Use of *in Vitro* HTS-Derived Concentration-Response Data as Biological Descriptors Improves the Accuracy of QSAR Models of *in Vivo* Toxicity

Alexander Sedykh[1], Hao Zhu[1], Hao Tang[1], Liying Zhang[1], Ann Richard[2], Ivan Rusyn[3] and Alexander Tropsha[1*]

[1]Laboratory for Molecular Modeling, Division of Medicinal Chemistry and Natural Products, and

[2]National Center for Computational Toxicology, US Environmental Protection Agency, RTP, NC 27711

[3]Department of Environmental Sciences and Engineering, University of North Carolina, Chapel Hill, NC 27599

*Corresponding author:

327 Beard Hall

University of North Carolina

Chapel Hill, NC 27599-7568

Telephone: (919) 966-2955

FAX: (919) 966-0204

Email: alex_tropsha@unc.edu

Supplemental Material

Table of content

Supplemental Methods……. pages 3-8

qHTS NTP dataset (369 compounds)……..….. page 3

Noise-filtering algorithm to process qHTS data…… pages 3-5

Down-sampling procedure…….pages 5-6

Modeling methods……..... pages 6-7

Descriptor frequency analysis……....pages 7-8

Reference…….……..…… page 8

Supplemental figures (1 through 5)……..……. pages 9-13

Supplemental table…….pages 14-21

qHTS NTP dataset (369 compounds):

See the Supplemental Material, Table 1 for the list.

Noise-filtering algorithm to process qHTS data:

Here we describe a non-parametric noise-filtering approach, which in contrast to conventional curve-fitting (e.g. Hill's equation of the canonical sigmoid curve), can process sigmoid shapes, their partial versions, as well constant response curves. It deals with normalized experimental data (e.g. that batch/plate effects have been addressed). Our algorithm is based on the assumption that a canonical (or ideal) concentration-response curve should feature monotonous behavior. Data points that violate this monotonic pattern (and thus are likely outliers) were replaced by new values that were calculated by extrapolation from the adjacent data points (Figure 2, main manuscript). The algorithm is controlled by following two user-defined parameters (both defined in terms of percentage of maximum measured response). Baseline threshold THR is defined as the largest deviation of the response from the baseline (i.e., no response) below which the measured signal is treated as baseline, i.e., considered as no response (cf. Figure 2A, main manuscript); THR is used to suppress signal noise near baseline. The curve deviation threshold, MXDV, is the largest difference in the signal value for two adjacent data points, below which the response of these two points is considered the same. The use of these parameters allowed to suppress local ruggedness and small deviations of a dose-response curve from its monotonic behavior and to streamline the dose-response curves (cf. Figure 2B, main manuscript).

The algorithm searches for and applies to each curve the smallest number of corrections that would restore its monotonic behavior. The pseudocode of the algorithm is as follows:

1. Concentration points with response < THR are set to base line (response = 0).
2. Calculate the difference of response between the lowest and the highest tested concentrations, if > THR then the curve is monotone decreasing, if < -THR then it is monotone increasing. If the absolute difference > THR, jump to step 7.
3. Calculate st.dev and average response over all concentration points.
4. If st.dev is less than MXDV then curve type is constant (jump to step 7).
5. Calculate the number of intervals (pairs of adjacent concentrations) with increasing response and subtract the number of intervals with decreasing response.
6. If the absolute value of the Step 4 result is less than 3 then the curve is constant. If the result is ≥ 3 then the curve is monotone increasing, if ≤ -3 - then it is monotone decreasing.
7. If the curve type is constant the response of all concentration points is replaced by the average response calculated in Step 3.
8. Scan from highest to lowest concentrations to detect and list those data points that violate curve's monotonicity by more than MXDV. Repeat the scan from lowest to highest concentrations, compare two resulting lists and keep the smallest.
9. Replace the response values of violating points by linear extrapolation from the adjacent points.
10. Calculate concentration-response descriptors and curve fingerprints, print out results.

The parameters THR and MXDV have to be adjusted so as to find the best trade-off between noise filtering and loss of meaningful data. We compared qHTS data before and after the processing using different combinations of THR and MXDV parameters. Baseline threshold

parameter, THR, was varied in 1% - 25% range of the response scale and ruggedness factor, MXDV, was varied in 1% - 10% range. Results of the qHTS data comparison are shown in Supplemental Material, Figure 2. As can be seen, THR turned out to be more important parameter, while different values of MXDV give similar results, the only exception being MXDV=1%. In this case too much useful information was removed from the original data since even small variations of response were treated as significant and resulted in a lot of corrections. Based on the above comparison MXDV was set to a constant value of 5% and three sets of qHTS descriptors were produced for THR values of 5%, 15% and 25%, respectively.

Supplemental Material, Figure 3 illustrates the effect of the noise-filtering procedure on the biological data variation. For example, the absence of x-values <0.5 in panel A shows that there are no *in vitro* profiles that are exactly the same, if original qHTS data is used.

Down-sampling procedure:

Using Dragon chemical descriptors and Euclidean distance as a metric, a group of inactive compounds that were nearest neighbors to the active compounds was selected. The similarity threshold was defined as Euclidean distance in Dragon descriptor space ($\langle d_{AI} \rangle - 0.5\sigma_{AI}$, where σ_{AI} is the standard deviation and $\langle d_{AI} \rangle$ is the average of distances between each active data point and its nearest inactive one). Thus, only interclass distances are considered, unlike the similar equation of Applicability Domain.

This procedure removed those inactive compounds that are easily explainable by simple similarity search and retained only those inactive compounds that are most difficult to differentiate from the active compounds. Such measure allows the subsequent modeling to reveal non-trivial features discriminating active and inactive classes of compounds.

Modeling methods:

Two modeling algorithms, kNN and Random Forest, have been employed as they represent conceptually different methodologies and thus are appropriate benchmarks. Though other modeling techniques are available, in this study we have applied only the described methods, as they provide sufficient proof-of-concept demonstration.

Random Forest: In the random forest modeling procedure, n samples are randomly drawn from the original data. These samples were used to construct n training sets and to build n trees. For each node of the tree, m variables were randomly chosen from the total set of descriptors. The best data split was calculated using these m variables for each training set. In this study, the default defined parameters (n = 500 and m = square root of the total number of descriptors) were used for the model development.

***k*NN:** Initially, a subset of *nvar* (number of selected variables) descriptors is selected randomly. The model developed with this set of descriptors is validated by leave-one-out cross-validation, where each compound is eliminated from the training set and its biological activity is predicted as the weighted average activity of its *k* (*k*= 1 to 9) nearest neighbors in the subspace of *nvar* descriptors (Eq. 1). The weights of neighbors, w_i, decrease with distance, thus closer neighbors contribute to the calculated activity more:

$$y_{pred} = \frac{\sum_{i=1}^{k} w_i y_i}{\sum_{i=1}^{k} w_i}; \qquad w_i = \exp(-d_i) \qquad [1]$$

Here y_{pred} is predicted activity; d_i , w_i and y_i are, respectively: Euclidean distance, weight and actual activity for the nearest neighbor *i*.

A genetic algorithm was used to optimize the variable selection (with population size of 500 solutions of *nvar* size from 5 to 40 descriptors). The predictive power of models was characterized by the CCR (correct classification rate) for the test and training sets.

$$CCR = \frac{Sensitivity + Specificity}{2} \qquad [2]$$

Here sensitivity (specificity) is the correctly predicted fraction of "toxic" ("non-toxic") compounds.

Individual models were considered acceptable if $CCR_{train}>0.65$ and $CCR_{test}>0.75$; otherwise they were discarded. The ensemble of all models that passed the above criteria was then used for consensus prediction of compounds from external validation set. Note, that we have applied stricter criterion for the test set ($CCR_{test}>0.75$) in order to suppress overtrained models.

Descriptor frequency analysis:

In total, in 5 modeling splits over 7000 individual *k*NN models were generated that satisfied the acceptability criteria (see *k*NN description). Each descriptor occurrence in one such individual model was normalized by its size (*nvar* parameter) and summed up over all generated *k*NN models. In general, individual *k*NN models were independent from each other (i.e. each model operates in a local subspace of *nvar* descriptors, and average degree of overlap between these subspaces, as measured by descriptor-based Jaccard similarity coefficient (Tan et al. 2006), was less than 0.1) and therefore no correction was applied to the descriptor occurrence values. Mean occurrence number averaged over all descriptors was equal to 3.3% of the theoretical

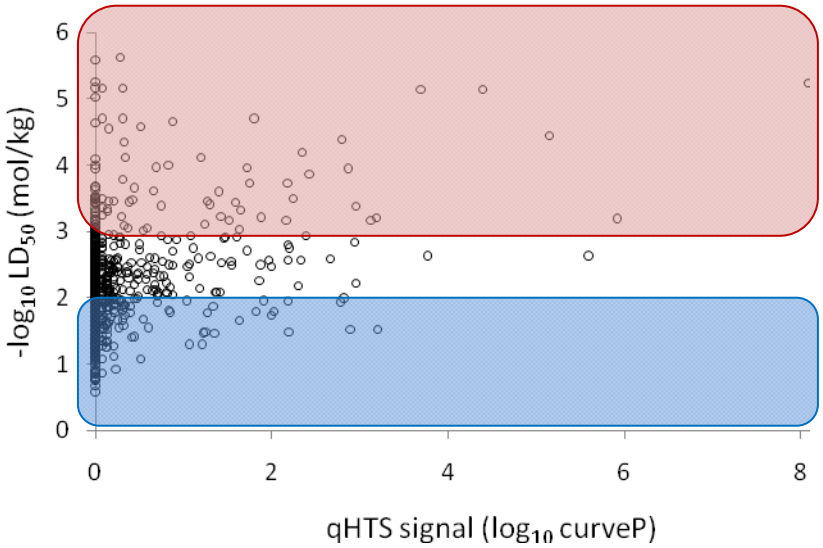
maximum (i.e., when a descriptor occurs in all individual *k*NN models). This value was used to select a group of most frequent descriptors (~90). Supplemental Material, Figure 4 demonstrates that there is even distribution of qHTS descriptors by frequency, i.e., the ratio of qHTS descriptors to chemical ones is almost the same in any fraction of most frequent descriptors.

Note that our *k*NN QSAR approach uses stochastic descriptor sampling algorithm to select the final set of descriptors for each model that gives the highest accuracy of prediction for the training set. Thus, the selection of most frequent descriptors is dictated by the sampling algorithm with no user intervention.

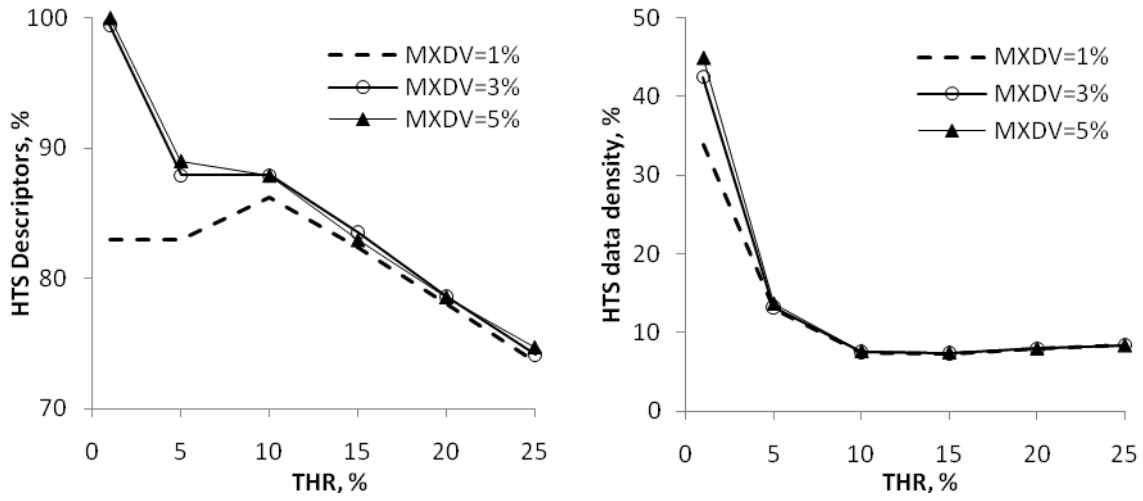
Supplemental references

Tan P-N, Steinbach M, Vipin K. 2006. Introduction to data mining. Boston:

Supplemental figures

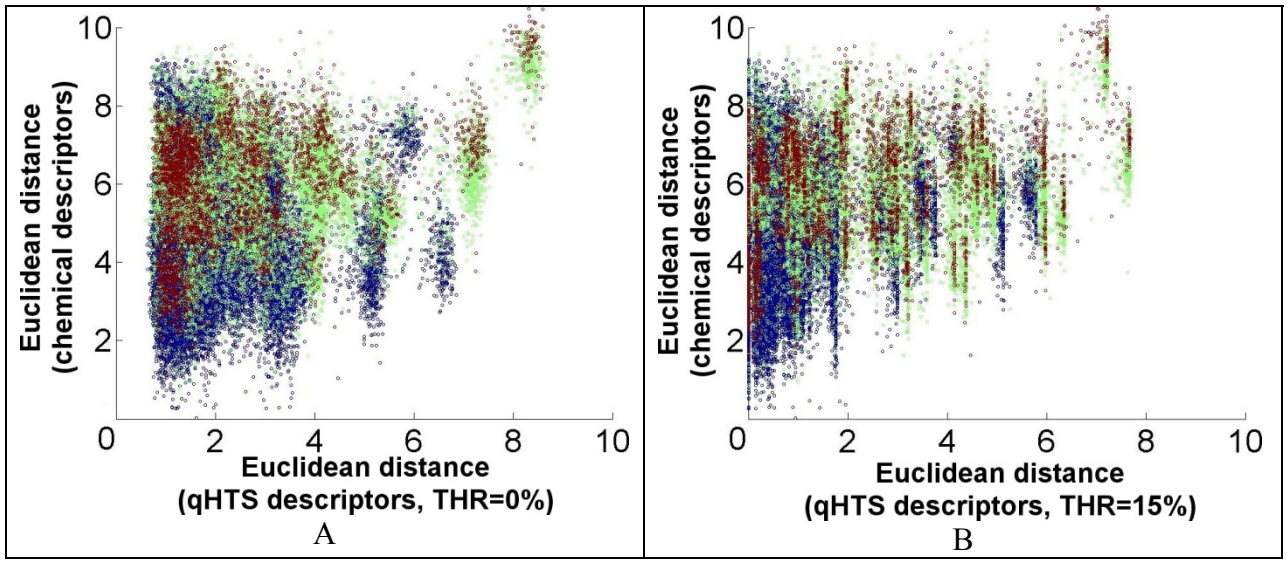


Supplemental Material, Figure 1. Distribution of LD_{50} values for the entire qHTS LD_{50} dataset (695 compounds) versus the qHTS signal represented by a concentration-response curve fingerprint, curveP (its logarithm averaged across 13 cell lines). Red and blue areas show 92 "toxic" and 277 "non-toxic" compounds used for modeling.

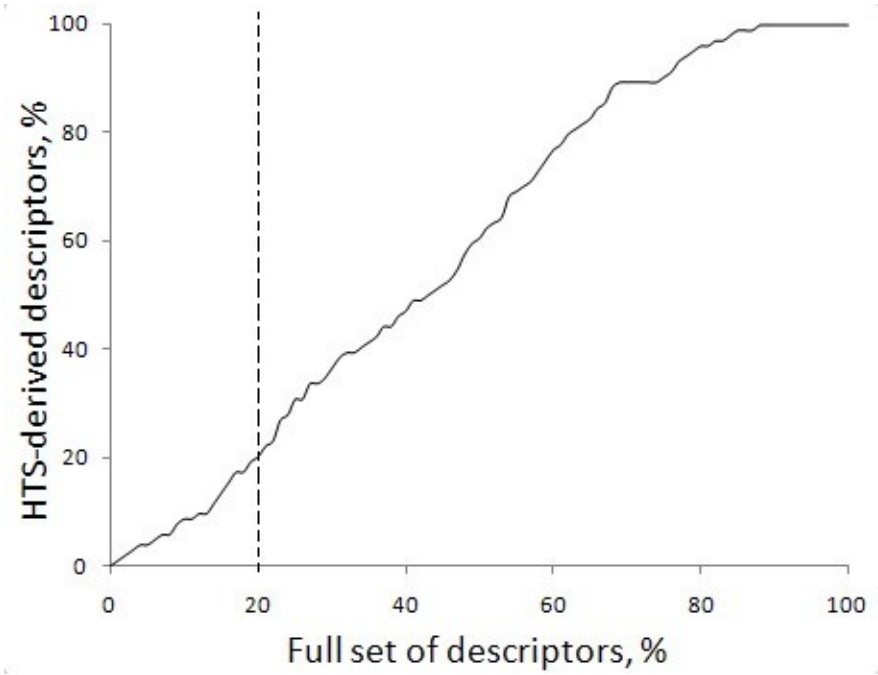


A B

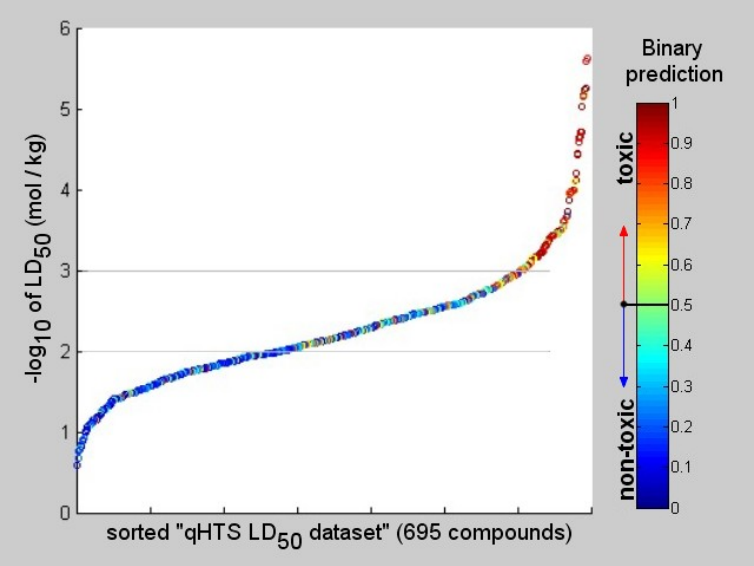
Supplemental Material, Figure 2. Comparison of qHTS data noise-filtered at various settings. A) non-redundant qHTS descriptors remaining after the noise-filtering (in percent of the initial set of 182 descriptors); B) average percent of non-zero values in the processed qHTS descriptors. Comparison is based on the entire qHTS LD_{50} dataset (695 compounds).



Supplemental Material, Figure 3. Pairwise Euclidean distances in the chemical (*y* axis) and biological (*x* axis) descriptor space for the qHTS LD_{50} dataset. Descriptors from original qHTS data (A) and noise-filtered qHTS data (B) were used to calculate distance-based similarity. Dots represent all pairs of compounds; colors reflect *in vivo* toxicity: blue, pairs of non-toxic; red, toxic; and green, toxic+non-toxic compounds.



Supplemental Material, Figure 4. Occurrence of qHTS descriptors in the Hybrid *k*NN (THR=15%) model. The full descriptor list is sorted by decreasing occurrence (*x*-axis). The fraction of most frequent descriptors selected by mean occurrence number is marked by the dashed line.



Supplemental Material, Figure 5. Predictions by Hybrid (THR=15%) model of the entire qHTS LD_{50} dataset (695 compounds). "Marginally toxic" compounds (between the gray solid lines at 2 and 3 of $-\log LD_{50}$) were not used in modeling (see also Figure 1, main manuscript).

CAS RN	Toxicity Class (1 - Toxic, 0 - Non-Toxic)	-log10 (LD50 in Mol per kg)	SMILES	NAME
107028	1	3.33	O=CC=C	Acrolein
50760	1	5.24	O7C(C)C(NC(=O)C4=C(N)C(=O)C(C)=C3Oc5c(C)ccc(C(NC2C(C)OC(=O)C(C(C)C)N(C)C(CN(C)C(=O)C1CCCN1C(=O)C(C(C)C)NC2=O)=O)=O)c5N=C34)C(NC(C(C)C)C(=O)N6CCCC6C(=O)N(C)CC(=O)N(C)C(C(C)C)C7=O)=O	Actinomycin D
116063	1	5.58	S(C)C(C)(C=NOC(=O)NC)C	Aldicarb
309002	1	3.97	ClC34C(Cl)(C(Cl)(C2C1CC(C=C1)C23)C(Cl)=C4Cl)Cl	Aldrin
106923	0	1.85	O1CC1COCC=C	Allyl glycidyl ether
99570	0	1.81	Oc1c(N)cc(N(=O)=O)cc1	2-Amino-4-nitrophenol
2465272	0	1.95	N(C)(C)c2ccc(C(=N)c1ccc(N(C)C)cc1)cc2	Auramine
86500	1	4.66	S(CN1N=Nc2ccccc2C1=O)P(OC)(=S)OC	Azinphosmethyl
30516871	0	1.94	O1C(CC(C1CO)N=N#N)N2C(NC(=O)C(C)=C2)=O	3'-Azido-3'-deoxythymidine (AIDS)
140114	0	1.78	O(Cc1ccccc1)C(C)=O	Benzyl acetate
100516	0	1.94	OCc1ccccc1	Benzyl alcohol
92524	0	1.86	c2ccccc2c1ccccc1	Biphenyl
111444	1	3.28	ClCCOCCCl	Bis(2-chloroethyl)ether
58082	1	3.01	O=C1N(C)C(=O)N(C)C2N=CN(C)C1=2	Caffeine
56235	0	1.82	ClC(Cl)(Cl)Cl	Carbon tetrachloride
133904	0	1.77	Clc1c(N)cc(Cl)cc1C(O)=O	Chloramben
305033	1	3.60	ClCCN(CCCl)c1ccc(CCCC(O)=O)cc1	Chlorambucil
95749	0	1.98	Clc1c(C)ccc(N)c1	3-Chloro-p-toluidine
532274	1	3.49	ClCC(=O)c1ccccc1	2-Chloroacetophenone (CN)
72548	1	3.45	Clc2ccc(C(C(Cl)Cl)c1ccc(Cl)cc1)cc2	Tetrachlorodiphenylethane
50293	1	3.61	Clc2ccc(C(C(Cl)(Cl)Cl)c1ccc(Cl)cc1)cc2	Dichlorodiphenyltrichloroethane (DDT)
333415	1	3.66	S=P(OCC)(OCC)Oc1cc(C)nc(C(C)C)n1	Diazinon
96128	1	3.14	BrC(CCl)CBr	1,2-Dibromo-3-chloropropane
106934	1	3.24	BrCCBr	1,2-Dibromoethane
99309	0	1.94	Clc1c(N)c(Cl)cc(N(=O)=O)c1	Dichloran
79436	0	1.66	ClC(C(=O)O)Cl	Dichloroacetic acid
120832	1	3.54	Clc1c(O)ccc(Cl)c1	2,4-Dichlorophenol
60571	1	4.00	ClC45C(Cl)(C(Cl)(C3C2CC(C1OC12)C34)C(Cl)=C5Cl)Cl	Dieldrin
2921882	1	3.63	Clc1cc(Cl)c(Cl)nc1OP(OCC)(OCC)=S	Chlorpyrifos (Dursban)
111466	0	0.93	O(CCO)CCO	Diethylene glycol
60117	1	3.05	N(C)(C)c2ccc(N=Nc1ccccc1)cc2	4-Dimethylaminoazobenzene
115297	1	3.48	ClC23C(Cl)(C(Cl)(C1COS(OCC12)=O)C(Cl)=C3Cl)Cl	Endosulfan
106898	1	3.01	ClCC1CO1	Epichlorhydrin
106887	0	2.16	O1C(CC)C1	1,2-Epoxybutane
64175	0	0.82	OCC	Ethanol
72560	0	1.67	ClC(Cl)C(c1ccc(CC)cc1)c2ccc(CC)cc2	Di(p-ethylphenyl)dichloroethane
103231	0	1.61	O(CC(CC)CCCC)C(CCCCC(OCC(CC)CCCC)=O)=O	Di(2-ethylhexyl)adipate
117817	0	1.12	O(CC(CC)CCCC)C(=O)c1ccccc1C(OCC(CC)CCCC)=O	Di(2-ethylhexyl) phthalate
470826	0	1.79	O2C1(C)CCC(CC1)C2(C)C	1,8-Cineol
55389	1	3.19	S(C)c1ccc(OP(OC)(=S)OC)cc1C	Fenthion
98011	1	3.17	O1C=CC=C1C=O	Furfural
77065	0	1.74	O5C24C=CC(O)C(C)(C4C(C13CC(=C)C(O)(C1)CCC23)C(O)=O)C5=O	Gibberellic acid
76448	1	3.97	ClC23C1C(Cl)C=CC1C(Cl)(C2(Cl)Cl)C(Cl)=C3Cl	Heptachlor
87683	1	3.50	ClC(C(Cl)=C(Cl)Cl)=C(Cl)Cl	Hexachloro-1,3-butadiene
67721	0	1.73	ClC(Cl)(C(Cl)(Cl)Cl)Cl	Hexachloroethane

70304	1	3.86	Clc2c(Cl)cc(Cl)c(O)c2Cc1c(Cl)c(Cl)cc(Cl)c1O	Hexachlorophene
53861	1	5.17	Clc3ccc(C(=O)N1C(C)=C(CC(O)=O)c2cc(OC)ccc12)cc3	Indomethacin
75478	1	3.05	IC(I)I	Iodoform
1634782	1	3.30	S(C(CC(OCC)=O)C(OCC)=O)P(OC)(OC)=O	Malaoxon
121755	1	3.06	S(C(CC(OCC)=O)C(OCC)=O)P(OC)(OC)=S	Malathion
148823	1	4.44	ClCCN(CCCl)c1ccc(CC(N)C(O)=O)cc1	Melphalan
70257	1	3.21	O=NN(C)C(=N)NN(=O)=O	1-Methyl-3-nitro-1-nitroso-guanidine
298000	1	4.64	S=P(OC)(OC)Oc1ccc(N(=O)=O)cc1	Methyl parathion
75092	0	1.73	ClCCl	Methylene chloride
90120	0	1.89	c12ccccc2cccc1C	1-Methylnaphthalene
91576	0	1.94	c1c(C)ccc2ccccc12	2-Methylnaphthalene
79118	1	3.24	ClCC(=O)O	Monochloroacetic acid
54115	1	3.51	n2cc(C1CCCN1C)ccc2	Nicotine
99592	0	1.87	O(C)c1ccc(N(=O)=O)cc1N	5-Nitro-o-anisidine
62237	0	1.93	OC(=O)c1ccc(N(=O)=O)cc1	p-Nitrobenzoic acid
79243	0	1.83	O=N(CC)=O	Nitroethane
75525	0	1.81	O=N(C)=O	Nitromethane
62759	1	3.30	O=NN(C)C	N-Nitrosodimethylamine
2058460	0	1.98	OC3C(N(C)C)C2C(O)C1C(C)(O)c4cccc(O)c4C(=O)C1=C(O)C2(O)C(=O)C=3C(N)=O	Oxytetracycline hydrochloride
87865	1	3.99	Clc1c(O)c(Cl)c(Cl)c(Cl)c1Cl	Pentachlorophenol, purified
89258	0	1.96	O=C1CC(C)=NN1c2ccccc2	1-Phenyl-3-methyl-5-pyrazolone
135886	0	1.40	N(c1ccccc1)c2ccc3ccccc3c2	N-Phenyl-2-naphthylamine
50339	1	3.10	O=C1C(CCCC)C(=O)N(N1c2ccccc2)c3ccccc3	Phenylbutazone
122601	0	1.59	O1C(C1)COc2ccccc2	Phenyl glycidyl ether
90437	0	1.93	Oc1ccccc1c2ccccc2	o-Phenylphenol
51036	0	1.74	O1COc2cc(COCCOCCOCCCC)c(CCC)cc12	Piperonyl butoxide
50555	1	3.16	O(C)c6c(OC)cc(C(OC4CC3CN5CCC1=C(Nc2cc(OC)ccc12)C5CC3C(C4OC)C(OC)=O)=O)cc6OC	Reserpine
96093	0	1.78	O1CC1c2ccccc2	Styrene oxide
97187	1	4.71	Clc2cc(Cl)cc(Sc1cc(Cl)cc(Cl)c1O)c2O	2,2'-Thiobis(4,6-dichlorophenol)
79196	1	4.00	S=C(N)NN	Thiosemicarbazide
538238	0	1.15	O(C(COC(CCCCCCC)=O)COC(CCCCCCC)=O)C(CCCCCCC)=O	Tricaprylin
634935	0	1.91	Clc1cc(Cl)cc(Cl)c1N	2,4,6-Trichloroaniline
79016	0	1.43	ClC(Cl)=CCl	Trichloroethylene
78422	0	1.07	P(OCC(CC)CCCC)(OCC(CC)CCCC)(OCC(CC)CCCC)=O	Tris(2-ethylhexyl)phosphate
67641	0	1.00	O=C(C)C	Acetone
79083	1	3.44	BrCC(=O)O	Bromoacetic acid
74975	0	1.41	BrCCl	Bromochloromethane
95578	1	3.51	Clc1ccccc1O	o-Chlorophenol
598914	1	3.41	BrC(N(=O)=O)Br	Dibromonitromethane (water disinfection byproducts)
58902	1	3.22	Clc1c(O)c(Cl)cc(Cl)c1Cl	2,3,4,6-Tetrachlorophenol
59507	0	1.89	Clc1c(C)cc(O)cc1	4-Chloro-m-cresol
71556	0	1.14	ClC(C)(Cl)Cl	1,1,1-Trichloroethane
78831	0	1.48	OCC(C)C	Isobutyl alcohol
84742	0	1.57	O(CCCC)C(=O)c1ccccc1C(OCCCC)=O	Dibutyl phthalate
104881	0	1.95	Clc1ccc(C=O)cc1	4-Chlorobenzaldehyde
105679	0	1.58	Oc1c(C)cc(C)cc1	2,4-Dimethylphenol
107120	1	3.15	N#CCC	Propionitrile
107197	1	3.45	OCC#C	Propargyl alcohol
108930	0	1.86	OC1CCCCC1	Cyclohexanol

110861	0	1.95	n1ccccc1	Pyridine
111159	0	1.69	O(CCOCC)C(C)=O	Ethylene glycol monoethyl ether acetate
117840	0	0.92	O(CCCCCCCC)C(=O)c1ccccc1C(OCCCCCCCC)=O	Dioctyl phthalate
118558	0	1.85	O(C(=O)c1c(O)cccc1)c2ccccc2	Phenyl salicylate
123864	0	1.03	O(CCCC)C(C)=O	n-Butyl acetate
126738	0	1.95	P(OCCCC)(OCCCC)(OCCCC)=O	Tributyl phosphate
141786	0	1.20	O(CC)C(C)=O	Ethyl acetate
298044	1	5.02	S(CC)CCSP(OCC)(OCC)=S	Disulfoton
608719	1	3.39	Brc1c(O)c(Br)c(Br)c(Br)c1Br	Pentabromophenol
4901513	1	3.22	Clc1c(Cl)c(O)cc(Cl)c1Cl	2,3,4,5-Tetrachlorophenol
13071799	1	5.26	S(CSP(OCC)(OCC)=S)C(C)(C)C	O,O-diethyl S-(((1,1-dimethylethyl)thio)methyl)phoshorodithioate
20830755	1	4.44	O8CC(C2CCC3(O)C4CCC1CC(CCC1(C)C4CC(O)C23C)OC5CC(O)C(C(C)O5)OC6CC(O)C(C(C)O6)OC7CC(O)C(O)C(C)O7)=CC8=O	Digoxin
56542	1	3.09	O(C)c3ccc4nccc(C(O)C2CC1CCN2CC1C=C)c4c3	Quinidine
112243	0	1.77	N(CCN)CCNCCN	Triethylenetetramine
131577	0	1.49	O(C)c2cc(O)c(C(=O)c1ccccc1)cc2	2-Hydroxy-4-methoxybenzophenone
821067	1	3.54	BrCC=CCBr	(E)-1,4-Dibromo-2-butene
87616	0	2.00	Clc1c(Cl)cccc1Cl	1,2,3-Trichlorobenzene
96231	1	3.07	ClCC(O)CCl	1,3-Dichloro-2-propanol
13048334	0	1.66	O(CCCCCCOC(=O)C=C)C(=O)C=C	1,6-Hexamethylene diacrylate
107040	1	3.35	BrCCCl	1-Chloro-2-bromoethane
90131	0	2.02	Clc1cccc2ccccc12	1-Chloronaphthalene
112561	1	3.35	S(CCOCCOCCCC)C#N	2-(2-Butoxyethoxy)ethyl thiocyanate
117180	0	1.54	Clc1c(Cl)cc(Cl)c(Cl)c1N(=O)=O	2,3,5,6-Tetrachloronitrobenzene
616239	1	3.16	ClC(CO)CCl	2,3-Dichloro-1-propanol
2213630	1	3.01	Clc1c(Cl)nc2ccccc2n1	2,3-Dichloroquinoxaline
13014181	0	1.92	Clc1cc(Cl)ccc1C(Cl)(Cl)Cl	2,4,a,a,a-Pentachlorotoluene
25168267	1	3.05	Clc1cc(Cl)ccc1OCC(OCC(CC)CCCC)=O	2,4-D, Isooctyl ester, 67%
584849	0	1.48	O=C=Nc1c(C)ccc(N=C=O)c1	2,4-Toluene diisocyanate
89985	0	1.81	Clc1ccccc1C=O	2-Chlorobenzaldehyde
91587	0	1.89	Clc1ccc2ccccc2c1	2-Chloronaphthalene
110805	0	1.63	O(CCO)CC	Ethylene glycol monoethyl ether (EGMEE)
2461156	0	1.38	O1CC1COCC(CC)CCCC	2-Ethylhexyl glycidyl ether
78820	1	3.14	N#CC(C)C	2-Methylpropanenitrile
100696	1	3.02	n1c(C=C)cccc1	2-Vinylpyridine
102363	1	3.32	Clc1cc(N=C=O)ccc1Cl	3,4-Dichlorophenyl isocyanate
542767	1	3.95	ClCCC#N	3-Chloropropionitrile
98566	0	1.14	Clc1ccc(C(F)(F)F)cc1	p-Chloro-a,a,a-trifluorotoluene
67470	0	1.70	O1C(CO)=CC=C1C=O	5-(Hydroxymethyl)-2-furfural
59143	0	1.56	BrC2C(=O)NC(=O)N(C1CC(O)C(CO)O1)C=2	5-Bromo-2'-deoxyuridine (BRDU)
1646873	1	5.62	S(C)(=O)C(C)(C=NOC(=O)NC)C	Aldicarb sulfoxide
106956	1	3.00	BrCC=C	Allyl bromide
98873	0	1.70	ClC(Cl)c1ccccc1	alpha,alpha-Dichlorotoluene
20562021	1	3.17	O9C(CO)C(O)C(OC1OC(CO)C(O)C(O)C1O)C(OC2OC(C)C(O)C(O)C2O)C9OC6CCC7(C)C8CCC4(C)C(CC3N5CC(C)CCC5C(C)C34)C8CC=C7C6	alpha-Solanine
17804352	0	1.46	O(C)C(=O)NC1N(C(NCCCC)=O)c2ccccc2N=1	Benomyl
98077	0	1.51	ClC(Cl)(Cl)c1ccccc1	Benzotrichloride
98088	0	0.99	FC(F)(F)c1ccccc1	Benzotrifluoride

2550267	0	1.67	O=C(C)CCc1ccccc1	benzylacetone
111911	1	3.43	ClCCOCOCCCl	Bis(2-chloroethoxy)methane
80057	0	1.85	Oc2ccc(C(C)(C)c1ccc(O)cc1)cc2	Bisphenol A
108861	0	1.82	Brc1ccccc1	Bromobenzene
12789036	1	3.31	ClC23C1C(Cl)C(Cl)CC1C(Cl)(C2(Cl)Cl)C(Cl)=C3Cl	Chlordane (technical grade)
931975	1	4.10	OC1(CCCCC1)C#N	Cyclohexanone cyanohydrin
66819	1	5.15	O=C2C(C)CC(C)CC2C(O)CC1CC(=O)NC(C1)=O	Cycloheximide
6055192	1	3.42	ClCCN(CCCl)P1(NCCCO1)=O	Cyclophosphamide monohydrate
23541506	1	3.20	O1C(C)C(O)C(N)CC1OC2CC(O)(Cc3c(O)c4C(=O)c5cccc(OC)c5C(=O)c4c(O)c23)C(C)=O	Daunomycin HCL
62737	1	4.11	ClC(Cl)=COP(OC)(=O)OC	Dichlorvos (Vapona)
3056175	0	1.75	O1C(C=CC1CO)N2C(NC(=O)C(C)=C2)=O	Dideoxydidehydrothymidine
78386	0	1.85	P(CC)(OCC)(OCC)=O	Diethyl ethylphosphonate
112732	0	1.75	O(CCCC)CCOCCOCCCC	Diethylene glycol dibutyl ether
629141	0	1.43	O(CC)CCOCC	Ethylene glycol diethyl ether
462066	0	1.34	Fc1ccccc1	Fluorobenzene
75127	0	0.91	O=CN	Formamide
111499	1	3.68	N1CCCCCC1	Hexamethyleneimine
123922	0	0.89	O(CCC(C)C)C(C)=O	Isoamyl acetate
4016142	0	1.44	O1C(COC(C)C)C1	Isopropyl glycidyl ether
143500	1	3.73	ClC45C1(Cl)C(Cl)(Cl)C3(Cl)C2(Cl)C1(Cl)C5(Cl)C(=O)C2(Cl)C3 4Cl	Kepone
78706	0	1.74	OC(C)(CCC=C(C)C)C=C	Linalool
624839	1	3.04	O=C=NC	Methyl isocyanate
556616	1	3.01	S=C=NC	Methane, isothiocyanato-
78944	1	3.48	O=C(C=C)C	Methyl vinyl ketone
6317186	1	3.37	S(CSC#N)C#N	Methylene bis(thiocyanate)
2365482	1	3.10	SCC(=O)OC	Methylthioglycolate
121891	0	1.71	O=C(C)c1cccc(N(=O)=O)c1	m-Nitroacetophenone
121904	0	1.68	ClC(=O)c1cccc(N(=O)=O)c1	m-Nitrobenzoyl chloride
793248	0	1.88	N(c1ccccc1)c2ccc(NC(C)CC(C)C)cc2	N-(1,3-Dimethylbutyl)-N'-phenyl-p-phenylenediamine
68122	0	1.42	O=CN(C)C	N,N-Dimethylformamide
99989	1	3.44	N(C)(C)c1ccc(N)cc1	N,N-Dimethyl-p-phenylenediamine
101962	1	3.17	N(C(C)CC)c1ccc(NC(C)CC)cc1	N,N'-Di-sec-butyl-p-phenyldiamine
55867	1	4.19	ClCCN(C)CCCl	Nitrogen mustard hydrochloride
56382	1	5.16	S=P(OCC)(OCC)Oc1ccc(N(=O)=O)cc1	Parathion
106434	0	1.78	Clc1ccc(C)cc1	p-Chlorotoluene
22224926	1	4.58	S(C)c1c(C)cc(OP(OCC)(NC(C)C)=O)cc1	Phenamiphos
103855	1	4.71	S=C(N)Nc1ccccc1	Phenylthiourea
122043	0	1.52	ClC(=O)c1ccc(N(=O)=O)cc1	p-Nitrobenzoyl chloride
99990	0	1.85	O=N(=O)c1ccc(C)cc1	p-Nitrotoluene
32809168	0	1.61	Clc3cc(Cl)cc(N2C(=O)C1(C)CC1(C)C2=O)c3	Procymidone
129000	0	1.88	c24c1ccc3cccc(ccc2ccc1)c34	Pyrene
127184	0	1.80	ClC(Cl)=C(Cl)Cl	Tetrachloroethylene
112572	0	1.68	N(CCNCCN)CCNCCN	Tetraethylenepentamine
1684408	1	3.45	n1c3CCCCc3c(N)c2ccccc12	1,2,3,4-Tetrahydro-9-acridinamine monohydrochloride
102089	1	3.66	S=C(Nc1ccccc1)Nc2ccccc2	Thiocarbanilide
2451629	1	3.20	O1CC1CN4C(=O)N(CC2OC2)C(=O)N(CC3OC3)C4=O	1,3,5-Triglycidyl isocyanurate
140089	1	3.43	ClCCOP(OCCCl)OCCCl	Tris(2-chloroethyl) phosphate
152114	1	3.45	O(C)c2cc(C(C(C)C)(CCCN(C)CCc1ccc(OC)c(OC)c1)C#N)ccc2 OC	Verapamil HCl

67970	1	3.96	OC1CCC(=C)C(C1)=CC=C2CCCC3(C)C(CCC23)C(C)CCCC(C)C	Vitamin D3
75070	0	1.82	O=CC	Acetaldehyde
103902	0	1.89	Oc1ccc(NC(C)=O)cc1	Acetaminophen (4-hydroxyacetanilide)
75058	0	1.22	N#CC	Acetonitrile
118923	0	1.40	OC(=O)c1ccccc1N	o-Anthranilic acid
50817	0	1.17	O1C(=O)C(O)=C(O)C1C(O)CO	L-Ascorbic acid
100527	0	1.91	O=Cc1ccccc1	Benzaldehyde
65850	0	1.86	OC(=O)c1ccccc1	Benzoic acid
119539	0	1.33	OC(C(=O)c1ccccc1)c2ccccc2	Benzoin
74964	0	1.91	BrCC	Bromoethane (ethyl bromide)
109693	0	1.54	ClCCCC	n-Butyl chloride
3068880	0	0.68	O1C(CC1C)=O	beta-Butyrolactone
2244168	0	1.63	O=C1CC(CC=C1C)C(C)=C	D-Carvone
118752	0	1.79	ClC1=C(Cl)C(=O)C(Cl)=C(Cl)C1=O	Chloranil
140498	0	1.99	ClCC(=O)c1ccc(NC(C)=O)cc1	4-(Chloroacetyl)acetanilide
77929	0	1.81	OC(CC(O)=O)(CC(O)=O)C(O)=O	1,2,3-Propanetricarboxylic acid, 2-hydroxy-
120718	0	1.98	O(C)c1ccc(C)cc1N	p-Cresidine
7585399	0	1.78	O1C(CO)C2OC8OC(CO)C(OC7OC(CO)C(OC6OC(CO)C(OC5OC(CO)C(OC4OC(CO)C(OC3OC(CO)C(OC1C(O)C2O)C(O)C3O)C(O)C4O)C(O)C5O)C(O)C6O)C(O)C7O)C(O)C8O	beta-Cyclodextrin
108941	0	1.81	O=C1CCCCC1	Cyclohexanone
1596845	0	1.28	OC(CCC(NN(C)C)=O)=O	Daminozide
78875	0	1.76	ClC(C)CCl	1,2-Dichloropropane (propylene dichloride)
828002	0	1.96	O1C(C)CC(OC1C)OC(C)=O	Dimethoxane
513371	0	1.31	ClC=C(C)C	Dimethylvinyl chloride (DMVC)
938738	0	1.88	O(CC)c1ccccc1C(N)=O	2-Eethoxybenzamide
77838	0	1.58	O1C(C1(C)c2ccccc2)C(OCC)=O	Ethyl-3-methyl-3-phenylglycidate
100414	0	1.48	c1ccccc1CC	Ethylbenzene
107211	0	1.12	OCCO	Ethylene glycol
96457	0	1.75	S=C1NCCN1	Ethylene thiourea (ETU)
106876	0	1.82	O3CC3C1CCC2C(C1)O2	4-Vinyl-1-cyclohexene diepoxide
104767	0	1.54	OCC(CC)CCCC	2-Ethylhexanol
97530	0	1.93	O(C)c1c(O)ccc(CC=C)c1	Eugenol
67630	0	1.08	OC(C)C	Isopropanol
123331	0	1.47	O=C1NNC(=O)C=C1	Maleic hydrazide
150765	0	1.89	O(C)c1ccc(O)cc1	Hydroquinone monomethyl ether
1634044	0	1.34	O(C)C(C)(C)C	Methyl-t-butyl ether
80626	0	1.10	O(C)C(=O)C(C)=C	Methyl methacrylate
56042	0	1.98	S=C1NC(C)=CC(N1)=O	6-Methyl-2-thiouracil
443481	0	1.76	OCCN1C(=CN=C1C)N(=O)=O	Metronidazole
5307142	0	1.86	O=N(=O)c1c(N)ccc(N)c1	2-Nitro-p-phenylenediamine
1116547	0	1.25	OCCN(CCO)N=O	N-Nitrosodiethanolamine
85449	0	1.99	O1C(=O)c2ccccc2C1=O	Phthalic anhydride
57556	0	0.58	OC(C)CO	1,2-Propylene glycol
94597	0	1.92	O1COc2ccc(CC=C)cc12	Safrole
599791	0	1.41	S(=O)(=O)(Nc1ccccn1)c3ccc(N=Nc2ccc(O)c(C(O)=O)c2)cc3	Salicylazosulfapyridine
110441	0	1.18	OC(=O)C=CC=CC	Sorbic acid
100425	0	1.59	c1ccccc1C=C	Styrene
108305	0	1.82	O1C(=O)CCC1=O	Succinic anhydride

127695	0	1.43	S(NC1ON=C(C)C=1C)(=O)(=O)c2ccc(N)cc2	Sulfisoxazole
109999	0	1.64	O1CCCC1	Tetrahydrofuran
148798	0	1.99	S1C=NC(=C1)C2Nc3ccccc3N=2	Thiabendazole
57136	0	0.85	O=C(N)N	Urea
108054	0	1.47	O(C(C)=O)C=C	Vinyl acetate
100403	0	1.63	C1CC=CCC1C=C	4-Vinylcyclohexene
123728	0	1.46	O=CCCC	Butyraldehyde
78933	0	1.42	O=C(C)CC	Methyl ethyl ketone
431038	0	1.74	O=C(C)C(C)=O	2,3-Butanedione
111717	0	1.55	O=CCCCCCC	n-Heptanal
111148	0	1.27	OC(CCCCCC)=O	n-Heptanoic acid
116530	0	1.77	OC(=O)C(C)CC	Butanoic acid, 2-methyl-
78842	0	1.88	O=CC(C)C	Isobutyraldehyde
110623	0	1.27	O=CCCCC	n-Pentanal
123386	0	1.62	O=CCC	Propionaldehyde
67561	0	0.76	OC	Methanol
71363	0	1.97	OCCCC	n-Butanol
71410	0	1.28	OCCCCC	1-Pentanol
78900	0	1.52	NC(C)CN	Propylenediamine
78922	0	1.53	OC(C)CC	2-Butanol
80466	0	1.95	Oc1ccc(C(C)(C)CC)cc1	p-tert-Pentylphenol
84662	0	1.41	O(CC)C(=O)c1ccccc1C(OCC)=O	Diethyl phthalate
96297	0	1.97	ON=C(C)CC	Methyl ethyl ketoxime
98828	0	1.93	c1ccccc1C(C)C	Cumene
100378	0	1.96	OCCN(CC)CC	(Diethylamino)ethanol
101848	0	1.84	O(c1ccccc1)c2ccccc2	Diphenyl oxide
106423	0	1.33	c1c(C)ccc(C)c1	p-Xylene
106638	0	1.31	O(C(C=C)=O)CC(C)C	Isobutyl acrylate
107153	0	1.70	NCCN	Ethylenediamine
108101	0	1.68	O=C(C)CC(C)C	Methyl isobutyl ketone
121335	0	1.98	O(C)c1c(O)ccc(C=O)c1	Vanillin
134623	0	1.99	O=C(N(CC)CC)c1cc(C)ccc1	N,N-Diethyl-m-toluamide
141435	0	1.55	OCCN	Ethanolamine
150787	0	1.58	O(C)c1ccc(OC)cc1	Hydroquinone dimethyl ether
529191	0	1.56	N#Cc1ccccc1C	o-Tolunitrile
529204	0	1.73	O=Cc1ccccc1C	o-Tolualdehyde
603509	0	1.92	O(C(C)=O)c3ccc(C(c1ccc(OC(C)=O)cc1)c2ccccn2)cc3	Bisacodyl
58946	0	1.47	Clc1cc2N=CNS(=O)(c2cc1S(N)(=O)=O)=O	Chlorothiazide
50215	0	1.41	OC(C(=O)O)C	Lactic acid
144821	0	1.89	S1C(C)=NN=C1NS(=O)(=O)c2ccc(N)cc2	Sulfamethizole
63741	0	1.65	S(N)(=O)(=O)c1ccc(N)cc1	Sulfanilamide
131113	0	1.46	O(C)C(=O)c1ccccc1C(OC)=O	Dimethyl phthalate
84695	0	1.27	O(CC(C)C)C(=O)c1ccccc1C(OCC(C)C)=O	Diisobutyl phthalate
103504	0	1.90	O(Cc1ccccc1)Cc2ccccc2	1,1-Oxybis methylene, bis benzene
147477	0	1.94	N1C(C)(C=C(C)c2ccccc12)C	1,2-Dihydro-2,2,4-trimethylquinoline (monomer)
110634	0	1.77	OCCCCO	1,4-Butanediol
625694	0	1.18	OC(C)CC(C)O	2,4-Pentanediol
95783	0	1.97	Nc1c(C)ccc(C)c1	2,5-Xylidine
1122629	0	1.73	O=C(C)c1ccccn1	2-Acetylpyridine
90415	0	1.86	Nc1ccccc1c2ccccc2	2-Biphenylamine
123057	0	1.69	O=CC(CC)CCCC	2-Ethylhexanal

149575	0	1.68	OC(=O)C(CC)CCCC	2-Ethylhexanoic acid
103117	0	1.51	O(CC(CC)CCCC)C(C=C)=O	2-Ethylhexyl acrylate
105306	0	1.86	OCC(C)CCC	1-Pentanol, 2-methyl-
121391	0	1.92	O1C(C1c2ccccc2)C(OCC)=O	Ethyl 3-phenylglycidate
101688	0	1.44	O=C=Nc2ccc(Cc1ccc(N=C=O)cc1)cc2	4,4'-Diphenylmethane diisocyanate
584134	0	0.79	N1=CN(N)C=N1	4-amino-1,2,4-triazole
5131588	0	1.53	O=N(=O)c1c(N)cc(N)cc1	4-Nitro-1,3-benzenediamine
2835952	0	1.53	Oc1c(C)ccc(N)c1	5-Amino-o-cresol
89576	0	1.74	Oc1ccc(N)cc1C(O)=O	5-Aminosalicylic acid
92488	0	1.98	O1C(=O)C=Cc2cc(C)ccc12	Methyl coumarin
91441	0	1.67	O1C(=O)C=C(C)c2ccc(N(CC)CC)cc12	7-Diethylamino-4-methylcoumarin
105577	0	1.41	O(CC)C(C)OCC	Acetal
108247	0	1.76	O(C(C)=O)C(C)=O	Acetic anhydride
102012	0	1.86	O=C(C)CC(=O)Nc1ccccc1	Acetoacetanilide
110134	0	1.74	O=C(C)CCC(C)=O	2,5-Hexanedione
98839	0	1.38	c1ccccc1C(C)=C	alpha-Methylstyrene
80568	0	1.57	C12CC(CC=C1C)C2(C)C	alpha-Pinene
71432	0	1.92	c1ccccc1	Benzene
98099	0	1.96	ClS(=O)(=O)c1ccccc1	Benzenesulfonyl chloride
501531	0	1.76	ClC(OCc1ccccc1)=O	Carbonochloridic acid, phenylmethyl ester
106229	0	1.66	OCCC(C)CCC=C(C)C	Citronellol
1675543	0	1.41	O1CC1COc4ccc(C(C)(C)c3ccc(OCC2CO2)cc3)cc4	Bisphenol A diglycidyl ether
106310	0	1.26	O(C(CCC)=O)C(CCC)=O	Butanoic acid, anhydride
133062	0	1.52	ClC(Cl)(SN2C(=O)C1CC=CCC1C2=O)Cl	Captan 90-concentrate (solid)
10605217	0	1.48	O(C)C(=O)NC1Nc2ccccc2N=1	Carbendazim
75150	0	1.80	S=C=S	Carbon disulfide
99489	0	1.71	OC1CC(CC=C1C)C(C)=C	Carveol
55561	0	1.74	Clc2ccc(NC(NC(NCCCCCCNC(=N)NC(=N)Nc1ccc(Cl)cc1)=N)=N)cc2	Chlorhexidine
14371109	0	1.78	O=CC=Cc1ccccc1	trans-Cinnamaldehyde
156592	0	-999.00	ClC=CCl	cis-1,2-Dichloroethylene
286204	0	1.97	O1C2CCCCC12	Cyclohexene oxide
287923	0	0.79	C1CCCC1	Cyclopentane
91178	0	1.52	C12CCCCC2CCCC1	Decalin
80433	0	1.82	O(OC(C)(C)c1ccccc1)C(C)(C)c2ccccc2	Dicumyl peroxide
101906	0	1.94	O1CC1COc3cccc(OCC2CO2)c3	Diglycidyl resorcinol ether (DGRE)
87252	0	1.64	O(CC)C(=O)c1ccccc1N	Ethyl anthranilate
105873	0	1.49	O(CC=C(C)CCC=C(C)C)C(C)=O	Geranyl acetate
151677	0	1.54	BrC(C(F)(F)F)Cl	Halothane
4098719	0	1.66	O=C=NC1CC(C)(CC(C)(C1)CN=C=O)C	Isophorone diisocyanate
108225	0	1.52	O(C(C)=C)C(C)=O	Isopropenyl acetate
108214	0	1.18	O(C(C)C)C(C)=O	Isopropyl acetate
143077	0	1.22	OC(CCCCCCCCCCC)=O	Lauric acid
118718	0	1.95	O1C=CC(=O)C(O)=C1C	Maltol
134203	0	1.72	O(C)C(=O)c1ccccc1N	Methyl anthranilate
99752	0	1.66	O(C)C(=O)c1ccc(C)cc1	Methyl-p-toluate
620224	0	1.59	N#Cc1cccc(C)c1	m-Tolunitrile
108383	0	1.33	c1c(C)cccc1C	m-Xylene
607910	0	1.65	O1COc2cc(CC=C)cc(OC)c12	Myristicin
85983	0	1.99	O=C(N(CC)c1ccccc1)N(CC)c2ccccc2	N,N'-Diethylcarbanilide
17557232	0	1.68	O1CC1COCC(C)(C)COCC2OC2	Neopentyl glycol diglycidyl ether
2039874	0	1.53	Clc1ccccc1C=C	o-Chlorostyrene

2210799	0	1.61	O1CC1COc2ccccc2C	o-Cresyl glycidyl ether
124072	0	1.16	OC(CCCCCCC)=O	Octanoic acid
112801	0	1.05	OC(CCCCCCCC=CCCCCCCCC)=O	Oleic acid
93834	0	1.47	OCCN(CCO)C(CCCCCCCC=CCCCCCCCC)=O	Oleic acid diethanolamine condensate
88744	0	1.94	O=N(=O)c1ccccc1N	o-Nitroaniline
121573	0	1.15	S(O)(=O)(=O)c1ccc(N)cc1	4-Aminobenzenesulfonic acid
122805	0	1.78	O=C(C)Nc1ccc(N)cc1	p-Amino acetanilide
123115	0	1.96	O(C)c1ccc(C=O)cc1	p-Anisaldehyde
104949	0	1.97	O(C)c1ccc(N)cc1	p-Anisidine
13684634	0	1.88	O(C(=O)Nc1cccc(C)c1)c2cccc(NC(OC)=O)c2	3-((Methoxycarbonyl)amino)phenyl N-(3-methylphenyl)carbamate (Phenmedipham)
7664382	0	1.81	P(=O)(O)(O)O	Phosphoric acid
1918021	0	1.47	Clc1c(C(O)=O)nc(Cl)c(Cl)c1N	Picloram
120570	0	1.75	O1COc2ccc(C=O)cc12	Piperonal
326614	0	1.97	O1COc2ccc(COC(C)=O)cc12	Piperonyl acetate
589184	0	1.50	OCc1ccc(C)cc1	p-Methyl benzyl alcohol
1569013	0	1.67	O(CC(C)O)CCC	2-Propanol, 1-propoxy-
98293	0	1.77	Oc1cc(C(C)(C)C)ccc1O	p-tert-Butylcatechol
98511	0	1.98	c1cc(C)ccc1C(C)(C)C	p-tert-Butyltoluene
104858	0	1.49	N#Cc1ccc(C)cc1	p-Tolunitrile
723466	0	1.61	S(NC1C=C(C)ON=1)(=O)(=O)c2ccc(N)cc2	Sulfamethoxazole
5329146	0	1.49	S(N)(=O)(O)=O	sulfamic acid
994058	0	1.81	O(C)C(C)(CC)C	Tertiary amyl methyl ether (TAME)
7665727	0	1.81	O1C(COC(C)(C)C)C1	t-Butyl glycidyl ether
119642	0	1.92	C1CCc2ccccc2C1	Tetralin
23564058	0	1.71	S=C(NC(OC)=O)Nc1ccccc1NC(NC(OC)=O)=S	Thiophanate M
110021	0	1.78	S1C=CC=C1	Thiophene
102761	0	1.86	O(C(COC(C)=O)COC(C)=O)C(C)=O	Triacetin
120149	0	1.92	O(C)c1cc(C=O)ccc1OC	Veratraldehyde
55268752	0	1.63	S1CC(COC(N)=O)=C(C(O)=O)N3C1C(NC(=O)C(=NOC)C2OC=CC=2)C3=O	Cefuroxime
36791045	0	1.96	O1C(CO)C(O)C(O)C1N2C=NC(C(N)=O)=N2	Ribavirin
40596698	0	1.09	O(C(=O)C=C(C)C=CCC(C)CCCC(C)(OC)C)C(C)C	Methoprene
80443410	0	1.96	Clc2ccc(CCC(O)(CN1N=CN=C1)C(C)(C)C)cc2	Tebuconazole