl sulfoxide	T	10.00	10.02	12
377	004170-80-3	Ethyl phenyl sulfoxide	R	9.92	10.02	12
378	004253-89-8	Diisopropyl disulfide	R	10.30	9.87	12
379	004253-91-2	Dipropyl sulfoxide	T	9.80	9.93	12
380	004567-18-4	2,5,5-trimethyl-1-pyrroline N-oxide	T	9.68	9.59	12
381	004756-30-3	p-chlorophenyl-β-D-glucopyranoside	T	9.70	9.86	12
382	005076-82-4	Sarcosine anhydride	T	9.41	9.56	12
383	005150-42-5	2,3-Dimethoxyphenol	T	10.30	9.99	12
384	005192-03-0	5-aminoindole	R	10.52	10.30	12
385	005327-44-6	3,5-Dinitroanisole	T	9.60	9.45	12
386	005340-95-4	melibiose	T	9.58	9.36	12
387	005391-88-8	1-(*p*-Bromophenyl)ethanol	T	9.79	9.66	12
388	005625-46-7	Alanine anhydride	T	9.26	9.36	12
389	006032-32-2	p-methoxyphenyl-β-D-glucopyranoside	T	10.00	10.01	12
390	006099-08-7	Glycyltyrosylglycine	T	9.86	10.07	12
391	006146-52-7	5-nitroindole	T	10.00	9.78	12
392	006359-38-2	benzoflavine	T	10.15	10.31	12
393	006556-12-3	glucuronic acid	T	9.11	9.24	12

ID	CAS No.	Names	Set[a]	Observed	Predicted	Ref[b]
394	006830-83-7	*n*-Methylpivalamide	T	9.38	9.49	12
395	006915-15-7	Malic acid	T	8.86	9.07	12
396	006976-37-0	2,2-Bis(hydroxymethyl)-2,2',2'-nitrilotriethanol	T	9.48	9.71	12
397	007600-08-0	2,3-dihydro-5-hydroxy-1,4-phthalazinedione	T	10.08	9.98	12
398	010075-50-0	5-Bromoindole	R	10.20	9.91	12
399	013803-39-9	5-phenylfurfural	R	9.77	9.94	12
400	014007-10-4	homocysteine thiolactone	R	9.46	9.53	12
401	014215-68-0	2-acetamido-2-deoxy-D-galactopyranose	R	9.20	9.40	12
402	014898-87-4	1-Phenyl-2-propanol	T	10.32	10.03	12
403	015861-24-2	5-cyanoindole	R	10.04	10.07	12
404	016484-86-9	Ethylene glycol diethyl ether	T	9.36	9.43	12
405	017113-33-6	2-phenylfuran	T	10.20	10.07	12
406	017422-32-1	5-chloroindole	T	10.30	10.13	12
407	017860-87-6	2-o-methyl-L-ascorbic acid	T	9.43	9.52	12
408	020710-99-0	Selenodicysteine	T	9.91	9.95	12
409	020762-30-5	Adenosine 5'-diphosphoribose	T	9.62	9.59	12
410	020838-44-2	m-nitrophenyl-β-D-glucopyranoside	R	9.60	9.65	12
411	024331-71-3	*N*,*N*-Dimethylpivalamide	T	9.59	9.55	12
412	025339-57-5	Butadiene	R	9.85	10.10	12
413	027025-41-8	glutathione,oxidised	T	9.98	9.81	12
414	027440-00-2	N-acetylglycylglycinamide	T	8.93	9.28	12
415	027876-94-4	Crocetin	T	10.36	10.06	12
416	028128-19-0	2-mercaptopurine	R	9.64	9.96	12
417	029883-15-6	D-amygdalin	T	9.62	9.69	12
418	032449-92-6	D-glucuronolactone	T	9.23	9.30	12
419	033577-16-1	Methyl methylthiomethyl sulfoxide	R	9.68	9.92	12
420	033967-18-9	1-(*p*-Ethylphenyl)ethanol	T	10.11	10.02	12
421	040085-12-9	4-*tert*-Butyl-1,2-dihydrobenzene	T	9.88	10.06	12
422	052067-39-7	1-Methoxy-2-methyl-1-phenylpropane	T	9.87	9.83	12
423	063447-38-1	4,6-Dihydroxy-5-methylpyrimidine	T	9.62	9.52	12
424	066893-81-0	α-(4-pyridyl 1-oxide)-N-tert-butylnitrone	R	9.58	9.68	12

ID	CAS No.	Names	Set[a]	Observed	Predicted	Ref[b]
425	069902-43-8	6-Hydroxy-1,4-dimethylcarbazole	T	10.18	10.29	12
426	071490-72-7	2,3-dihydro-5-methyl-1,4-phthalazinedione	R	9.95	9.83	12
427	865467-09-0	*n*-Methylpropionamide	T	9.15	9.50	12
428	000500-38-9	4,4'-(2,3-dimethyltetramethylene)dicatechol	T	10.18	10.36	12
429	035790-48-8	glycylglycylglycylglycinamide	T	9.18	9.32	12
430	000365-07-1	Thymidine 5’-monophosphate	R	9.72	9.65	12
431	040117-29-1	4-methylphenyl-N-tert-butylnitrone	T	9.67	9.84	12
432	040117-28-0	4-methoxyphenyl-N-tert-butylnitrone	R	9.81	9.94	12
433	031448-40-5	m-bromophenyl-β-D-glucopyranoside	T	9.57	9.62	12
434	028217-28-9	m-chlorophenyl-β-D-glucopyranoside	T	9.57	9.80	12
435	028217-36-9	p-cyanophenyl-β-D-glucopyranoside	R	9.61	9.78	12
436	031539-55-6	2,4-dimethylphenyl-β-D-glucopyranoside	T	9.59	9.76	12
437	031448-41-6	3,4-dimethylphenyl-β-D-glucopyranoside	R	9.62	9.79	12
438	031098-42-7	2-(3-aminopropylamino)ethanethiol	T	9.92	10.05	12
439	010465-78-8	1,1’-azobis(N,N-dimethylformide)(diamide)	T	9.71	9.52	12
440	000064-17-5	Ethanol	T	9.30	9.28	12
441	000071-23-8	1-Propanol	T	9.44	9.42	12
442	000079-09-4	Propionic acid	T	8.67	8.85	12
443	000056-81-5	1,2,3-Propanetriol	T	9.28	9.39	12
444	000057-55-6	1,2-Propanediol	T	9.23	9.31	12
445	000069-65-8	D-Mannitol	T	9.22	9.34	12
446	000107-21-1	Ethylene glycol	T	9.24	9.30	12
447	000110-63-4	1,4-Butanediol	T	9.53	9.51	12
448	000111-29-5	1,5-Pentanediol	T	9.60	9.61	12
449	000149-32-6	Erythritol	T	9.23	9.26	12
450	000504-63-2	1,3-Propanediol	T	9.41	9.39	12
451	000107-07-3	2-Chloroethanol	T	8.96	9.24	12
452	000115-20-8	2,2,2-Trichloroethanol	T	8.45	8.29	12
453	000872-50-4	N-methylpyrolidone	R	9.48	9.55	12
454	000109-73-9	*n*-Butylamine	T	9.89	9.89	12
455	000075-64-9	*tert*-Butylamine	T	9.78	9.69	12

ID	CAS No.	Names	Set[a]	Observed	Predicted	Ref[b]
456	000067-66-3	Trichloromethane	R	6.87	6.70	12
457	000074-89-5	Methylamine	T	9.61	9.59	12
458	000075-15-0	Carbon disulfide	R	9.90	9.74	12
459	000076-22-2	Camphor	T	9.61	9.44	12
460	000097-97-2	2-Chloro-1,1-dimethoxyethane	T	9.18	9.39	12
461	000107-92-6	*n*-Butyric acid	T	9.34	9.10	12
462	000108-90-7	Chlorobenzene	T	9.74	9.82	12
463	000120-80-9	Catechol	R	10.04	10.11	12
464	000123-72-8	Butyraldehyde	T	9.59	9.35	12
465	001634-04-4	*tert*-Butyl methyl ether	T	9.20	9.45	12
466	027178-34-3	*tert*-Butylphenol	T	10.28	10.06	12
467	000058-08-2	Caffeine	R	9.89	9.72	12
468	042835-25-6	flumequine	T	9.92	9.87	13
469	113617-63-3	orbifloxacin	T	9.84	10.08	13
470	115550-35-1	marbofloxacin	T	9.96	10.18	13
471	000298-04-4	disulfoton	T	10.05	10.19	14
472	000333-41-5	diazinon	T	9.89	9.61	14
473	000994-22-9	dyfonate	T	10.08	10.13	14
474	001610-18-0	prometon	T	9.43	9.48	14
475	005902-51-2	terbacil	R	9.91	9.66	14
476	013071-79-9	terbufos	T	10.01	9.99	14
477	000330-55-2	linuron	R	9.81	9.94	14
478	000359-13-7	2,2-Difluoroethanol	R	8.45	8.11	15
479	000598-38-9	2,2-Dichloroethanol	T	8.59	8.93	15
480	000371-62-0	2-Fluoroethanol	R	8.73	9.07	15
481	000882-09-7	Clofibric Acid	T	9.67	9.78	16
482	015687-27-1	Ibuprofen	T	9.81	9.92	16
483	022204-53-1	Naproxen	T	9.98	10.21	16
484	252880-06-1	2-chloro-4-ethylimino-6-isopropylamino-s-triazine	T	9.23	9.32	17
485	083364-15-2	4-acetamido-2-chloro-6-isopropylamino-s-triazine	T	9.08	9.31	17
486	014277-97-5	Domoic acid	T	9.96	10.02	18
487	000294-93-9	12-crown-4	T	9.86	9.70	19
488	033100-27-5	15-crown-5	T	9.91	9.66	19
489	000123-38-6	Propanal	T	9.40	9.07	20

ID	CAS No.	Names	Set[a]	Observed	Predicted	Ref[b]
490	000066-25-1	n-Hexanal	T	9.40	9.56	21
491	000096-17-3	2-Methylbutanal	T	9.49	9.33	21
492	000112-12-9	2-Undecanone	T	9.82	9.94	21
493	000122-78-1	Phenyl acetaldehyde	T	9.77	9.70	21
494	000590-86-3	3-Methylbutanal	T	9.46	9.34	21
495	002445-69-4	2-Methylbutyl isobutyrate	R	9.56	9.42	21
496	000060-54-8	Tetracycline	T	9.89	9.69	22
497	000154-21-2	Lincomycin	T	9.93	10.04	22
498	000738-70-5	Trimethoprim	T	9.84	9.95	22
499	001404-90-6	Vancomycin	T	9.91	9.80	22
500	021312-10-7	N(4)-acetyl-sulfamethoxazole	T	9.83	9.59	22
501	037517-28-5	Amikacin	R	9.86	9.65	22
502	080214-83-1	Roxithromycin	T	9.73	9.64	22
503	083905-01-5	Azithromycin	T	9.46	9.65	22
504	000061-72-3	Cloxacillin	T	9.80	9.94	23
505	000069-53-4	Ampicillin	T	9.91	10.08	23
506	025953-19-9	Cefazolin	T	9.81	9.91	23
507	061477-96-1	Piperacillin	T	9.89	10.05	23
508	000087-62-7	2,6-dimethyl-aniline	T	10.23	10.16	24
509	000094-75-7	2,4-dichlorophenoxyacetic acid	T	9.82	9.96	25
510	000120-83-2	2,4-dichlorophenol	T	9.85	10.07	25
511	000122-59-8	phenoxyacetic acid	R	10.00	9.91	25
512	000108-10-1	methyl isobutyl ketone	T	9.53	9.28	26
513	000114-07-8	erythromycin	T	9.48	9.54	27
514	000134-62-3	N,N'-diethyl-m-toluamide	T	9.69	9.76	27
515	000389-08-2	nalidixic acid	T	9.83	9.64	27
516	025812-30-0	gemfibrozil	T	10.00	9.99	27
517	066108-95-0	Iohexol	R	9.51	9.72	27
518	093413-69-5	venlafaxine	R	9.93	10.36	27
519	019092-01-4	methylethylnitramine	T	8.88	8.95	28
520	085721-33-1	ciprofloxacin	T	10.33	10.19	29
521	098079-51-7	lomefloxacin	R	9.91	9.98	29
522	000074-95-3	dibromomethane	T	8.00	8.15	30
523	000077-47-4	hexachlorocyclopentadiene	R	9.34	9.18	30
524	000096-12-8	1,2-dibromo-3-chloropropane	R	8.51	8.83	30
525	000541-73-1	m-dichlorobenzene	T	9.70	9.83	30

ID	CAS No.	Names	Set[a]	Observed	Predicted	Ref[b]
526	000626-86-8	adipic acid monoethyl ester	T	9.48	9.42	30

[a] T represents the training set and R represents the validation set.

[b] Ref.:

[1] Monod, A., Poulain, L., Grubert, S., Voisin, D., Wortham, H., 2005. Kinetics of OH-initiated oxidation of oxygenated organic compounds in the aqueous phase: new rate constants, structure-activity relationships and atmospheric implications. Atmos. Environ. 39, 7667-7688.

[2] Acero, J.L., Haderlein, S.B., Schmidt, T.C., Suter, M.J.F., Von Gunten, U., 2001. MTBE oxidation by conventional ozonation and the combination ozone/hydrogen peroxide: Efficiency of the processes and bromate formation. Environ. Sci. Technol. 35, 4252-4259.

[3] Boreen, A.L., Arnold, W.A., McNeill, K., 2004. Photochemical fate of sulfa drugs in the aquatic environment: Sulfa drugs containing five-membered heterocyclic groups. Environ. Sci. Technol. 38, 3933-3940.

[4] Boreen, A.L., Arnold, W.A., McNeill, K., 2005. Triplet-sensitized photodegradation of sulfa drugs containing six-membered heterocyclic groups: identification of an SO2 extrusion photoproduct. Environ. Sci. Technol. 39, 3630-3638.

[5] Ervens, B., Gligorovski, S., Herrmann, H., 2003. Temperature-dependent rate constants for hydroxyl radical reactions with organic compounds in aqueous solutions. Phys. Chem. Chem. Phys. 5, 1811-1824.

[6] Bell, J.A., Grunwald, E., Hayon, E., 1975. Kinetics of deprotonation of organic free radicals in water. Reaction of HOCHC02-, HOCHCONH2, and HOCCH3CONH2 with various bases. J. Am. Chem. Soc. 97, 2995-3000.

[7] Wolfenden, B.S., Willson, R.L., 1982. Radical-cations as reference chromogens in kinetic studies of one-electron transfer reactions: Pulse radiolysis studies of 2,2'-azinobis-(3-ethyl benzthiazoline-6-sulphonate). J. Chem. Soc. Perkin. Trans. II. 7, 805-812.

[8] Davis, J.E., Sangster, D.F., Senogles, E., 1981. Pulse radiolysis of aqueous solutions of N-vinylpyrrolidin-2-one and poly(N-vinylpyrrolidin-2-one). Aust. J. Chem. 34, 1423-1431.

[9] Hoffmann, D., Weigert, B., Barzaghi, P., Herrmann, H., 2009. Reactivity of poly-alcohols towards OH, NO3 and SO4- in aqueous solution. Phys. Chem. Chem. Phys. 11, 9351-9363.

[10] Bledzka, D., Miller, J.S., Ledakowicz, S., 2012. Kinetic studies of n-butylparaben degradation in H2O2/UV system. Ozone-Sci. Eng. 34, 354-358.

[11] Wen, G., Ma, J., Liu, Z.Q., Zhao, L., 2011. Ozonation kinetics for the degradation of phthalate esters in water and the reduction of toxicity in the process of O3/H2O2. J. Hazard. Mater. 195, 371-377.

[12] Buxton, G.V., Greenstock, C.L., Helman, W.P., Ross, A.B., 1988. Critical review of rate constants for reactions of hydrated electrons, hydrogen atoms and hydroxyl radicals (·OH/·O-) in aqueous solution. J. Phys. Chem. Ref. Data. 17, 513-886.

[13] Santoke, H., Song, W.H., Cooper, W.J., Greaves, J., Miller, G.E., 2009. Free-radical-induced oxidative and reductive degradation of fluoroquinolone pharmaceuticals: Kinetic studies and degradation mechanism. J. Phys. Chem. A. 113, 7846-7851.

[14] Shemer, H., Sharpless, C.M., Elovitz, M.S., Linden, K.G., 2006. Relative rate constants of contaminant candidate list pesticides with hydroxyl radicals. Environ. Sci. Technol. 40, 4460-4466.

[15] Morozov, I., Gligorovski, S., Barzaghi, P., Hoffmann, D., Lazarou, Y.G., Vasiliev, E., Herrmann, H., 2008. Hydroxyl radical reactions with halogenated ethanols in aqueous solution: kinetics and thermochemistry. Int. J. Chem. Kinet. 40, 174-188.

[16] Packer, J.L., Werner, J.J., Latch, D.E., McNeill, K., Arnold, W.A., 2003. Photochemical fate of pharmaceuticals in the environment: Naproxen, diclofenac, clofibric acid, and ibuprofen. Aquat. Sci. 65, 342-351.
[17] Acero, J.L., Stemmler, K., Von Gunten, U., 2000. Degradation kinetics of atrazine and its degradation products with ozone and OH radicals: A predictive tool for drinking water treatment. Environ. Sci. Technol. 34, 591-597.
[18] Jones, K.G., Cooper, W.J., Mezyk, S.P., 2009. Bimolecular rate constant determination for the reaction of hydroxyl radicals with domoic and kainic acid in aqueous solution. Environ. Sci. Technol. 43, 6764-6768.
[19] Wan, L.K., Peng, J., Lin, M.Z., Muroya, Y., Katsumura, Y., Fu, H.Y., 2012. Hydroxyl radical, sulfate radical and nitrate radical reactivity towards crown ethers in aqueous solutions. Radiat. Phys. Chem. 81, 524-530.
[20] Chin, M., Wine, P.H., 1994. A temperature-dependent competitive kinetics study of the aqueous-phase reactions of OH radicals with formate, formic acid, acetate, acetic acid, and hydrated formaldehyde. Aquatic and Surface Photochemistry. 85-96.
[21] Jurgens, M., Jacob, F., Ekici, P., Friess, A., Parlar, H., 2007. Determination of direct photolysis rate constants and OH radical reactivity of representative odour compounds in brewery broth using a continuous flow-stirred photoreactor. Atmos. Environ. 41, 4571-4584.
[22] Dodd, M.C., Buffle, M.O., Von Gunten, U., 2006. Oxidation of antibacterial molecules by aqueous ozone: Moiety-specific reaction kinetics and application to ozone-based wastewater treatment. Environ. Sci. Technol. 40, 1969-1977.
[23] Dail, M.K., Mezyk, S.P., 2010. Hydroxyl-radical-induced degradative oxidation of β-lactam antibiotics in water:absolute rate constant measurements. J. Phys. Chem. A. 114, 8391-8395.
[24] Boonrattanakij, N., Lu, M.C., Anotai, J., 2009. Kinetics and mechanism of 2,6-dimethyl-aniline degradation by hydroxyl radicals. J. Hazard. Mater. 172, 952-957.
[25] Zona, R., Solar, S., Sehested, K., Holcman, J., Mezyk, S.P., 2002. OH-radical induced oxidation of phenoxyacetic acid and 2,4-dichlorophenoxyacetic acid. Primary radical steps and products. J. Phys. Chem. A. 106, 6743-6749.
[26] Gligorovski, S., Rousse, D., George, C.H., Herrmann, H., 2009. Rate constants for the OH reactions with oxygenated organic compounds in aqueous solution. Int. J. Chem. Kinet. 41, 309-326.
[27] Ben Abdelmelek, S., Greaves, J., Ishida, K.P., Cooper, W.J., Song, W.H., 2011. Removal of pharmaceutical and personal care products from reverse osmosis retentate using advanced oxidation processes. Environ. Sci. Technol. 45, 3665-3671.
[28] Mezyk, S.P., Ewing, D.B., Kiddle, J.J., Madden, K.P., 2006. Kinetics and mechanisms of the reactions of hydroxyl radicals and hydrated electrons with nitrosamines and nitramines in water. J. Phys. Chem. A. 110, 4732-4737.
[29] An, T.C., Yang, H., Li, G.Y., Song, W.H., Cooper, W.J., Nie, X.P., 2010. Kinetics and mechanism of advanced oxidation processes (AOPs) in degradation of ciprofloxacin in water. Appl. Catal. B-Environ. 94, 288-294.
[30] Haag, W.R., Yao, C.C.D., 1992. Rate constants for reaction of hydroxyl radicals with several drinking water contaminants. Environ. Sci. Technol. 26, 1005-1013.

Table S2 Quantum chemical descriptors calculated in this study

Descriptor	Definition of the descriptors	Type[a]

Descriptor	Definition of the descriptors	Type[a]
μ	Molecular dipole moment	I
α	Molecular polarizability	I
E_{CCR}	Core-core repulsion energy	II
E_{HOMO}	The highest occupied molecular orbital of the solute	II
E_{LUMO}	The lowest unoccupied molecular orbital of the solute	II
E_{SCF}	Self-consistent field (SCF) Done Energy	II
q_C^+	Most positive charge on C atom	II
q_C^-	Most negative charge on C atom	II
q_H^+	Most positive charge on H atom	II, III
q_F^-	Most negative charge on F atom	II, III
q_{Br}^-	Most negative charge on Br atom	II, III
q_{Cl}^-	Most negative charge on Cl atom	II, III
q_I^-	Most negative charge on I atom	II, III
q_O^-	Most negative charge on O atom	II, III
q_P^-	Most negative charge on P atom	II, III
q_N^-	Most negative charge on N atom	II, III
q_S^-	Most negative charge on S atom	II, III

[a] Type I represents the descriptors used for characterizing the Van der Waals forces; Type II represents the descriptors used for characterizing the intermolecular electrostatic interactions, and the charge type for the charge descriptors is Mulliken charge; Type III represents the descriptors used for characterizing hydrogen bond interactions.

Table S3 Compounds with $h > h^*$ and $|\delta| < 3$

No.	Compound	δ	h
1'	2,2-Difluoroethanol	1.352	0.225
2	1-(p-Bromophenyl)ethanol	0.727	0.101
3	m-bromophenyl-β-D-glucopyranoside	-0.292	0.104
4'	1,2-dibromo-3-chloropropane	-1.307	0.379
5	dibromomethane	-0.945	0.400
6'	5-Bromoindole	1.183	0.104
7	3-Bromopyridine	-1.321	0.104

No.	Compound	δ	h
8	2-Bromopyridine	0.786	0.103
9	Dichloromethane	-1.118	0.291
10'	Trichloromethane	0.687	1.086
11	2,2,2-Trichloroethanol	0.956	0.204
12'	hexachlorocyclopentadiene	0.649	0.108
13	Trichloroethylene	0.327	0.182
14	Tetrachloroethylene	-0.351	0.213
15'	trans-1,2-Dichloroethylene	0.272	0.182
16'	Butadiene	-1.015	0.173
17	Ethylene	-0.890	0.175
18	1,4-Pentadien-3-ol	0.720	0.182
19'	Carbon disulfide	0.640	0.104
20	glyoxylic acid	0.541	0.123
21	disulfoton	-0.822	0.115
22	terbufos	0.109	0.125
23'	Amikacin	0.815	0.101
24	Cefazolin	-0.575	0.211
25	Azithromycin	-1.106	0.131
26	erythromycin	-0.369	0.132
27	Tetracycline	1.193	0.156
28	Roxithromycin	0.578	0.157
29	Vancomycin	0.634	0.411
30	glutathione,oxidised	1.000	0.142
31	Nicotinamide-N-oxide	0.483	0.159
32'	2-Hydroxypropionamide	0.303	0.159
33	Isobutyramide	1.051	0.161
34'	Trimethylacetamide	0.509	0.166
35	Propionamide	-0.442	0.172
36'	Glycinamide	1.750	0.179
37'	Glycolamide	0.940	0.185

No.	Compound	δ	h
38'	Acetamide	-1.830	0.211
39	N-acetylglycylglycinamide	-2.061	0.169
40	glycylglycylglycylglycinamide	-0.860	0.172

* The numbers without superscript represent the compounds in the training set; The numbers with a comma as superscript represent the compounds in the validation set.

Table S4 Descriptors and corresponding coefficients of the model for the 224 nitrogen compounds

Descriptor	Definition of the descriptors	Coefficient
HOMO Energy	Energy of the highest occupied molecular orbital	11.47
JGT	Global topological charge index	-0.577
nN(CO)2	Number of imides (-thio)	0.366
O-057	Number of alcohol	0.161
MATS1i	Moran autocorrelation of lag 1 weighted by ionization potential	0.199
HATS0s	Leverage-weighted autocorrelation of lag 0 / weighted by I-state	-0.014
SssssC	Sum of ssssC E-states	0.192
Mor12v	Signal 12 / weighted by van der Waals volume	-0.203
nArNO2	Number of nitro groups (aromatic)	0.225
B02[C-O]	Presence/absence of C - O at topological distance 2	-0.123
HATS2e	Leverage-weighted autocorrelation of lag 2 / weighted by Sanderson electronegativity	-0.214