Electronic Supplementary Material (ESI) for Environmental Science: Processes & Impacts.
This journal is © The Royal Society of Chemistry 2024

Supporting Information

In silico approaches for the prediction of the breakthrough of organic contaminants in wastewater treatment plants

*Nicola Chirico[a, *], Michael S. McLachlan[b], Zhe Li[b] and Ester Papa[a]*

[a] QSAR Research Unit in Environmental Chemistry and Ecotoxicology, Department of Theoretical and Applied Sciences, University of Insubria, via J. H. Dunant 3, 21100, Varese, Italy.

[b] Department of Environmental Science (ACES), Stockholm University, 106 91 Stockholm, Sweden.

*Corresponding author: nicola.chirico@uninsubria.it

Table of contents

Figures

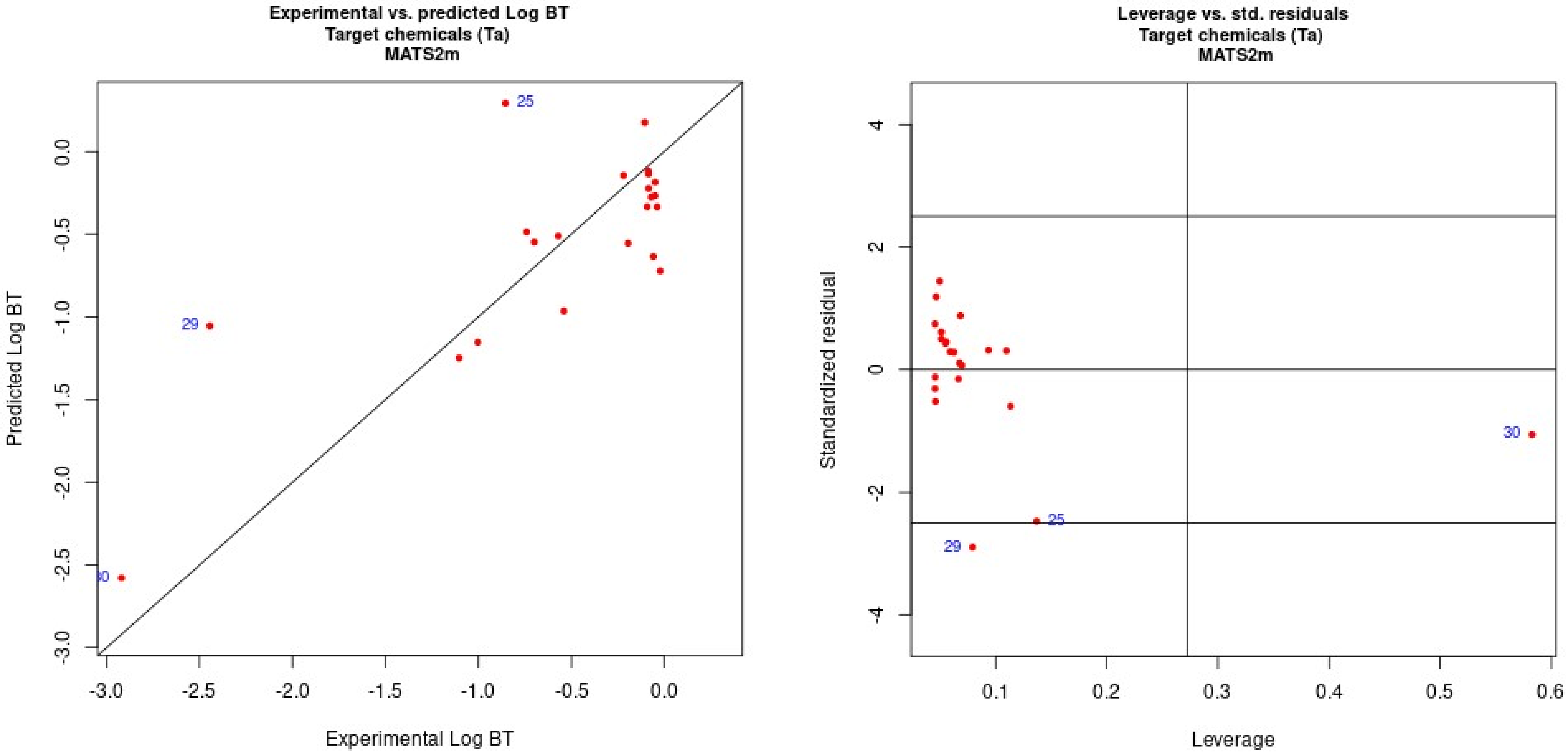


Figure S1. Performances of the target chemicals QSPR (Ta dataset).

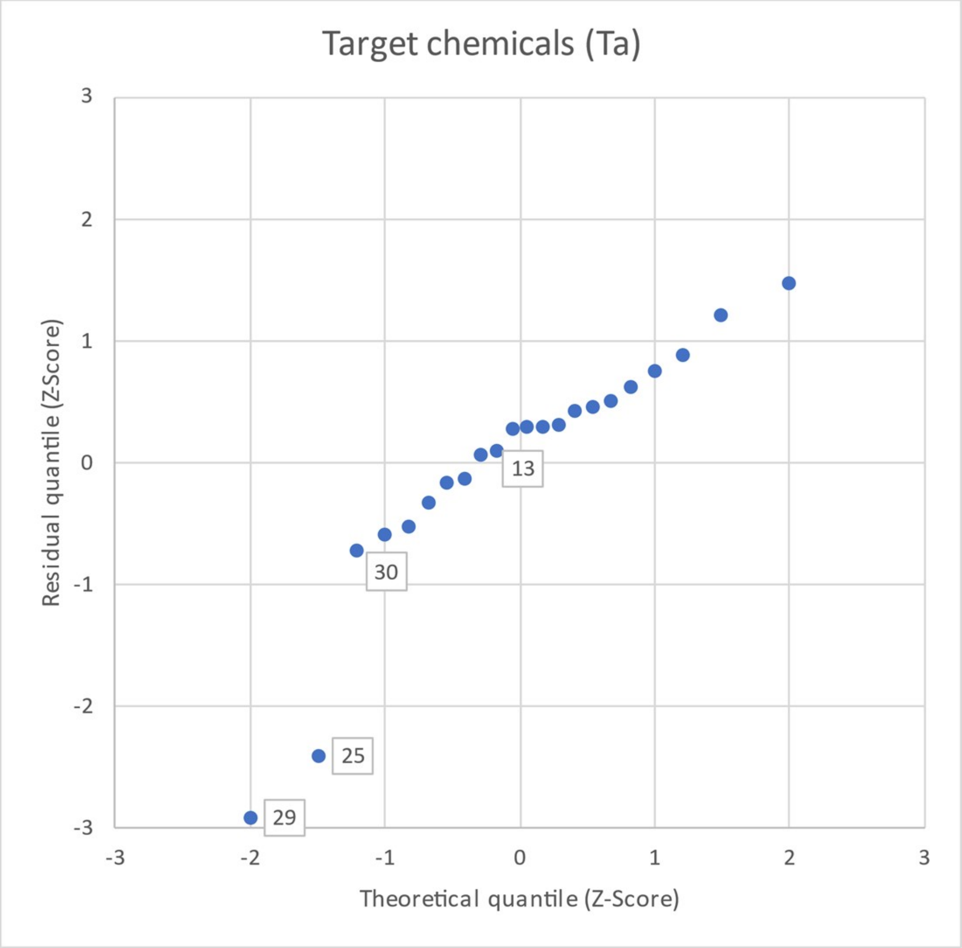


Figure S1. QQ chart of the target chemicals QSPR (Ta dataset).

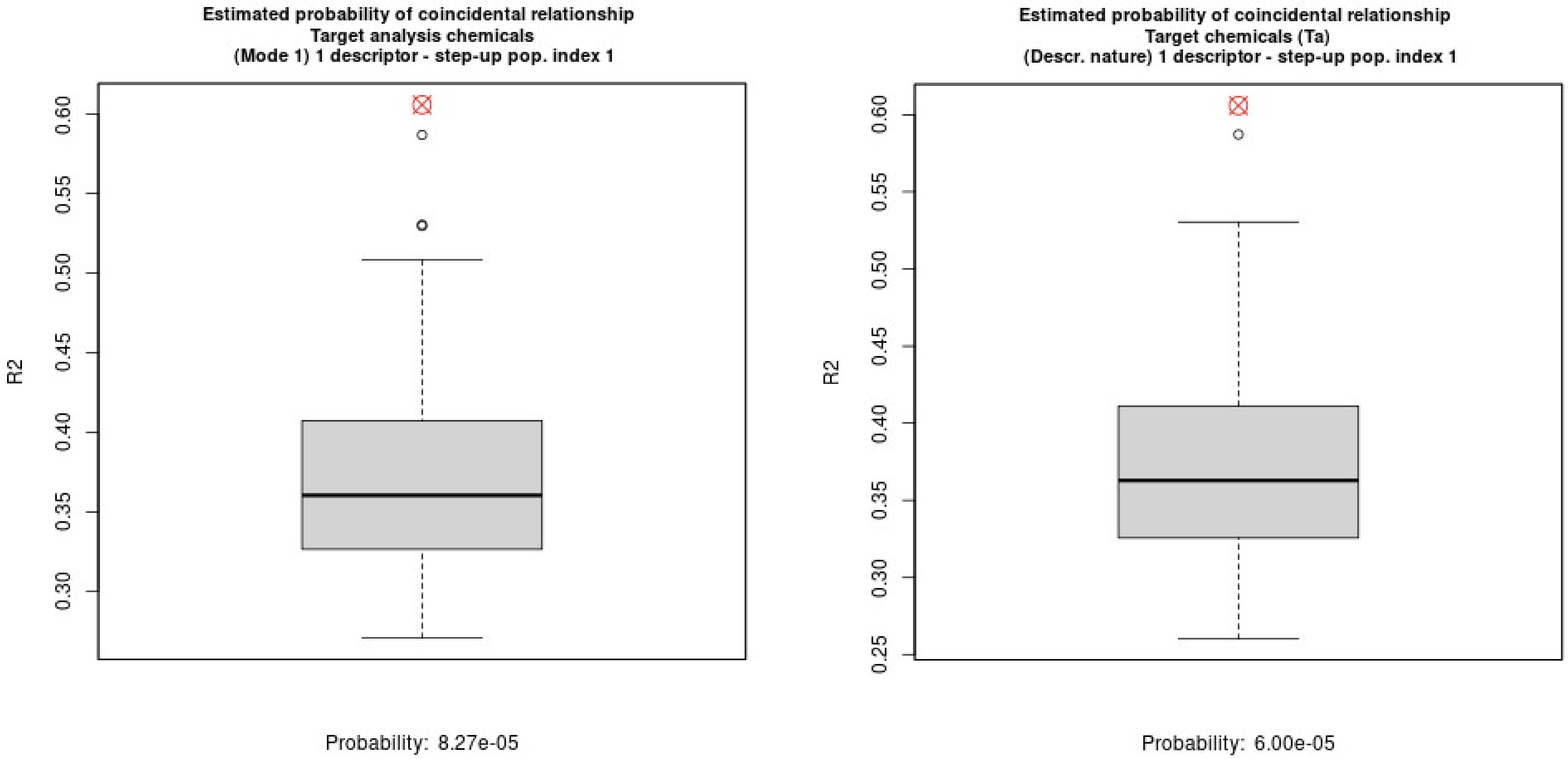


Figure S1. Probability of coincidental relationships of the target chemicals QSPR (Ta dataset).

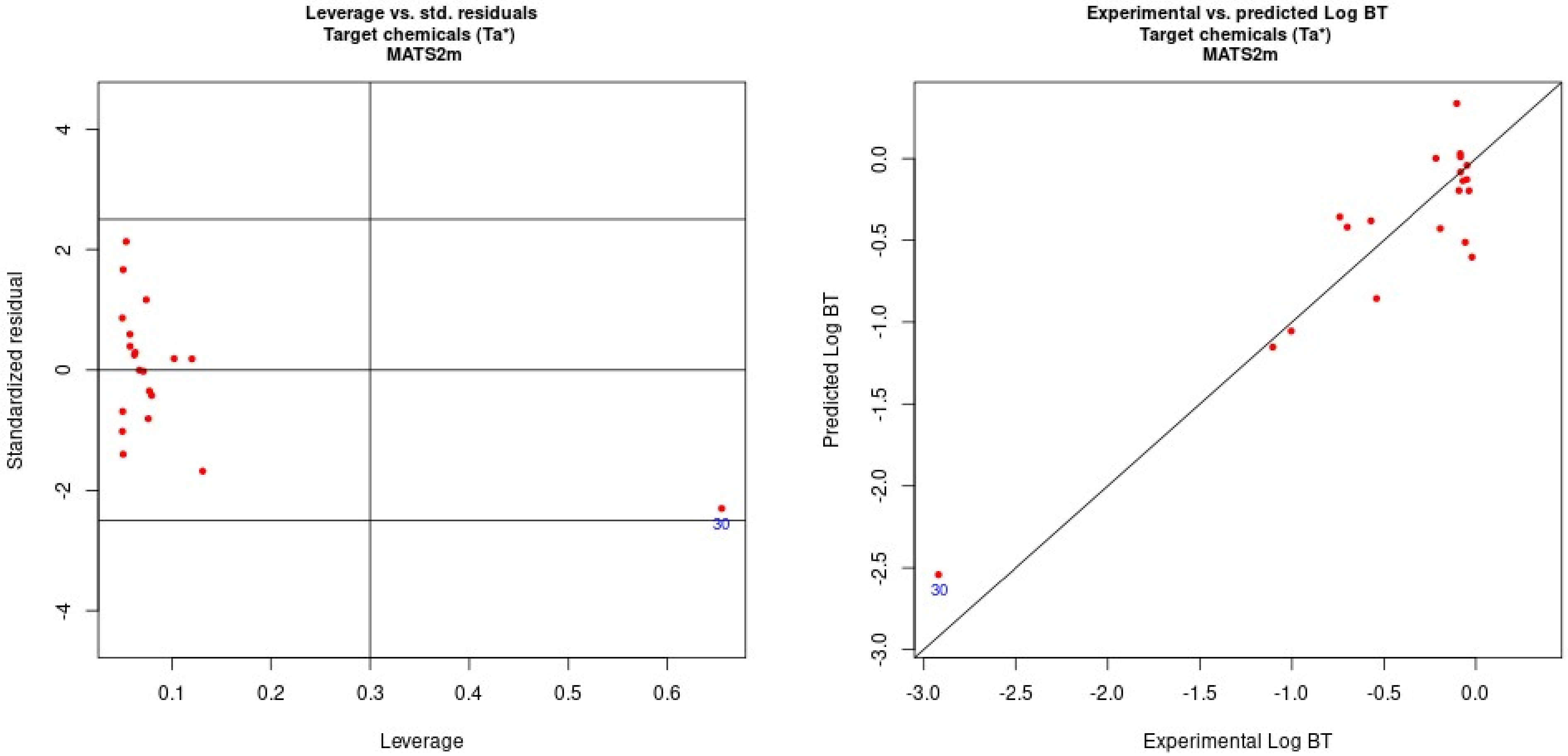


Figure S2. Performances of the target chemicals QSPR (Ta* dataset).

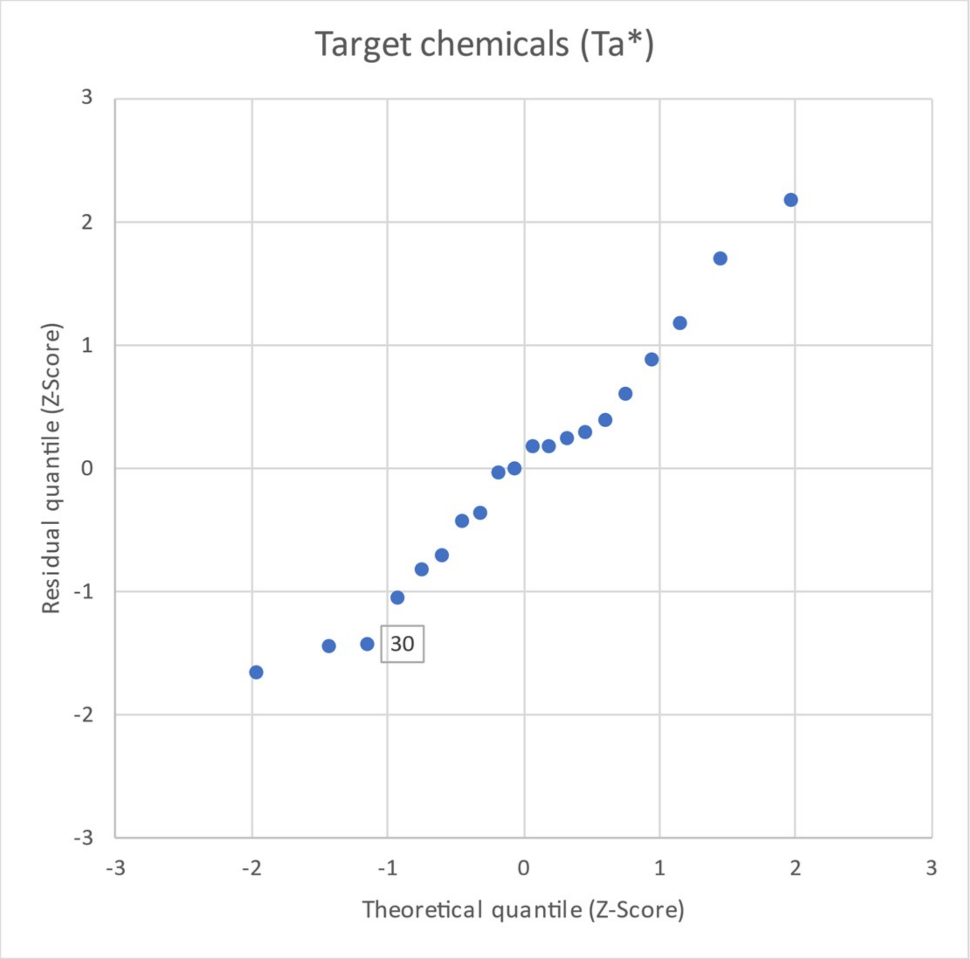
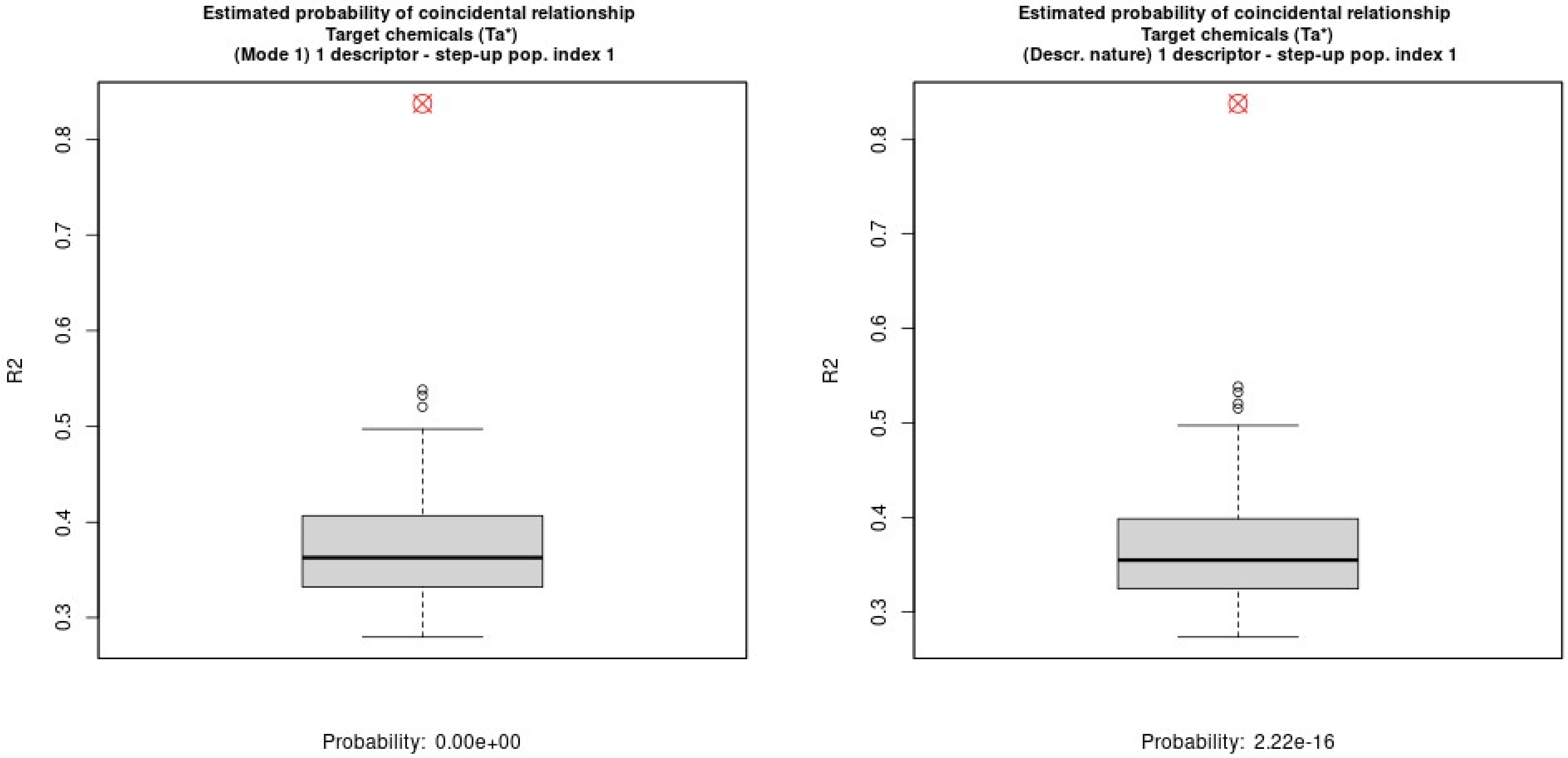


Figure S2. QQ chart of the target chemicals QSPR (Ta* dataset).



Estimated probability of coincidental relationship
Target chemicals (Ta*)
(Mode 1) 1 descriptor - step-up pop. index 1

Estimated probability of coincidental relationship
Target chemicals (Ta*)
(Descr. nature) 1 descriptor - step-up pop. index 1

Figure S2. Probability of coincidental relationships of the target chemicals, by removing endpoint outliers (Ta* dataset).

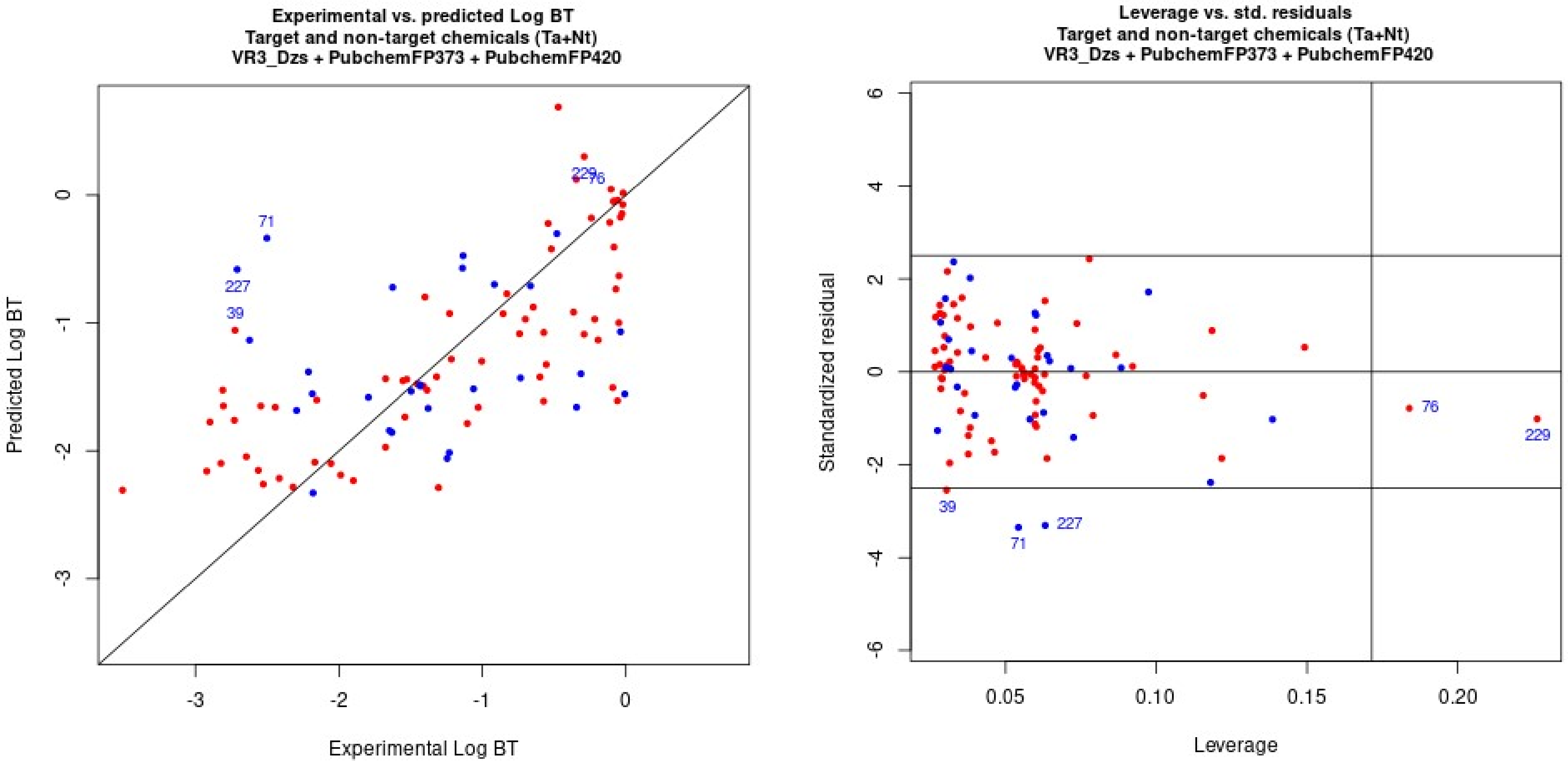
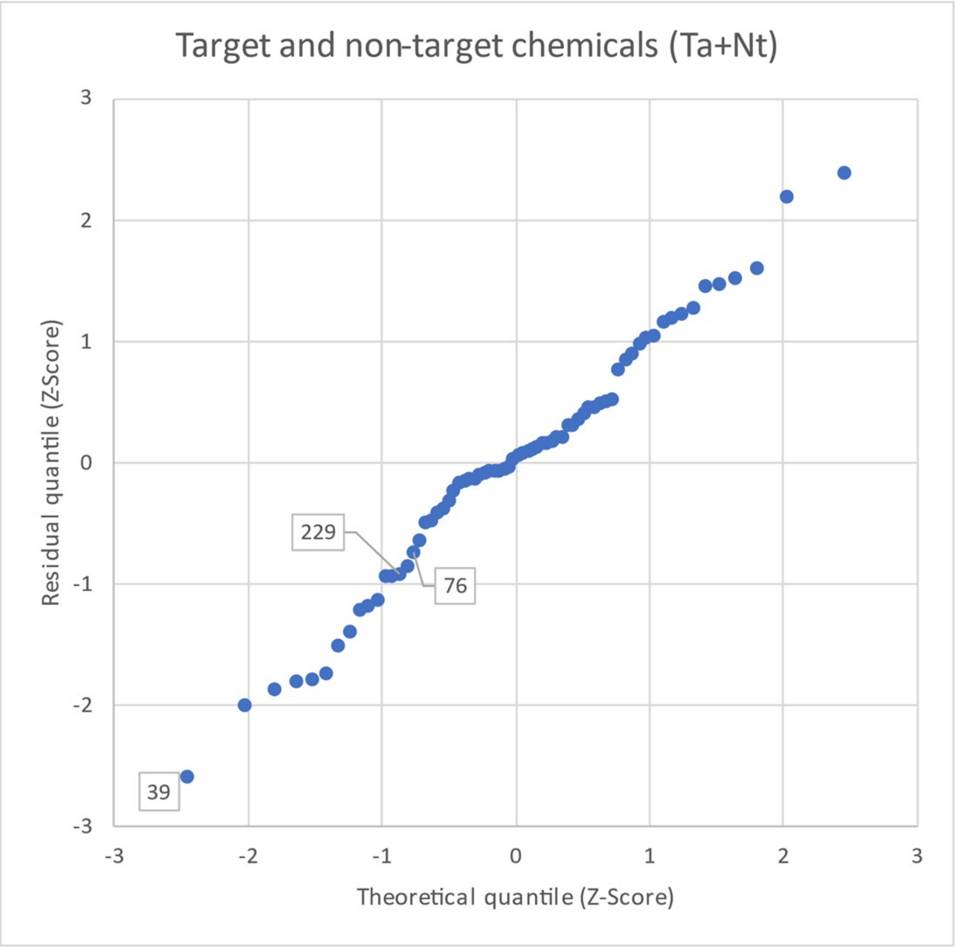


Figure S3. Performances of the target and non-target chemicals QSPR (Ta+Nt dataset).


Target and non-target chemicals (Ta+Nt)
Residual quantile (Z-Score)
3
2
1
0
-1
-2
-3
Theoretical quantile (Z-Score)
-3
-2
-1
0
1
2
3
229
76
39

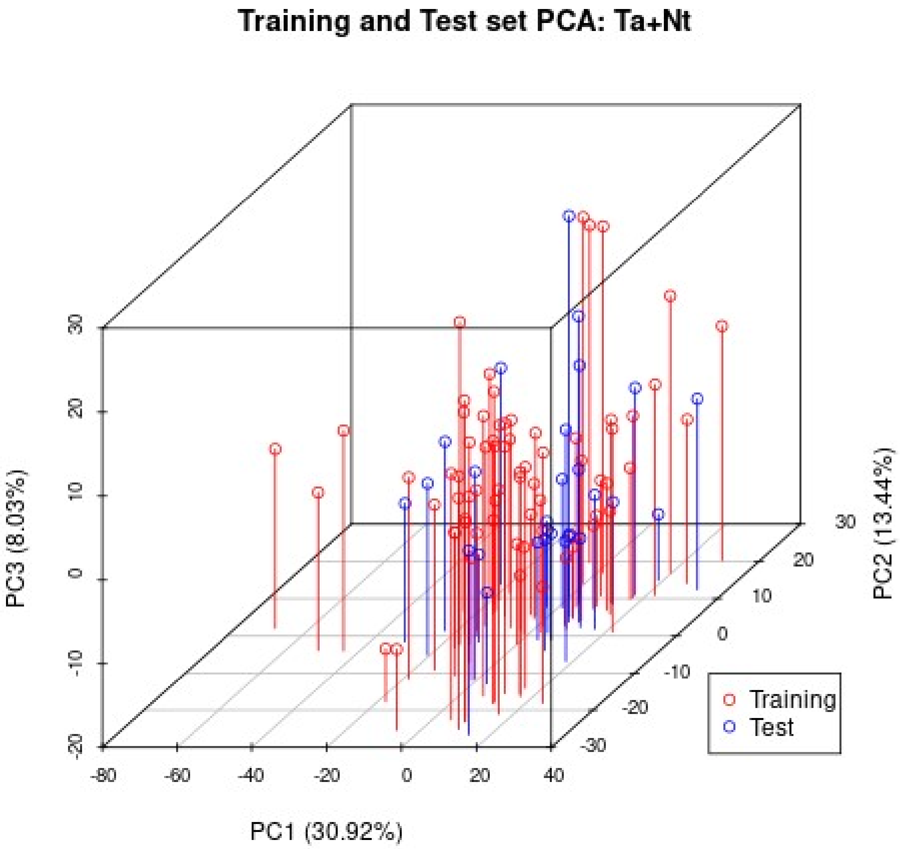
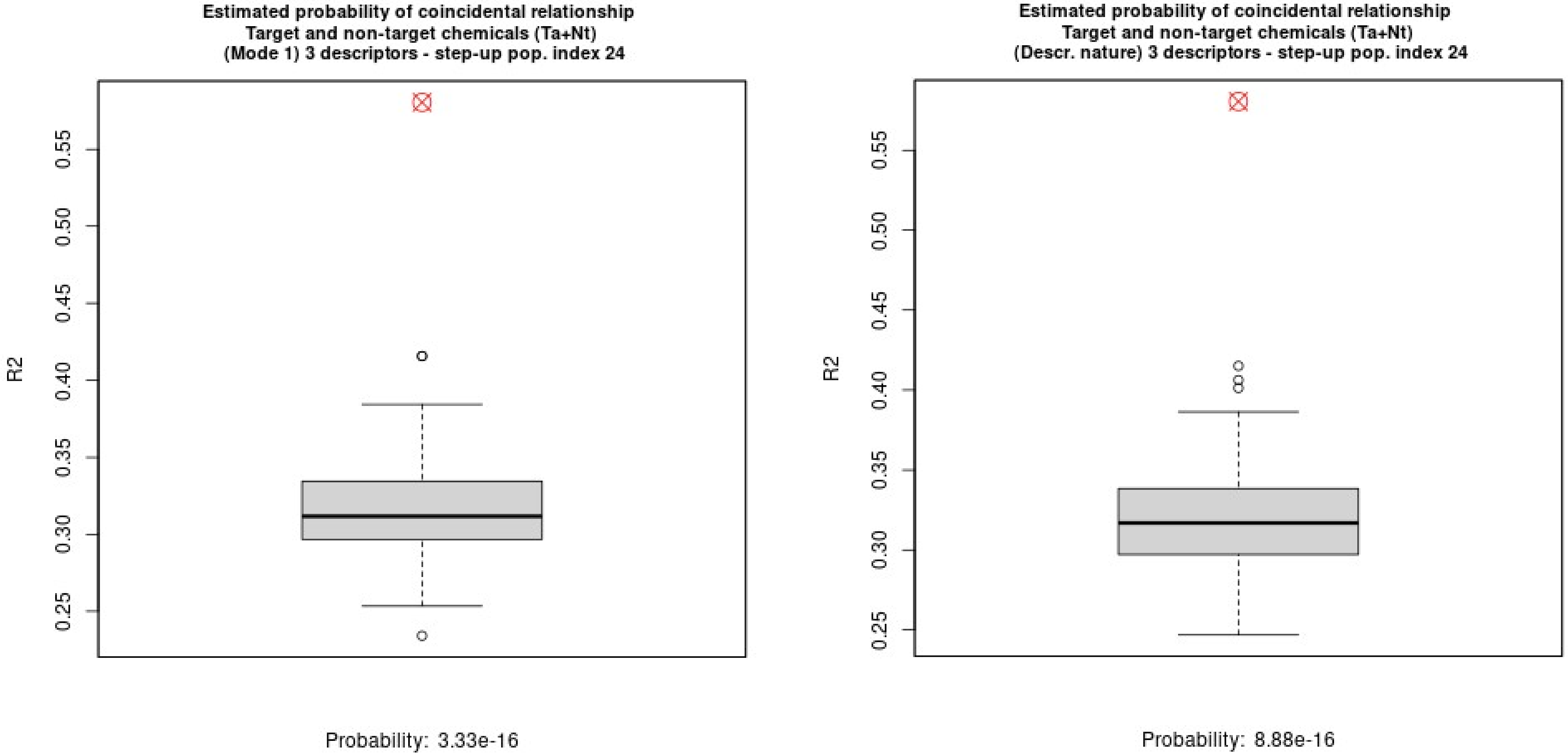

Training and Test set PCA: Ta+Nt
PC3 (8.03%)
30
20
10
0
-10
-20
PC1 (30.92%)
-80
-60
-40
-20
0
20
40
PC2 (13.44%)
30
20
10
0
-10
-20
-30
○ Training
○ Test

Figure S3. QQ chart of the target and non-target chemicals QSPR (Ta+Nt dataset).



Estimated probability of coincidental relationship
Target and non-target chemicals (Ta+Nt)
(Mode 1) 3 descriptors - step-up pop. index 24

Estimated probability of coincidental relationship
Target and non-target chemicals (Ta+Nt)
(Descr. nature) 3 descriptors - step-up pop. index 24

Figure S3. Probability of coincidental relationships of the target and non-target chemicals QSPR (Ta+Nt dataset).

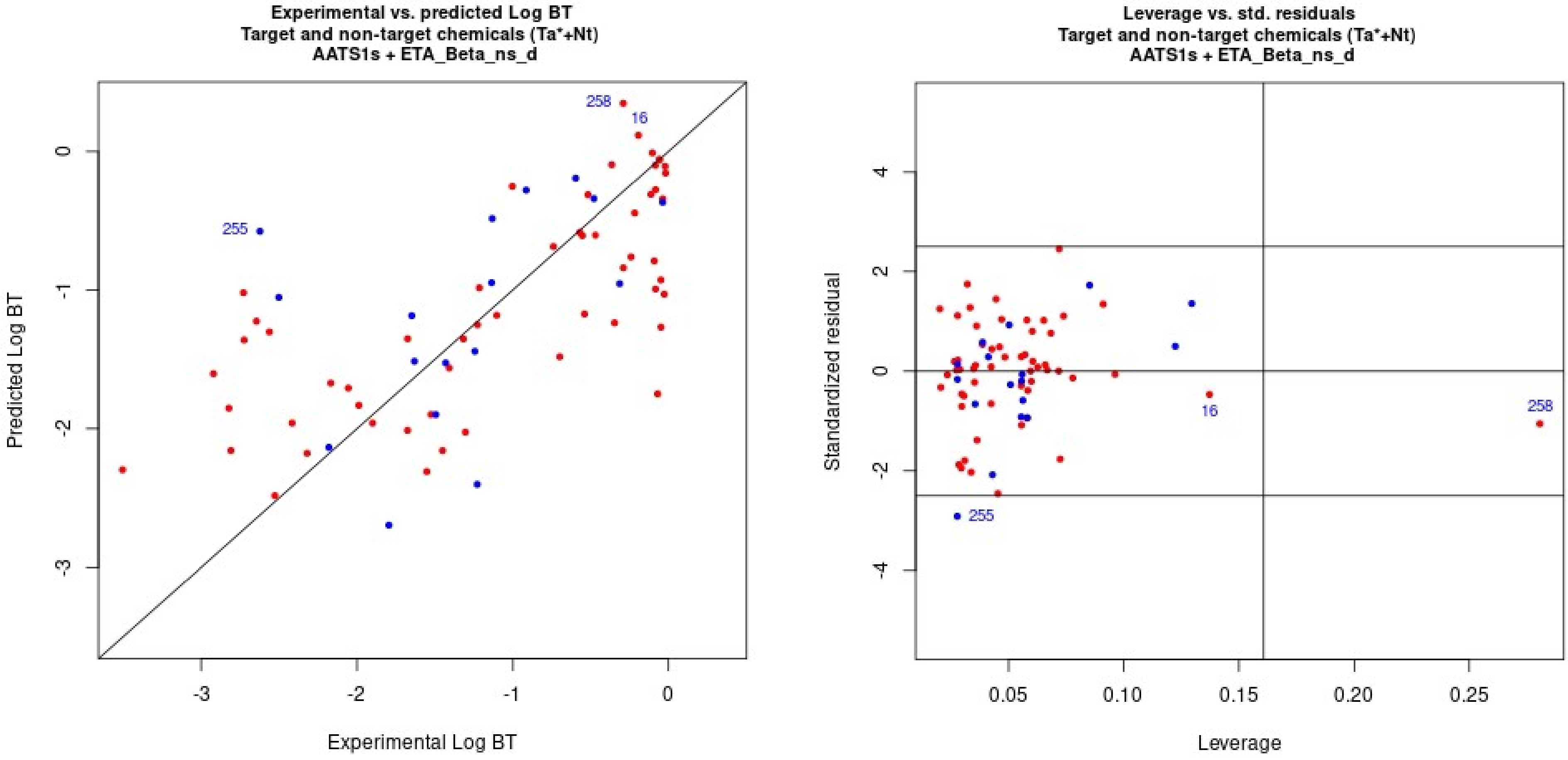


Figure S4. Performances of the target chemicals, by removing endpoint outliers, and non-target chemicals QSPR (Ta*+Nt dataset).

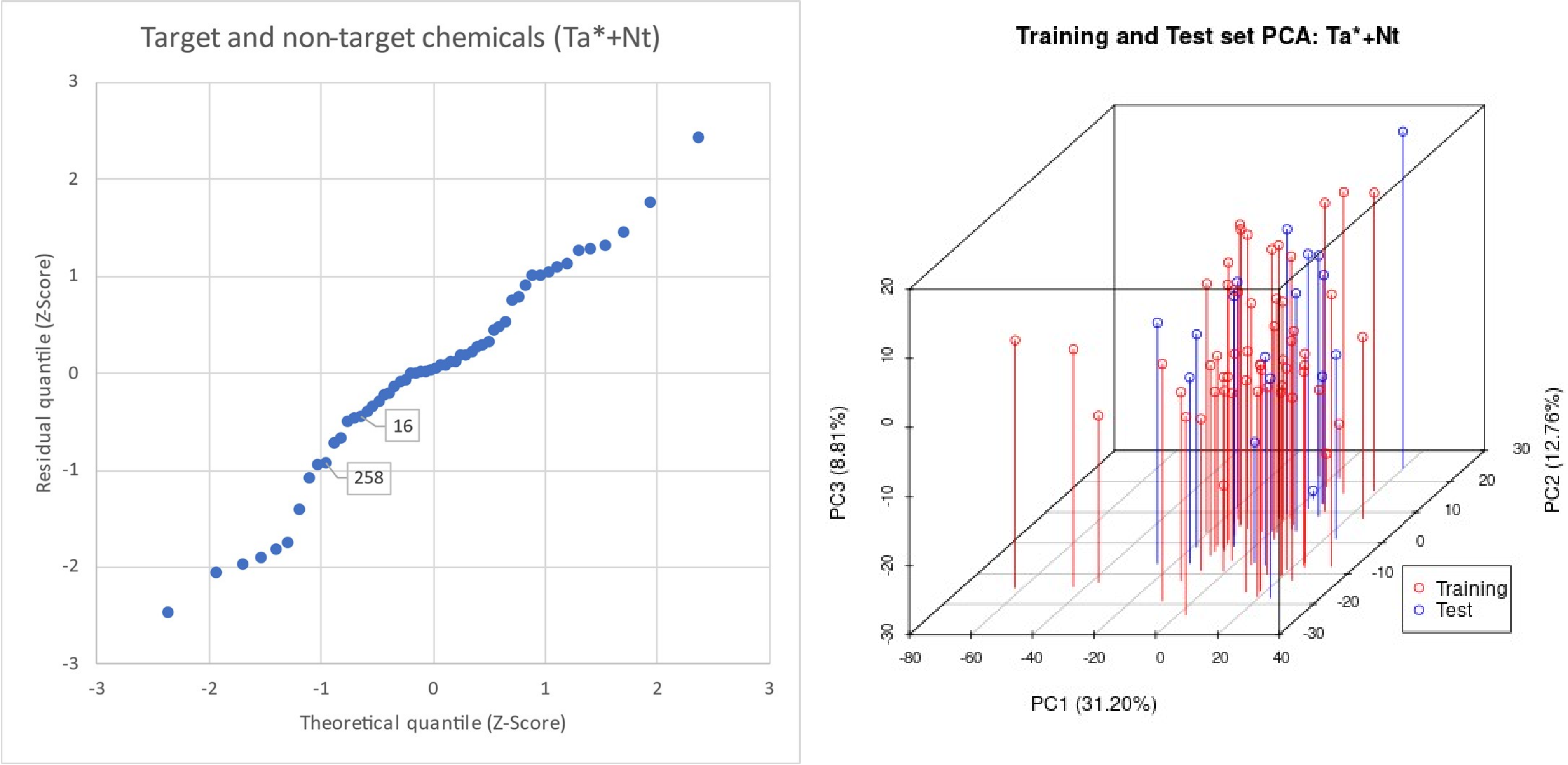


Figure S4. QQ chart and PCA for structural AD of training and test chemical of the target chemicals, by removing endpoint outliers, and non-target chemicals QSPR (Ta*+Nt dataset).

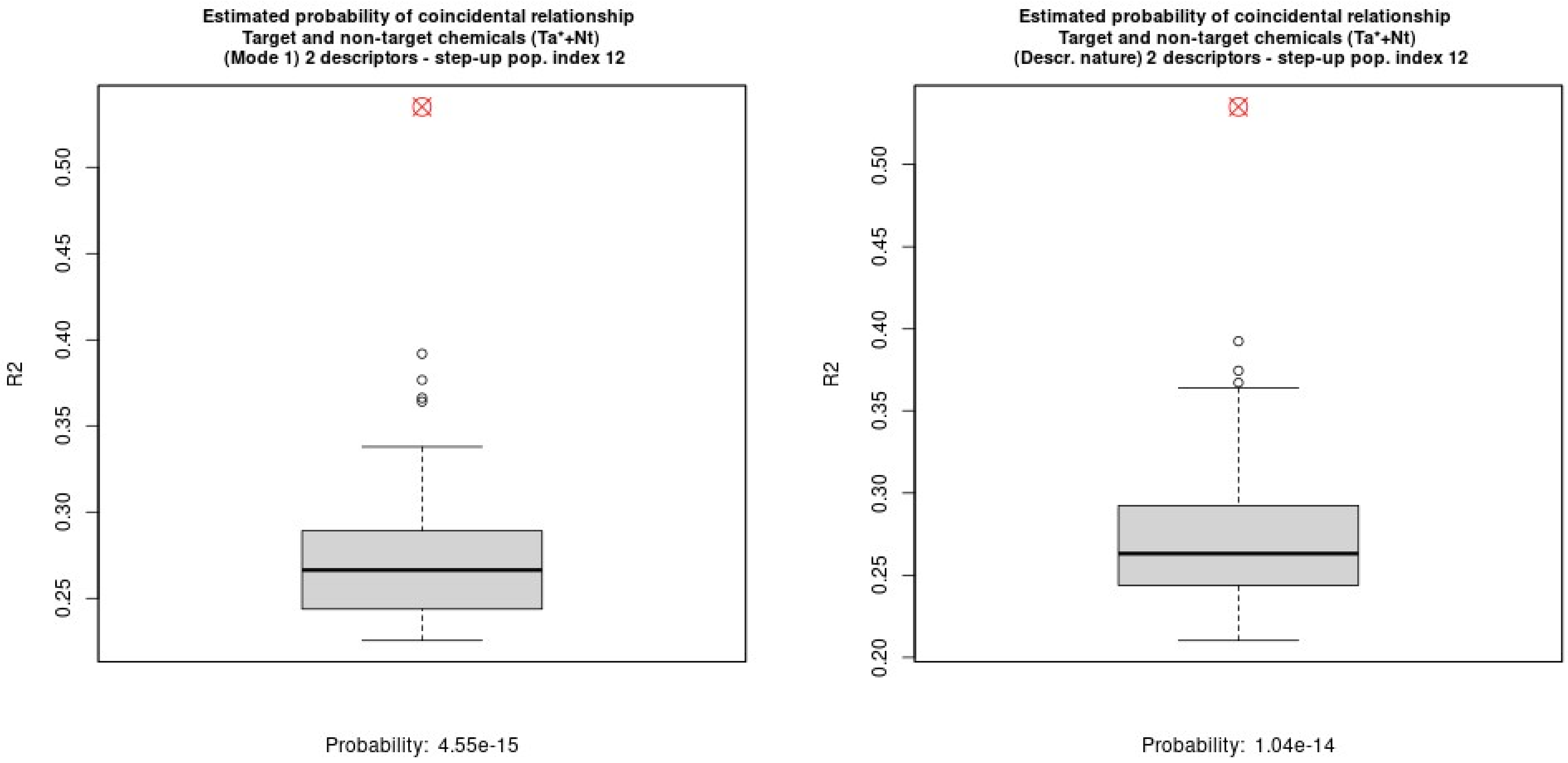


Figure S4. Probability of coincidental relationships of the target chemicals, by removing endpoint outliers, and non-target chemicals QSPR (Ta*+Nt dataset).

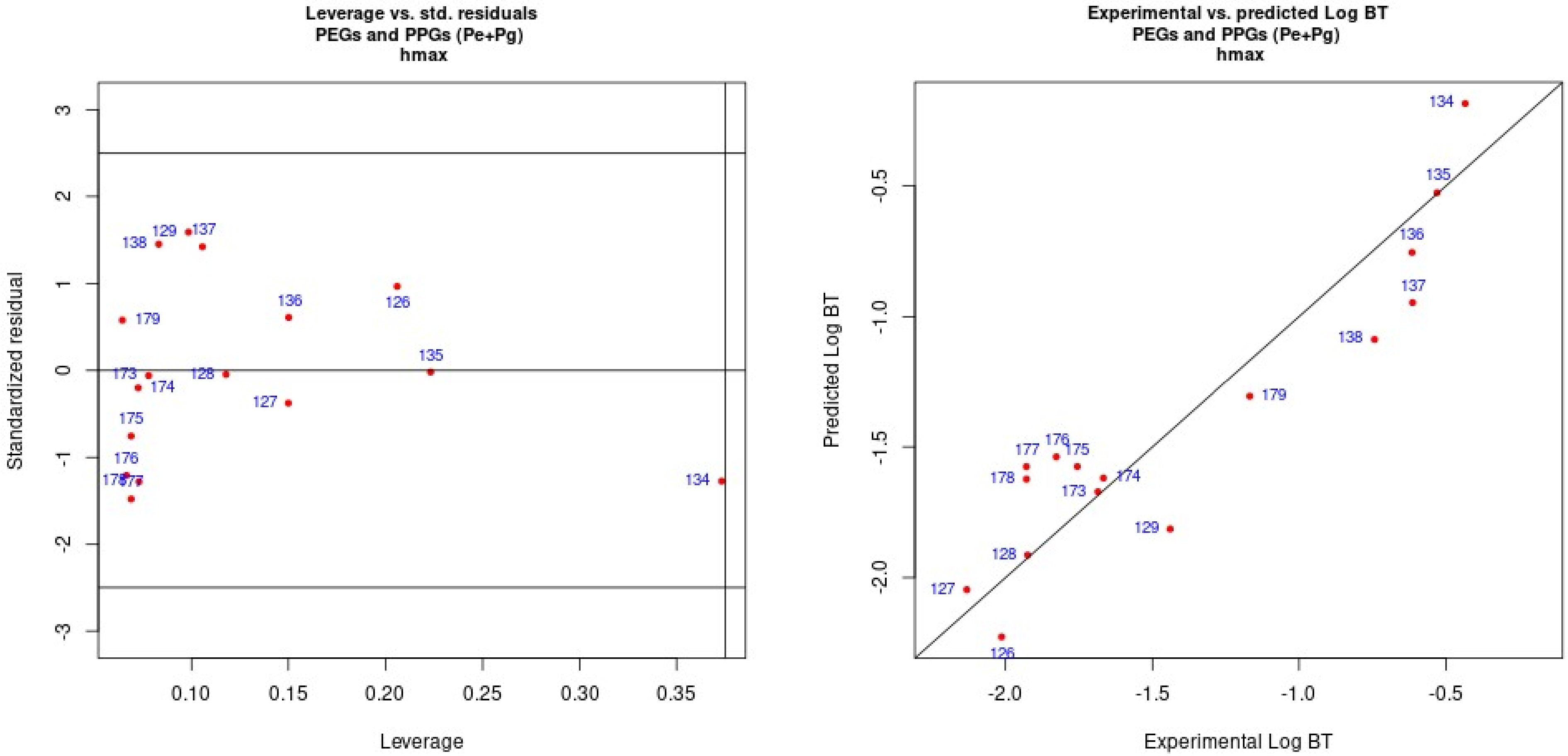
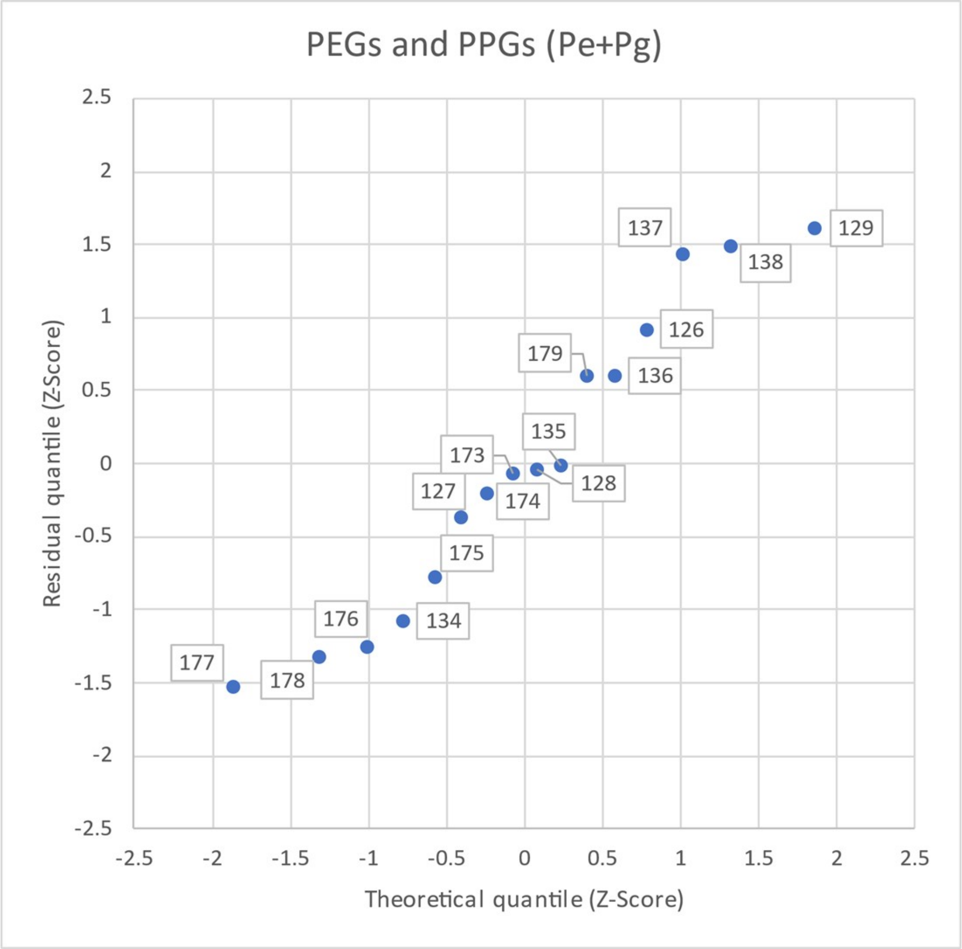


Figure S5. Performances of the PEGs and PPGs QSPR (Pe+Pg dataset).


PEGs and PPGs (Pe+Pg)
Residual quantile (Z-Score)
Theoretical quantile (Z-Score)
2.5
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-2.5
-2.5
-2
-1.5
-1
-0.5
0
0.5
1
1.5
2
2.5
137
129
138
126
179
136
135
173
128
127
174
175
176
134
177
178

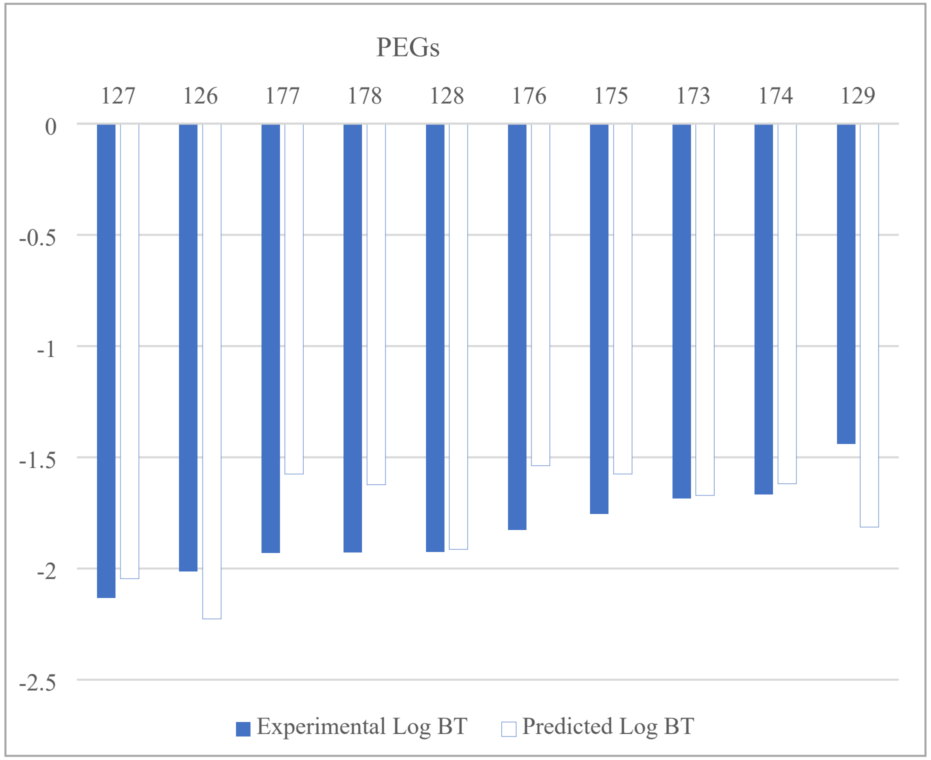

PEGs
127
126
177
178
128
176
175
173
174
129
0
-0.5
-1
-1.5
-2
-2.5
■ Experimental Log BT
□ Predicted Log BT

Figure S5. QQ chart of the PEGs and PPGs QSPR and Log BT distribution (Pe + Pg dataset).

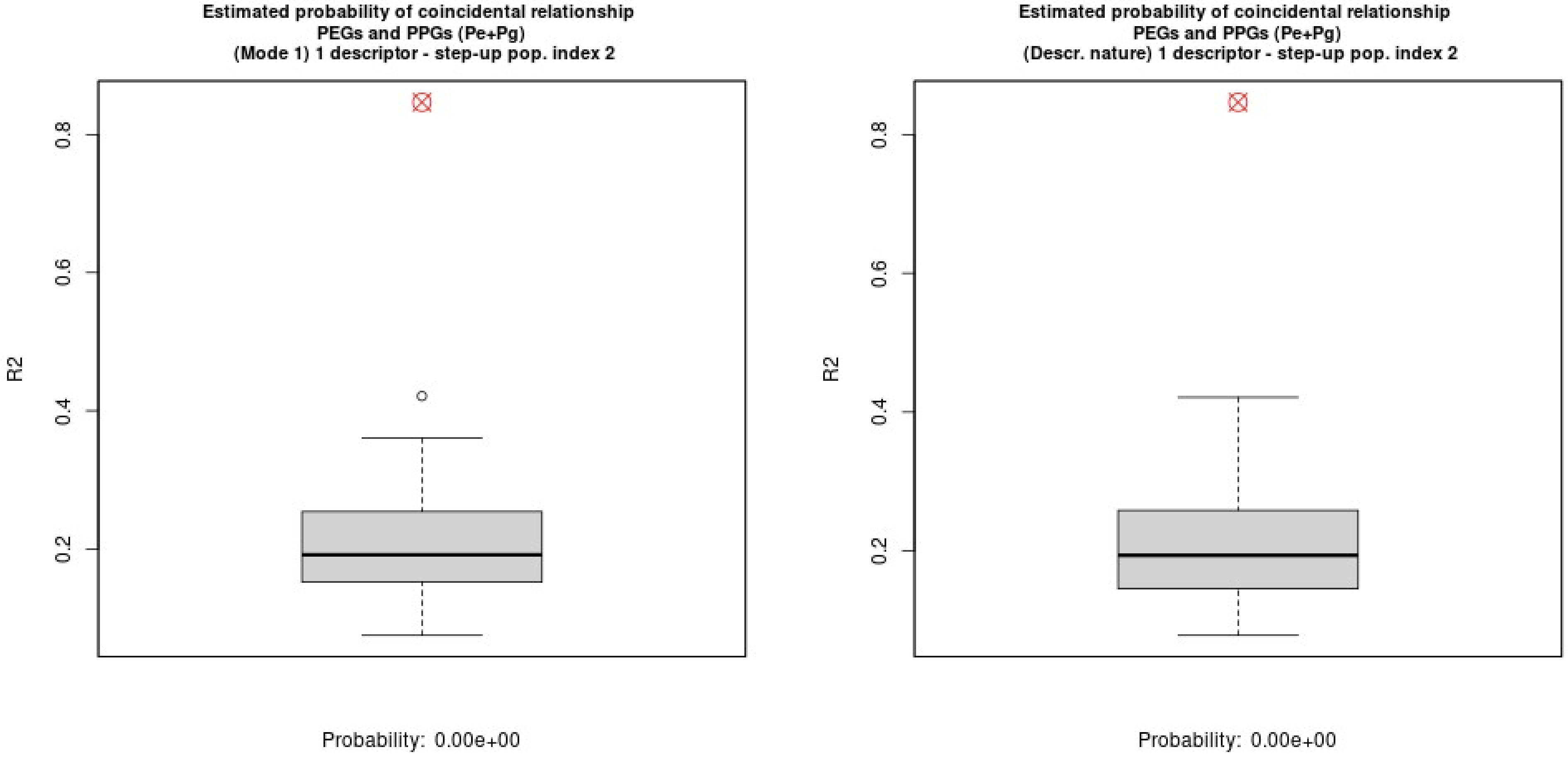


Figure S5. Probability of coincidental relationships of the PEGs and PPGs QSPR (Pe+Pg dataset).

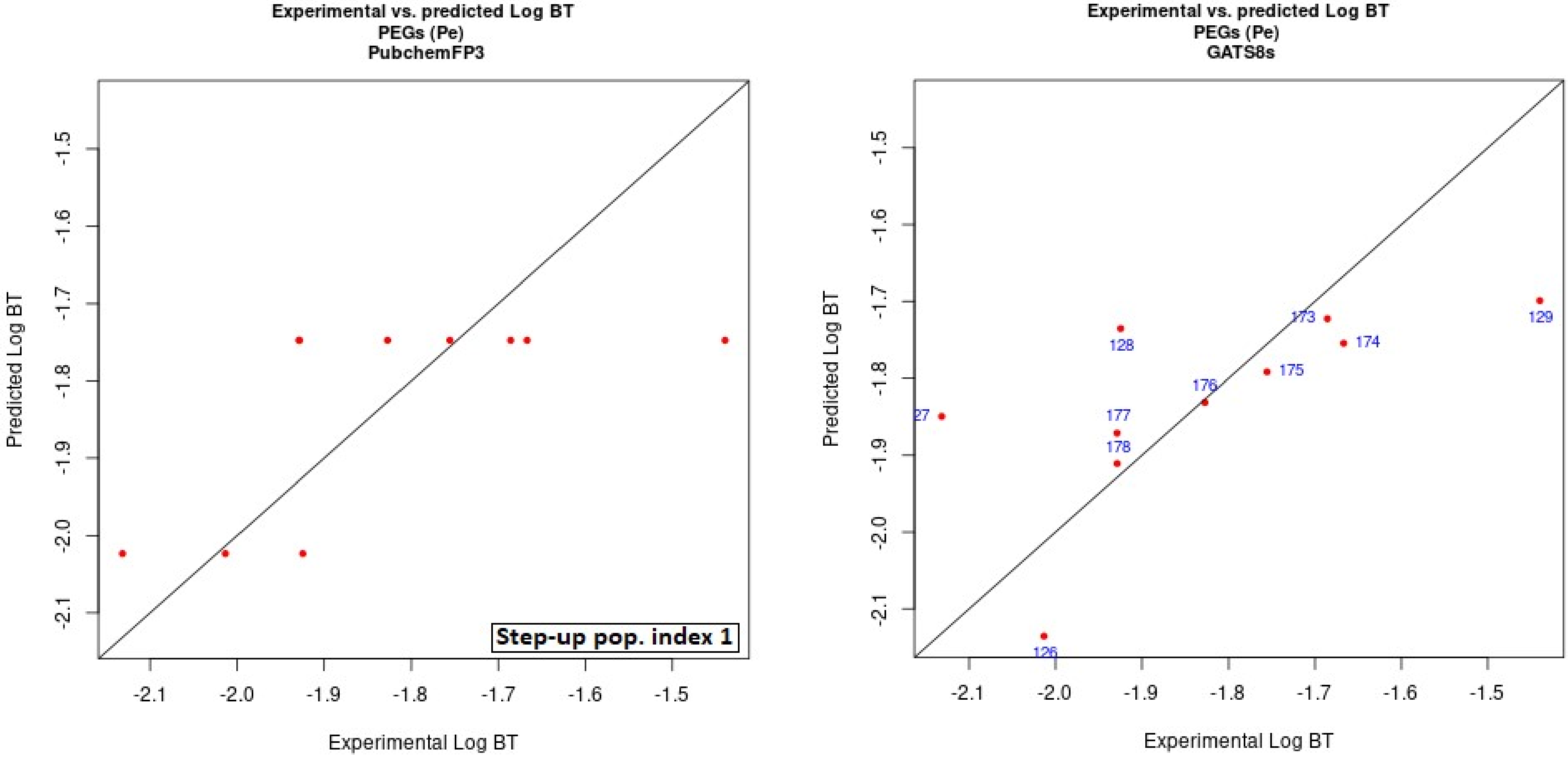


Figure S6. Performances of the PEGs QSPRs (Pe dataset).

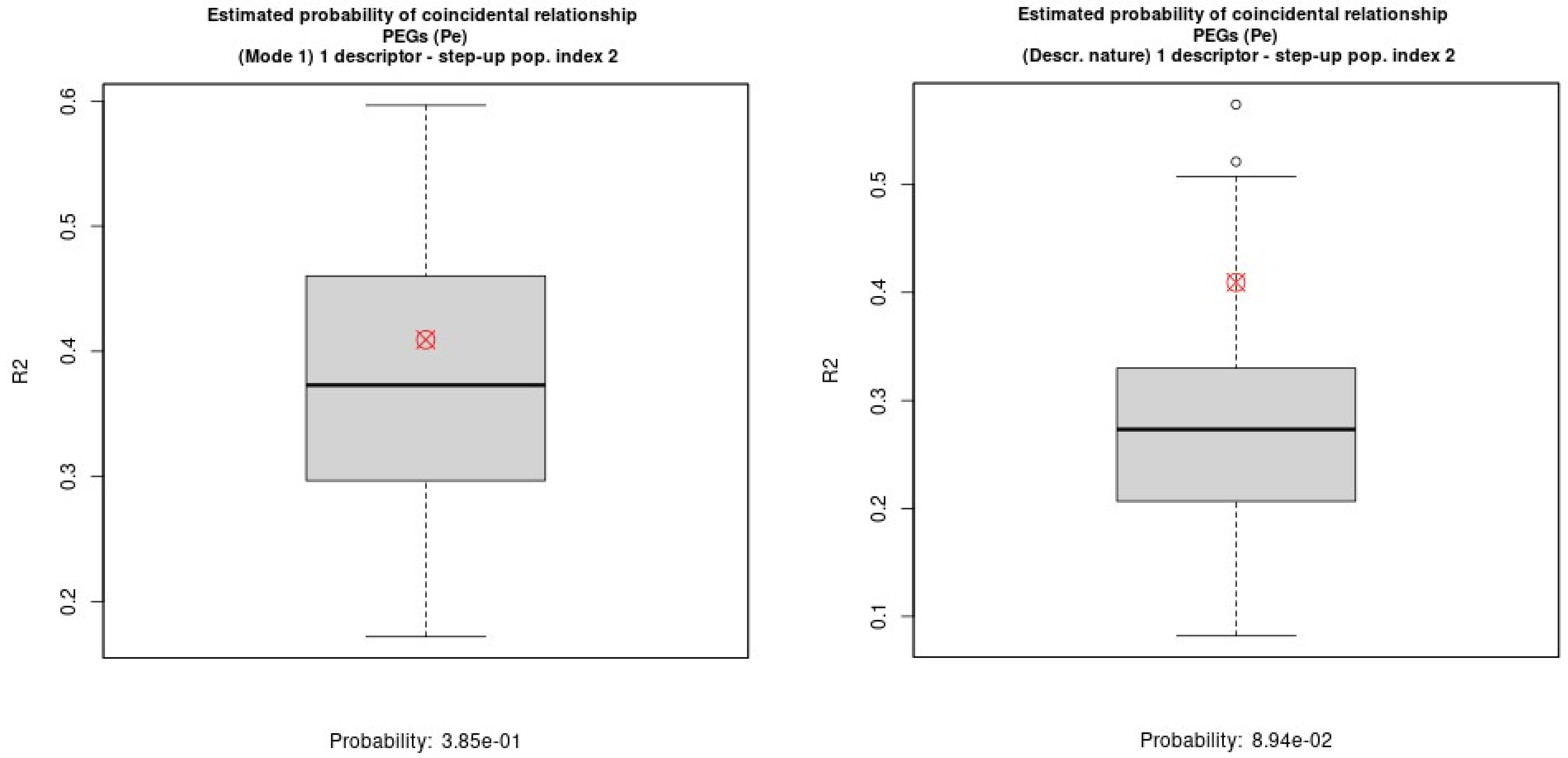


Figure S6. Probability of coincidental relationships of the PEGs selected QSPR (Pe dataset).

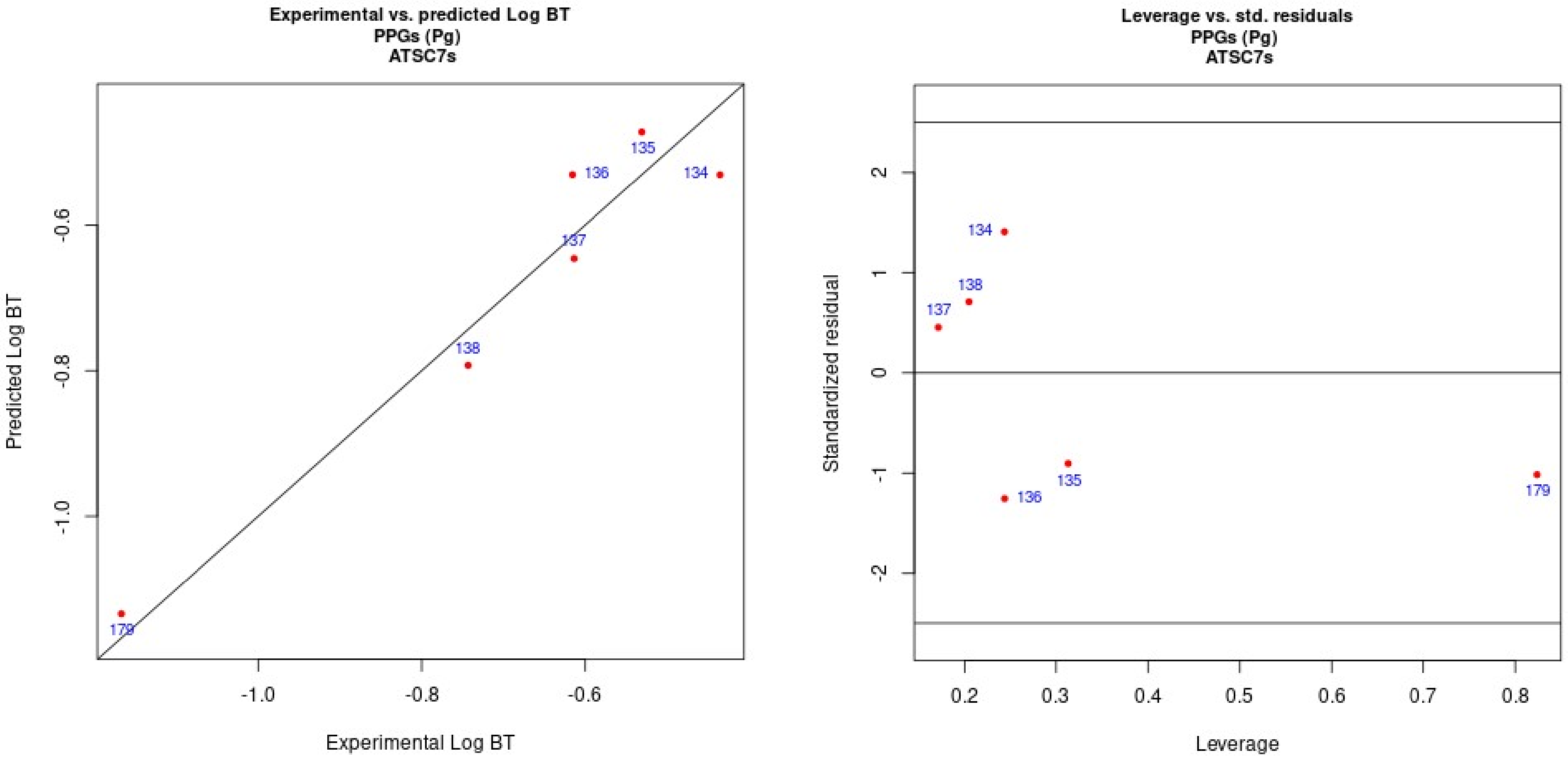


Figure S7. Performances and probability of coincidental relationships of the PPGs QSPR (Pg dataset).

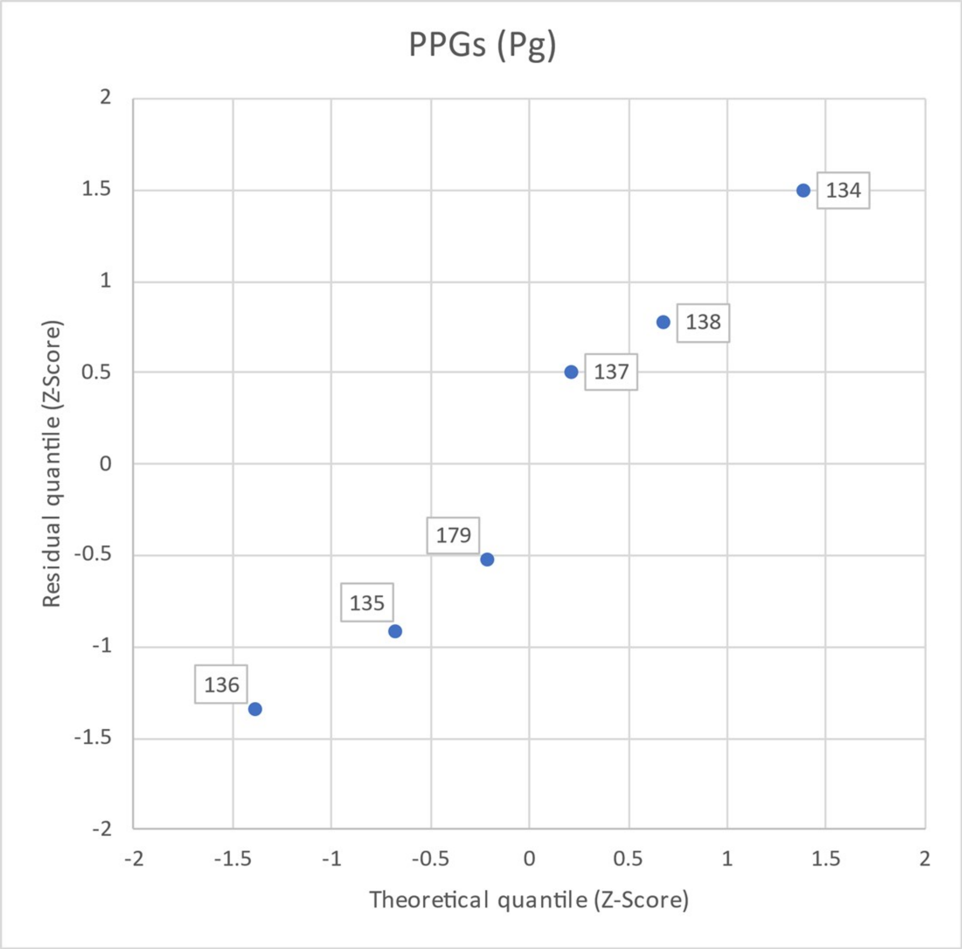


Figure S7. QQ chart of the PPGs QSPR (Pg dataset).

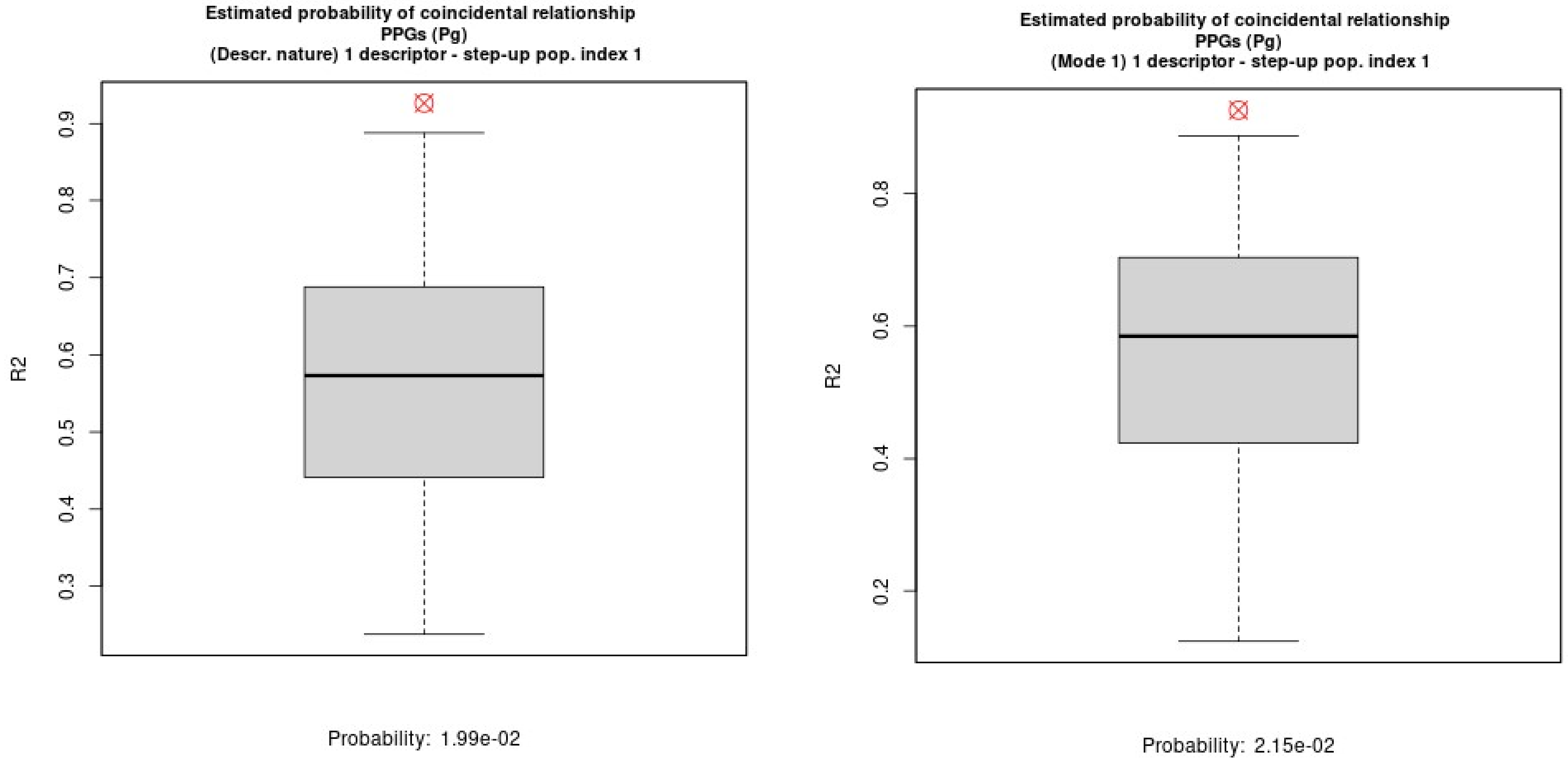


Figure S7. Performances and probability of coincidental relationships of the PPGs QSPR (Pg dataset).

Statistics

```
Residuals:
    Min      1Q  Median      3Q     Max 
-1.3885 -0.1321  0.1390  0.2377  0.7024 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -0.4400     0.1083  -4.062 0.000608 ***
MATS2m       -6.1773     1.1140  -5.545 1.99e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5001 on 20 degrees of freedom
Multiple R-squared:  0.6059,  Adjusted R-squared:  0.5862 
F-statistic: 30.75 on 1 and 20 DF,  p-value: 1.989e-05


-----------------------
 ADDITIONAL STATISTICS
-----------------------

MAE training: 0.3359
MAE bootstrap: 0.6782±0.0102

Q2: 0.4836
Y-scrambled R2: 0.04742

---------
 DATASET
---------

Filtered descriptors: 491
Training set: 22
```

Statistics S1. Performances the target chemicals QSPR (Ta dataset).

```
Residuals:
     Min       1Q   Median       3Q      Max 
-0.43975 -0.19601  0.02392  0.12064  0.58269 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.30779    0.06425  -4.790 0.000147 ***
MATS2m      -6.45112    0.66984  -9.631 1.59e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2811 on 18 degrees of freedom
Multiple R-squared:  0.8375,	Adjusted R-squared:  0.8284 
F-statistic: 92.75 on 1 and 18 DF,  p-value: 1.588e-08

-----------------------
 ADDITIONAL STATISTICS
-----------------------

MAE training: 0.2104
MAE bootstrap: 0.4389±0.01051

Q2: 0.6935
Y-Scrambled R2: 0.04037

---------
 DATASET
---------

Filtered descriptors: 529
Training set: 20
```

Statistics S2. Performances of the target chemicals, by removing endpoint outliers, QSPR (Ta* dataset).

```
Residuals:
     Min       1Q   Median       3Q      Max 
-1.66713 -0.31507  0.03371  0.33011  1.55018 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -1.18414    0.23608  -5.016 4.23e-06 ***
VR3_Dzs       0.11155    0.02097   5.319 1.34e-06 ***
PubchemFP373 -0.62832    0.19197  -3.273   0.0017 ** 
PubchemFP420 -1.01904    0.18236  -5.588 4.71e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6651 on 66 degrees of freedom
Multiple R-squared:  0.5805,	Adjusted R-squared:  0.5614 
F-statistic: 30.44 on 3 and 66 DF,  p-value: 1.794e-12


-------------------------
 DESCRIPTORS CORRELATION 
-------------------------



             VR3_Dzs PubchemFP373 PubchemFP420
VR3_Dzs        1.000        -0.18       -0.062
PubchemFP373  -0.180         1.00        0.200
PubchemFP420  -0.062         0.20        1.000


-----------------------
 ADDITIONAL STATISTICS 
-----------------------

MAE training: 0.4879
MAE bootstrap: 0.7457±0.0033
MAE test: 0.6921

Y-scrambled R2: 0.04461
Q2: 0.5303

Standardized coefficients:
VR3_Dzs: 0.4314
PubchemFP373: -0.2702
PubchemFP420: -0.4545


---------
 DATASET 
---------

Filtered descriptors: 517
Training set: 70
Test set: 28
```

Statistics S3. Performances of the target and non-target chemicals QSPR (Ta+Nt dataset).

```
Residuals:
     Min       1Q   Median       3Q      Max 
-1.70780 -0.31362  0.03023  0.41034  1.68073 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)     1.7916     0.6444   2.780   0.0075 ** 
AATS1s         -0.9174     0.1759  -5.216 3.09e-06 ***
ETA_Beta_ns_d   0.9881     0.2140   4.617 2.51e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7119 on 53 degrees of freedom
Multiple R-squared:  0.535,   Adjusted R-squared:  0.5175 
F-statistic: 30.49 on 2 and 53 DF,  p-value: 1.538e-09


-------------------------
 DESCRIPTORS CORRELATION
-------------------------


              AATS1s ETA_Beta_ns_d
AATS1s          1.00         -0.21
ETA_Beta_ns_d  -0.21          1.00


-----------------------
 ADDITIONAL STATISTICS
-----------------------

MAE training: 0.5145
MAE bootstrap: 0.7502±0.0031
MAE test: 0.581

Y-Scrambled R2: 0.04709
Q2: 0.482

Standardized coefficients:
AATS1s: -0.4992
ETA_Beta_ns_d: 0.4419


---------
 DATASET
---------

Filtered descriptors: 538
Training set: 56
Test set: 17
```

Statistics S4. Performances of the target chemicals, by removing endpoint outliers, and non-target chemicals QSPR (Ta*+Nt dataset).

```
Residuals:
     Min       1Q   Median       3Q      Max 
-0.35346 -0.19784 -0.01252  0.15756  0.37396 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   17.629      2.163   8.149 1.10e-06 ***
hmax         -27.639      3.141  -8.800 4.44e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2478 on 14 degrees of freedom
Multiple R-squared:  0.8469,	Adjusted R-squared:  0.8359 
F-statistic: 77.43 on 1 and 14 DF,  p-value: 4.442e-07


-----------------------
 ADDITIONAL STATISTICS
-----------------------

MAE training: 0.1927
MAE bootstrap: 0.2481±0.0034

Q2: 0.7987
Y-scrambled R2: 0.07912

Standardized coefficients:
hmax: -0.9203


---------
 DATASET
---------

Filtered descriptors: 19
Training set: 16
```

Statistics S5. Performances of the PEGs and PPGs QSPR (Pe+Pg dataset).

```
Residuals:
     Min       1Q   Median       3Q      Max 
-0.28239 -0.04762  0.02026  0.07508  0.25938 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)  
(Intercept)    4.992      2.899   1.722   0.1233  
GATS8s        -6.497      2.760  -2.354   0.0464 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1627 on 8 degrees of freedom
Multiple R-squared:  0.4092,	Adjusted R-squared:  0.3353 
F-statistic: 5.541 on 1 and 8 DF,  p-value: 0.0464


-----------------------
 ADDITIONAL STATISTICS
-----------------------

MAE training: 0.1094

Q2: -0.3731
Y-Scrambled R2: 0.1143

Standardized coefficients:
GATS8s: -0.6397

---------
 DATASET
---------

Filtered descriptors: 27
Training set: 10
```

Statistics S6. Performances of the PEGs QSPR (Pe dataset).

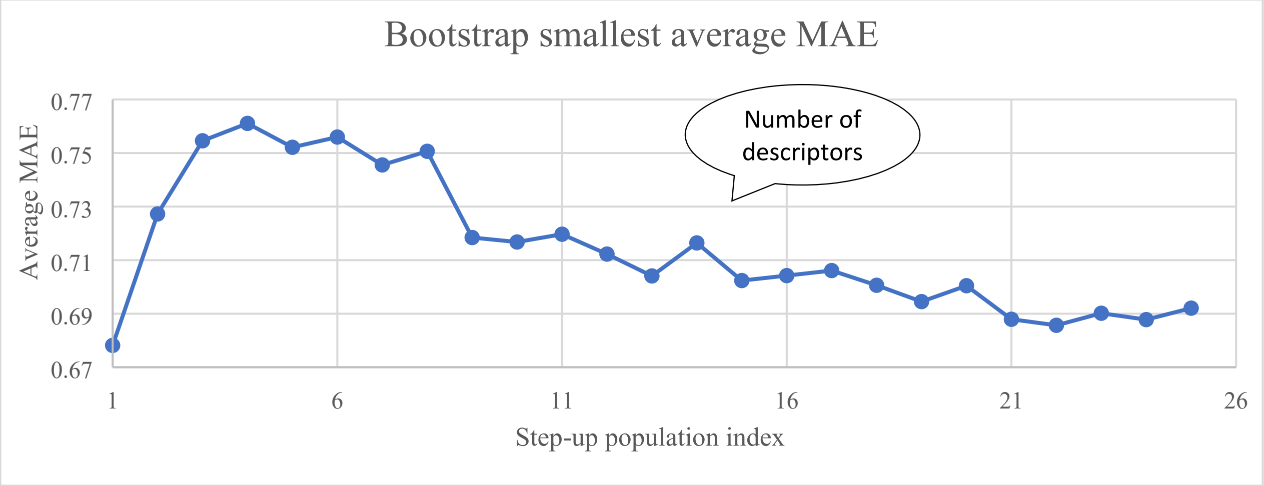
```
Residuals:
     134      135      136      137      138      179 
 0.09569 -0.05865 -0.08537  0.03228  0.04941 -0.03336 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)   
(Intercept) -0.07418    0.09168  -0.809  0.46387   
ATSC7s       0.70693    0.09959   7.099  0.00208 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.07823 on 4 degrees of freedom
Multiple R-squared:  0.9265,	Adjusted R-squared:  0.9081 
F-statistic: 50.39 on 1 and 4 DF,  p-value: 0.00208


-----------------------
 ADDITIONAL STATISTICS
-----------------------

MAE training: 0.05913

Q2: 0.7681
Y-scrambled R2: 0.1913

Standardized coefficients:
ATSC7s: 0.9625


---------
 DATASET
---------

Filtered descriptors: 10
Training set: 6
```

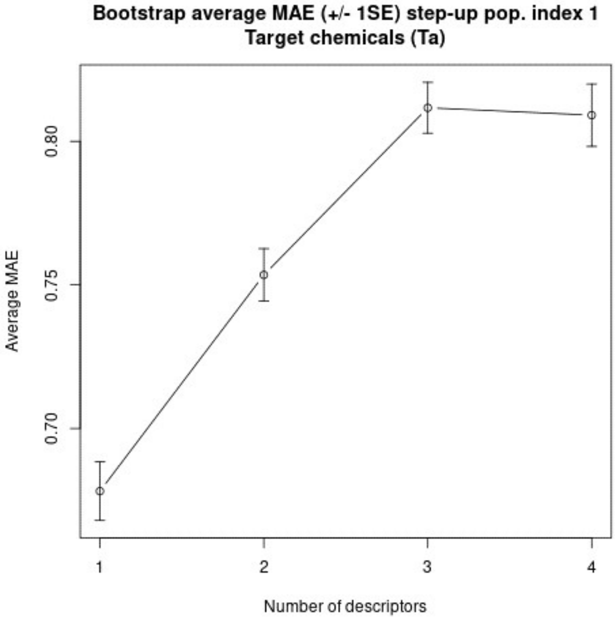
Statistics S7. Performances of the PPGs QSPR (Pg dataset).

Bootstrap analysis

Generated bootstrap folds per iteration

Run	No. folds	Run	No. folds
1	38	14	42
2	38	15	44
3	35	16	43
4	38	17	36
5	32	18	37
6	44	19	45
7	41	20	36
8	44	21	59
9	40	22	55
10	36	23	44
11	48	24	35
12	39	25	48
13	37		

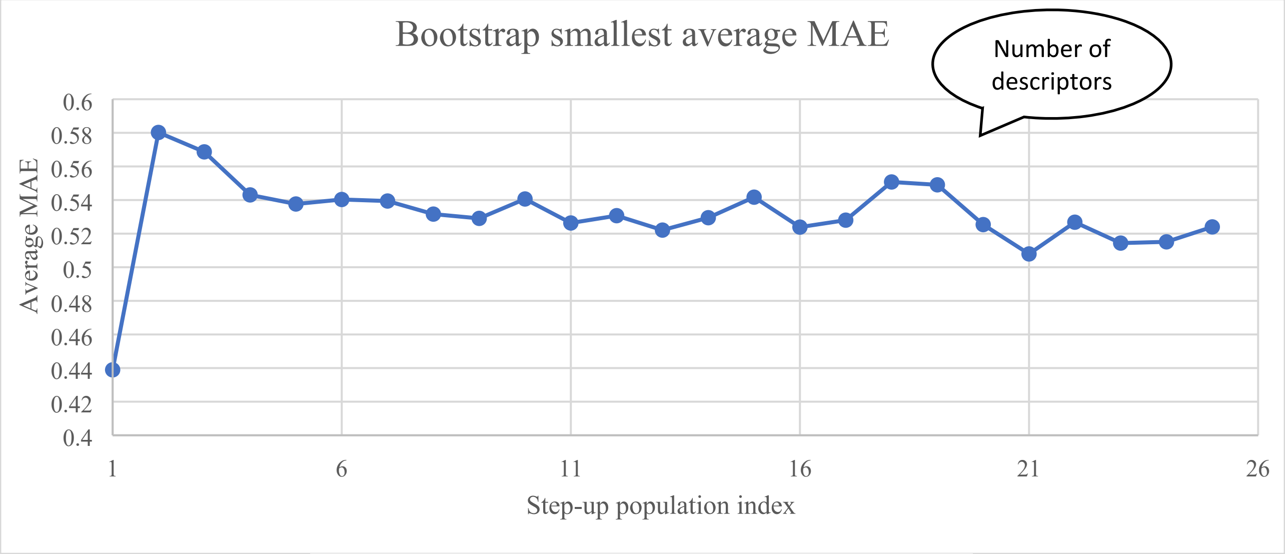


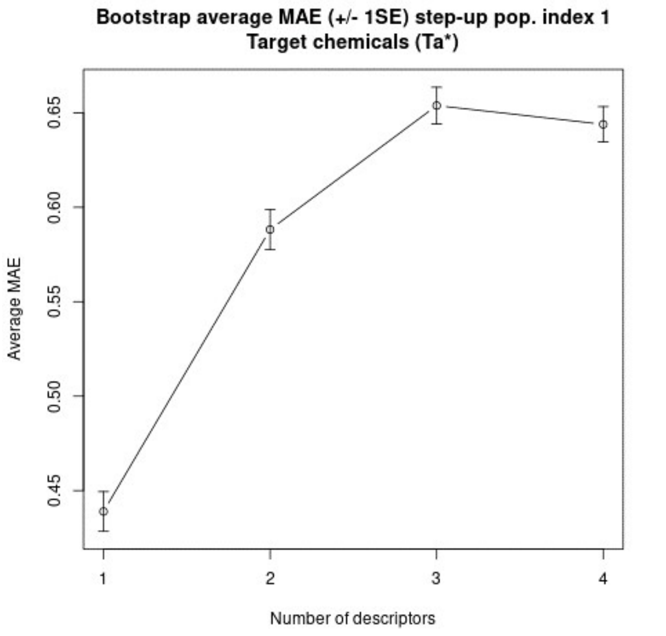


Bootstrap analysis S1. Selection of the best candidate of target chemicals QSPR (Ta dataset).

Generated bootstrap folds per iteration

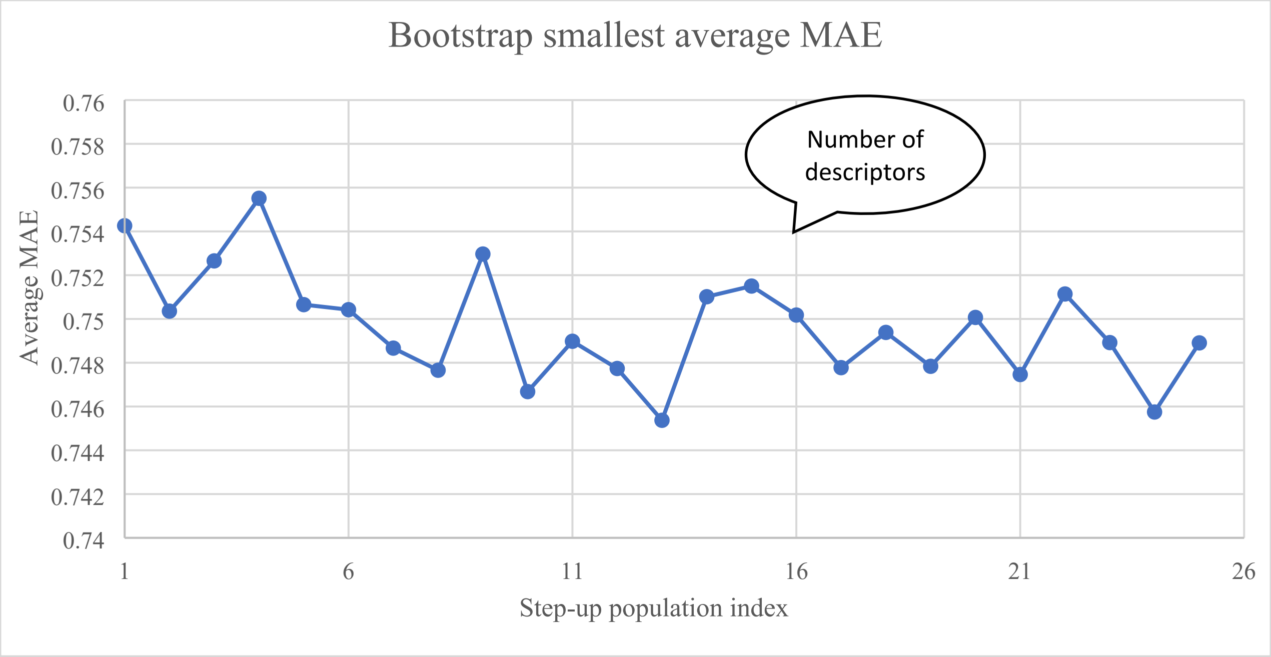
Run	No. folds	Run	No. folds
1	39	14	45
2	30	15	27
3	37	16	55
4	36	17	33
5	32	18	44
6	44	19	38
7	33	20	36
8	44	21	38
9	47	22	43
10	36	23	41
11	35	24	37
12	36	25	36
13	48		

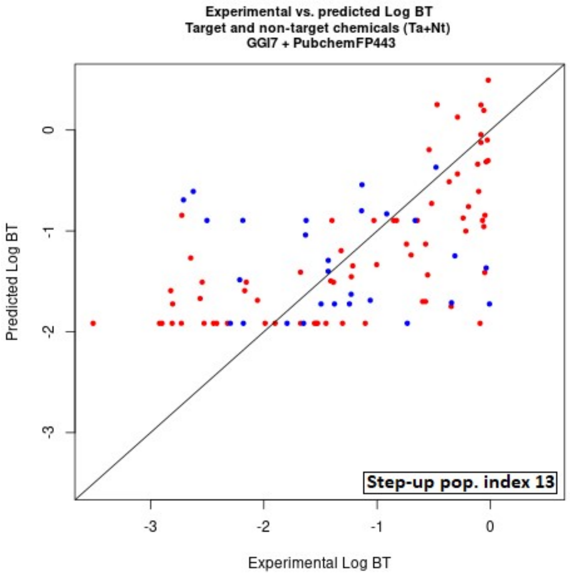


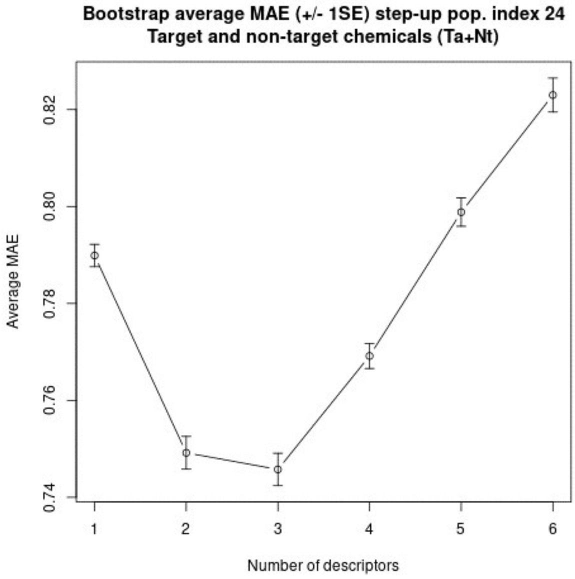


Bootstrap analysis S2. Selection of the best candidate of target chemicals, by removing endpoint outliers, QSPR (Ta* dataset).

Run	No. Folds	Run	No. Folds
1	45	14	64
2	64	15	69
3	55	16	50
4	55	17	46
5	47	18	47
6	45	19	59
7	58	20	47
8	58	21	63
9	53	22	55
10	67	23	54
11	46	24	60
12	67	25	55
13	52		

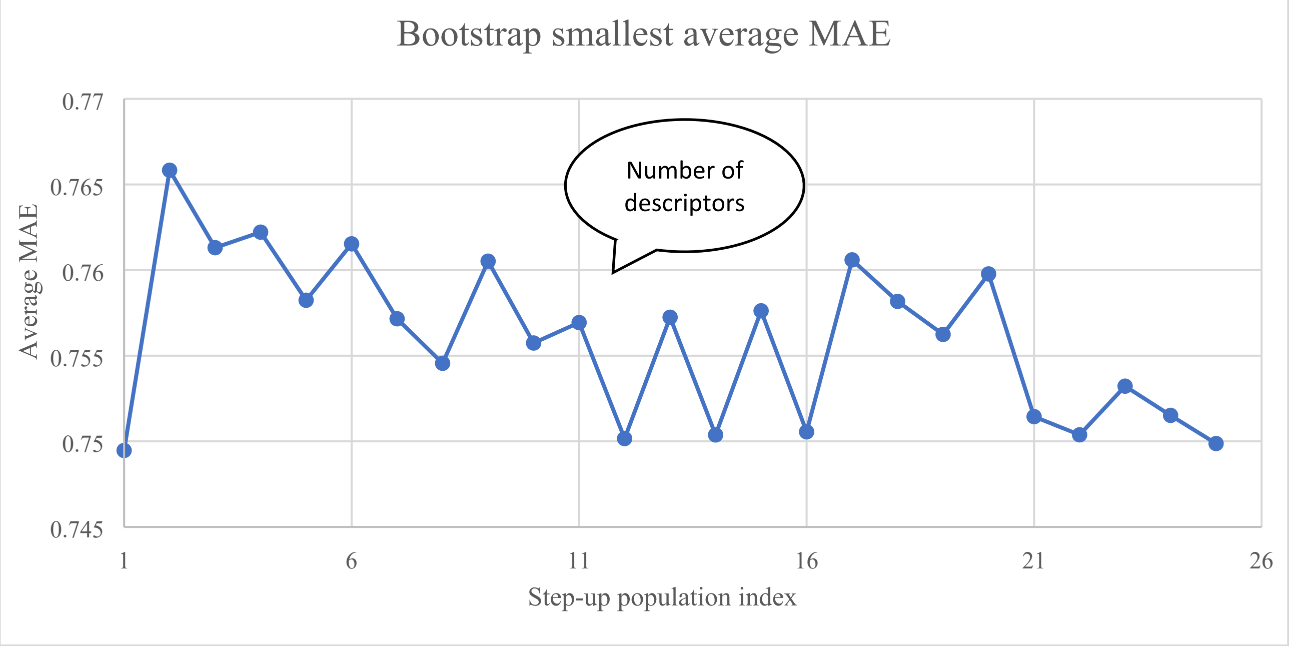


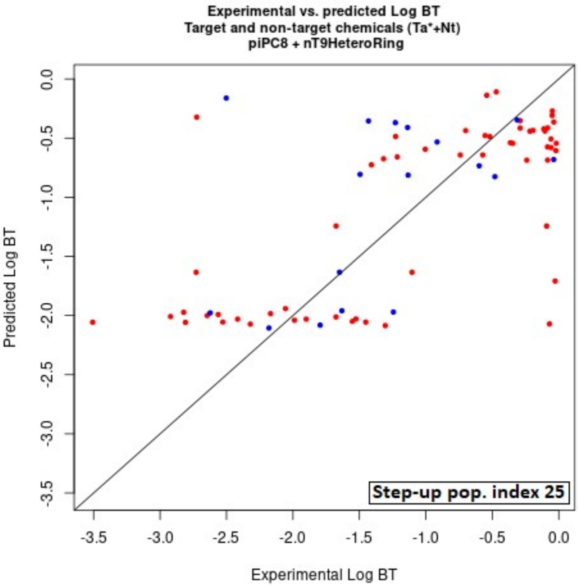


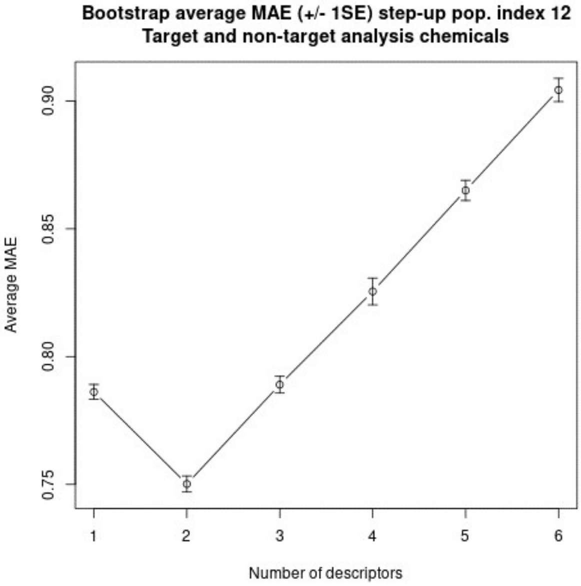


Bootstrap analysis S3. Selection of the best candidate of target and non-target chemicals QSPR (Ta+Nt dataset).

Run	No. Folds	Run	No. Folds
1	62	14	49
2	53	15	50
3	47	16	52
4	47	17	50
5	44	18	45
6	49	19	49
7	54	20	47
8	52	21	52
9	51	22	47
10	52	23	54
11	45	24	52
12	52	25	49
13	53		

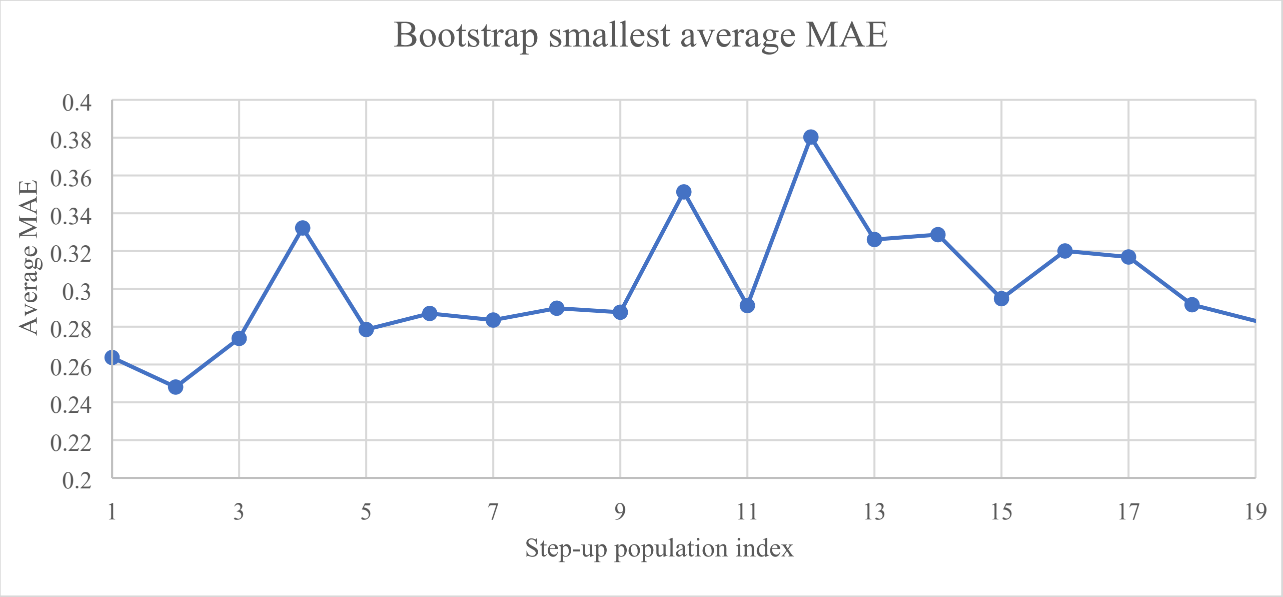


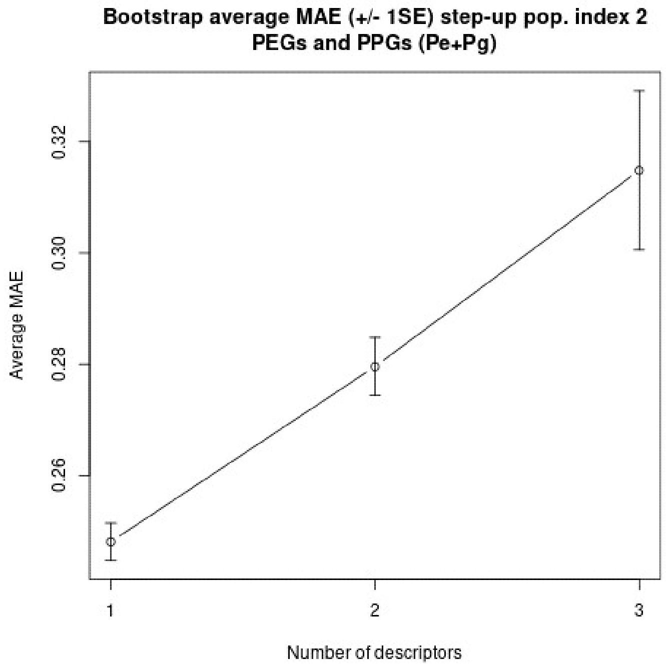




Bootstrap analysis S4. Selection of the best candidate of target chemicals, by removing endpoint outliers, and non-target chemicals QSPR (Ta*+Nt dataset).

Run	No. Folds	Run	No. Folds
1	27	14	26
2	43	15	40
3	46	16	38
4	32	17	29
5	36	18	39
6	40	19	32
7	47	20	41
8	41	21	39
9	33	22	31
10	38	23	49
11	35	24	43
12	43	25	38
13	43		





Bootstrap analysis S5. Selection of the best candidate of PEGs and PPGs QSPR (Pe+Pg dataset).

Methods

Method S1. Leave-one-out bootstrap.

Each bootstrap training set was composed by sampling with replacement from the original training set, keeping the same size. The endpoint of the chemicals not included in the bootstrap training set i.e., the left-out ones, was predicted by applying the QSPR with the best objective function value (R^2) from the step-up population, and the residuals collected. The bootstrap procedure, and the corresponding variable selection procedure, was reiterated until all chemicals were left out at least once. To improve representativeness of the left-out chemicals (some may be left out many times while other only one time), the whole procedure was repeated 5 times without resetting the random seed. The mean absolute error (MAE) was calculated using the collected residuals and later associated to the corresponding QSPR developed using the original training set.

The whole procedure, here referred as a run, was repeated 25 times starting from different random seeds.

See **Appendix** for algorithms.

Method S2. Selection of the best QSPRs from the step-up population.

In the following part the default step-up population of 25 QSPRs is assumed. Once the step-up procedure has been completed, it results in a population of QSPRs for each number of descriptors i.e., 25 one-descriptor QSPRs, 25 two-descriptors QSPRs and so on. Each population contains the 25 best fitting QSPRs, indexed from 1 (highest fitting) to 25 (lowest fitting). See the exemplificative table below, assuming that the step-up procedure has been run up to 6 descriptors. Let us call it "table of candidate QSPRs".

Index	1-descriptor QSPR R^2	2-descriptors QSPR R^2	…	6-descriptors QSPR R^2
1	0.70	0.84	…	0.94
2	0.67	0.82	…	0.92
…	…	…	…	
25	0.58	0.71	…	0.87

Table 1

To detect which one has the lowest chance of being overfitted, for each step-up index (see Index, Table 1) the step-up procedure has been run 25 times by using training sets built using the leave-one-out bootstrap procedure, then the average of the smallest MAEs, and the corresponding one standard error, are calculated and collected, as exemplified in the table below (Table 2).

Index	1-descriptor QSPR bootstrap	2-descriptors QSPR boostrap	…	6-descriptors QSPR bootstrap
1	average MAE and 1SE	average MAE and 1SE	…	average MAE and 1SE
2	average MAE and 1SE	average MAE and 1SE	…	average MAE and 1SE
…	…	…	…	
25	average MAE and 1SE	average MAE and 1SE	…	average MAE and 1SE

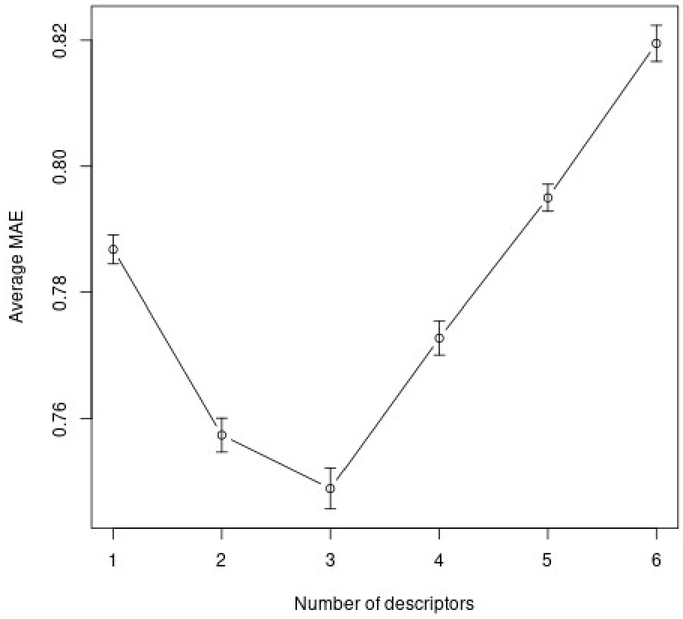
Table 2

For each step-up index (see Table 2), the smallest average MAE has been collected as in the example of the table below.

Index	Smallest average MAE	Corresponding QSPR
1	0.88	4-descriptors QSPR
2	0.75	3-descriptors QSPR
…	…	…
25	1.02	4-descriptors QSPR

Table 3

The smallest average MAE value found in Table 3 is then used to select the step-up population index of Table 2. Let us assume that the smallest average MAE is for the 3-descriptors QSPR, index 2 in Table 3. Then, from Table 2, index 2, the corresponding average MAEs and corresponding SEs, are plotted as in the figure below example.



The one standard error rule[1] states that the most parsimonious model (QSPR) is the one using the smallest number of predictors (descriptors) within one standard error from the model (QSPR) having the smallest estimated prediction error (in this case the average MAE of the 3-descriptors

[1] Breiman L. et al., Stone C.J. Classification and regression trees. Chapman & Hall. New York. 1984.

QSPR). Since, from the chart above, the 2-descriptors QSPR is outside the one standard error of the 3-descriptors QSPR, the latter is finally selected (in case the 2-descriptors, or 1-descriptor, QSPR should fall within the one standard error, they must be chosen instead, accordingly the smallest falling within the one standard error). So, in this example, the 3-descriptors QSPR from the step-up population index 2 of Table 2 ("table of candidate QSPRs"), is deemed as the one with the smallest chance of being overfitted. In case the final candidate is not deemed appropriate for any reason, the procedure restart from Figure 3, choosing the second smallest average MAE population index.

Method S3. Estimation of the probability of coincidental relationship among the descriptors and the endpoint.

From the randomization techniques proposed by Rücker et al.[2], the one called mode 1, which randomizes the descriptors while keeping the endpoint untouched, was considered the most appropriate in this work. Since this technique does not consider the nature of the descriptors, it could be prone to be more permissive than it should, since randomness is unbounded (for example, using whatever value for a fingerprint, which cannot assume, by definition, different values from 0 or 1, raised questions to us). For this reason, random values were bounded to the range of each descriptor, using integer numbers for fingerprints/counters, and real numbers for the remaining. This randomization technique has been used alongside mode 1 to help evaluating the probability of coincidental relationships.

[2] Rücker et al. y-Randomization and Its Variants in QSPR/QSAR, 2007, J. Chem. Inf. Model, 47, 2345-2357

Appendix

A1. Descriptors filtering

R version: 4.1.2
R caret package version: 6.0.90

Algorithm

1) Select one dataset e.g., Ta (all datasets can be found in supplemental_datasets.xlsx).

2) Exclude descriptors having more than 80 percent same value.

3) Exclude the remaining descriptors by correlation, using the *findCorrelation* function of the *caret* package. Set *x* as the correlation matrix of the filtered descriptors from point 2 and *cutoff* as 0.95.

4) Exclude the remaining descriptors spanning more than 2 orders of magnitude.

A2. Descriptors selection

R version: 4.1.2

Setup

Step-up maximum number of descriptors: Ta = 4, Ta* = 4, Ta+Nt = 6, Ta*+Nt = 6, Pe+Pg = 3.
Step-up population size: Ta = 25, Ta* = 25, Ta+Nt = 25, Ta*+Nt = 25, Pe+Pg = 19.

Algorithm

1) Filter a training dataset as explained in **A1. Descriptors filtering**.

2) Set the step-up maximum number of descriptors and the population size for a dataset according to the **Setup** specifications.

3) Run the step-up procedure using *lm* for regression over the filtered dataset and R^2 as the cost function.

A3. Randomization

R version: 4.1.2

Algorithm

1) Filter a training dataset as explained in **A1. Descriptors filtering**. Let denote the filtered training set as ORIG_DATA.

2) Let denote STORED_R2 the vector that will contain the highest R^2 values found after each randomization round. These values will be used for the calculation of probability of coincidental relationships.
3) Let denote s RAND_DATA a matrix, that will contain the random descriptors and the endpoint of ORIG_DATA.

4) Set the random seed to 1 using the *set.seed* function.

5) For i = 1, 2, ..., total randomization rounds (set to 100).
   a) If randomization is range-based fill the descriptors of RAND_DATA using **Randomization 1**, otherwise fill the descriptors of RAND_DATA using **Randomization 2** (keep the nature of the descriptors).
   b) Run the step-up procedure using *lm* for regression over RAND_DATA, and R^2 as the cost function.
   c) Store the highest R^2 value to the i^{th} index of STORED_R2.

Randomization 1

1) For i = 1, 2, ..., number of ORIG_DATA descriptors.
   a) Generate a vector of random numbers using the *runif* function, set *n* as the number of chemicals, *min* as -1 and *max* as 1.
   b) Copy the randomized vector of step a to the i^{th} descriptor of RAND_DATA.

Randomization 2

1) For i = 1, 2, ..., number of ORIG_DATA descriptors.
   a) Generate a vector of random numbers using the *runif* function, set *n* as the number of chemicals, *min* and *max* respectively as the minimum and the maximum value of the descriptor of ORIG_DATA.

b) Detect whether the ith descriptor from ORIG_DATA is filled with integers by using the *all.equal* function by setting *target* as the ith descriptor, *current* as *as.integer(*ith descriptor*)* and *check.attributes* as *FALSE*.

c) If all values of the ith descriptor are integers, round the values of the randomized vector using the *round* function, then copy the rounded randomized vector to the ith descriptor of RAND_DATA, otherwise (since not all values are integers) copy the randomized vector as it is to the ith descriptor of RAND_DATA.

A4. Leave-one-out bootstrap

R version: 4.1.2
R caret package version: 6.0.90

Setup

Step-up maximum number of descriptors: Ta = 4, Ta* = 4, Ta+Nt = 6, Ta*+Nt = 6, Pe+Pg = 3.
Step-up population size: Ta = 25, Ta* = 25, Ta+Nt = 25, Ta* + Nt = 25, Pe+Pg = 19.

Algorithm

1) Filter a training dataset as explained in **A1. Descriptors filtering**. Let denote the filtered training set as ORIG_DATA.

2) Let denote IS_TEST_IDX a boolean vector containing the chemicals that has been at least once in the test set of any of the bootstrapped datasets.

3) Let denote BOOT_TRAIN_IDX a vector containing the bootstrap training set indexes.

4) Let denote BOOT_TRAIN the bootstrap training set matrix filled by selecting rows (chemicals) from ORIG_DATA, according to BOOT_TRAIN_IDX.

5) Let denote BOOT_TEST_IDX the bootstrap test set indexes.

6) Let denote BOOT_TEST the bootstrap test set matrix filled by selecting rows (chemicals) from ORIG_DATA, according to BOOT_TEST_IDX.

7) Let denote SUM_ABS_RES a matrix of step-up population size x step-up maximum number of descriptors, that will contain the sum of the absolute value of the residuals, calculated by applying the bootstrapped step-up models to BOOT_TEST.

8) Let denote BOOT_ROUNDS = 25 the number of bootstrap rounds.

9) Let denote BOOT_MAE a group of BOOT_ROUNDS matrices of step-up population size x step-up maximum number of descriptors. The matrices elements will contain the MAE values calculated by the bootstrap procedure.

10) For i = 1, 2, ..., BOOT_ROUNDS.
    a) Set the random seed to i using the *set.seed* function.
    b) Set all elements of SUM_ABS_RES matrix to 0.
    c) For j = 1, 2, ..., total bootstrap repeats (set to 5).
        i) Set all elements of IS_TEST_IDX to *FALSE*.
        ii) Fill BOOT_TRAIN_IDX using the *createResample* function of the *caret* package, by setting *y* as a vector containing values from 1 to the number of training chemicals, *times* = 1 and *list* = *FALSE*.
        iii) Fill BOOT_TRAIN according to BOOT_TRAIN_IDX.
        iv) Fill BOOT_TEST_IDX by the training set indexes not included in BOOT_TRAIN_IDX.
        v) Fill BOOT_TEST according to BOOT_TEST_IDX.
        vi) Run the step-up procedure using *lm* for regression over BOOT_TRAIN and R^2 as the cost function.
        vii) Apply BOOT_TEST to each *lm* model of the step-up population.
        viii) Add absolute residuals values, calculated from vii, to SUM_ABS_RES.
        ix) By keeping all previous values, flag IS_TEST_IDX elements as *TRUE* according to BOOT_TEST_IDX.
        x) If any of IS_TEST_IDX elements is FALSE go to step ii.
        xi) Calculate the MAE values from SUM_ABS_RES and fill the i^th BOOT_MAE matrix accordingly.

Tables

Table S1. Canonical SMILES.

ID	Name	SMILES
2	Tramadol	COc1cccc(c1)C1(O)CCCCC1CN(C)C
4	Oxazepam	O=C1Nc2ccc(cc2C(=NC1O)c1ccccc1)Cl
6	Carbamazepine-10,11-epoxide	NC(=O)N1c2ccccc2C2C(c3c1cccc3)O2
7	4-hydroxy-1H-benzotriazole	Oc1cccc2c1[nH]nn2
8	Sotalol	CC(NCC(c1ccc(cc1)NS(=O)(=O)C)O)C
9	Propranolol	OC(COc1cccc2c1cccc2)CNC(C)C
10	Hydrochlorothiazide	Clc1cc2NCNS(=O)(=O)c2cc1S(=O)(=O)N
11	Fluconazole	Fc1ccc(c(c1)F)C(Cn1cncn1)(Cn1cncn1)O
12	Venlafaxine	COc1ccc(cc1)C(C1(O)CCCCC1)CN(C)C
13	Metoprolol	COCCc1ccc(cc1)OCC(CNC(C)C)O
14	Gabapentin	NCC1(CCCCC1)CC(=O)O
15	Furosemide	OC(=O)c1cc(c(cc1NCc1ccco1)Cl)S(=O)(=O)N
16	Diclofenac	OC(=O)Cc1ccccc1Nc1c(Cl)cccc1Cl
18	Atenolol	OC(COc1ccc(cc1)CC(=O)N)CNC(C)C
20	Valsartan	CCCCC(=O)N(C(C(=O)O)C(C)C)Cc1ccc(cc1)c1ccccc1c1n[nH]nn1
21	Ketoprofen	OC(=O)C(c1cccc(c1)C(=O)c1ccccc1)C
22	Metoprolol acid	OC(COc1ccc(cc1)CC(=O)O)CNC(C)C
24	Sulfamethoxazole	Nc1ccc(cc1)S(=O)(=O)Nc1noc(c1)C
25	Aniline	Nc1ccccc1
27	Acesulfame	O=C1C=C(C)OS(=O)(=O)N1
29	Acetaminophen	CC(=O)Nc1ccc(cc1)O
30	Caffeine	Cn1cnc2c1c(=O)n(C)c(=O)n2C
31	(-)-Erythromycin	CCC1OC(=O)C(C)C(OC2OC(C)C(C(C2)(C)OC)O)C(C)C(OC2OC(C)CC(C2O)N(C)C)C(CC(C(=O)C(C(C1(C)O)O)C)C)(C)O
32	(±)-Abscisic acid	OC(=O)C=C(C=CC1(O)C(=CC(=O)CC1(C)C)C)C

Table S1. Canonical SMILES.

33	(S)-Nicotine	CN1CCCC1c1cccnc1
34	1-(2-Morpholinophenyl)dihydro-1H-pyrrole-2,5-dione	O=C1CCC(=O)N1c1ccccc1N1CCOCC1
35	1-Aminocyclohexanecarboxylic acid	OC(=O)C1(N)CCCCC1
36	1-Methyluric acid	O=c1[nH]c2c([nH]1)c(=O)n(c(=O)[nH]2)C
37	1,2-Benzisothiazolin-3-one	O=c1[nH]sc2c1cccc2
38	1,7-Dimethyluric acid	Cn1c(=O)[nH]c2c1c(=O)n(c(=O)[nH]2)C
39	10-Hydroxycarbazepine	OC1Cc2ccccc2N(c2c1cccc2)C(=O)N
40	15-Deoxy-Δ12,14-prostaglandin A1	CCCCCC=CC=C1C=CC(=O)C1CCCCCCC(=O)O
41	16α-Hydroxyestrone	Oc1ccc2c(c1)CCC1C2CCC2(C1CC(C2=O)O)C
42	17α-Hydroxyprogesterone	O=C1CCC2(C(=C1)CCC1C2CCC2(C1CCC2(O)C(=O)C)C)C
43	2-[(Dimethylamino)methylidene]indan-1-one	CN(C=C1Cc2c(C1=O)cccc2)C
44	2-[4-(3-Amino-2-hydroxypropoxy)phenyl]acetamide	NCC(COc1ccc(cc1)CC(=O)N)O
45	2-Methoxy-5-methylaniline	COc1ccc(cc1N)C
46	2-Phenylbenzimidazole-5-sulfonic acid	OS(=O)(=O)c1ccc2c(c1)[nH]c(n2)c1ccccc1
47	2,2,6,6-Tetramethyl-1-piperidinol (TEMPO)	[O]N1C(C)(C)CCCC1(C)C
48	2,2,6,6-Tetramethyl-4-piperidinol	OC1CC(C)(C)NC(C1)(C)C
49	2,3,5,6-Tetramethylpyrazine	Cc1nc(C)c(nc1C)C
50	2,4-Diaminotoluene	Nc1ccc(c(c1)N)C
51	3-Aminosalicylic acid	OC(=O)c1cccc(c1O)N
52	3-Hydroxy-2-methylpyridine	Oc1cccnc1C
53	3,4-Dimethoxycinnamic acid	COc1cc(C=CC(=O)O)ccc1OC
54	3,5-di-tert-Butyl-4-hydroxybenzoic acid	OC(=O)c1cc(c(c(c1)C(C)(C)C)O)C(C)(C)C
55	3,5-Dimethyl-1-phenylpyrazole	Cc1nn(c(c1)C)c1ccccc1
56	4-Acetamidobenzaldehyde	O=Cc1ccc(cc1)NC(=O)C
57	4-Amino-3-hydroxybenzoic acid	OC(=O)c1ccc(c(c1)O)N
58	4-Aminophenol	Nc1ccc(cc1)O
59	4-Hydroxycoumarin	O=c1cc(O)c2c(o1)cccc2
60	4-Methyl-5-thiazoleethanol	Cc1ncsc1CCO
61	4-tert-Butylcyclohexyl acetate	CC(=O)OC1CCC(CC1)C(C)(C)C
62	6-(3,4,5-Trimethoxystyryl)-2,3,4,5-tetrahydropyridazin-3-one	COc1cc(C=CC2=NNC(=O)CC2)cc(c1OC)OC

Table S1. Canonical SMILES.

63	6-Aminocaproic acid	NCCCCCC(=O)O
64	6-Methyl[1,2,4]triazolo[4,3-b]pyridazin-8-ol	Cc1cc(=O)c2n([nH]1)cnn2
65	6,7-Dihydroxy-4-methylcoumarin	O=c1cc(C)c2c(o1)cc(c(c2)O)O
66	7-Methylguanine	Nc1nc(=O)c2c([nH]1)ncn2C
67	7-Methylxanthine	O=c1[nH]c(=O)c2c([nH]1)ncn2C
68	7α-Hydroxytestosterone	O=C1CCC2(C(=C1)CC(C1C2CCC2(C1CCC2O)C)O)C
69	Acetanilide	CC(=O)Nc1ccccc1
70	Acetylcholine	CC(=O)OCC[N+](C)(C)C
71	Acridine	c1ccc2c(c1)nc1c(c2)cccc1
72	Acycloguanosine	Nc1nc(=O)c2c([nH]1)n(COCCO)cn2
73	Androstenedione	O=C1CCC2(C(=C1)CCC1C2CCC2(C1CCC2=O)C)C
75	Azobenzene	c1ccc(cc1)N=Nc1ccccc1
76	Benzophenone	O=C(c1ccccc1)c1ccccc1
77	Benzoylecgonine	CN1C2CCC1C(C(C2)OC(=O)c1ccccc1)C(=O)O
78	Bis(2-butoxyethyl) ether	CCCCOCCOCCOCCCC
79	Bis(2-ethylhexyl) amine	CCCCC(CNCC(CCCC)CC)CC
80	Cafestol	OCC1(O)CC23CC1CCC3C1(C(CC2)c2ccoc2CC1)C
81	Caprolactam	O=C1CCCCCN1
82	Carbendazim	COC(=O)Nc1nc2c([nH]1)cccc2
83	Citroflex 2	CCOC(=O)C(CC(=O)OCC)(CC(=O)OCC)O
84	Citroflex 4	CCCCOC(=O)C(CC(=O)OCCCC)(CC(=O)OCCCC)O
85	Clarithromycin	CCC1OC(=O)C(C)C(OC2CC(C)(OC)C(C(O2)C)O)C(C)C(OC2OC(C)CC(C2O)N(C)C)C(CC(C(=O)C(C(C1(C)O)O)C)C)(C)OC
86	Climbazole	Clc1ccc(cc1)OC(C(=O)C(C)(C)C)n1cncc1
87	Codeine	COc1ccc2c3c1OC1C43CCN(C(C2)C4C=CC1O)C
88	Cotinine	O=C1CCC(N1C)c1cccnc1
89	D-Sphingosine	CCCCCCCCCCCCCC=CC(C(CO)N)O
90	Decanamide	CCCCCCCCCC(=O)N
91	DEET	CCN(C(=O)c1cccc(c1)C)CC
92	Dehydroepiandrosterone (DHEA)	OC1CCC2(C(=CCC3C2CCC2(C3CCC2=O)C)C1)C
93	Dibenzylamine	N(Cc1ccccc1)Cc1ccccc1

Table S1. Canonical SMILES.

94	Dibutyl phosphate	CCCCOP(=O)(OCCCC)O
95	Diethyl phosphate	CCOP(=O)(OCC)O
96	Diethyl phthalate	CCOC(=O)c1ccccc1C(=O)OCC
97	Diglyme	COCCOCCOC
98	Diketo-Metribuzin	CC(c1n[nH]c(=O)n(c1=O)N)(C)C
99	DL-Carnitine	OC(C[N+](C)(C)C)CC(=O)[O-]
100	Ecgonine	OC1CC2CCC(C1C(=O)O)N2C
101	Ethyl paraben	CCOC(=O)c1ccc(cc1)O
102	Ferulic acid	COc1cc(C=CC(=O)O)ccc1O
103	Galaxolidone	O=C1OCC(c2c1cc1c(c2)C(C(C1(C)C)C)(C)C)C
104	Guaifenesin	OCC(COc1ccccc1OC)O
105	Guanine	Nc1nc(=O)c2c([nH]1)nc[nH]2
106	Histamine	NCCc1cnc[nH]1
107	Ibuprofen	CC(Cc1ccc(cc1)C(C(=O)O)C)C
108	Icaridin	OCCC1CCCCN1C(=O)OC(CC)C
109	Indole-3-butyric acid	OC(=O)CCCc1c[nH]c2c1cccc2
110	Indole-3-pyruvic acid	OC(=O)C(=O)Cc1c[nH]c2c1cccc2
111	Isoprenaline	CC(NCC(c1ccc(c(c1)O)O)O)C
112	Isotretinoin	CC(=CC=CC(=CC(=O)O)C)C=CC1=C(C)CCCC1(C)C
113	Kahweol	OCC1(O)CC23CC1CCC3C1(C(CC2)c2ccoc2C=C1)C
114	L-threo-3-Phenylserine	OC(C(C(=O)O)N)c1ccccc1
115	Losartan	CCCCc1nc(c(n1Cc1ccc(cc1)c1ccccc1c1n[nH]nn1)CO)Cl
116	Mephedrone	CNC(C(=O)c1ccc(cc1)C)C
117	Metamfepramone	CC(C(=O)c1ccccc1)N(C)C
118	Methyl indole-3-acetate	COC(=O)Cc1c[nH]c2c1cccc2
119	Morphine	OC1C=CC2C34C1Oc1c4c(CC2N(CC3)C)ccc1O
120	N-(2,4-Dimethylphenyl)formamide	O=CNc1ccc(cc1C)C
121	N,N-Dimethylaniline	CN(c1ccccc1)C
122	Nootkatone	O=C1CC(C)C2(C(=C1)CCC(C2)C(=C)C)C
123	Norfenefrine	NCC(c1cccc(c1)O)O
124	Oxybenzone	COc1ccc(c(c1)O)C(=O)c1ccccc1

Table S1. Canonical SMILES.

126	PEG n5	OCCOCCOCCOCCOCCO
127	PEG n6	OCCOCCOCCOCCOCCOCCO
128	PEG n7	OCCOCCOCCOCCOCCOCCOCCO
129	PEG n8	OCCOCCOCCOCCOCCOCCOCCOCCO
130	Perillartine	ON=CC1=CCC(CC1)C(=C)C
131	Phenacetin	CCOc1ccc(cc1)NC(=O)C
132	Pilocarpine	CCC1C(=O)OCC1Cc1cncn1C
133	Polygodial	O=CC1C(=CCC2C1(C)CCCC2(C)C)C=O
134	PPG n4	OCC(OCC(OCC(OCC(O)C)C)C)C
135	PPG n5	OCC(OCC(OCC(OCC(OCC(O)C)C)C)C)C
136	PPG n6	OCC(OCC(OCC(OCC(OCC(OCC(O)C)C)C)C)C)C
137	PPG n7	OCC(OCC(OCC(OCC(OCC(OCC(OCC(O)C)C)C)C)C)C)C
138	PPG n8	OCC(OCC(OCC(OCC(OCC(OCC(OCC(OCC(O)C)C)C)C)C)C)C)C
139	Pregabalin	NCC(CC(=O)O)CC(C)C
140	PV9	CCCCCCC(C(=O)c1ccccc1)N1CCCC1
141	Pyridostigmine	C[n+]1cccc(c1)OC(=O)N(C)C
142	Pyroquilon	O=C1CCc2c3N1CCc3ccc2
143	Rhodamine 6G	CCOC(=O)c1ccccc1c1c2cc(C)c(=NCC)cc2oc2c1cc(C)c(c2)NCC
144	Ricinine	N#Cc1c(OC)ccn(c1=O)C
145	Serotonin	NCCc1c[nH]c2c1cc(O)cc2
146	Sulfapyridine	Nc1ccc(cc1)S(=O)(=O)Nc1ccccn1
147	Theobromine	Cn1cnc2c1c(=O)[nH]c(=O)n2C
148	Tranexamic acid	NCC1CCC(CC1)C(=O)O
149	Triethyl phosphate	CCOP(=O)(OCC)OCC
150	Triisopropanolamine	CC(CN(CC(O)C)CC(O)C)O
151	Trilostane	N#CC1=C(O)C2C3(C(C1)(C)C1CCC4(C(C1CC3)CCC4O)C)O2
152	Trimethoprim	COc1cc(Cc2cnc(nc2N)N)cc(c1OC)OC
153	Tropinone	CN1C2CCC1CC(=O)C2
154	Venlafaxine N-Oxide	COc1ccc(cc1)C(C1(O)CCCCC1)CN(=O)(C)C
155	(+/-)12(13)-DiHOME	CCCCCC(C(CC=CCCCCCCCC(=O)O)O)O
156	2,4-Dimethylbenzaldehyde	O=Cc1ccc(cc1C)C

Table S1. Canonical SMILES.

157	2'-Deoxyadenosine	OCC1OC(CC1O)n1cnc2c1ncnc2N
158	3-Hydroxypyridine	Oc1cccnc1
159	Acetyl-β-methylcholine	CC(C[N+](C)(C)C)OC(=O)C
160	Alfuzosin	COc1cc2nc(nc(c2cc1OC)N)N(CCCNC(=O)C1CCCO1)C
161	Bezafibrate	Clc1ccc(cc1)C(=O)NCCc1ccc(cc1)OC(C(=O)O)(C)C
162	Cocaine	COC(=O)C1C(CC2N(C1CC2)C)OC(=O)c1ccccc1
163	D-Panthenol	OCCCNC(=O)C(C(CO)(C)C)O
164	D,L-Camphor	O=C1CC2C(C1(C)CC2)(C)C
165	Dodecylamine	CCCCCCCCCCCCN
166	Ethylenediaminetetraacetic acid (EDTA)	OC(=O)CN(CC(=O)O)CCN(CC(=O)O)CC(=O)O
167	Indole-3-acrylic acid	OC(=O)C=Cc1c[nH]c2c1cccc2
168	Isoamylamine	NCCC(C)C
169	Methionine	CSCCC(C(=O)O)N
170	Methylimidazoleacetic acid	Cn1cc(nc1)CC(=O)O
171	N,N'-Dicyclohexylurea	O=C(NC1CCCCC1)NC1CCCCC1
172	Paraxanthine	Cn1cnc2c1c(=O)n(c(=O)[nH]2)C
173	PEG n10	OCCOCCOCCOCCOCCOCCOCCOCCOCCOCCO
174	PEG n11	OCCOCCOCCOCCOCCOCCOCCOCCOCCOCCOCCO
175	PEG n12	OCCOCCOCCOCCOCCOCCOCCOCCOCCOCCOCCOCCO
176	PEG n13	OCCOCCOCCOCCOCCOCCOCCOCCOCCOCCOCCOCCOCCO
177	PEG n14	OCCOCCOCCOCCOCCOCCOCCOCCOCCOCCOCCOCCOCCOCCO
178	PEG n15	OCCOCCOCCOCCOCCOCCOCCOCCOCCOCCOCCOCCOCCOCCOCCO
179	PPG n10	OCC(OCC(OCC(OCC(OCC(OCC(OCC(OCC(OCC(OCC(O)C)C)C)C)C)C)C)C)C)C
180	Thymine	Cc1c[nH]c(=O)[nH]c1=O
181	α-Eleostearic acid	CCCCC=CC=CC=CCCCCCCCC(=O)O
183	1-(2-Furylmethyl)-5-oxopyrrolidine-3-carboxylic acid	OC(=O)C1CC(=O)N(C1)Cc1ccco1
184	1-(Carboxymethyl)cyclohexanecarboxylic acid	OC(=O)C1(CCCCC1)CC(=O)O
185	1-Methylguanine	Cn1c(N)nc2c(c1=O)nc[nH]2
186	1,3,7-Trimethyluric acid	Cn1c(=O)[nH]c2c1c(=O)n(c(=O)n2C)C
188	10-Hydroxydecanoic acid	OCCCCCCCCCC(=O)O
189	12-Hydroxydodecanoic acid	OCCCCCCCCCCCC(=O)O

Table S1. Canonical SMILES.

190	2-Amino-6-methylmercaptopurine	CSc1nc(N)nc2c1[nH]cn2
191	2-Deoxyribose 5-phosphate	OC1CC(C(O1)COP(=O)(O)O)O
192	2-Hydroxycinnamic acid	OC(=O)C=Cc1ccccc1O
193	2-Hydroxyhippuric acid	OC(=O)CNC(=O)c1ccccc1O
194	2-Naphthalenesulfonic acid	OS(=O)(=O)c1ccc2c(c1)cccc2
195	2,5-di-tert-Butylhydroquinone	CC(c1cc(O)c(cc1O)C(C)(C)C)(C)C
196	3-(4-Hydroxyphenyl)propionic acid	OC(=O)CCc1ccc(cc1)O
197	3-Anisic acid	COc1cccc(c1)C(=O)O
198	3-Hydroxydecanoic acid	CCCCCCCC(CC(=O)O)O
199	3-Phenoxybenzoic acid	OC(=O)c1cccc(c1)Oc1ccccc1
200	3-Phenyllactic acid	OC(C(=O)O)Cc1ccccc1
201	3-tert-Butyladipic acid	OC(=O)CCC(C(C)(C)C)CC(=O)O
202	3,3'-Dinitro(1,1'-biphenyl)-4,4'-diamine	O=N(=O)c1cc(ccc1N)c1ccc(c(c1)N(=O)=O)N
203	3,4-Dihydroxybenzenesulfonic acid	Oc1ccc(cc1O)S(=O)(=O)O
204	3,7-Dimethyluric acid	O=c1[nH]c(=O)c2c(n1C)[nH]c(=O)n2C
205	4-Acetamidobenzoic acid	CC(=O)Nc1ccc(cc1)C(=O)O
206	4-Hydroxy-3- methoxyphenylglycol sulfate	OCC(c1ccc(c(c1)OC)OS(=O)(=O)O)O
207	4-Hydroxyphenylpyruvic acid	Oc1ccc(cc1)CC(=O)C(=O)O
208	4-Oxo-6-(3-pyridyl)-2-thioxo-1,2,3,4-tetrahydropyrimidine-5-carbonitrile	N#Cc1c(=O)[nH]c(=S)[nH]c1c1cccnc1
209	4-Pyridoxic acid	OCc1cnc(c(c1C(=O)O)O)C
210	5-Hydroxyindole-3-acetic acid	OC(=O)Cc1c[nH]c2c1cc(O)cc2
211	5,7-Dihydroxy-4-methylcoumarin	Oc1cc(O)c2c(c1)oc(=O)cc2C
212	6-Methoxysalicylic acid	COc1cccc(c1C(=O)O)O
214	8-(4-Sulfophenyl) octanoic acid	OC(=O)CCCCCCCc1ccc(cc1)S(=O)(=O)O
215	8-Iso-15-keto-prostaglandin-F2β	CCCCCC(=O)C=CC1C(O)CC(C1CC=CCCCC(=O)O)O
216	9-Methyluric acid	O=c1[nH]c(=O)c2c([nH]1)n(C)c(=O)[nH]2
218	Azelaic acid	OC(=O)CCCCCCCC(=O)O
219	Biotin	OC(=O)CCCCC1SCC2C1NC(=O)N2
220	Capryloylglycine	CCCCCCCC(=O)NCC(=O)O
221	Cholic acid	OC1CCC2(C(C1)CC(C1C2CC(O)C2(C1CCC2C(CCC(=O)O)C)C)O)C
222	Cyclamic acid	OS(=O)(=O)NC1CCCCC1

Table S1. Canonical SMILES.

223	DL-Mandelic acid	OC(c1ccccc1)C(=O)O
224	Dodecanedioic acid	OC(=O)CCCCCCCCCCC(=O)O
225	Dodecyl sulfate	CCCCCCCCCCCCOS(=O)(=O)O
226	Equol	Oc1ccc(cc1)C1COc2c(C1)ccc(c2)O
227	Epinephrine	CNCC(c1ccc(c(c1)O)O)O
229	Fexofenadine	OC(=O)C(c1ccc(cc1)C(CCCN1CCC(CC1)C(c1ccccc1)(c1ccccc1)O)O)(C)C
230	Hippuric acid	O=C(c1ccccc1)NCC(=O)O
232	Mesalamine	Nc1ccc(c(c1)C(=O)O)O
233	Mono(2-ethylhexyl) phthalate (MEHP)	CCCCC(COC(=O)c1ccccc1C(=O)O)CC
234	Monobutyl phthalate	CCCCOC(=O)c1ccccc1C(=O)O
235	Myristyl sulfate	CCCCCCCCCCCCCCOS(=O)(=O)O
236	N-(2-Morpholinoethyl)-4-(1H-pyrazol-1-yl)benzamide	O=C(c1ccc(cc1)n1cccn1)NCCN1CCOCC1
237	N-(4,6-Dimethyl-2-pyrimidinyl)-4-[(E)-(2-hydroxybenzylidene)amino]benzenesulfonamide	Cc1cc(C)nc(n1)NS(=O)(=O)c1ccc(cc1)N=Cc1ccccc1O
238	N-Acetyl-4-aminosalicylic acid	CC(=O)Nc1ccc(c(c1)O)C(=O)O
239	N-Acetyl-DL-tryptophan	CC(=O)NC(C(=O)O)Cc1c[nH]c2c1cccc2
240	N-Acetyl-L-phenylalanine	OC(=O)C(Cc1ccccc1)NC(=O)C
241	N-Acetyl-L-tyrosine	OC(=O)C(Cc1ccc(cc1)O)NC(=O)C
242	N2-[2-(2-Pyridyl)ethyl]-4-hydroxyquinazoline-2-carboxamide	O=C(c1nc(=O)c2c([nH]1)cccc2)NCCc1ccccn1
245	Porphobilinogen	NCc1[nH]cc(c1CC(=O)O)CCC(=O)O
246	Propylparaben	CCCOC(=O)c1ccc(cc1)O
247	Saccharin	O=C1NS(=O)(=O)c2c1cccc2
248	Tetradecanedioic acid	OC(=O)CCCCCCCCCCCCC(=O)O
249	Theophylline	Cn1c(=O)n(C)c2c(c1=O)[nH]cn2
250	Xylenesulfonate	Cc1ccc(cc1C)S(=O)(=O)O
251	β-D-Glucopyranuronic acid	OC1OC(C(=O)O)C(C(C1O)O)O
252	16-Hydroxyhexadecanoic acid	OCCCCCCCCCCCCCCCC(=O)O
253	2'-Deoxyuridine	OCC1OC(CC1O)n1ccc(=O)[nH]c1=O
254	2'-O-Methylguanosine	COC1C(O)C(OC1n1cnc2c1[nH]c(N)nc2=O)CO
255	3-Indoxyl sulphate	OS(=O)(=O)Oc1c[nH]c2c1cccc2
256	3-Methylxanthine	O=c1[nH]c(=O)c2c(n1C)nc[nH]2

Table S1. Canonical SMILES.

257	4-Acetamidobutanoic acid	OC(=O)CCCNC(=O)C
258	4'-Hydroxydiclofenac	OC(=O)Cc1ccccc1Nc1c(Cl)cc(cc1Cl)O
260	D-(-)-Quinic acid	OC1C(O)CC(CC1O)(O)C(=O)O
261	Desthiobiotin	OC(=O)CCCCCC1NC(=O)NC1C
262	Glycoursodeoxycholic acid	OC1CCC2(C(C1)CC(C1C2CCC2(C1CCC2C(CCC(=O)NCC(=O)O)C)C)O)C
263	Glycyl-L-leucine	NCC(=O)NC(C(=O)O)CC(C)C
264	Guanosine	OCC1OC(C(C1O)O)n1cnc2c1[nH]c(N)nc2=O
265	Hypoxanthine	O=c1[nH]cnc2c1[nH]cn2
266	Indole-3-lactic acid	OC(=O)C(Cc1c[nH]c2c1cccc2)O
267	L-Tyrosine	OC(=O)C(Cc1ccc(cc1)O)N
268	Leucylproline	CC(CC(C(=O)N1CCCC1C(=O)O)N)C
269	Methylmalonic acid	CC(C(=O)O)C(=O)O
271	N-Acetylanthranilic acid	CC(=O)Nc1ccccc1C(=O)O
272	Pantothenic acid	OCC(C(C(=O)NCCC(=O)O)O)(C)C
273	Probenecid	CCCN(S(=O)(=O)c1ccc(cc1)C(=O)O)CCC
274	Thymidine	OCC1OC(CC1O)n1cc(C)c(=O)[nH]c1=O
275	Uric acid	O=c1[nH]c2c([nH]1)c(=O)n(c(=O)n2C)C
276	Uridine	OCC1OC(C(C1O)O)n1ccc(=O)[nH]c1=O

Table S1. Canonical SMILES.

Table S2. Descriptor classes calculated using the PaDEL-Descriptor software.

AcidicGroupCount
ALOGP
APol
AromaticAtomsCount
AromaticBondsCount
AtomCount
Autocorrelation
BaryszMatrix
BasicGroupCount
BCUT
BondCount
BPol
BurdenModifiedEigenvalues
CarbonTypes
ChiChain
ChiCluster
ChiPathCluster
ChiPath
Constitutional
Crippen
DetourMatrix
EccentricConnectivityIndex
EStateAtomType
ExtendedTopochemicalAtom
FMF
FragmentComplexity
HBondAcceptorCount
HBondDonorCount
HybridizationRatio
InformationContent
KappaShapeIndices
LargestChain
LargestPiSystem
LongestAliphaticChain
MannholdLogP
McGowanVolume
MDE
MLFER
PathCount
PetitjeanNumber
RingCount
RotatableBondsCount
RuleOfFive
Topological

Table S2. Descriptor classes calculated using the PaDEL-Descriptor software. This is the full list of the descriptors classes before applying the filtering and then the step-up procedure.

TopologicalCharge
TopologicalDistanceMatrix
TPSA
VABC
VAdjMa
WalkCount
Weight
WeightedPath
WienerNumbers
XLogP
ZagrebIndex

Fingerprints

PubchemFingerprinter
SubstructureFingerprinter
SubstructureFingerprinterCount

Table S2. Descriptor classes calculated using the PaDEL-Descriptor software. This is the full list of the descriptors classes before applying the filtering and then the step-up procedure.

Table S3. Experimental and predicted Log BT values.

ID	Name	Log BT	Ta	Ta*	Ta+Nt	Ta*+Nt	Pe+Pg	Pe	Pg
2	Tramadol	-0.02	-0.72	-0.60	-0.07	-0.11			
4	Oxazepam	-0.05	-0.27	-0.13	-1.00	-0.93			
6	Carbamazepine-10,11-epoxide	-0.05	-0.19	-0.04	-0.63	-1.27			
7	4-hydroxy-1H-benzotriazole	-0.07	-0.28	-0.14	-0.73	-1.75			
8	Sotalol	-0.08	-0.12	0.03	-0.05	-0.10			
9	Propranolol	-0.04	-0.34	-0.20	-0.17	-0.34			
10	Hydrochlorothiazide	-0.10	0.18	0.34	0.05	-0.01			
11	Fluconazole	-0.08	-0.22	-0.08	-0.41	-0.99			
12	Venlafaxine	-0.06	-0.64	-0.51	-0.04	-0.06			
13	Metoprolol	-0.08	-0.13	0.01	-0.05	-0.28			
14	Gabapentin	-0.09	-0.34	-0.20	-1.50	-0.79			
15	Furosemide	-0.22	-0.14	0.00	-0.97	-0.45			
16	Diclofenac	-0.19	-0.56	-0.43	-1.13	0.12			
18	Atenolol	-0.57	-0.51	-0.38	-1.07	-0.59			
20	Valsartan	-0.54	-0.97	-0.86	-0.22	-1.17			
21	Ketoprofen	-0.70	-0.55	-0.42	-0.97	-1.48			
22	Metoprolol acid	-0.74	-0.49	-0.36	-1.08	-0.69			
24	Sulfamethoxazole	-1.00	-1.15	-1.05	-1.30	-0.25			
25	Aniline	-0.85	0.29		-0.93				
27	Acesulfame	-1.10	-1.25	-1.15	-1.79	-1.18			
29	Acetaminophen	-2.44	-1.06		-1.66				
30	Caffeine	-2.92	-2.58	-2.54	-2.16	-1.60			
31	(-)-Erythromycin	-0.30							
32	(±)-Abscisic acid	-2.29							
33	(S)-Nicotine	-1.87							
34	1-(2-Morpholinophenyl)dihydro-1H-pyrrole-2,5-dione	-1.82							

Table S3. Experimental and predicted Log BT values. Log BT: experimental value. Datasets: Ta = target chemicals, Ta* = target chemicals -> endpoint outliers removed, Ta+Nt = Target and non-target chemicals, Ta*+Nt = target chemicals-> endpoint outliers removed and non-target chemicals, Pe+Pg = PEGs and PPGs, Pe = PEGs, Pg = PPGs. Row colors -> gray and dark gray: target analysis chemicals (gray: used for Ta*), orange and dark orange: non-target analysis chemicals (dark orange: used for Ta*+Nt), green: PEGs (Pe), dark green: PPGs (Pg). Cell colors: red: training set, blue: test set.

ID	Name	Log BT	Ta	Ta*	Ta+Nt	Ta*+Nt	Pe+Pg	Pe	Pg
35	1-Aminocyclohexanecarboxylic acid	-2.73			-1.76	-1.02			
36	1-Methyluric acid	-1.80			-1.58	-2.70			
37	1,2-Benzisothiazolin-3-one	-0.41							
38	1,7-Dimethyluric acid	-1.55			-1.45	-2.31			
39	10-Hydroxycarbazepine	-2.72			-1.06	-1.36			
40	15-Deoxy-Δ12,14-prostaglandin A1	-2.30							
41	16α-Hydroxyestrone	-1.73							
42	17α-Hydroxyprogesterone	-2.29							
43	2-[(Dimethylamino)methylidene]indan-1-one	-0.67							
44	2-[4-(3-Amino-2-hydroxypropoxy)phenyl]acetamide	-1.22			-1.28	-0.98			
45	2-Methoxy-5-methylaniline	-1.63			-0.72				
46	2-Phenylbenzimidazole-5-sulfonic acid	-0.03			-0.15	-1.03			
47	2,2,6,6-Tetramethyl-1-piperidinol (TEMPO)	-1.47							
48	2,2,6,6-Tetramethyl-4-piperidinol	-0.01							
49	2,3,5,6-Tetramethylpyrazine	-1.49							
50	2,4-Diaminotoluene	-1.40			-0.80				
51	3-Aminosalicylic acid	-1.47							
52	3-Hydroxy-2-methylpyridine	-1.31							
53	3,4-Dimethoxycinnamic acid	-1.85							
54	3,5-di-tert-Butyl-4-hydroxybenzoic acid	-0.02							
55	3,5-Dimethyl-1-phenylpyrazole	-1.14			-0.57	-0.95			
56	4-Acetamidobenzaldehyde	-0.57			-1.61				
57	4-Amino-3-hydroxybenzoic acid	-0.90							
58	4-Aminophenol	-0.64			-0.88				
59	4-Hydroxycoumarin	-0.78							
60	4-Methyl-5-thiazoleethanol	-2.45							
61	4-tert-Butylcyclohexyl acetate	-0.79							
62	6-(3,4,5-Trimethoxystyryl)-2,3,4,5-tetrahydropyridazin-3-one	-2.20							

Table S3. Experimental and predicted Log BT values. Log BT: experimental value. Datasets: Ta = target chemicals, Ta* = target chemicals -> endpoint outliers removed, Ta+Nt = Target and non-target chemicals, Ta*+Nt = target chemicals-> endpoint outliers removed and non-target chemicals, Pe+Pg = PEGs and PPGs, Pe = PEGs, Pg = PPGs. Row colors -> gray and dark gray: target analysis chemicals (gray: used for Ta*), orange and dark orange: non-target analysis chemicals (dark orange: used for Ta*+Nt), green: PEGs (Pe), dark green: PPGs (Pg). Cell colors: red: training set, blue: test set.

ID	Name	Log BT	Ta	Ta*	Ta+Nt	Ta*+Nt	Pe+Pg	Pe	Pg
63	6-Aminocaproic acid	-1.65			-1.84	-1.19			
64	6-Methyl[1,2,4]triazolo[4,3-b]pyridazin-8-ol	-2.18			-2.33	-2.14			
65	6,7-Dihydroxy-4-methylcoumarin	-1.51							
66	7-Methylguanine	-2.32			-2.28	-2.18			
67	7-Methylxanthine	-2.53			-2.26	-2.48			
68	7α-Hydroxytestosterone	-0.49							
69	Acetanilide	-2.90			-1.77				
70	Acetylcholine	-2.48							
71	Acridine	-2.50			-0.34	-1.05			
72	Acycloguanosine	-1.68			-1.97	-2.01			
73	Androstenedione	-1.74							
75	Azobenzene	-0.52			-0.42	-0.32			
76	Benzophenone	-0.35			0.13	-1.24			
77	Benzoylecgonine	-1.27							
78	Bis(2-butoxyethyl) ether	-1.79							
79	Bis(2-ethylhexyl) amine	-1.82							
80	Cafestol	-1.80							
81	Caprolactam	-0.51							
82	Carbendazim	0.00							
83	Citroflex 2	-0.80							
84	Citroflex 4	-1.68							
85	Clarithromycin	-0.02							
86	Climbazole	-0.06			-1.61	-0.06			
87	Codeine	-0.83							
88	Cotinine	-1.57							
89	D-Sphingosine	-1.39							
90	Decanamide	-1.31							
91	DEET	-0.06							

Table S3. Experimental and predicted Log BT values. Log BT: experimental value. Datasets: Ta = target chemicals, Ta* = target chemicals -> endpoint outliers removed, Ta+Nt = Target and non-target chemicals, Ta*+Nt = target chemicals-> endpoint outliers removed and non-target chemicals, Pe+Pg = PEGs and PPGs, Pe = PEGs, Pg = PPGs. Row colors -> gray and dark gray: target analysis chemicals (gray: used for Ta*), orange and dark orange: non-target analysis chemicals (dark orange: used for Ta*+Nt), green: PEGs (Pe), dark green: PPGs (Pg). Cell colors: red: training set, blue: test set.

ID	Name	Log BT	Ta	Ta*	Ta+Nt	Ta*+Nt	Pe+Pg	Pe	Pg
92	Dehydroepiandrosterone (DHEA)	-1.73							
93	Dibenzylamine	-0.24			-0.18	-0.76			
94	Dibutyl phosphate	-0.60							
95	Diethyl phosphate	-0.10							
96	Diethyl phthalate	-0.10							
97	Diglyme	-0.76							
98	Diketo-Metribuzin	-0.66							
99	DL-Carnitine	-1.67							
100	Ecgonine	-0.86							
101	Ethyl paraben	-0.34			-1.66				
102	Ferulic acid	-1.43			-1.49				
103	Galaxolidone	-1.96							
104	Guaifenesin	-1.13			-0.47	-0.49			
105	Guanine	-1.21							
106	Histamine	-2.47							
107	Ibuprofen	-2.74							
108	Icaridin	-1.04							
109	Indole-3-butyric acid	-2.65			-2.05	-1.23			
110	Indole-3-pyruvic acid	-2.82			-2.10	-1.85			
111	Isoprenaline	-0.48			-0.30	-0.34			
112	Isotretinoin	-2.71							
113	Kahweol	-1.45							
114	L-threo-3-Phenylserine	-1.49			-1.53	-1.90			
115	Losartan	-0.47			0.69	-0.61			
116	Mephedrone	-0.52							
117	Metamfepramone	-0.01			-1.56				
118	Methyl indole-3-acetate	-2.56			-2.15	-1.30			
119	Morphine	-1.42							

Table S3. Experimental and predicted Log BT values. Log BT: experimental value. Datasets: Ta = target chemicals, Ta* = target chemicals -> endpoint outliers removed, Ta+Nt = Target and non-target chemicals, Ta*+Nt = target chemicals-> endpoint outliers removed and non-target chemicals, Pe+Pg = PEGs and PPGs, Pe = PEGs, Pg = PPGs. Row colors -> gray and dark gray: target analysis chemicals (gray: used for Ta*), orange and dark orange: non-target analysis chemicals (dark orange: used for Ta*+Nt), green: PEGs (Pe), dark green: PPGs (Pg). Cell colors: red: training set, blue: test set.

ID	Name	Log BT	Ta	Ta*	Ta+Nt	Ta*+Nt	Pe+Pg	Pe	Pg
120	N-(2,4-Dimethylphenyl)formamide	-1.03			-1.66				
121	N,N-Dimethylaniline	-0.83			-0.77				
122	Nootkatone	-0.78							
123	Norfenefrine	-0.66			-0.71				
124	Oxybenzone	-0.55			-1.33	-0.61			
126	PEG n5	-2.01					-2.23	-2.01	
127	PEG n6	-2.13					-2.05	-2.02	
128	PEG n7	-1.92					-1.99	-2.04	
129	PEG n8	-1.44					-1.81	-1.54	
130	Perillartine	-0.91							
131	Phenacetin	-0.60			-1.42	-0.19			
132	Pilocarpine	-0.24							
133	Polygodial	-1.29							
134	PPG n4	-0.43					-0.18		-0.53
135	PPG n5	-0.53					-0.53		-0.47
136	PPG n6	-0.62					-0.75		-0.53
137	PPG n7	-0.61					-0.95		-0.65
138	PPG n8	-0.74					-1.09		-0.79
139	Pregabalin	-1.01							
140	PV9	-0.04			-1.07	-0.37			
141	Pyridostigmine	-1.34							
142	Pyroquilon	-0.26							
143	Rhodamine 6G	-2.57							
144	Ricinine	-0.53							
145	Serotonin	-0.61							
146	Sulfapyridine	-0.36			-0.92	-0.10			
147	Theobromine	-1.90			-2.23	-1.96			
148	Tranexamic acid	-1.14							

Table S3. Experimental and predicted Log BT values. Log BT: experimental value. Datasets: Ta = target chemicals, Ta* = target chemicals -> endpoint outliers removed, Ta+Nt = Target and non-target chemicals, Ta*+Nt = target chemicals-> endpoint outliers removed and non-target chemicals, Pe+Pg = PEGs and PPGs, Pe = PEGs, Pg = PPGs. Row colors -> gray and dark gray: target analysis chemicals (gray: used for Ta*), orange and dark orange: non-target analysis chemicals (dark orange: used for Ta*+Nt), green: PEGs (Pe), dark green: PPGs (Pg). Cell colors: red: training set, blue: test set.

ID	Name	Log BT	Ta	Ta*	Ta+Nt	Ta*+Nt	Pe+Pg	Pe	Pg
149	Triethyl phosphate	-0.10							
150	Triisopropanolamine	-0.06							
151	Trilostane	-0.95							
152	Trimethoprim	-0.89							
153	Tropinone	-2.38							
154	Venlafaxine N-Oxide	-0.02			0.02	-0.16			
155	(+/-)12(13)-DiHOME	-2.70							
156	2,4-Dimethylbenzaldehyde	-0.29							
157	2'-Deoxyadenosine	-0.97							
158	3-Hydroxypyridine	-2.19			-1.55				
159	Acetyl-β-methylcholine	-2.50							
160	Alfuzosin	-0.18							
161	Bezafibrate	-0.92			-0.70	-0.28			
162	Cocaine	-1.78							
163	D-Panthenol	-2.17							
164	D,L-Camphor	-1.97							
165	Dodecylamine	-2.13							
166	Ethylenediaminetetraacetic acid (EDTA)	-0.58							
167	Indole-3-acrylic acid	-2.06			-2.10	-1.71			
168	Isoamylamine	-2.05							
169	Methionine	-1.60							
170	Methylimidazoleacetic acid	-2.64							
171	N,N'-Dicyclohexylurea	-0.06							
172	Paraxanthine	-2.42			-2.21	-1.96			
173	PEG n10	-1.69					-1.67	-1.61	
174	PEG n11	-1.67					-1.62	-1.67	
175	PEG n12	-1.76					-1.58	-1.73	
176	PEG n13	-1.83					-1.54	-1.80	

Table S3. Experimental and predicted Log BT values. Log BT: experimental value. Datasets: Ta = target chemicals, Ta* = target chemicals -> endpoint outliers removed, Ta+Nt = Target and non-target chemicals, Ta*+Nt = target chemicals-> endpoint outliers removed and non-target chemicals, Pe+Pg = PEGs and PPGs, Pe = PEGs, Pg = PPGs. Row colors -> gray and dark gray: target analysis chemicals (gray: used for Ta*), orange and dark orange: non-target analysis chemicals (dark orange: used for Ta*+Nt), green: PEGs (Pe), dark green: PPGs (Pg). Cell colors: red: training set, blue: test set.

ID	Name	Log BT	Ta	Ta*	Ta+Nt	Ta*+Nt	Pe+Pg	Pe	Pg
177	PEG n14	-1.93					-1.58	-1.89	
178	PEG n15	-1.93					-1.62	-1.99	
179	PPG n10	-1.17					-1.31		-1.13
180	Thymine	-1.64							
181	α-Eleostearic acid	-2.20							
183	1-(2-Furylmethyl)-5-oxopyrrolidine-3-carboxylic acid	-1.06							
184	1-(Carboxymethyl)cyclohexanecarboxylic acid	-1.68			-1.44	-1.35			
185	1-Methylguanine	-1.30			-2.29	-2.03			
186	1,3,7-Trimethyluric acid	-1.53			-1.44	-1.90			
188	10-Hydroxydecanoic acid	-1.99							
189	12-Hydroxydodecanoic acid	-1.12							
190	2-Amino-6-methylmercaptopurine	-0.02							
191	2-Deoxyribose 5-phosphate	-2.80							
192	2-Hydroxycinnamic acid	-1.38			-1.67				
193	2-Hydroxyhippuric acid	-2.80							
194	2-Naphthalenesulfonic acid	-0.56							
195	2,5-di-tert-Butylhydroquinone	-0.02							
196	3-(4-Hydroxyphenyl)propionic acid	-2.54			-1.65				
197	3-Anisic acid	-2.30			-1.68				
198	3-Hydroxydecanoic acid	-1.61							
199	3-Phenoxybenzoic acid	-1.23			-0.93	-1.25			
200	3-Phenyllactic acid	-2.80			-1.65				
201	3-tert-Butyladipic acid	-1.96							
202	3,3'-Dinitro(1,1'-biphenyl)-4,4'-diamine	-2.93							
203	3,4-Dihydroxybenzenesulfonic acid	-1.61							
204	3,7-Dimethyluric acid	-1.45			-1.48	-2.16			
205	4-Acetamidobenzoic acid	-2.21			-1.38				
206	4-Hydroxy-3- methoxyphenylglycol sulfate	-1.25							

Table S3. Experimental and predicted Log BT values. Log BT: experimental value. Datasets: Ta = target chemicals, Ta* = target chemicals -> endpoint outliers removed, Ta+Nt = Target and non-target chemicals, Ta*+Nt = target chemicals-> endpoint outliers removed and non-target chemicals, Pe+Pg = PEGs and PPGs, Pe = PEGs, Pg = PPGs. Row colors -> gray and dark gray: target analysis chemicals (gray: used for Ta*), orange and dark orange: non-target analysis chemicals (dark orange: used for Ta*+Nt), green: PEGs (Pe), dark green: PPGs (Pg). Cell colors: red: training set, blue: test set.

ID	Name	Log BT	Ta	Ta*	Ta+Nt	Ta*+Nt	Pe+Pg	Pe	Pg
207	4-Hydroxyphenylpyruvic acid	-2.15			-1.60				
208	4-Oxo-6-(3-pyridyl)-2-thioxo-1,2,3,4-tetrahydropyrimidine-5-carbonitrile	-1.23			-2.01	-2.40			
209	4-Pyridoxic acid	-2.30							
210	5-Hydroxyindole-3-acetic acid	-1.24			-2.06	-1.44			
211	5,7-Dihydroxy-4-methylcoumarin	-1.00							
212	6-Methoxysalicylic acid	-0.25							
214	8-(4-Sulfophenyl) octanoic acid	-2.33							
215	8-Iso-15-keto-prostaglandin-F2β	-3.24							
216	9-Methyluric acid	-2.77							
218	Azelaic acid	-0.56							
219	Biotin	-3.22							
220	Capryloylglycine	-1.77							
221	Cholic acid	-3.93							
222	Cyclamic acid	-1.63							
223	DL-Mandelic acid	-1.54			-1.74				
224	Dodecanedioic acid	-1.66							
225	Dodecyl sulfate	-1.41							
226	Equol	-0.11			-0.21	-0.31			
227	Epinephrine	-2.71			-0.58				
229	Fexofenadine	-0.29			0.30	-0.84			
230	Hippuric acid	-1.06			-1.52				
232	Mesalamine	-0.34							
233	Mono(2-ethylhexyl) phthalate (MEHP)	-0.03							
234	Monobutyl phthalate	-0.04							
235	Myristyl sulfate	-3.39							
236	N-(2-Morpholinoethyl)-4-(1H-pyrazol-1-yl)benzamide	-0.31			-1.40	-0.96			
237	N-(4,6-Dimethyl-2-pyrimidinyl)-4-[(E)-(2-hydroxybenzylidene)amino]benzenesulfonamide	-1.95							

Table S3. Experimental and predicted Log BT values. Log BT: experimental value. Datasets: Ta = target chemicals, Ta* = target chemicals -> endpoint outliers removed, Ta+Nt = Target and non-target chemicals, Ta*+Nt = target chemicals-> endpoint outliers removed and non-target chemicals, Pe+Pg = PEGs and PPGs, Pe = PEGs, Pg = PPGs. Row colors -> gray and dark gray: target analysis chemicals (gray: used for Ta*), orange and dark orange: non-target analysis chemicals (dark orange: used for Ta*+Nt), green: PEGs (Pe), dark green: PPGs (Pg). Cell colors: red: training set, blue: test set.

ID	Name	Log BT	Ta	Ta*	Ta+Nt	Ta*+Nt	Pe+Pg	Pe	Pg
238	N-Acetyl-4-aminosalicylic acid	-1.83							
239	N-Acetyl-DL-tryptophan	-1.63			-1.86	-1.52			
240	N-Acetyl-L-phenylalanine	-1.41			-1.49	-1.56			
241	N-Acetyl-L-tyrosine	-1.32			-1.42	-1.35			
242	N2-[2-(2-Pyridyl)ethyl]-4-hydroxyquinazoline-2-carboxamide	-1.43			-1.49	-1.53			
245	Porphobilinogen	-1.34							
246	Propylparaben	-0.78							
247	Saccharin	-2.39							
248	Tetradecanedioic acid	-1.64							
249	Theophylline	-1.99			-2.19	-1.83			
250	Xylenesulfonate	-1.77							
251	β-D-Glucopyranuronic acid	-3.57							
252	16-Hydroxyhexadecanoic acid	-2.17							
253	2'-Deoxyuridine	-1.29							
254	2'-O-Methylguanosine	-1.66							
255	3-Indoxyl sulphate	-2.62			-1.13	-0.58			
256	3-Methylxanthine	-3.51			-2.31	-2.30			
257	4-Acetamidobutanoic acid	-2.21							
258	4'-Hydroxydiclofenac	-0.29			-1.09	0.35			
260	D-(-)-Quinic acid	-1.39							
261	Desthiobiotin	-1.63							
262	Glycoursodeoxycholic acid	-2.19							
263	Glycyl-L-leucine	-2.28							
264	Guanosine	-1.86							
265	Hypoxanthine	-2.75							
266	Indole-3-lactic acid	-2.17			-2.09	-1.67			
267	L-Tyrosine	-1.38			-1.53				
268	Leucylproline	-2.32							

Table S3. Experimental and predicted Log BT values. Log BT: experimental value. Datasets: Ta = target chemicals, Ta* = target chemicals -> endpoint outliers removed, Ta+Nt = Target and non-target chemicals, Ta*+Nt = target chemicals-> endpoint outliers removed and non-target chemicals, Pe+Pg = PEGs and PPGs, Pe = PEGs, Pg = PPGs. Row colors -> gray and dark gray: target analysis chemicals (gray: used for Ta*), orange and dark orange: non-target analysis chemicals (dark orange: used for Ta*+Nt), green: PEGs (Pe), dark green: PPGs (Pg). Cell colors: red: training set, blue: test set.

ID	Name	Log BT	Ta	Ta*	Ta+Nt	Ta*+Nt	Pe+Pg	Pe	Pg
269	Methylmalonic acid	-0.99							
271	N-Acetylanthranilic acid	-0.73			-1.43				
272	Pantothenic acid	-2.38							
273	Probenecid	-0.03							
274	Thymidine	-1.11							
275	Uric acid	-2.81			-1.52	-2.16			
276	Uridine	-1.50							

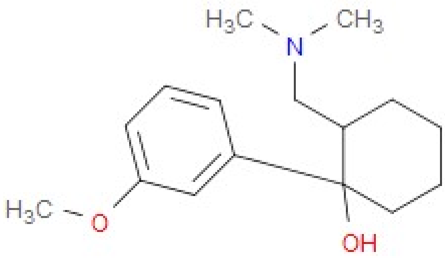
Table S3. Experimental and predicted Log BT values. Log BT: experimental value. Datasets: Ta = target chemicals, Ta* = target chemicals -> endpoint outliers removed, Ta+Nt = Target and non-target chemicals, Ta*+Nt = target chemicals-> endpoint outliers removed and non-target chemicals, Pe+Pg = PEGs and PPGs, Pe = PEGs, Pg = PPGs. Row colors -> gray and dark gray: target analysis chemicals (gray: used for Ta*), orange and dark orange: non-target analysis chemicals (dark orange: used for Ta*+Nt), green: PEGs (Pe), dark green: PPGs (Pg). Cell colors: red: training set, blue: test set.

Name	QSPR	Pattern	Description
MATS2m	Ta Ta*		Moran autocorrelation - lag 2 / weighted by mass
VR3_Dzs	Ta+Nt		Logarithmic Randic-like eigenvector-based index from Barysz matrix / weighted by I-state
PubchemFP373	Ta+Nt	C(~H)(:N)	Simple atom nearest neighbours - These bits test for the presence of atom nearest neighbour patterns, regardless of bond order (denoted by "~") or count, but where bond aromaticity (denoted by ":") is significant.
PubchemFP420	Ta+Nt	C=O	Detailed atom neighbourhoods - These bits test for the presence of detailed atom neighbourhood patterns, regardless of count, but where bond orders are specific, bond aromaticity matches both single and double bonds, and where "-", "=", and "#" matches a single bond, double bond, and triple bond order, respectively.
AATS1s	Ta*+Nt		Average Broto-Moreau autocorrelation - lag 1 / weighted by I-state
ETA_Beta_ns_d	Ta*+Nt		A measure of lone electrons entering into resonance
GATS8s	Pe		Geary autocorrelation - lag 8 / weighted by I-state
ATSC7s	Pg		Centered Broto-Moreau autocorrelation - lag 7 / weighted by I-state
hmax	Pe+Pg		Maximum H E-State

Table S4. Descriptors of the QSPRs. Ta = target chemicals, Ta* = target chemicals -> endpoint outliers removed, Ta+Nt = Target and non-target chemicals, Ta*+Nt = target chemicals-> endpoint outliers removed and non-target chemicals, Pe+Pg = PEGs and PPGs, Pe = PEGs, Pg = PPGs.

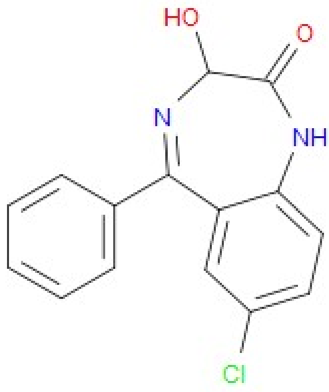
Chemical structures

Chemical structures S1. Target chemicals.



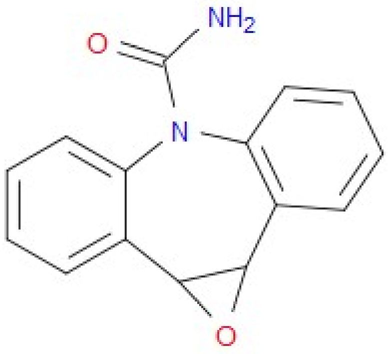
[ID 2] Tramadol

Ta Ta* Ta+Nt Ta*+Nt



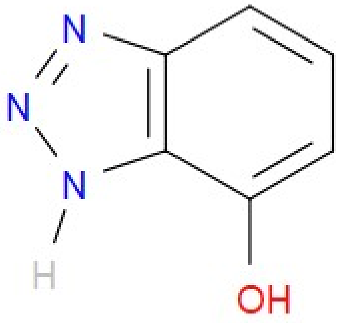
[ID 4] oxazepam

Ta Ta* Ta+Nt Ta*+Nt



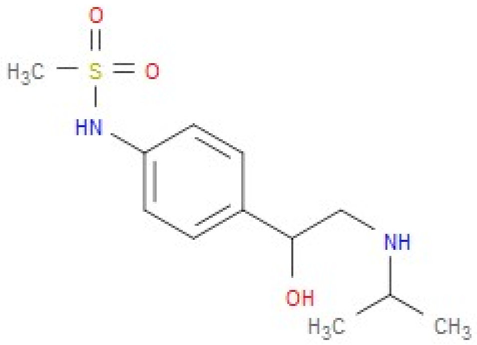
[ID 6] carbamazepine-10,11-epoxide

Ta Ta* Ta+Nt Ta*+Nt



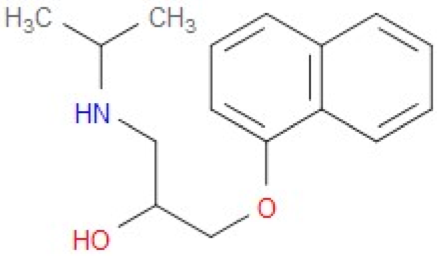
[ID 7] 4-hydroxy-1H-benzotriazole

Ta Ta* Ta+Nt Ta*+Nt



[ID 8] sotalol

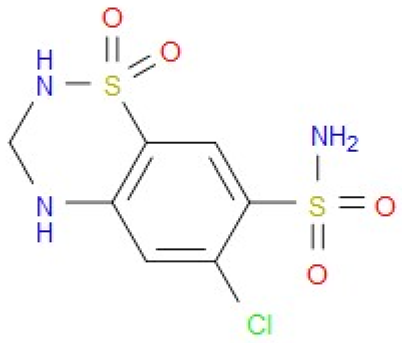
Ta Ta* Ta+Nt Ta*+Nt



[ID 9] propranolol

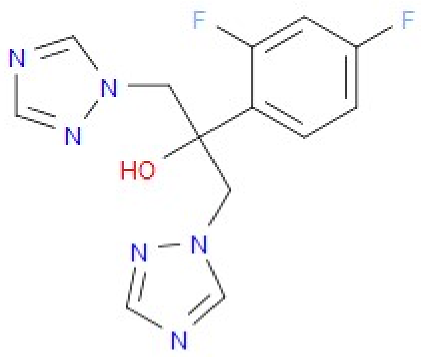
Ta Ta* tant Ta*+Nt

Chemical structures S1. Target chemicals. Datasets: Ta = target chemicals, Ta* = target chemicals -> endpoint outliers removed, Ta+Nt = target and non-target chemicals, Ta*+Nt = target chemicals -> endpoint outliers removed and non-target analysis chemicals. Red: training set.



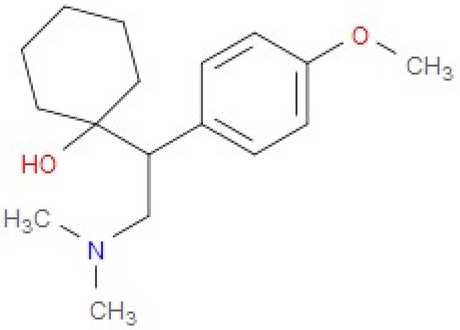
[ID 10] hydrochlorothiazide

Ta Ta* Ta+Nt Ta*+Nt



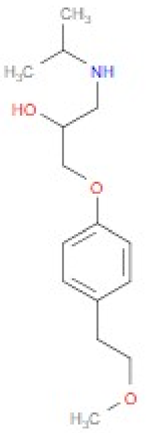
[ID 11] fluconazole

Ta Ta* Ta+Nt Ta*+Nt



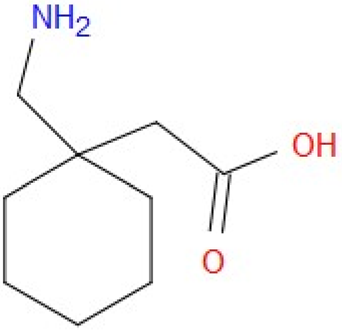
[ID 12] venlafaxine

Ta Ta* Ta+Nt Ta*+Nt



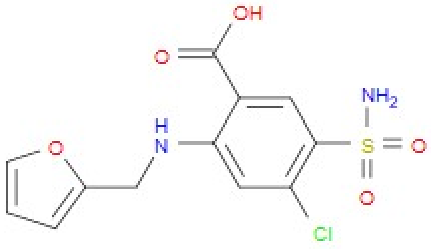
[ID 13] metoprolol

Ta Ta* Ta+Nt Ta*+Nt



[ID 14] gabapentin

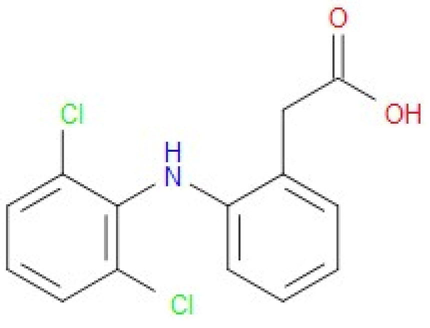
Ta Ta* Ta+Nt Ta*+Nt



[ID 15] furosemide

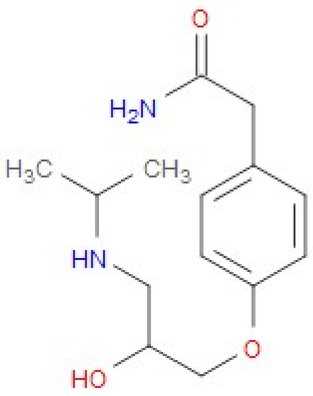
Ta Ta* Ta+Nt Ta*+Nt

Chemical structures S1. Target chemicals. Datasets: Ta = target chemicals, Ta* = target chemicals -> endpoint outliers removed, Ta+Nt = target and non-target chemicals, Ta*+Nt = target chemicals -> endpoint outliers removed and non-target analysis chemicals. Red: training set.



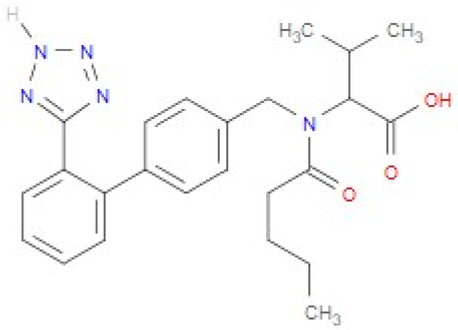
[ID 16] diclofenac

Ta Ta* Ta+Nt Ta*+Nt



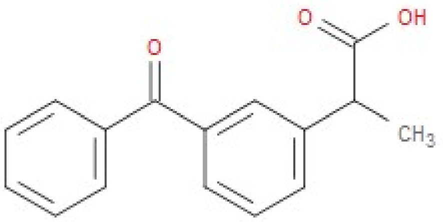
[ID 18] atenolol

Ta Ta* Ta+Nt Ta*+Nt



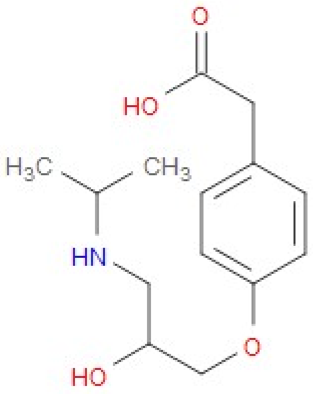
[ID 20] valsartan

Ta Ta* Ta+Nt Ta*+Nt



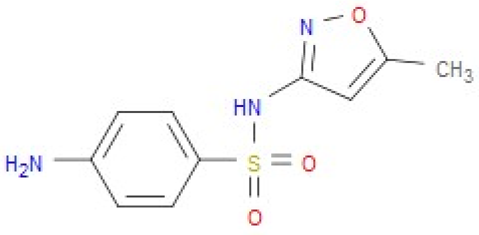
[ID 21] ketoprofen

Ta Ta* Ta+Nt Ta*+Nt



[ID 22] metoprolol acid

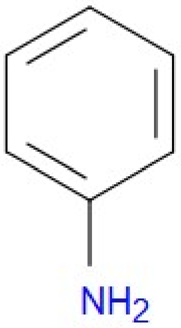
Ta Ta* Ta+Nt Ta*+Nt



[ID 24] sulfamethoxazole

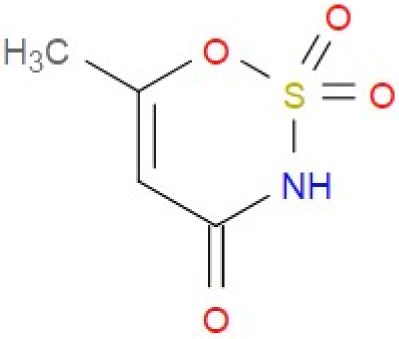
Ta Ta* Ta+Nt Ta*+Nt

Chemical structures S1. Target chemicals. Datasets: Ta = target chemicals, Ta* = target chemicals -> endpoint outliers removed, Ta+Nt = target and non-target chemicals, Ta*+Nt = target chemicals -> endpoint outliers removed and non-target analysis chemicals. Red: training set.



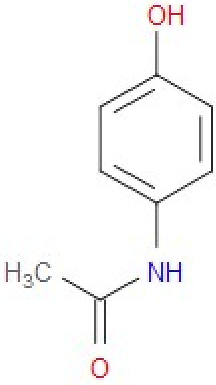
[ID 25] aniline

Ta Ta+Nt



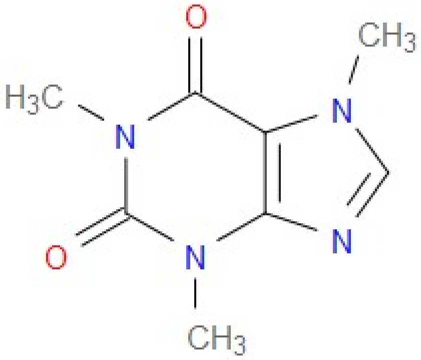
[ID 27] acesulfame

Ta Ta* Ta+Nt Ta*+Nt



[ID 29] acetaminophen

Ta Ta+Nt

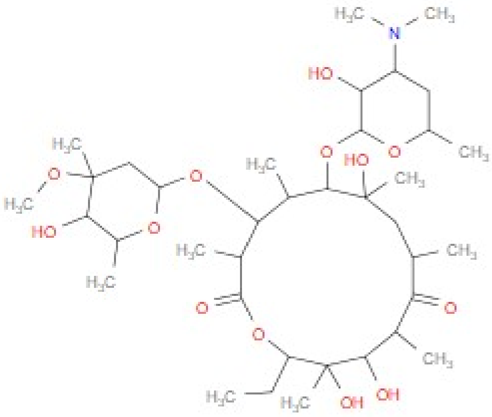


[ID 30] caffeine

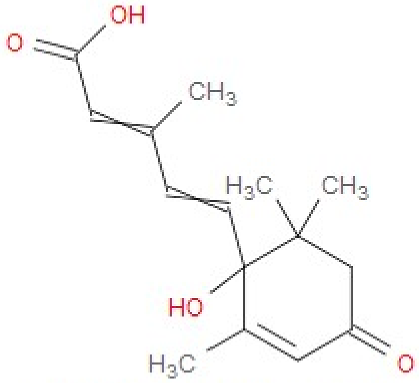
Ta Ta* Ta+Nt Ta*+Nt

Chemical structures S1. Target chemicals. Datasets: Ta = target chemicals, Ta* = target chemicals -> endpoint outliers removed, Ta+Nt = target and non-target chemicals, Ta*+Nt = target chemicals -> endpoint outliers removed and non-target analysis chemicals. Red: training set.

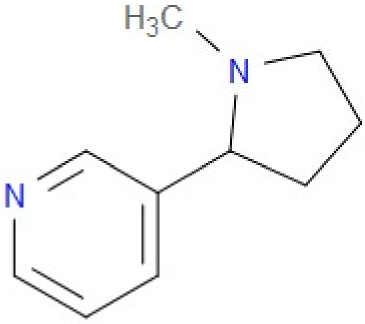
Chemical structures S2. Non-target chemicals.



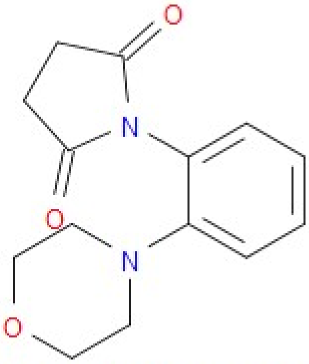
[ID 31] (-)-Erythromycin



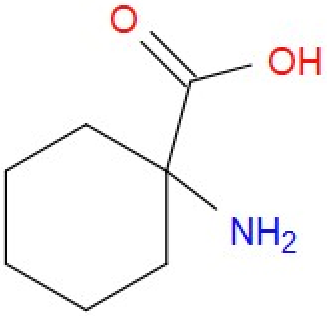
[ID 32] (±)-Abscisic acid



[ID 33] (S)-Nicotine

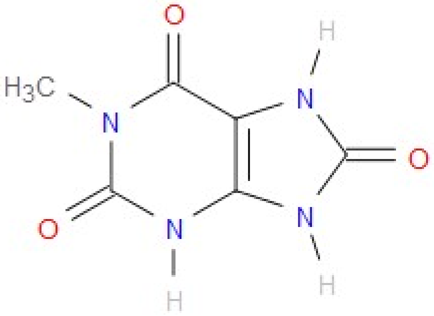


[ID 34] 1-(2-Morpholinophenyl) dihydro-1H-pyrrole-2,5-dione



[ID 35] 1-Aminocyclohexanecarboxylic acid

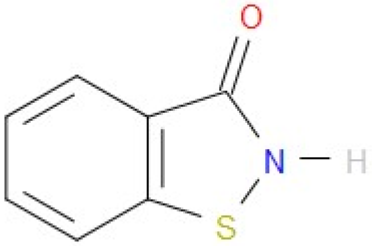
Ta+Nt Ta*+Nt



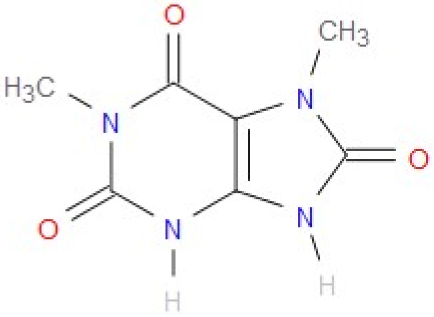
[ID 36] 1-Methyluric acid

Ta+Nt Ta*+Nt

Chemical structures S2. Non-target chemicals. Datasets: Ta+Nt = target and non-target chemicals, Ta*+Nt = target -> endpoint outliers removed and non-target chemicals, Pe = PEGs, Pg = PPGs, Pe+Pg = PEGs and PPGs. Red: training set, blue: test set.

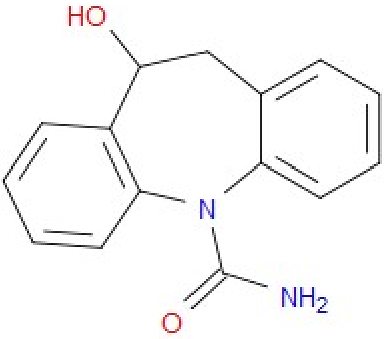


[ID 37] 1,2-Benzisothiazolin-3-one



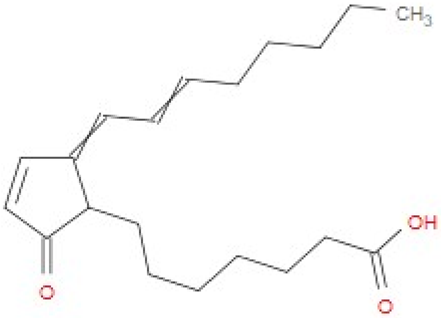
[ID 38] 1,7-Dimethyluric acid

Ta+Nt Ta*+Nt

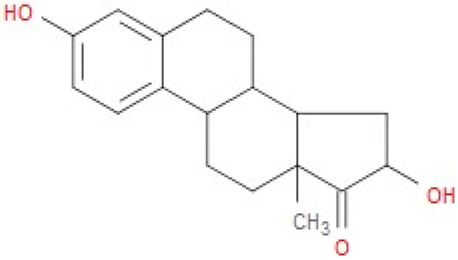


[ID 39] 10-Hydroxycarbazepine

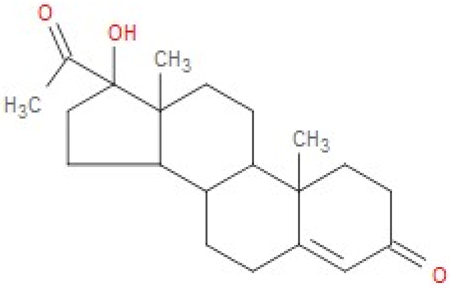
Ta+Nt Ta*+Nt



[ID 40] 15-Deoxy-Δ12,14-prostaglandin A1

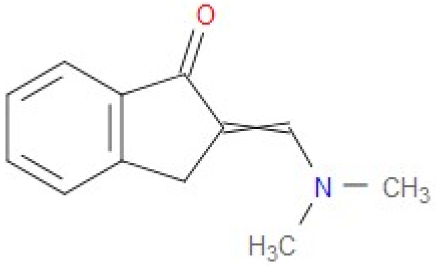


[ID 41] 16α-Hydroxyestrone

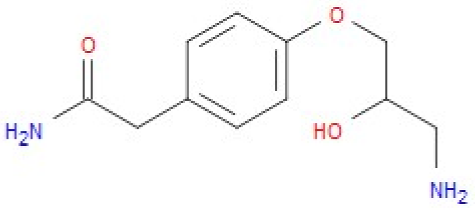


[ID 42] 17α-Hydroxyprogesterone

Chemical structures S2. Non-target chemicals. Datasets: Ta+Nt = target and non-target chemicals, Ta*+Nt = target -> endpoint outliers removed and non-target chemicals, Pe = PEGs, Pg = PPGs, Pe+Pg = PEGs and PPGs. Red: training set, blue: test set.

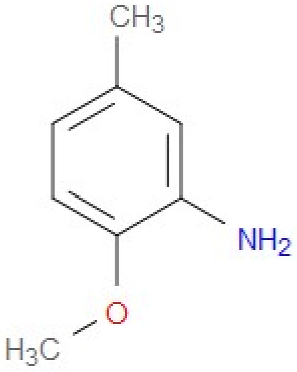


[ID 43] 2-[(Dimethylamino) methylidene]indan-1-one



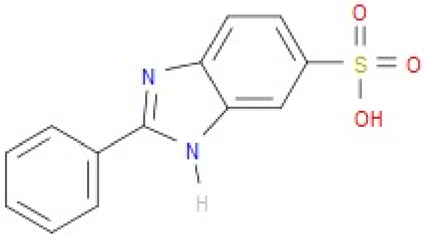
[ID 44] 2-[4-(3-Amino-2-hydroxypropoxy)phenyl]acetamide

Ta+Nt Ta*+Nt



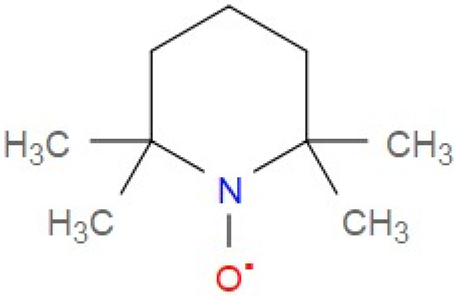
[ID 45] 2-Methoxy-5-methylaniline

Ta+Nt Ta*+Nt

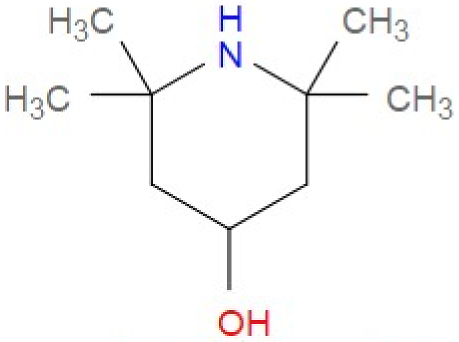


[ID 46] 2-Phenylbenzimidazole-5-sulfonic acid

Ta+Nt Ta*+Nt

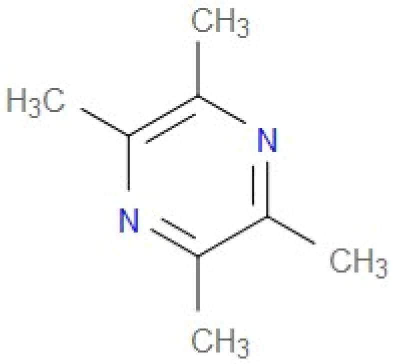


[ID 47] 2,2,6,6-Tetramethyl-1-piperidinol (TEMPO)

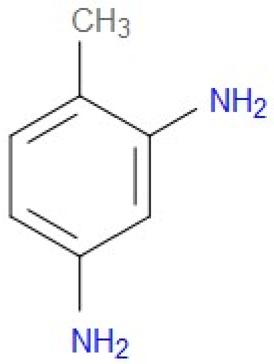


[ID 48] 2,2,6,6-Tetramethyl-4-piperidinol

Chemical structures S2. Non-target chemicals. Datasets: Ta+Nt = target and non-target chemicals, Ta*+Nt = target -> endpoint outliers removed and non-target chemicals, Pe = PEGs, Pg = PPGs, Pe+Pg = PEGs and PPGs. Red: training set, blue: test set.

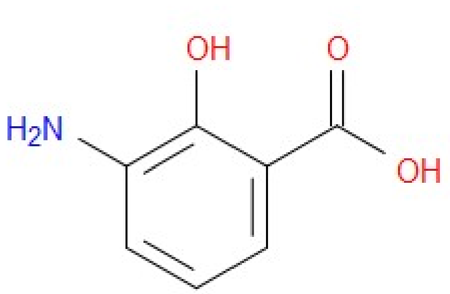


[ID 49] 2,3,5,6-Tetramethylpyrazine

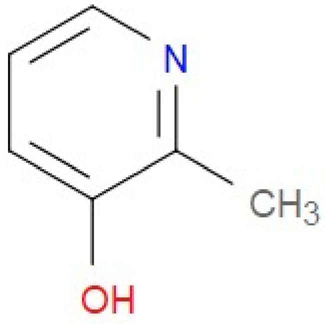


[ID 50] 2,4-Diaminotoluene

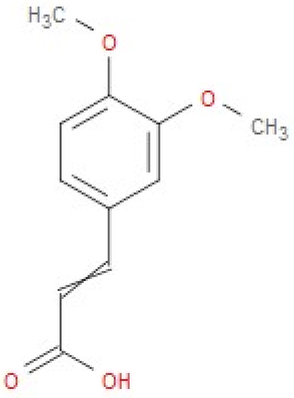
Ta+Nt



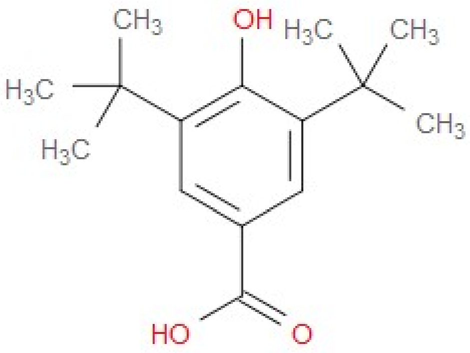
[ID 51] 3-Aminosalicylic acid



[ID 52] 3-Hydroxy-2-methylpyridine

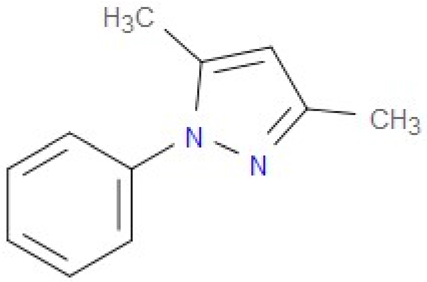


[ID 53] 3,4-Dimethoxycinnamic acid



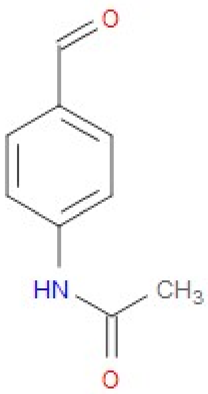
[ID 54] 3,5-di-tert-Butyl-4-hydroxybenzoic acid

Chemical structures S2. Non-target chemicals. Datasets: Ta+Nt = target and non-target chemicals, Ta*+Nt = target -> endpoint outliers removed and non-target chemicals, Pe = PEGs, Pg = PPGs, Pe+Pg = PEGs and PPGs. Red: training set, blue: test set.



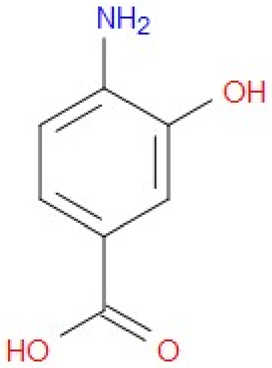
[ID 55] 3,5-Dimethyl-1-phenylpyrazole

Ta+Nt Ta*+Nt

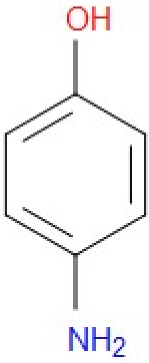


[ID 56] 4-Acetamidobenzaldehyde

Ta+Nt

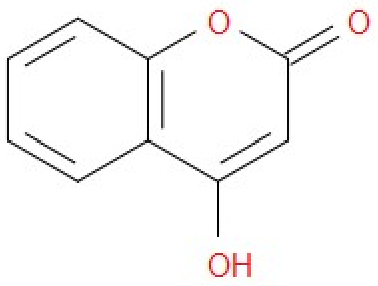


[ID 57] 4-Amino-3-hydroxybenzoic acid

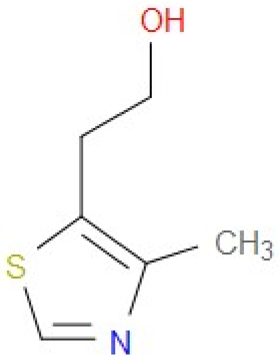


[ID 58] 4-Aminophenol

Ta+Nt

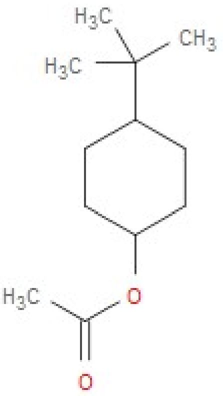


[ID 59] 4-Hydroxycoumarin

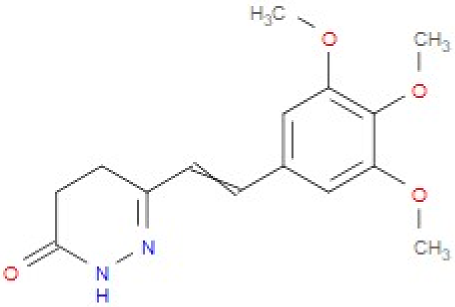


[ID 60] 4-Methyl-5-thiazoleethanol

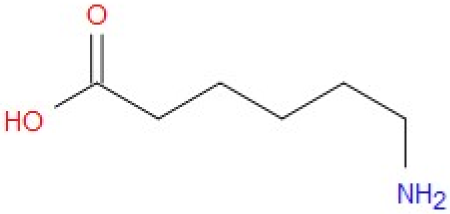
Chemical structures S2. Non-target chemicals. Datasets: Ta+Nt = target and non-target chemicals, Ta*+Nt = target -> endpoint outliers removed and non-target chemicals, Pe = PEGs, Pg = PPGs, Pe+Pg = PEGs and PPGs. Red: training set, blue: test set.



[ID 61] 4-tert-Butylcyclohexyl acetate

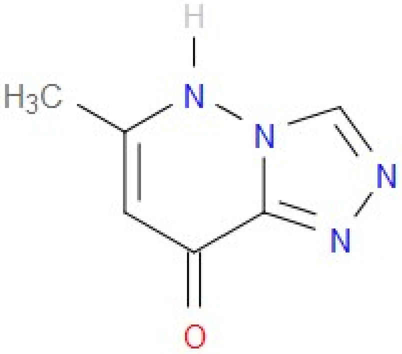


[ID 62] 6-(3,4,5-Trimethoxystyryl)-2,3,4,5-tetrahydropyridazin-3-one



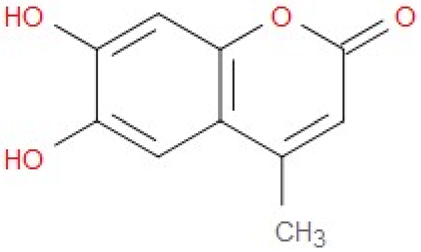
[ID 63] 6-Aminocaproic acid

Ta+Nt Ta*+Nt

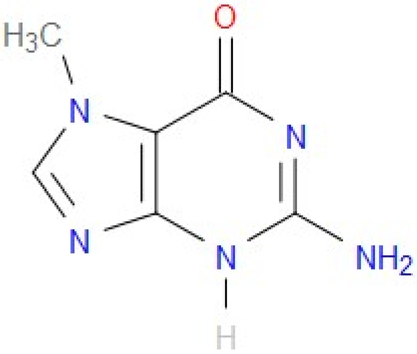


[ID 64] 6-Methyl[1,2,4]triazolo[4,3-b]pyridazin-8-ol

Ta+Nt Ta*+Nt



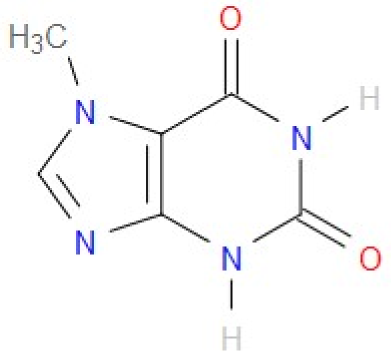
[ID 65] 6,7-Dihydroxy-4-methylcoumarin



[ID 66] 7-Methylguanine

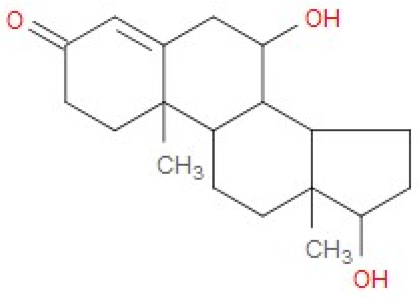
Ta+Nt Ta*+Nt

Chemical structures S2. Non-target chemicals. Datasets: Ta+Nt = target and non-target chemicals, Ta*+Nt = target -> endpoint outliers removed and non-target chemicals, Pe = PEGs, Pg = PPGs, Pe+Pg = PEGs and PPGs. Red: training set, blue: test set.

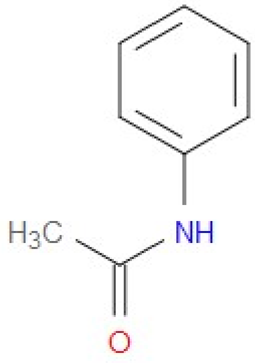


[ID 67] 7-Methylxanthine

Ta+Nt Ta*+Nt

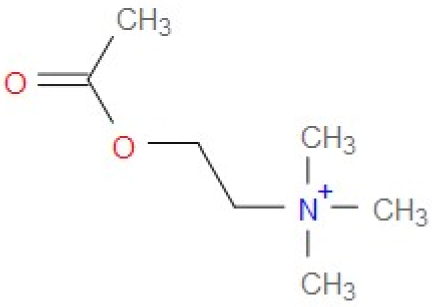


[ID 68] 7α-Hydroxytestosterone

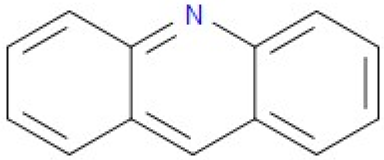


[ID 69] Acetanilide

Ta+Nt

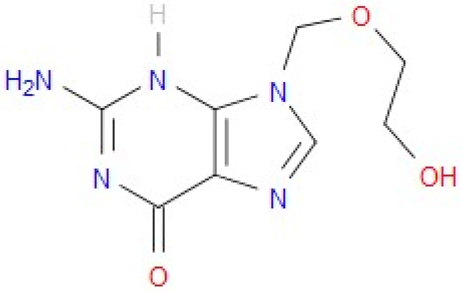


[ID 70] Acetylcholine



[ID 71] Acridine

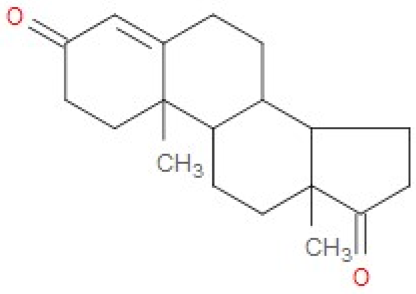
Ta+Nt   Ta*+Nt



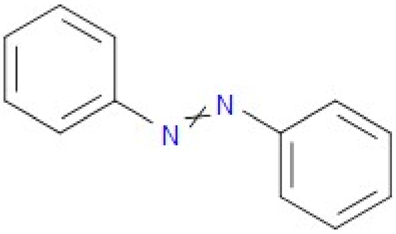
[ID 72] Acycloguanosine

Ta+Nt Ta*+Nt

Chemical structures S2. Non-target chemicals. Datasets: Ta+Nt = target and non-target chemicals, Ta*+Nt = target -> endpoint outliers removed and non-target chemicals, Pe = PEGs, Pg = PPGs, Pe+Pg = PEGs and PPGs. Red: training set, blue: test set.

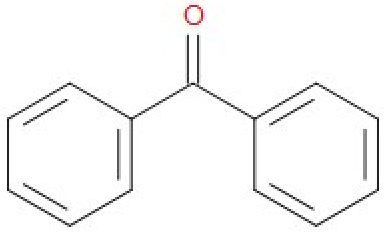


[ID 73] Androstenedione



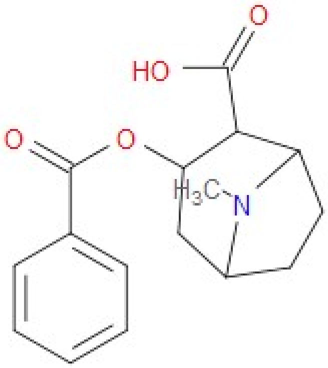
[ID 75] Azobenzene

Ta+Nt Ta*+Nt

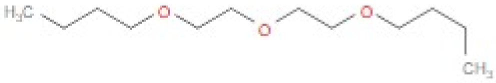


[ID 76] Benzophenone

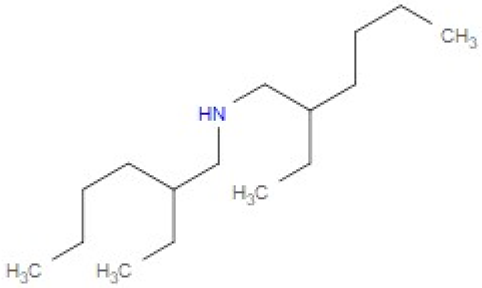
Ta+Nt Ta*+Nt



[ID 77] Benzoylecgonine

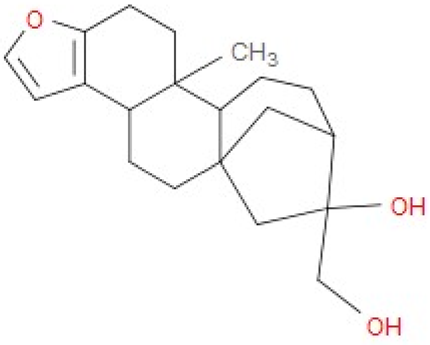


[ID 78] Bis(2-butoxyethyl) ether

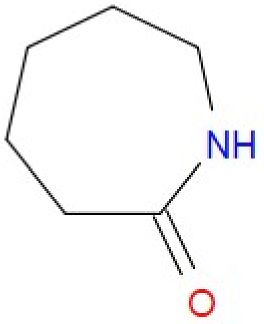


[ID 79] Bis(2-ethylhexyl) amine

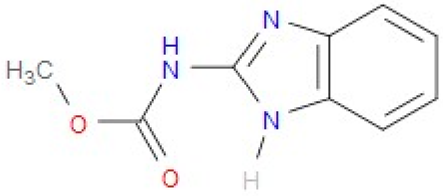
Chemical structures S2. Non-target chemicals. Datasets: Ta+Nt = target and non-target chemicals, Ta*+Nt = target -> endpoint outliers removed and non-target chemicals, Pe = PEGs, Pg = PPGs, Pe+Pg = PEGs and PPGs. Red: training set, blue: test set.



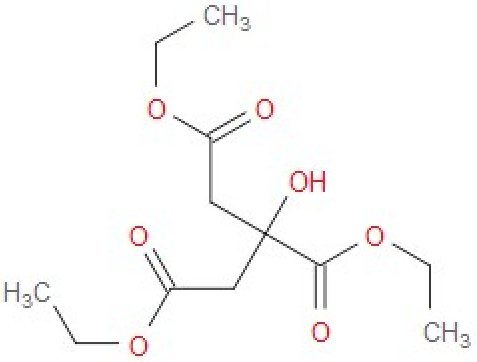
[ID 80] Cafestol



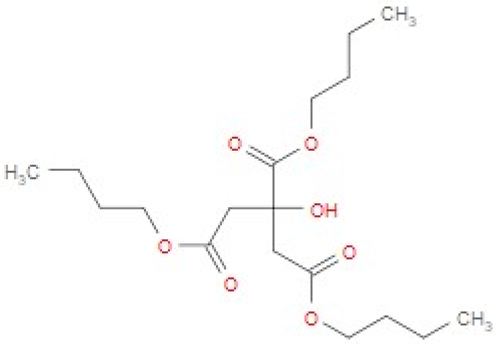
[ID 81] Caprolactam



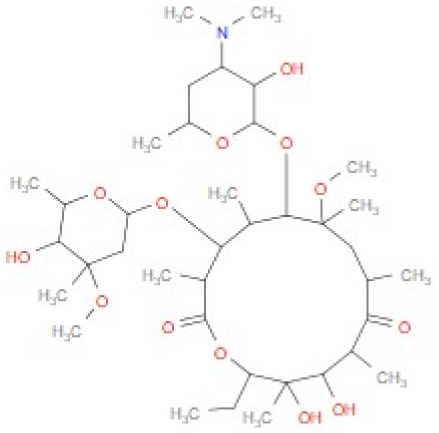
[ID 82] Carbendazim



[ID 83] Citroflex 2

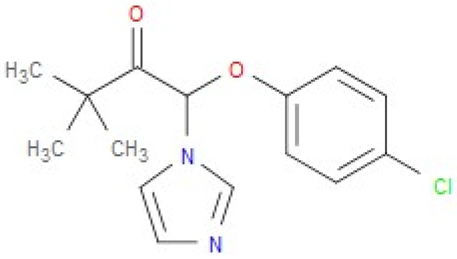


[ID 84] Citroflex 4



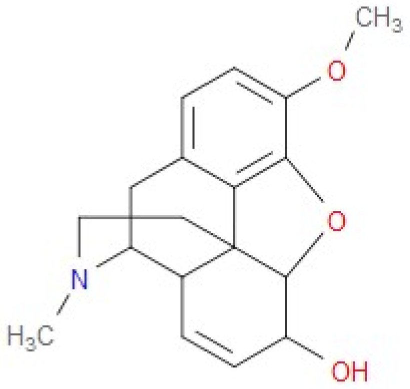
[ID 85] Clarithromycin

Chemical structures S2. Non-target chemicals. Datasets: Ta+Nt = target and non-target chemicals, Ta*+Nt = target -> endpoint outliers removed and non-target chemicals, Pe = PEGs, Pg = PPGs, Pe+Pg = PEGs and PPGs. Red: training set, blue: test set.

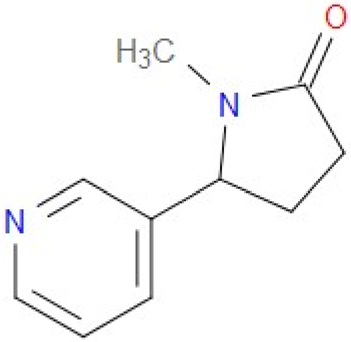


[ID 86] Climbazole

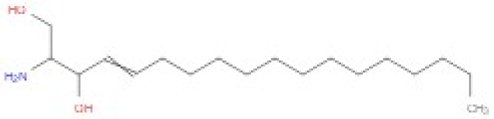
Ta+Nt Ta*+Nt



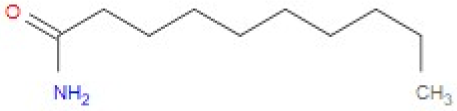
[ID 87] Codeine



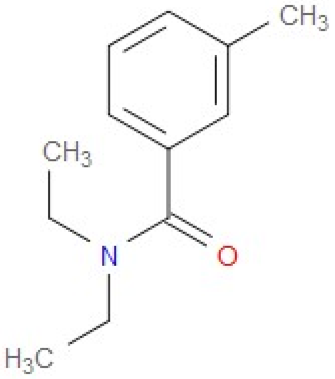
[ID 88] Cotinine



[ID 89] D-Sphingosine

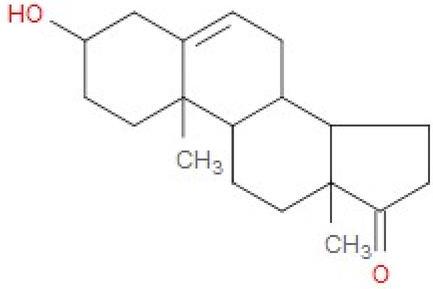


[ID 90] Decanamide

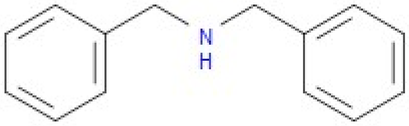


[ID 91] DEET

Chemical structures S2. Non-target chemicals. Datasets: Ta+Nt = target and non-target chemicals, Ta*+Nt = target -> endpoint outliers removed and non-target chemicals, Pe = PEGs, Pg = PPGs, Pe+Pg = PEGs and PPGs. Red: training set, blue: test set.

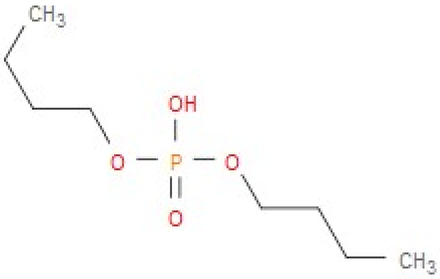


[ID 92] Dehydroepiandrosterone (DHEA)

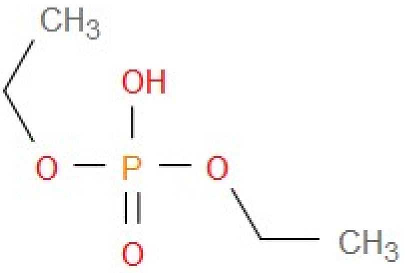


[ID 93] Dibenzylamine

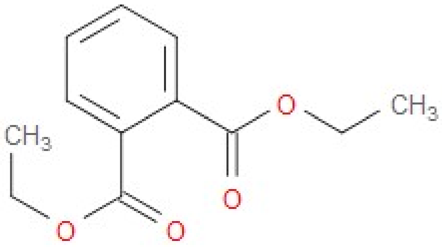
Ta+Nt Ta*+Nt



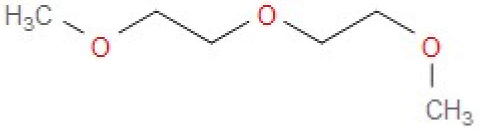
[ID 94] Dibutyl phosphate



[ID 95] Diethyl phosphate

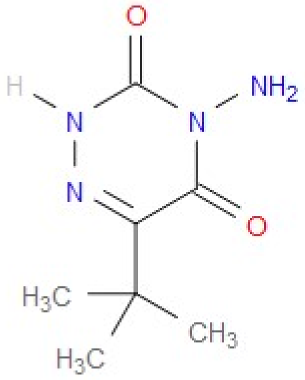


[ID 96] Diethyl phthalate

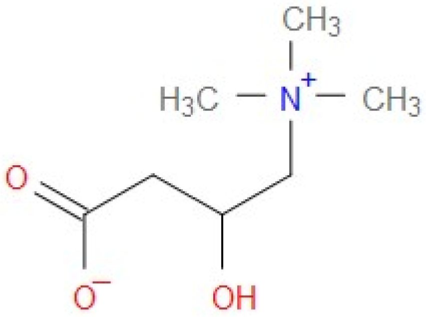


[ID 97] Diglyme

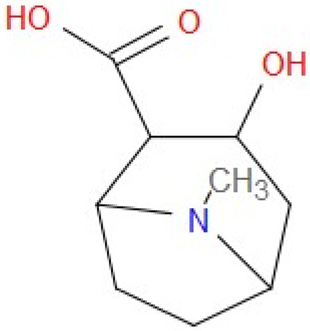
Chemical structures S2. Non-target chemicals. Datasets: Ta+Nt = target and non-target chemicals, Ta*+Nt = target -> endpoint outliers removed and non-target chemicals, Pe = PEGs, Pg = PPGs, Pe+Pg = PEGs and PPGs. Red: training set, blue: test set.



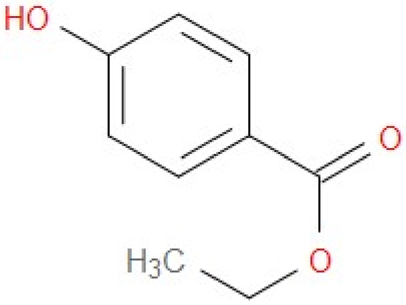
[ID 98] Diketo-Metribuzin



[ID 99] DL-Carnitine

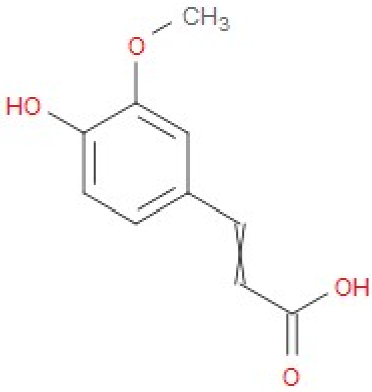


[ID 100] Ecgonine



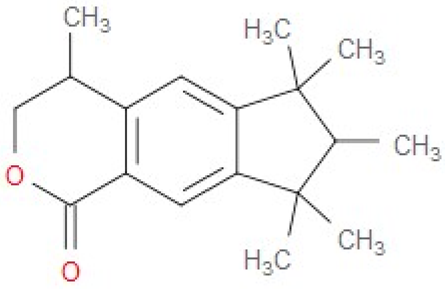
[ID 101] Ethyl paraben

Ta+Nt



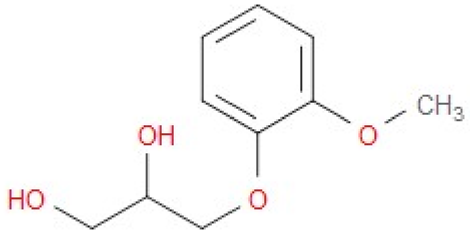
[ID 102] Ferulic acid

Ta+Nt



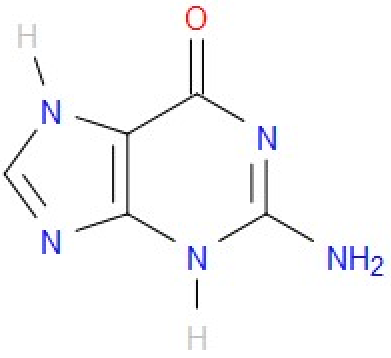
[ID 103] Galaxolidone

Chemical structures S2. Non-target chemicals. Datasets: Ta+Nt = target and non-target chemicals, Ta*+Nt = target -> endpoint outliers removed and non-target chemicals, Pe = PEGs, Pg = PPGs, Pe+Pg = PEGs and PPGs. Red: training set, blue: test set.

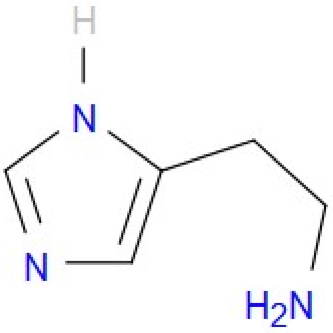


[ID 104] Guaifenesin

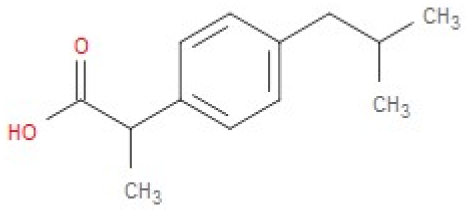
Ta+Nt Ta*+Nt



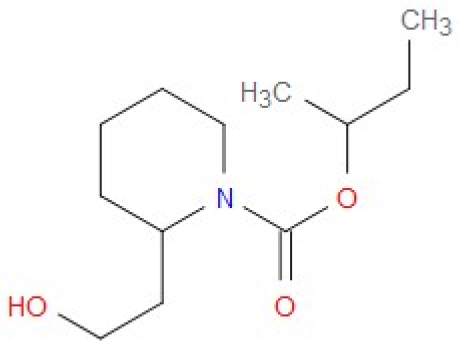
[ID 105] Guanine



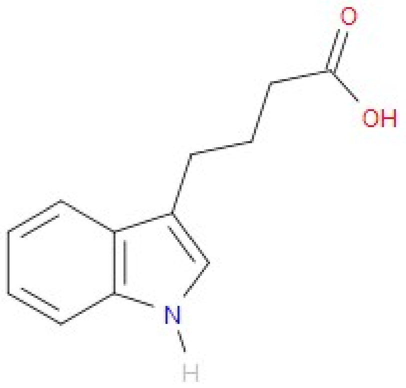
[ID 106] Histamine



[ID 107] Ibuprofen



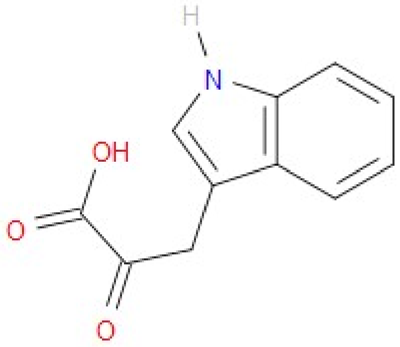
[ID 108] Icaridin



[ID 109] Indole-3-butyric acid

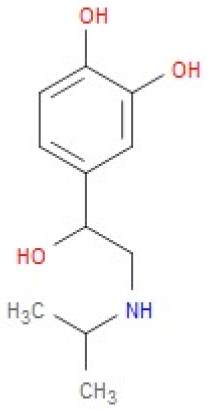
Ta+Nt Ta*+Nt

Chemical structures S2. Non-target chemicals. Datasets: Ta+Nt = target and non-target chemicals, Ta*+Nt = target -> endpoint outliers removed and non-target chemicals, Pe = PEGs, Pg = PPGs, Pe+Pg = PEGs and PPGs. Red: training set, blue: test set.



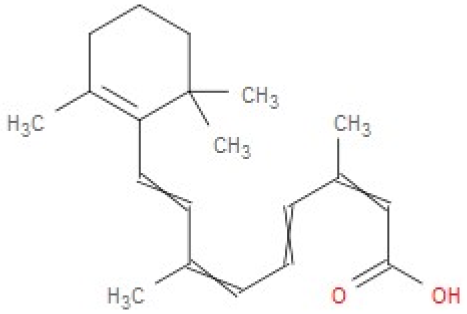
[ID 110] Indole-3-pyruvic acid

Ta+Nt Ta*+Nt

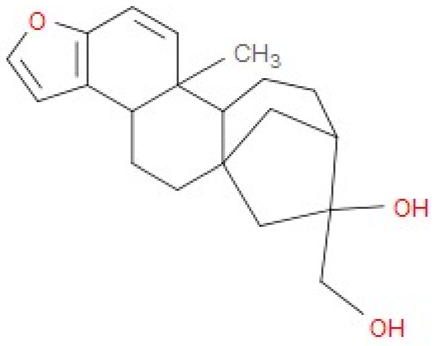


[ID 111] Isoprenaline

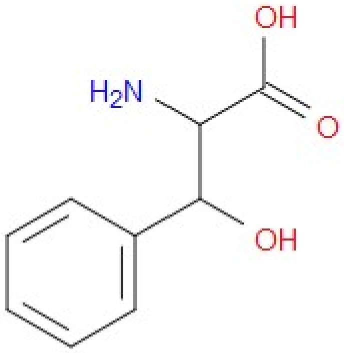
Ta+Nt Ta*+Nt



[ID 112] Isotretinoin

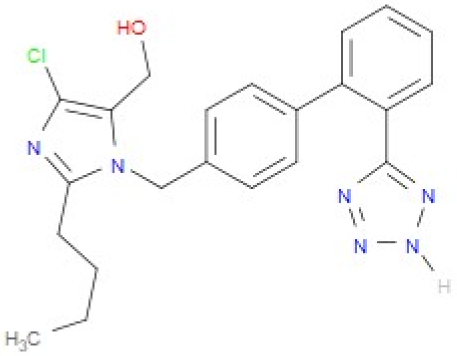


[ID 113] Kahweol



[ID 114] L-threo-3-Phenylserine

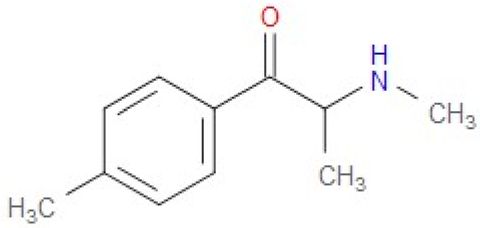
Ta+Nt   Ta*+Nt



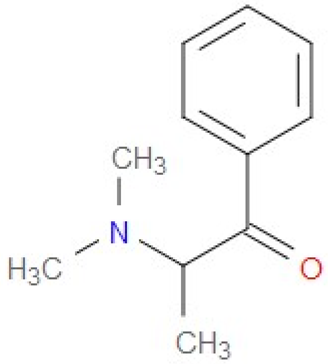
[ID 115] Losartan

Ta+Nt Ta*+Nt

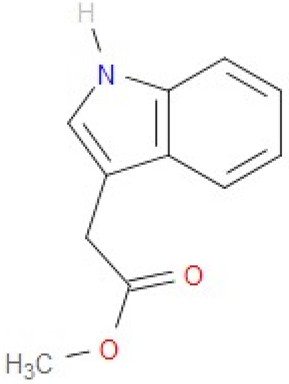
Chemical structures S2. Non-target chemicals. Datasets: Ta+Nt = target and non-target chemicals, Ta*+Nt = target -> endpoint outliers removed and non-target chemicals, Pe = PEGs, Pg = PPGs, Pe+Pg = PEGs and PPGs. Red: training set, blue: test set.



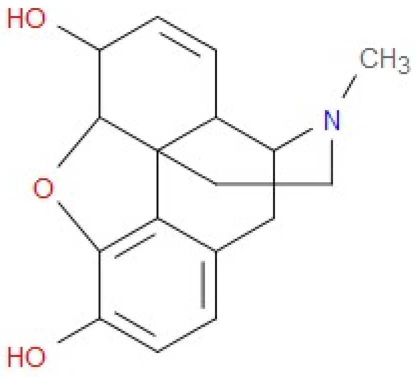
[ID 116] Mephedrone



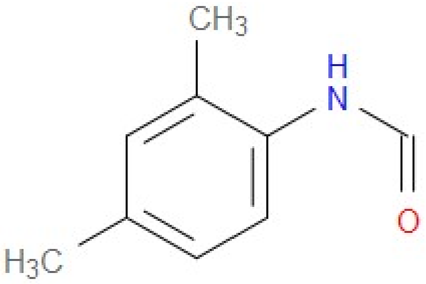
[ID 117] Metamfepramone
Ta+Nt



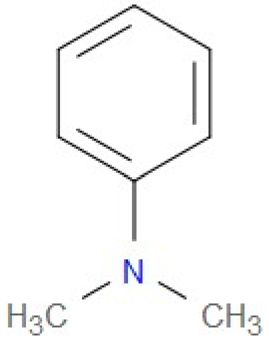
[ID 118] Methyl indole-3-acetate
Ta+Nt Ta*+Nt



[ID 119] Morphine

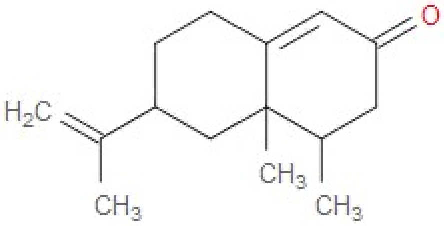


[ID 120] N-(2,4-Dimethylphenyl) formamide
Ta+Nt

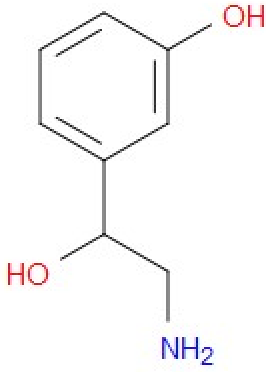


[ID 121] N,N-Dimethylaniline
Ta+Nt

Chemical structures S2. Non-target chemicals. Datasets: Ta+Nt = target and non-target chemicals, Ta*+Nt = target -> endpoint outliers removed and non-target chemicals, Pe = PEGs, Pg = PPGs, Pe+Pg = PEGs and PPGs. Red: training set, blue: test set.

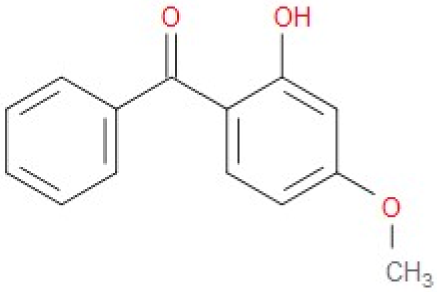


[ID 122] Nootkatone



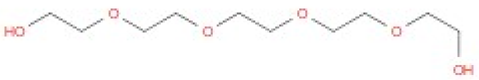
[ID 123] Norfenefrine

Ta+Nt



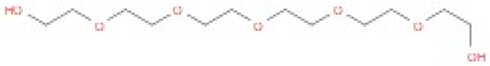
[ID 124] Oxybenzone

Ta+Nt Ta*+Nt



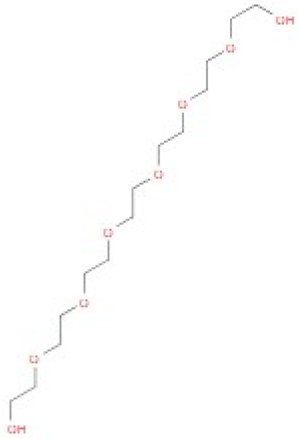
[ID 126] PEG n5

Pe Pe+Pg



[ID 127] PEG n6

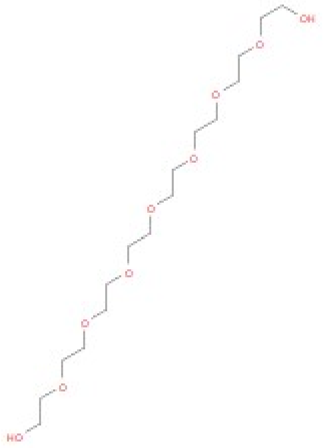
Pe Pe+Pg



[ID 128] PEG n7

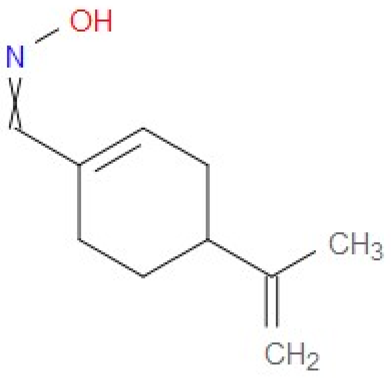
Pe Pe+Pg

Chemical structures S2. Non-target chemicals. Datasets: Ta+Nt = target and non-target chemicals, Ta*+Nt = target -> endpoint outliers removed and non-target chemicals, Pe = PEGs, Pg = PPGs, Pe+Pg = PEGs and PPGs. Red: training set, blue: test set.

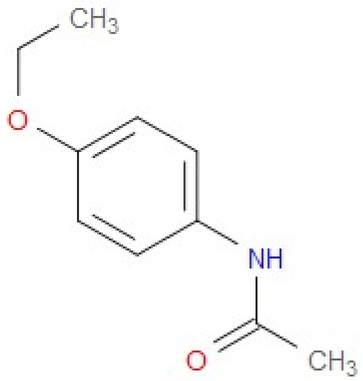


[ID 129] PEG n8

Pe Pe+Pg

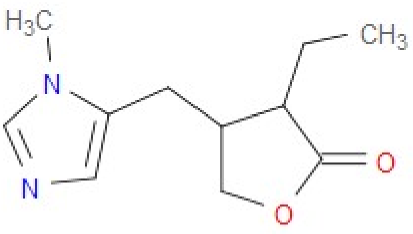


[ID 130] Perillartine

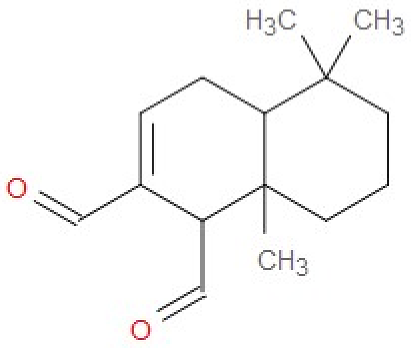


[ID 131] Phenacetin

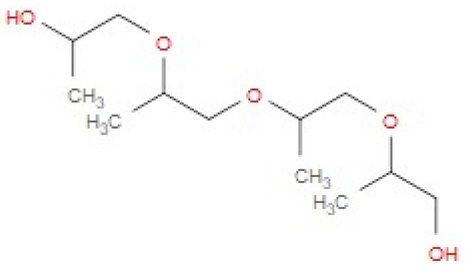
Ta+Nt Ta*+Nt



[ID 132] Pilocarpine



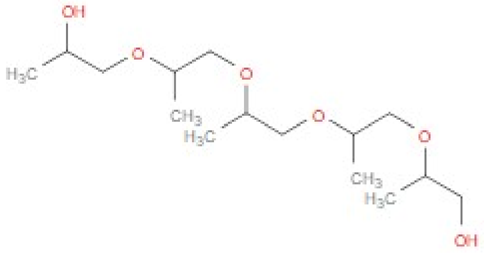
[ID 133] Polygodial



[ID 134] PPG n4

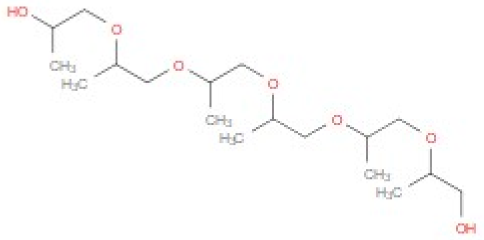
Pg Pe+Pg

Chemical structures S2. Non-target chemicals. Datasets: Ta+Nt = target and non-target chemicals, Ta*+Nt = target -> endpoint outliers removed and non-target chemicals, Pe = PEGs, Pg = PPGs, Pe+Pg = PEGs and PPGs. Red: training set, blue: test set.



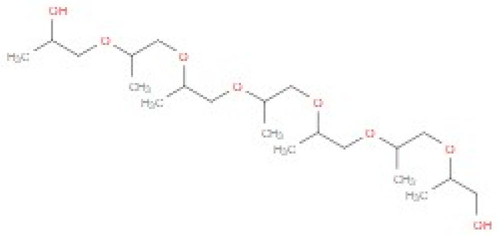
[ID 135] PPG n5

Pg Pe+Pg



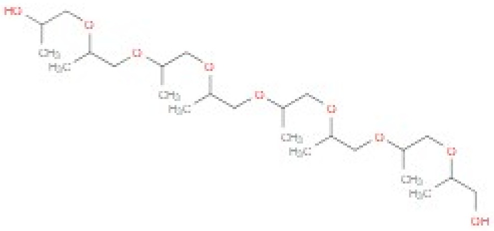
[ID 136] PPG n6

Pg Pe+Pg



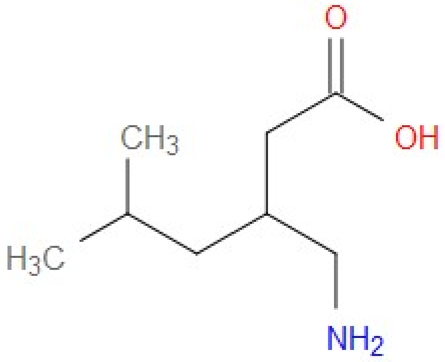
[ID 137] PPG n7

Pg Pe+Pg

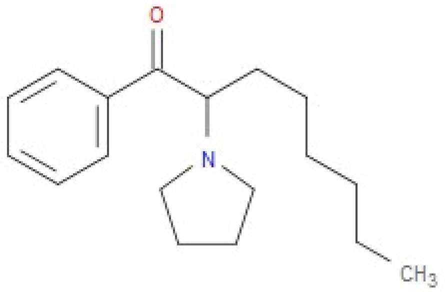


[ID 138] PPG n8

Pg Pe+Pg



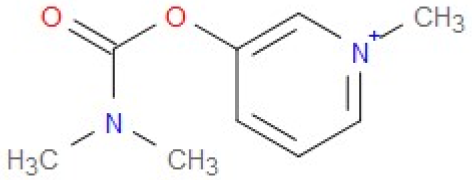
[ID 139] Pregabalin



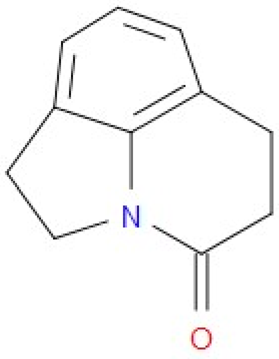
[ID 140] PV9

Ta+Nt Ta*+Nt

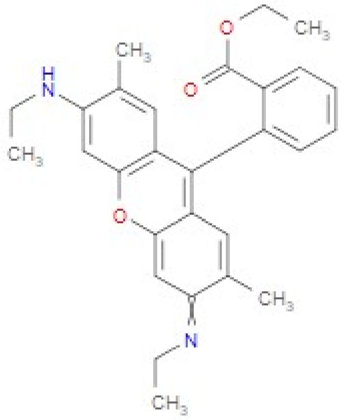
Chemical structures S2. Non-target chemicals. Datasets: Ta+Nt = target and non-target chemicals, Ta*+Nt = target -> endpoint outliers removed and non-target chemicals, Pe = PEGs, Pg = PPGs, Pe+Pg = PEGs and PPGs. Red: training set, blue: test set.



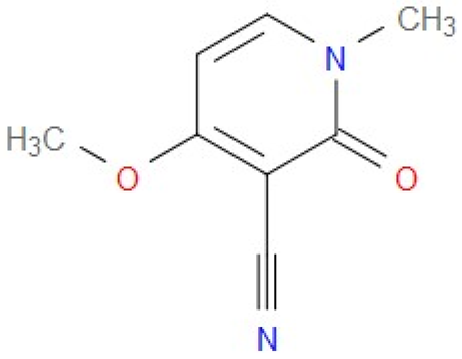
[ID 141] Pyridostigmine



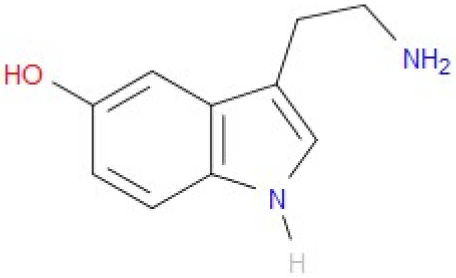
[ID 142] Pyroquilon



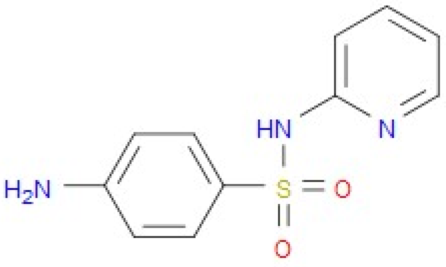
[ID 143] Rhodamine 6G



[ID 144] Ricinine



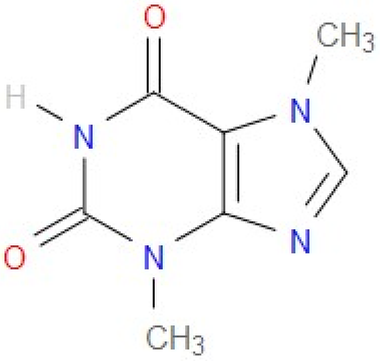
[ID 145] Serotonin



[ID 146] Sulfapyridine

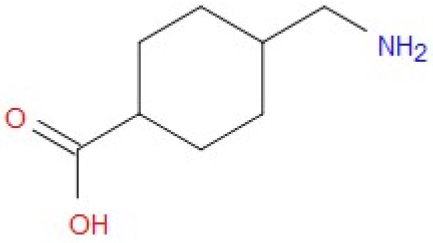
Ta+Nt Ta*+Nt

Chemical structures S2. Non-target chemicals. Datasets: Ta+Nt = target and non-target chemicals, Ta*+Nt = target -> endpoint outliers removed and non-target chemicals, Pe = PEGs, Pg = PPGs, Pe+Pg = PEGs and PPGs. Red: training set, blue: test set.

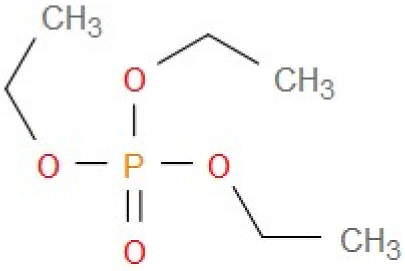


[ID 147] Theobromine

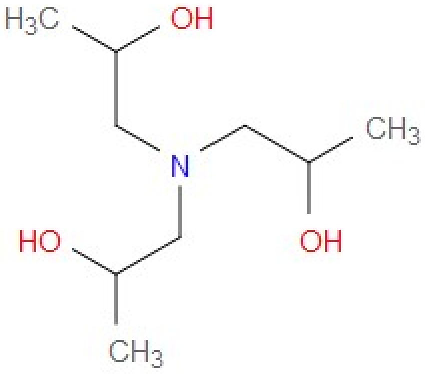
Ta+Nt Ta*+Nt



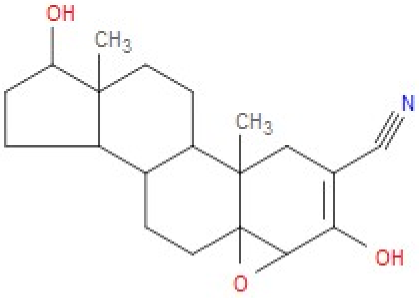
[ID 148] Tranexamic acid



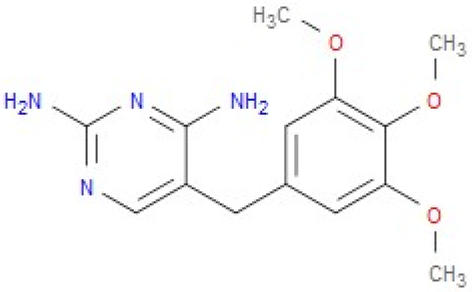
[ID 149] Triethyl phosphate



[ID 150] Triisopropanolamine

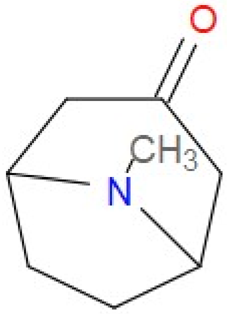


[ID 151] Trilostane

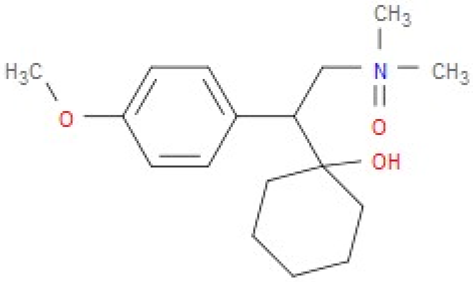


[ID 152] Trimethoprim

Chemical structures S2. Non-target chemicals. Datasets: Ta+Nt = target and non-target chemicals, Ta*+Nt = target -> endpoint outliers removed and non-target chemicals, Pe = PEGs, Pg = PPGs, Pe+Pg = PEGs and PPGs. Red: training set, blue: test set.

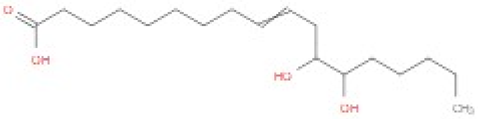


[ID 153] Tropinone

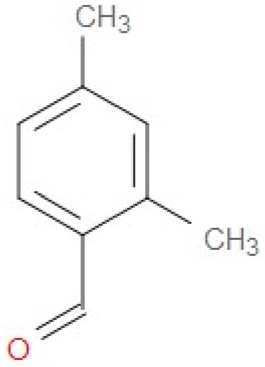


[ID 154] Venlafaxine N-Oxide

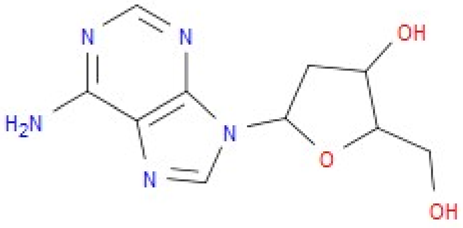
Ta+Nt Ta*+Nt



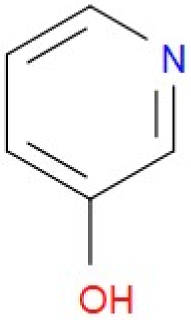
[ID 155] (+/-)12(13)-DiHOME



[ID 156] 2,4-Dimethylbenzaldehyde



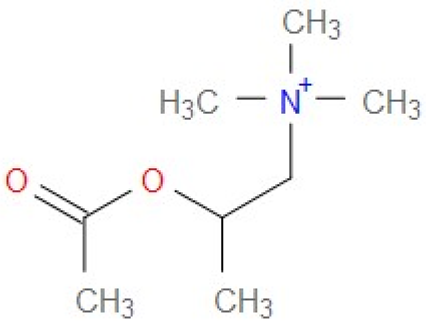
[ID 157] 2'-Deoxyadenosine



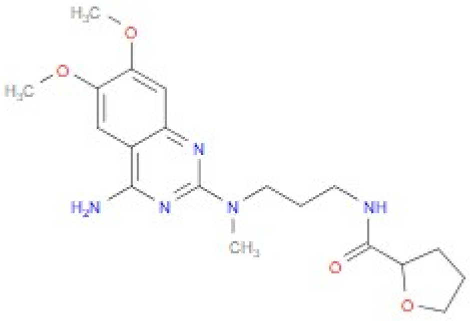
[ID 158] 3-Hydroxypyridine

Ta+Nt

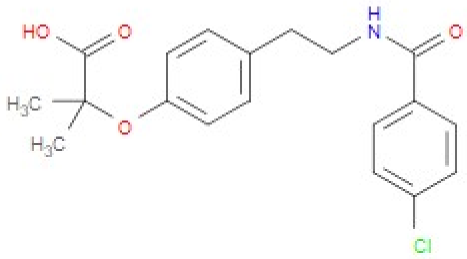
Chemical structures S2. Non-target chemicals. Datasets: Ta+Nt = target and non-target chemicals, Ta*+Nt = target -> endpoint outliers removed and non-target chemicals, Pe = PEGs, Pg = PPGs, Pe+Pg = PEGs and PPGs. Red: training set, blue: test set.



[ID 159] Acetyl-β-methylcholine

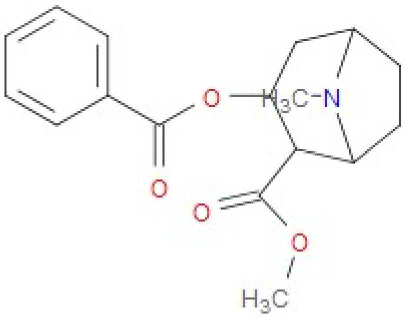


[ID 160] Alfuzosin

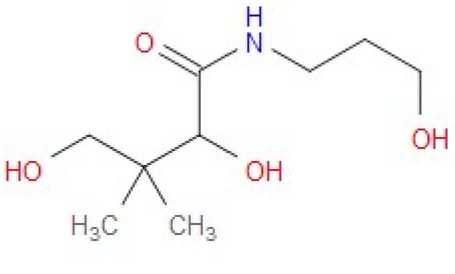


[ID 161] Bezafibrate

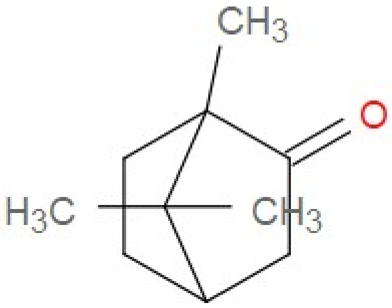
Ta+Nt Ta*+Nt



[ID 162] Cocaine

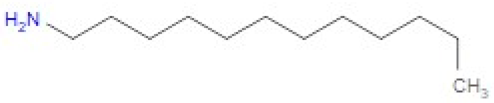


[ID 163] D-Panthenol

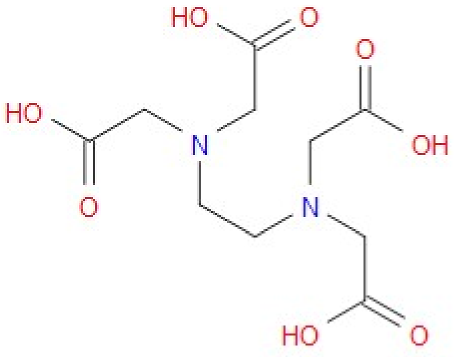


[ID 164] D,L-Camphor

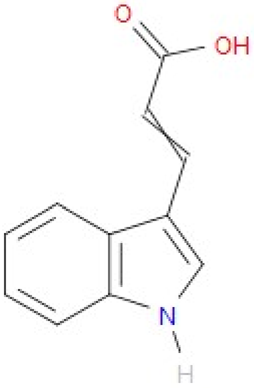
Chemical structures S2. Non-target chemicals. Datasets: Ta+Nt = target and non-target chemicals, Ta*+Nt = target -> endpoint outliers removed and non-target chemicals, Pe = PEGs, Pg = PPGs, Pe+Pg = PEGs and PPGs. Red: training set, blue: test set.



[ID 165] Dodecylamine

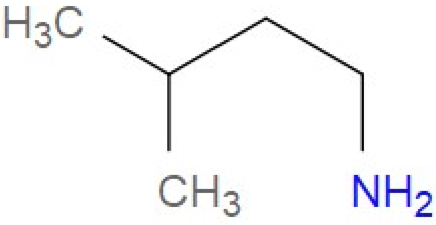


[ID 166] Ethylenediaminetetraacetic acid (EDTA)

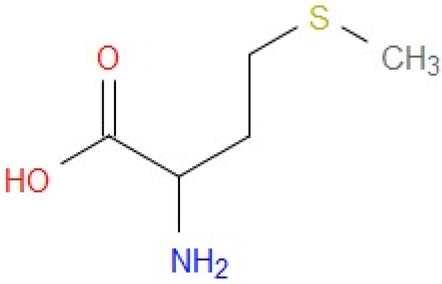


[ID 167] Indole-3-acrylic acid

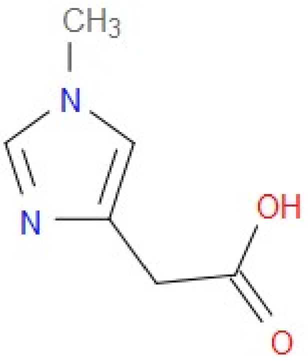
Ta+Nt Ta*+Nt



[ID 168] Isoamylamine

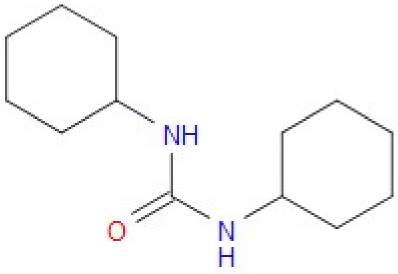


[ID 169] Methionine

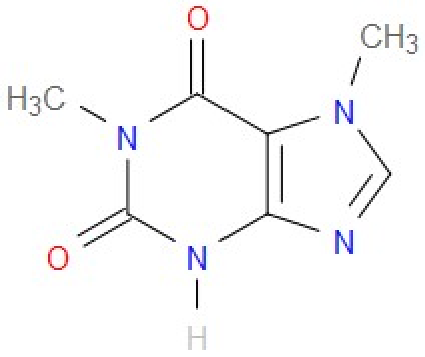


[ID 170] Methylimidazoleacetic acid

Chemical structures S2. Non-target chemicals. Datasets: Ta+Nt = target and non-target chemicals, Ta*+Nt = target -> endpoint outliers removed and non-target chemicals, Pe = PEGs, Pg = PPGs, Pe+Pg = PEGs and PPGs. Red: training set, blue: test set.

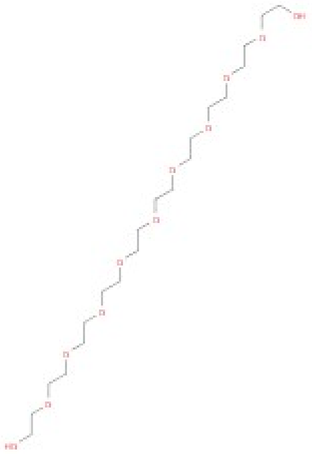


[ID 171] N,N'-Dicyclohexylurea



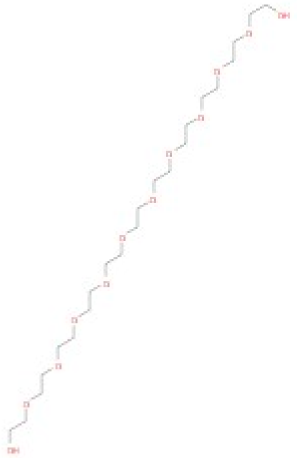
[ID 172] Paraxanthine

Ta+Nt Ta*+Nt



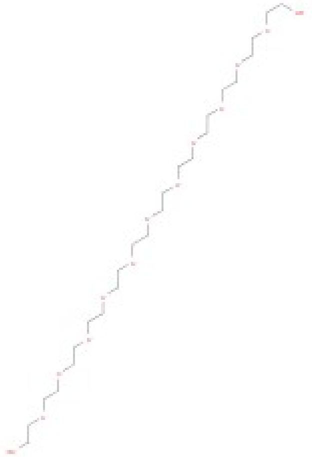
[ID 173] PEG n10

Pe Pe+Pg



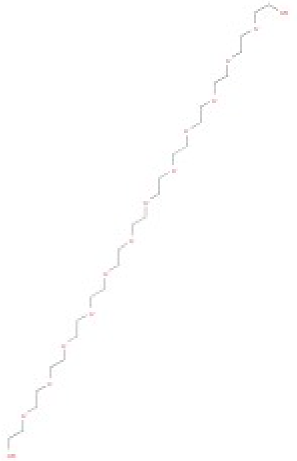
[ID 174] PEG n11

Pe Pe+Pg



[ID 175] PEG n12

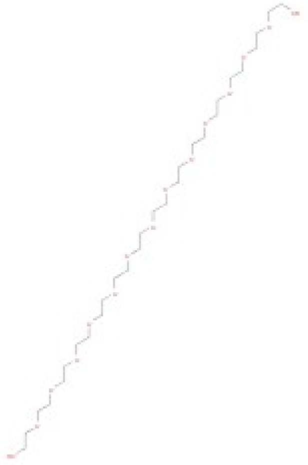
Pe Pe+Pg



[ID 176] PEG n13

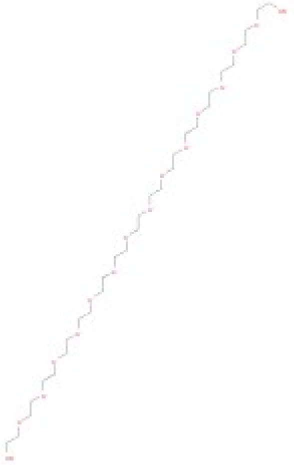
Pe Pe+Pg

Chemical structures S2. Non-target chemicals. Datasets: Ta+Nt = target and non-target chemicals, Ta*+Nt = target -> endpoint outliers removed and non-target chemicals, Pe = PEGs, Pg = PPGs, Pe+Pg = PEGs and PPGs. Red: training set, blue: test set.



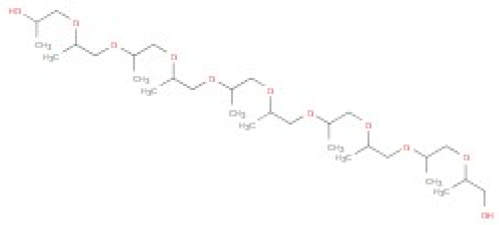
[ID 177] PEG n14

Pe Pe+Pg



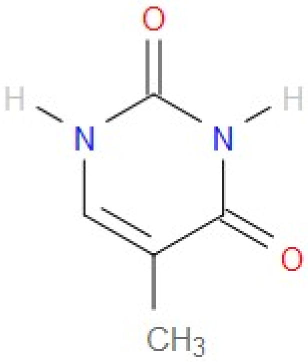
[ID 178] PEG n15

Pe Pe+Pg

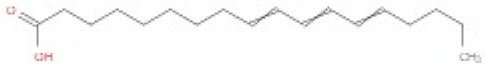


[ID 179] PPG n10

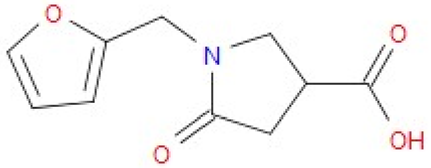
Pg Pe+Pg



[ID 180] Thymine

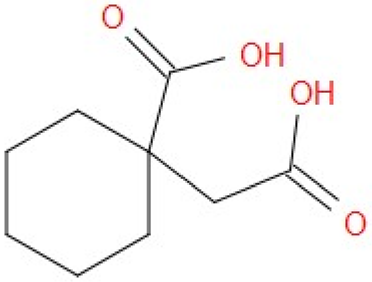


[ID 181] α-Eleostearic acid



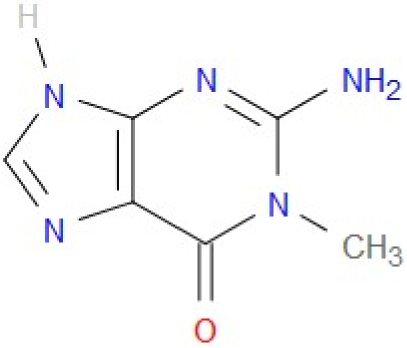
[ID 183] 1-(2-Furylmethyl)-5-oxopyrrolidine-3-carboxylic acid

Chemical structures S2. Non-target chemicals. Datasets: Ta+Nt = target and non-target chemicals, Ta*+Nt = target -> endpoint outliers removed and non-target chemicals, Pe = PEGs, Pg = PPGs, Pe+Pg = PEGs and PPGs. Red: training set, blue: test set.



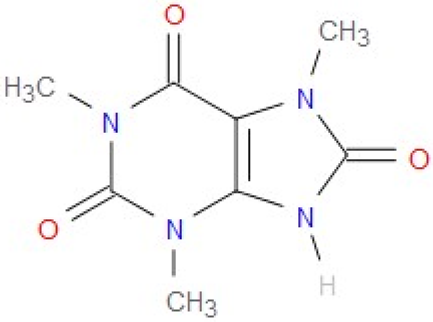
[ID 184] 1-(Carboxymethyl) cyclohexanecarboxylic acid

Ta+Nt Ta*+Nt



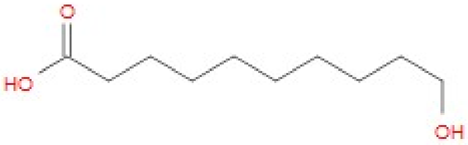
[ID 185] 1-Methylguanine

Ta+Nt Ta*+Nt

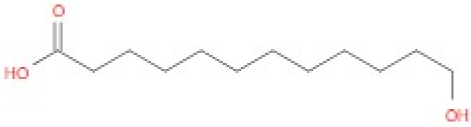


[ID 186] 1,3,7-Trimethyluric acid

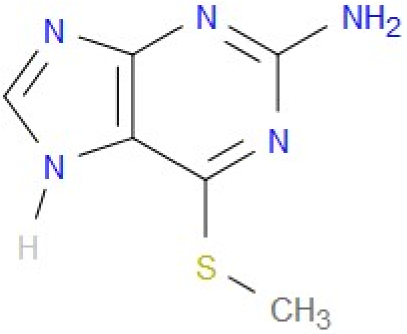
Ta+Nt Ta*+Nt



[ID 188] 10-Hydroxydecanoic acid

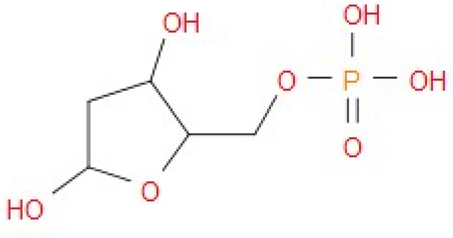


[ID 189] 12-Hydroxydodecanoic acid

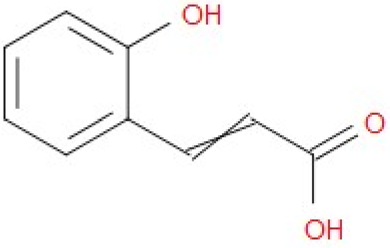


[ID 190] 2-Amino-6-methylmercaptopurine

Chemical structures S2. Non-target chemicals. Datasets: Ta+Nt = target and non-target chemicals, Ta*+Nt = target -> endpoint outliers removed and non-target chemicals, Pe = PEGs, Pg = PPGs, Pe+Pg = PEGs and PPGs. Red: training set, blue: test set.

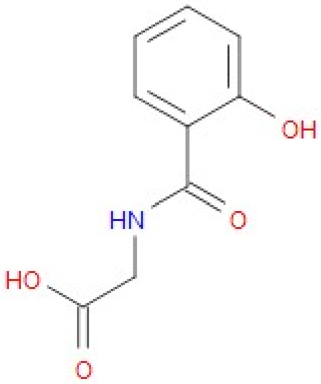


[ID 191] 2-Deoxyribose 5-phosphate

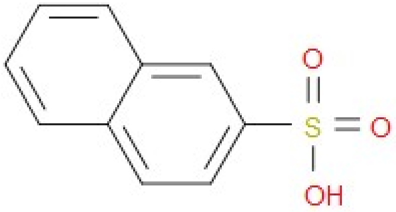


[ID 192] 2-Hydroxycinnamic acid

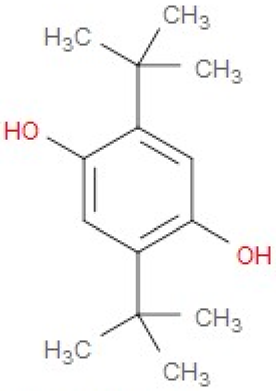
Ta+Nt



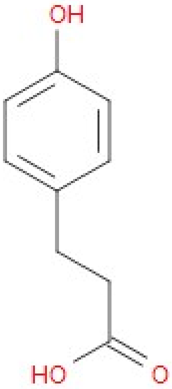
[ID 193] 2-Hydroxyhippuric acid



[ID 194] 2-Naphthalenesulfonic acid



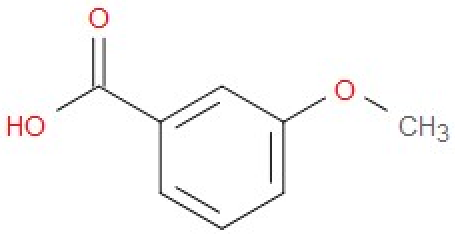
[ID 195] 2,5-di-tert-Butylhydroquinone



[ID 196] 3-(4-Hydroxyphenyl) propionic acid

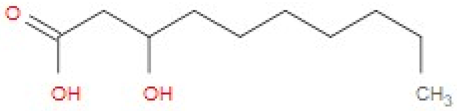
Ta+Nt

Chemical structures S2. Non-target chemicals. Datasets: Ta+Nt = target and non-target chemicals, Ta*+Nt = target -> endpoint outliers removed and non-target chemicals, Pe = PEGs, Pg = PPGs, Pe+Pg = PEGs and PPGs. Red: training set, blue: test set.

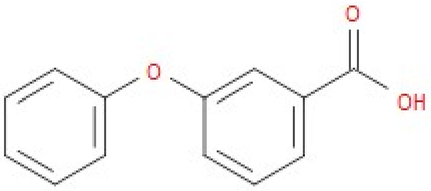


[ID 197] 3-Anisic acid

Ta+Nt

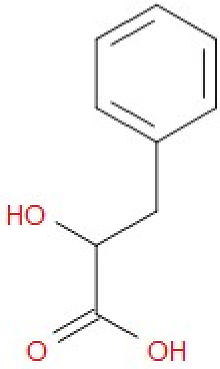


[ID 198] 3-Hydroxydecanoic acid



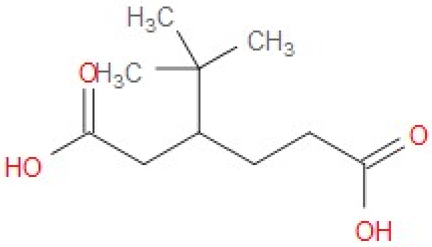
[ID 199] 3-Phenoxybenzoic acid

Ta+Nt Ta*+Nt

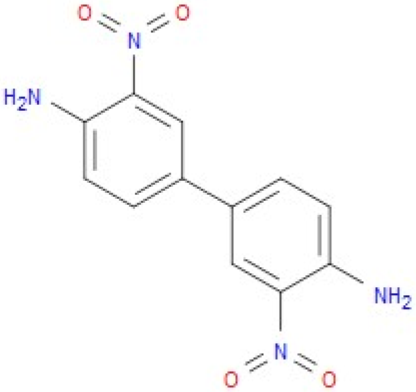


[ID 200] 3-Phenyllactic acid

Ta+Nt

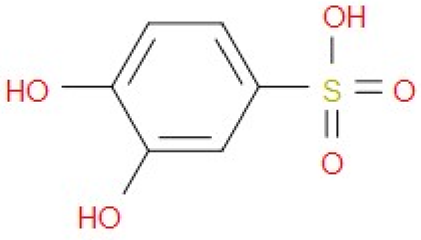


[ID 201] 3-tert-Butyladipic acid

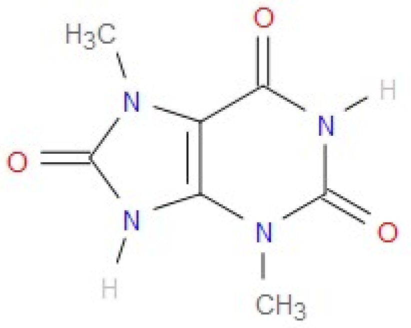


[ID 202] 3,3'-Dinitro(1,1'-biphenyl)-4,4'-diamine

Chemical structures S2. Non-target chemicals. Datasets: Ta+Nt = target and non-target chemicals, Ta*+Nt = target -> endpoint outliers removed and non-target chemicals, Pe = PEGs, Pg = PPGs, Pe+Pg = PEGs and PPGs. Red: training set, blue: test set.

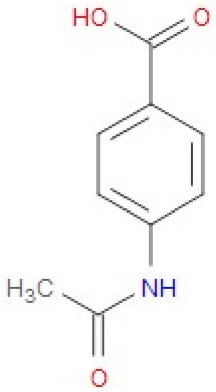


[ID 203] 3,4-Dihydroxybenzenesulfonic acid



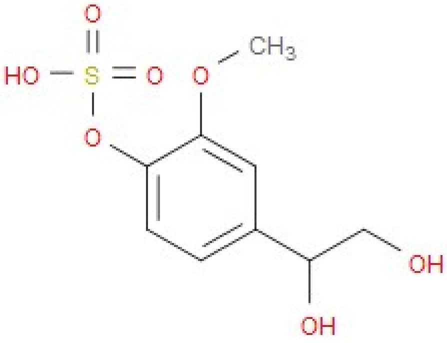
[ID 204] 3,7-Dimethyluric acid

Ta+Nt Ta*+Nt

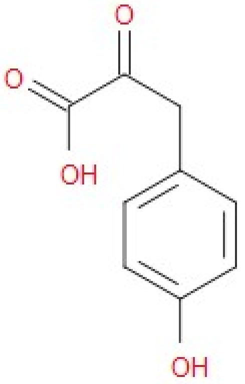


[ID 205] 4-Acetamidobenzoic acid

Ta+Nt

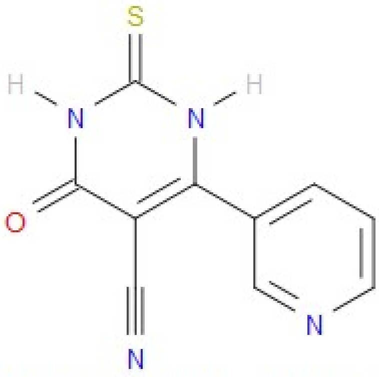


[ID 206] 4-Hydroxy-3-methoxyphenylglycol sulfate



[ID 207] 4-Hydroxyphenylpyruvic acid

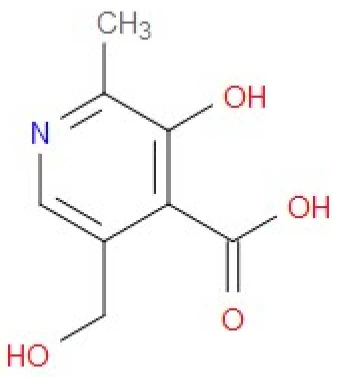
Ta+Nt



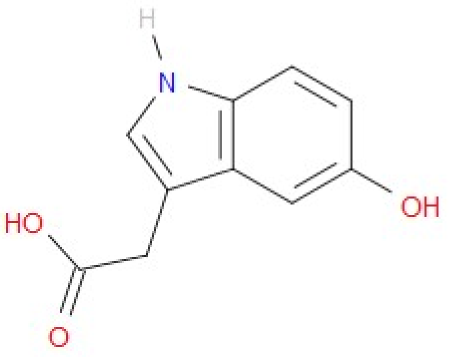
[ID 208] 4-Oxo-6-(3-pyridyl)-2-thioxo-1,2,3,4-tetrahydropyrimidine-5-carbonitrile

Ta+Nt Ta*+Nt

Chemical structures S2. Non-target chemicals. Datasets: Ta+Nt = target and non-target chemicals, Ta*+Nt = target -> endpoint outliers removed and non-target chemicals, Pe = PEGs, Pg = PPGs, Pe+Pg = PEGs and PPGs. Red: training set, blue: test set.

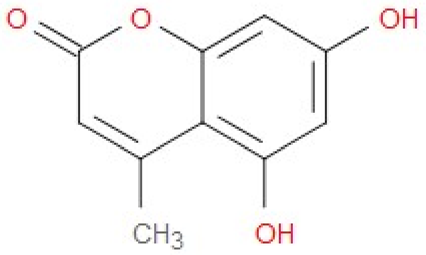


[ID 209] 4-Pyridoxic acid

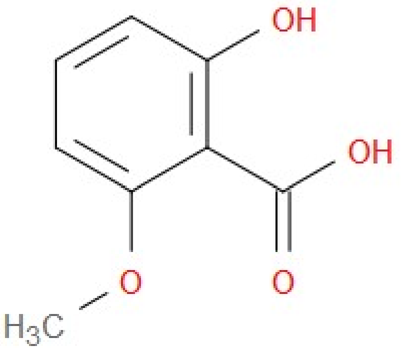


[ID 210] 5-Hydroxyindole-3-acetic acid

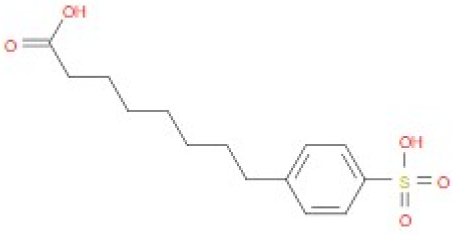
Ta+Nt Ta*+Nt



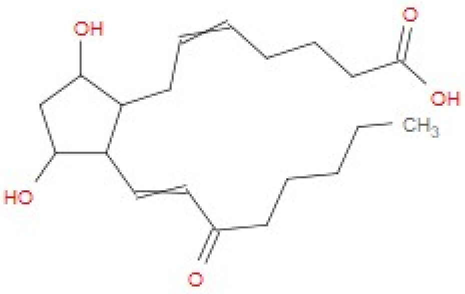
[ID 211] 5,7-Dihydroxy-4-methylcoumarin



[ID 212] 6-Methoxysalicylic acid

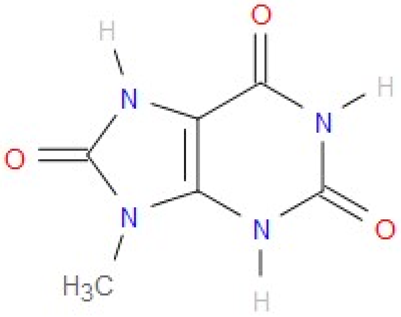


[ID 214] 8-(4-Sulfophenyl) octanoic acid

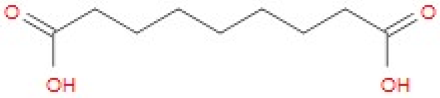


[ID 215] 8-Iso-15-keto-prostaglandin-F2β

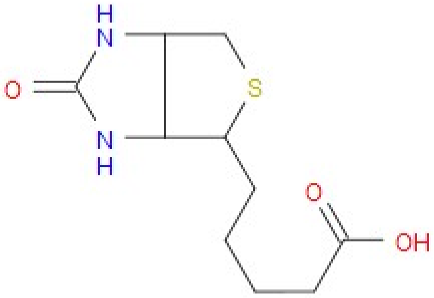
Chemical structures S2. Non-target chemicals. Datasets: Ta+Nt = target and non-target chemicals, Ta*+Nt = target -> endpoint outliers removed and non-target chemicals, Pe = PEGs, Pg = PPGs, Pe+Pg = PEGs and PPGs. Red: training set, blue: test set.



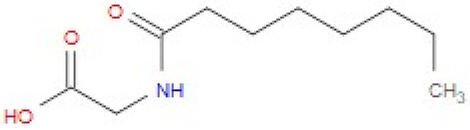
[ID 216] 9-Methyluric acid



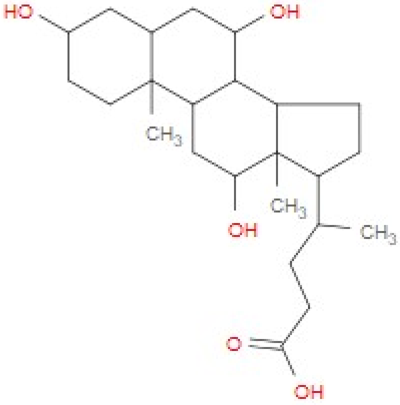
[ID 218] Azelaic acid



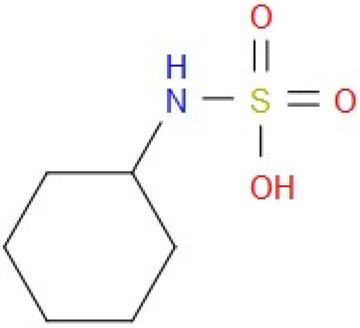
[ID 219] Biotin



[ID 220] Capryloylglycine

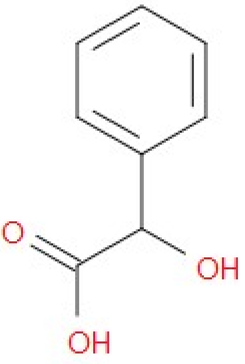


[ID 221] Cholic acid



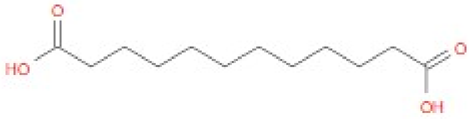
[ID 222] Cyclamic acid

Chemical structures S2. Non-target chemicals. Datasets: Ta+Nt = target and non-target chemicals, Ta*+Nt = target -> endpoint outliers removed and non-target chemicals, Pe = PEGs, Pg = PPGs, Pe+Pg = PEGs and PPGs. Red: training set, blue: test set.

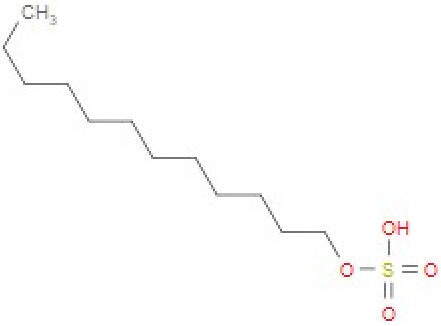


[ID 223] DL-Mandelic acid

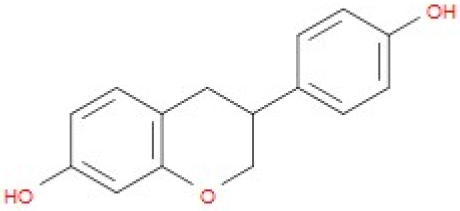
Ta+Nt



[ID 224] Dodecanedioic acid

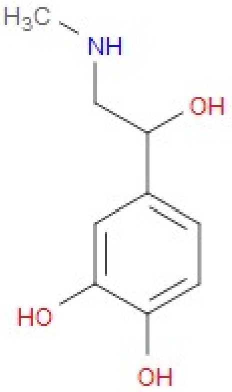


[ID 225] Dodecyl sulfate



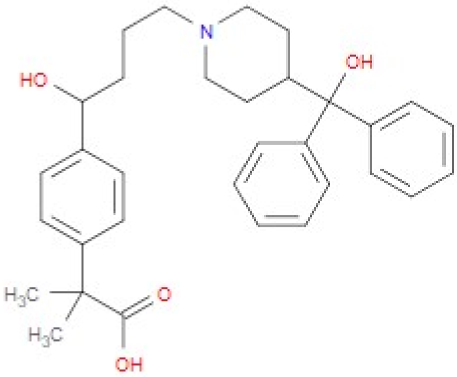
[ID 226] Equol

Ta+Nt Ta*+Nt



[ID 227] Epinephrine

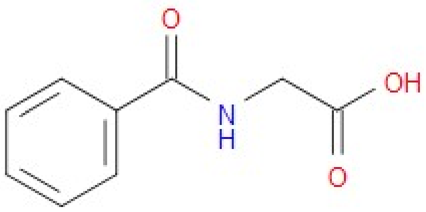
Ta+Nt



[ID 229] Fexofenadine

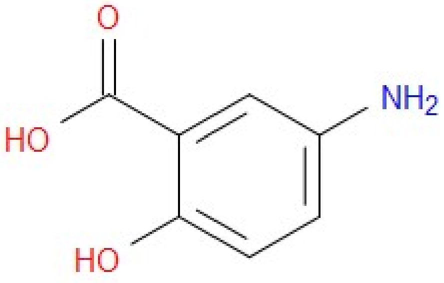
Ta+Nt Ta*+Nt

Chemical structures S2. Non-target chemicals. Datasets: Ta+Nt = target and non-target chemicals, Ta*+Nt = target -> endpoint outliers removed and non-target chemicals, Pe = PEGs, Pg = PPGs, Pe+Pg = PEGs and PPGs. Red: training set, blue: test set.

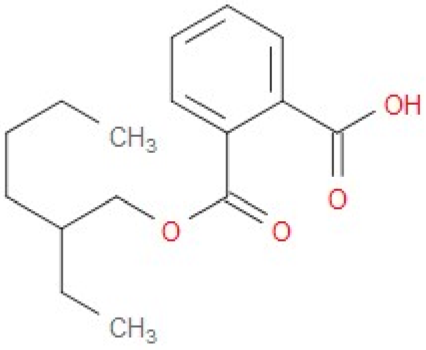


[ID 230] Hippuric acid

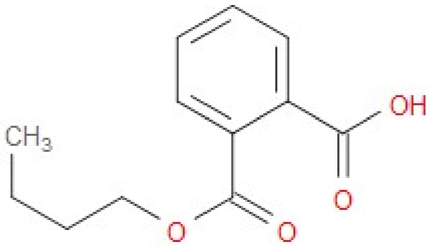
Ta+Nt



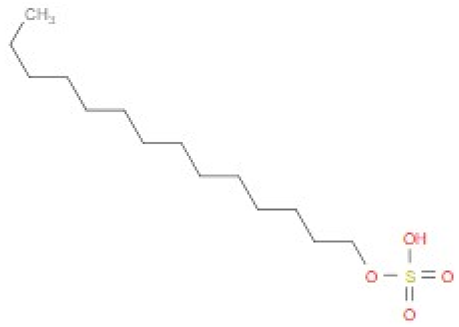
[ID 232] Mesalamine



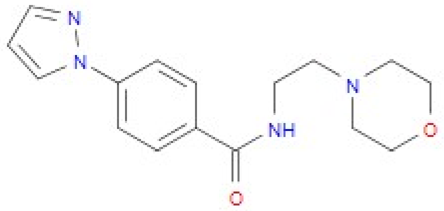
[ID 233] Mono(2-ethylhexyl) phthalate (MEHP)



[ID 234] Monobutyl phthalate



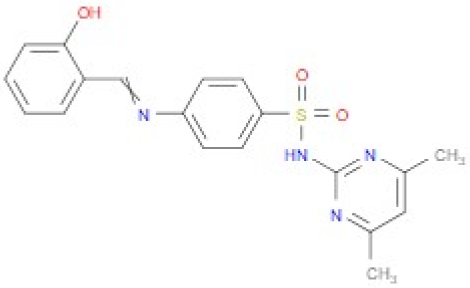
[ID 235] Myristyl sulfate



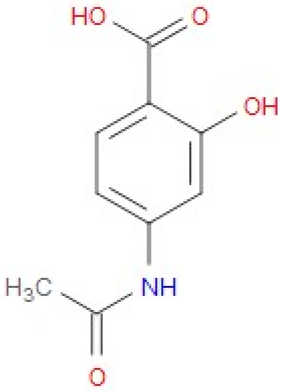
[ID 236] N-(2-Morpholinoethyl)-4-(1H-pyrazol-1-yl)benzamide

Ta+Nt Ta*+Nt

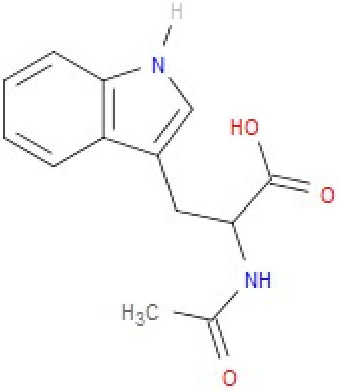
Chemical structures S2. Non-target chemicals. Datasets: Ta+Nt = target and non-target chemicals, Ta*+Nt = target -> endpoint outliers removed and non-target chemicals, Pe = PEGs, Pg = PPGs, Pe+Pg = PEGs and PPGs. Red: training set, blue: test set.



[ID 237] N-(4,6-Dimethyl-2-pyrimidinyl)-4-[(E)-(2-hydroxybenzylidene)amino] benzenesulfonamide

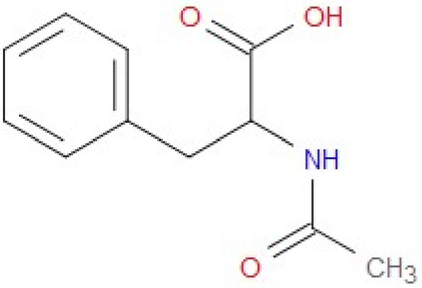


[ID 238] N-Acetyl-4-aminosalicylic acid



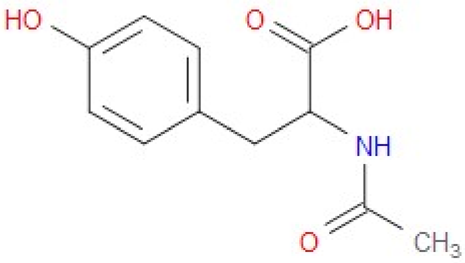
[ID 239] N-Acetyl-DL-tryptophan

Ta+Nt Ta*+Nt



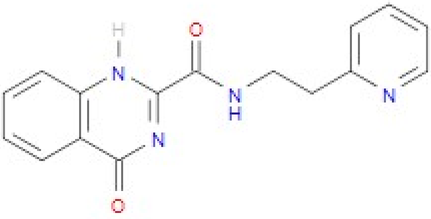
[ID 240] N-Acetyl-L-phenylalanine

Ta+Nt Ta*+Nt



[ID 241] N-Acetyl-L-tyrosine

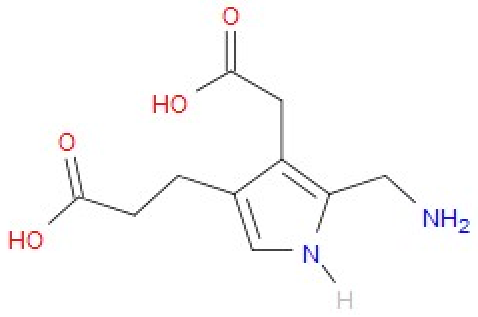
Ta+Nt Ta*+Nt



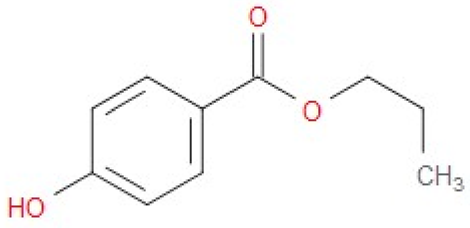
[ID 242] N2-[2-(2-Pyridyl)ethyl]-4-hydroxyquinazoline-2-carboxamide

Ta+Nt Ta*+Nt

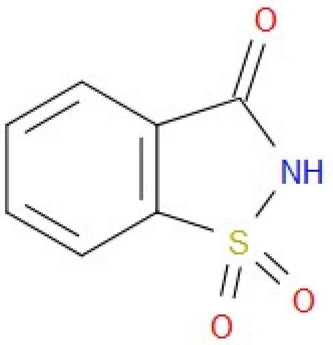
Chemical structures S2. Non-target chemicals. Datasets: Ta+Nt = target and non-target chemicals, Ta*+Nt = target -> endpoint outliers removed and non-target chemicals, Pe = PEGs, Pg = PPGs, Pe+Pg = PEGs and PPGs. Red: training set, blue: test set.



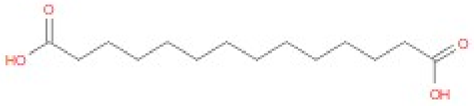
[ID 245] Porphobilinogen



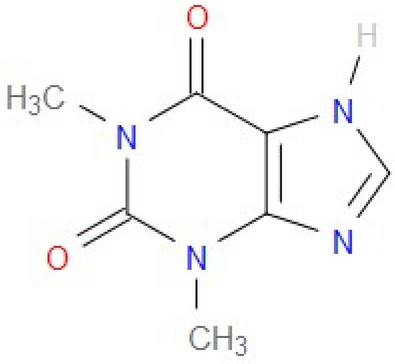
[ID 246] Propylparaben



[ID 247] Saccharin

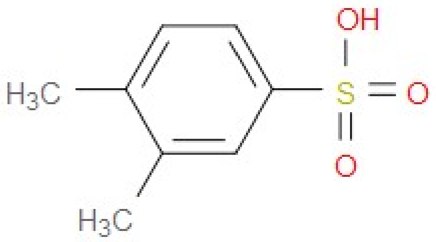


[ID 248] Tetradecanedioic acid



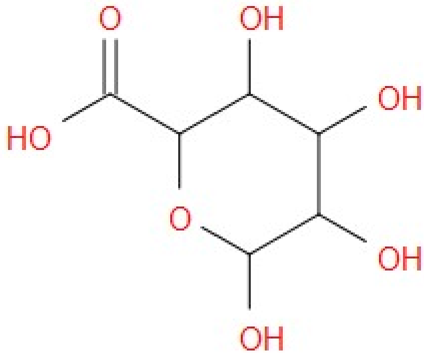
[ID 249] Theophylline

Ta+Nt Ta*+Nt

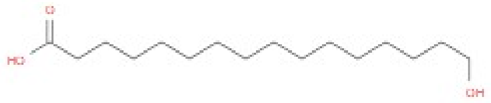


[ID 250] Xylenesulfonate

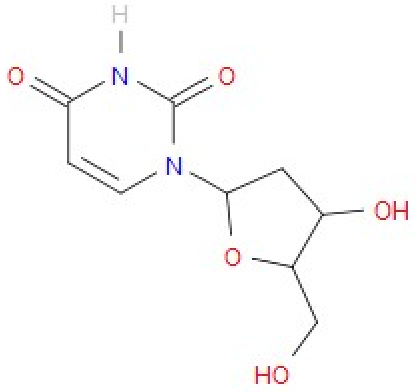
Chemical structures S2. Non-target chemicals. Datasets: Ta+Nt = target and non-target chemicals, Ta*+Nt = target -> endpoint outliers removed and non-target chemicals, Pe = PEGs, Pg = PPGs, Pe+Pg = PEGs and PPGs. Red: training set, blue: test set.



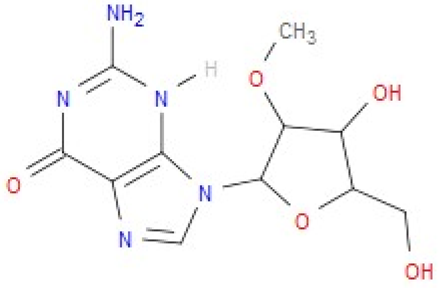
[ID 251] β-D-Glucopyranuronic acid



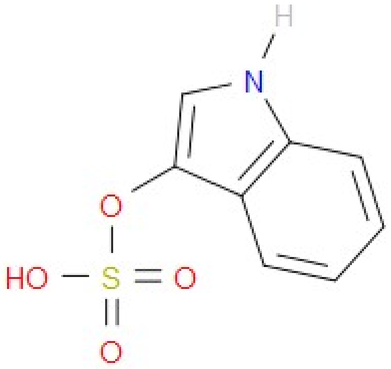
[ID 252] 16-Hydroxyhexadecanoic acid



[ID 253] 2'-Deoxyuridine

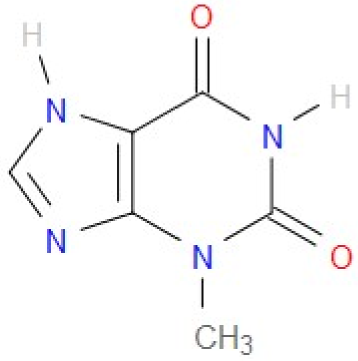


[ID 254] 2'-O-Methylguanosine



[ID 255] 3-Indoxyl sulphate

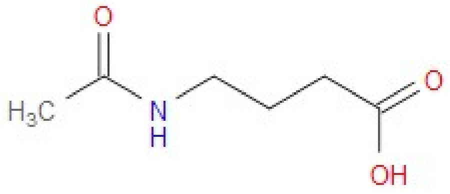
Ta+Nt Ta*+Nt



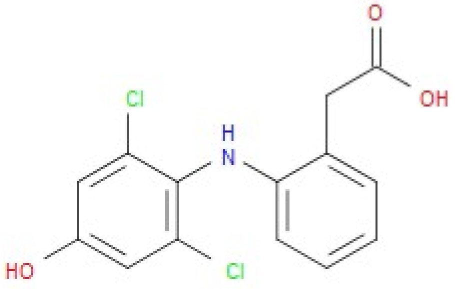
[ID 256] 3-Methylxanthine

Ta+Nt Ta*+Nt

Chemical structures S2. Non-target chemicals. Datasets: Ta+Nt = target and non-target chemicals, Ta*+Nt = target -> endpoint outliers removed and non-target chemicals, Pe = PEGs, Pg = PPGs, Pe+Pg = PEGs and PPGs. Red: training set, blue: test set.

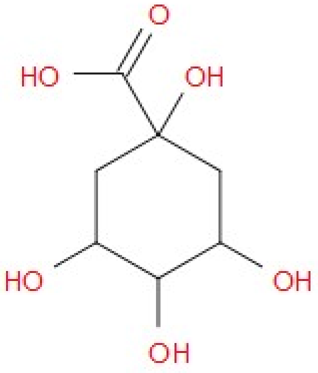


[ID 257] 4-Acetamidobutanoic acid

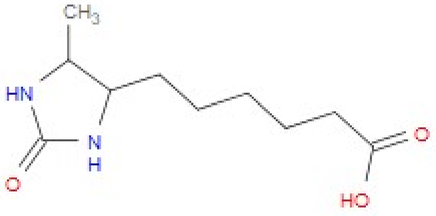


[ID 258] 4'-Hydroxydiclofenac

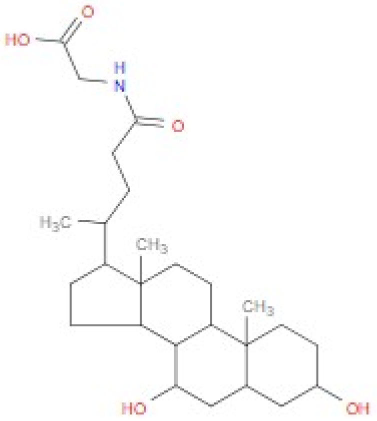
Ta+Nt Ta*+Nt



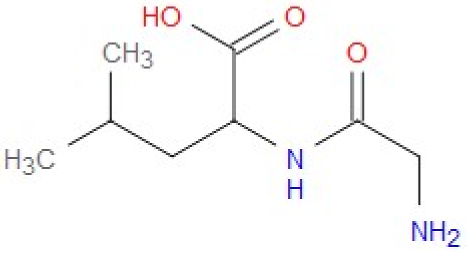
[ID 260] D-(-)-Quinic acid



[ID 261] Desthiobiotin

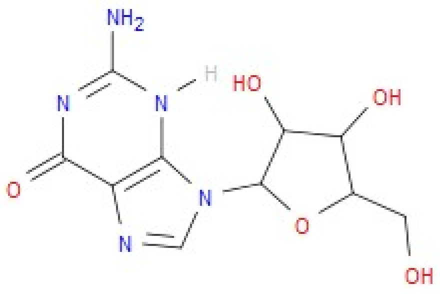


[ID 262] Glycoursodeoxycholic acid

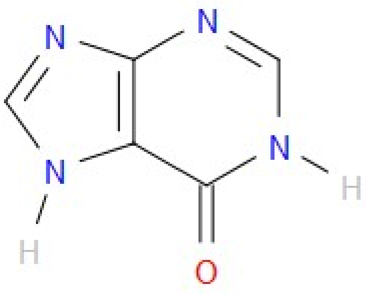


[ID 263] Glycyl-L-leucine

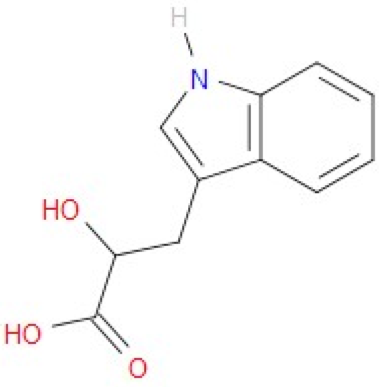
Chemical structures S2. Non-target chemicals. Datasets: Ta+Nt = target and non-target chemicals, Ta*+Nt = target -> endpoint outliers removed and non-target chemicals, Pe = PEGs, Pg = PPGs, Pe+Pg = PEGs and PPGs. Red: training set, blue: test set.



[ID 264] Guanosine

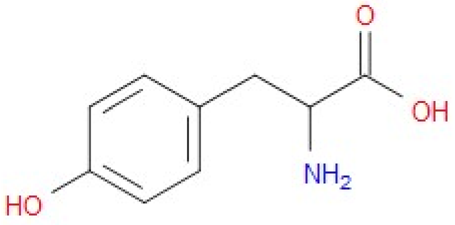


[ID 265] Hypoxanthine



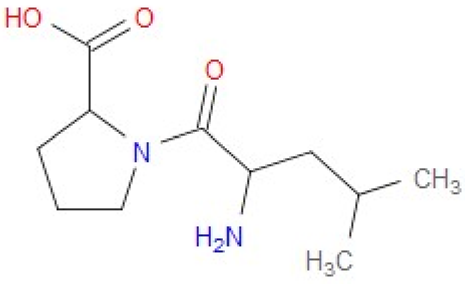
[ID 266] Indole-3-lactic acid

Ta+Nt Ta*+Nt

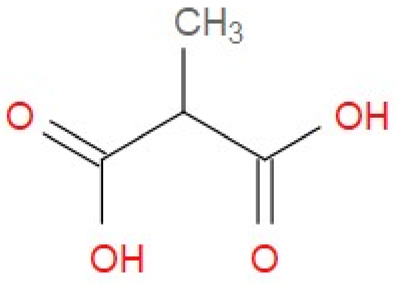


[ID 267] L-Tyrosine

Ta+Nt

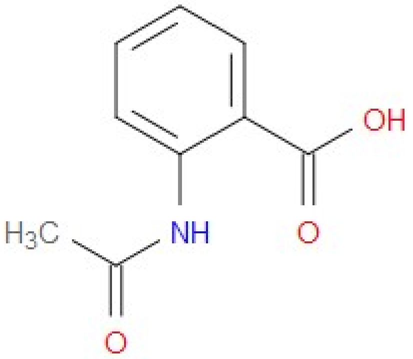


[ID 268] Leucylproline



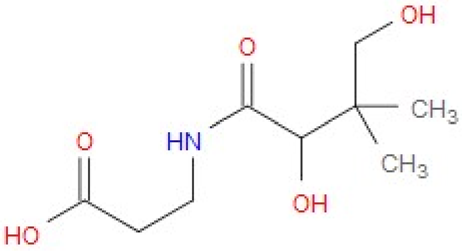
[ID 270] Methylmalonic acid

Chemical structures S2. Non-target chemicals. Datasets: Ta+Nt = target and non-target chemicals, Ta*+Nt = target -> endpoint outliers removed and non-target chemicals, Pe = PEGs, Pg = PPGs, Pe+Pg = PEGs and PPGs. Red: training set, blue: test set.

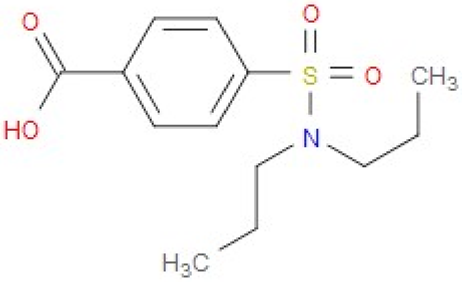


[ID 271] N-Acetylanthranilic acid

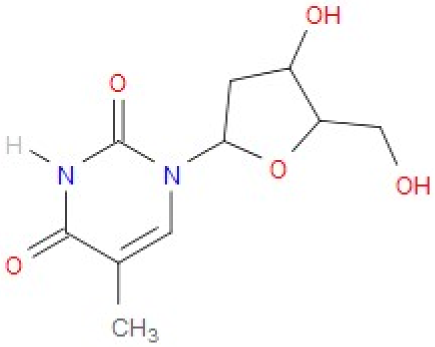
Ta+Nt



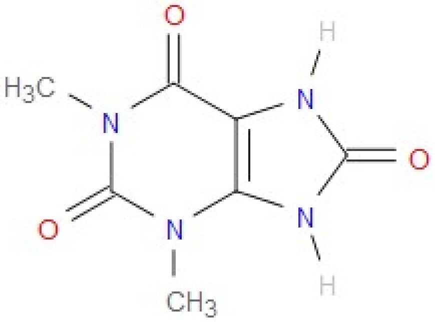
[ID 272] Pantothenic acid



[ID 273] Probenecid

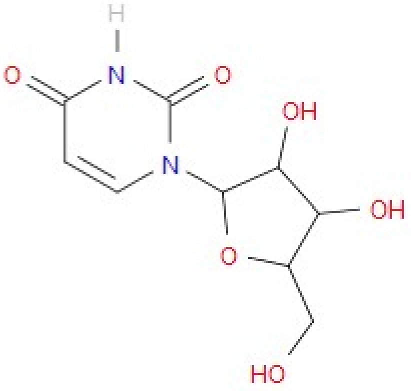


[ID 274] Thymidine



[ID 275] Uric acid

Ta+Nt Ta*+Nt



[ID 276] Uridine

Chemical structures S2. Non-target chemicals. Datasets: Ta+Nt = target and non-target chemicals, Ta*+Nt = target -> endpoint outliers removed and non-target chemicals, Pe = PEGs, Pg = PPGs, Pe+Pg = PEGs and PPGs. Red: training set, blue: test set.