Quinoline-based hydrazones for biocide detection: machine learning-aided design of new TBT chemosensors

*Rui P. C. L. Sousa, Filipe Teixeira, Susana P. G. Costa, Rita B. Figueira, M. Manuela M. Raposo**

*mfox@quimica.uminho.pt

Centre of Chemistry, University of Minho, Campus of Gualtar, 4710-057 Braga, Portugal

SUPPORTING INFORMATION

1. Synthesis of HZ **4a-e** and TSC **5a-b**

1.1. 2-((2-(Pyridin-2-yl)hydrazineylidene)methyl)quinoline 4a

Yellow solid (67%). Mp: 211 °C. ^{1}H NMR (DMSO-d_6, 400 MHz): δ = 6.84 (dq, J = 5.2 and 1.2 Hz, 1H, H5), 7.39 (dd, J = 8.4 and 1.2 Hz, 1H, H3), 7.56 (dt, J = 7.2 and 1.2 Hz, 1H, H6'), 7.68-7.76 (m, 2H, H4 and H7'), 7.95 (dt, J = 7.2 and 1.2 Hz, 2H, H5' and H8'), 8.13-8.16 (m, 2H, H6 and H3'), 8.20 (s, 1H, N=CH), 8.32 (d, J = 8.8 Hz, 1H, H4'), 11.34 (s, 1H, NH) ppm. ^{13}C NMR (DMSO-d_6, 100.6 MHz): δ = 106.9 (C3), 116.0 (C5), 117.3

(C3'), 126.7 (C6'), 127.5 (C4a'), 128.0 (C5'), 128.6 (C8'), 130.0 (C7'), 136.4 (C4'), 138.3 (C4), 139.3 (N=CH), 147.5 (C8a'), 148.0 (C6), 154.8 (C2'), 156.6 (C2) ppm. IR (solid): *v* 3193, 3062, 2923, 2852, 2160, 1968, 1648, 1610, 1596, 1575, 1545, 1504, 1467, 1457, 1444, 1425, 1357, 1336, 1300, 1260, 1241, 1176, 1150, 1134, 1115, 1095, 1050, 993, 954, 915, 898, 865, 832, 769, 745, 697, 648, 628, 609, 543, 512, 481, 475 cm^{-1}. HRMS: (ESI, positive mode) *m/z* for $C_{15}H_{12}N_4$, calcd 249.1135; found 249.1130.

1.2. 2-((2-(Pyridin-2-yl)hydrazineylidene)methyl)quinolin-8-ol 4b

Colorless solid (76%). Mp: 239 °C. ^{1}H NMR (DMSO-d_6, 400 MHz): δ = 6.84 (dq, *J* = 5.2 and 1.2 Hz, 1H, H5), 7.07 (dd, *J* = 7.2 and 2.0 Hz, 1H, H7'), 7.34-7.41 (m, 3H, H3 and H5' and H6'), 7.70 (dt, *J* = 7.2 and 2.0 Hz, 1H, H4), 8.12 (d, *J* = 8.8 Hz, 1H, H3'), 8.16 (dd, *J* = 7.2 and 1.2 Hz, 1H, H6), 8.24 (s, 1H, N=CH), 8.25 (d, *J* = 8.4 Hz, 1H, H4'), 9.67 (s, 1H, OH), 11.40 (s, 1H, NH) ppm. ^{13}C NMR (DMSO-d_6, 100.6 MHz): δ = 106.8 (C3), 112.0 (C7'), 115.9 (C5), 117.5 (C3'), 117.8 (C5'), 127.5 (C6'), 128.3 (C4a'), 136.2 (C4'), 138.1 (C8a'), 138.2 (C4), 139.2 (N=CH), 147.9 (C6), 152.7 (C2'), 153.1 (C8'), 156.7 (C2) ppm. IR (solid): *v* 3196, 3148, 3052, 2980, 2136, 2076, 1908, 1755, 1615, 1598, 1582, 1564, 1544, 1512, 1480, 1460, 1443, 1430, 1378, 1363, 1328, 1298, 1279, 1247, 1197, 1162, 1148, 1129, 1098, 1084, 1059, 1048, 989, 971, 954, 929, 900, 890, 874, 858, 832, 807, 767, 756, 718, 649, 625, 577, 555, 543, 536, 511, 495, 464 cm^{-1}. HRMS: (ESI, positive mode) *m/z* for $C_{15}H_{12}N_4O$, calcd 265.1084; found 265.1082.

1.3. 2-((2,2-Diphenylhydrazineylidene)methyl)quinoline 4c

Additionally to the general procedure, 0.318 mmol of triethylamine were added since the hydrazine precursor was in its hydrochloride form. Colorless solid (55%). Mp: 184 °C. ^{1}H NMR (DMSO-d_6, 400 MHz): δ = 7.22 (s, 1H, N=CH), 7.27-7.34 (m, 6H, H2 and H4 and

H6), 7.52-7.58 (m, 5H, H3 and H5 and H6'), 7.71 (dt, *J* = 7.6 and 1.2 Hz, 1H, H7'), 7.86 (dd, *J* = 8.0 and 0.8 Hz, 1H, H8'), 7.96 (dd, *J* = 8.0 and 1.2 Hz, 1H, H5'), 8.25 (d, *J* = 8.8 Hz, 1H, H3'), 8.37 (d, *J* = 8.8 Hz, 1H, H4') ppm. ^{13}C NMR (DMSO-d_6, 100.6 MHz): δ = 117.4 (C3'), 122.2 (C2 and C6), 125.4 (C4), 126.5 (C6'), 127.3 (C4a'), 127.9 (C5'), 128.4 (C8'), 129.9 (C7'), 130.2 (C3 and C5), 135.5 (N=CH), 136.4 (C4'), 142.4 (C1), 147.4 (C8a'), 154.7 (C2') ppm. IR (solid): *v* 3060, 2997, 2176, 2116, 1996, 1922, 1614, 1600, 1587, 1566, 1490, 1456, 1426, 1382, 1369, 1337, 1326, 1312, 1304, 1287, 1239, 1208, 1174, 1146, 1136, 1112, 1105, 1074, 1025, 1006, 994, 980, 957, 947, 903, 893, 882, 853, 828, 820, 807, 786, 777, 769, 746, 704, 687, 637, 618, 544, 524, 507, 477, 458 cm^{-1}. HRMS: (ESI, positive mode) *m/z* for $C_{22}H_{17}N_3$, calcd 324.1495; found 324.1485.

1.4. 2-((2,2-Diphenylhydrazineylidene)methyl)quinolin-8-ol 4d

Brown solid (53%). Mp: 165 °C. ^{1}H NMR (DMSO-d_6, 400 MHz): δ = 7.01 (dd, *J* = 6.4 and 2.4 Hz, 1H, H7'), 7.25-7.31 (m, 6H, H2 and H4 and H6), 7.33-7.38 (m, 2H, H5' and H6'), 7.36 (s, 1H, N=CH), 7.50-7.54 (m, 4H, H3 and H5), 8.23 (d, *J* = 8.4 Hz, 1H, H3'), 8.27 (d, *J* = 8.8 Hz, 1H, H4'), 9.56 (s, 1H, OH) ppm. ^{13}C NMR (DMSO-d_6, 100.6 MHz): δ = 112.0 (C7'), 117.7 (C5'), 117.7 (C3'), 122.2 (C2 and C6), 125.4 (C4), 127.4 (C6'), 128.2 (C4a'), 130.3 (C3 and C5), 135.7 (N=CH), 136.3 (C4'), 138.0 (C8a'), 142.6 (C1), 152.8 (C2'), 153.1 (C8') ppm. IR (solid): *v* 3737, 3429, 2989, 2194, 2108, 1918, 1632, 1599, 1588, 1564, 1505, 1495, 1469, 1455, 1433, 1378, 1335, 1320, 1306, 1284, 1253, 1230, 1208, 1195, 1174, 1134, 1115, 1085, 1074, 1024, 1002, 953, 909, 897, 880, 873, 834, 820, 804, 774, 756, 745, 730, 717, 707, 685, 636, 619, 579, 555, 546, 524, 504, 486, 457 cm^{-1}. HRMS: (ESI, positive mode) *m/z* for $C_{22}H_{17}N_3O$, calcd 340.1444; found 340.1444.

1.5. Methyl 3-(2-(quinolin-2-ylmethylene)hydrazineyl)thiophene-2-carboxylate 4e

Yellow solid (52%). Mp: 163 °C. ^{1}H NMR (DMSO-d_6, 400 MHz): δ = 3.82 (s, 3H, OCH_3), 7.50 (d, J = 5.6 Hz, 1H, H5), 7.58 (dt, J = 6.8 and 1.2 Hz, 1H, H6'), 7.75 (dt, J = 6.8 and 1.6 Hz, 1H, H7'), 7.83 (dd, J = 5.6 and 0.4 Hz, 1H, H4), 7.97 (t, J = 8.4 Hz, 2H, H8' and H5'), 8.14 (d, J = 8.8 Hz, 1H, H3'), 8.33 (d, J = 8.8 Hz, 1H, H4'), 8.42 (s, 1H, N=CH), 10.81 (s, 1H, NH) ppm. ^{13}C NMR (DMSO-d_6, 100.6 MHz): δ = 51.7 (OCH_3), 101.0 (C2), 117.5 (C3'), 119.1 (C5), 126.8 (C6'), 127.5 (C4a'), 128.0 (C5'), 128.7 (C8'), 129.9 (C7'), 133.5 (C4), 136.3 (C4'), 141.1 (N=CH), 147.4 (C8a'), 150.7 (C3), 154.4 (C2'), 163.2 (C=O) ppm. IR (solid): v 3738, 3347, 3280, 3114, 3081, 3040, 3012, 2959, 2178, 2107, 1958, 1849, 1781, 1673, 1603, 1592, 1575, 1558, 1501, 1472, 1458, 1447, 1421, 1402, 1379, 1350, 1323, 1308, 1283, 1250, 1237, 1206, 1158, 1115, 1096, 1083, 1011, 973, 954, 946, 927, 903, 892, 864, 838, 825, 803, 784, 771, 749, 730, 673, 646, 621, 600, 578, 524, 473 cm^{-1}. HRMS: (ESI, positive mode) m/z for $C_{16}H_{13}N_3O_2S$, calcd 312.0801; found 312.0804.

1.6. *N*-phenyl-2-(quinolin-2-ylmethylene)hydrazine-1-carbothioamide 5a

Following the general procedure, the temperature was raised to 60 °C, and stirred for another 6 h. Brown solid (17%). Mp: 160 °C. ^{1}H NMR (DMSO-d_6, 400 MHz): δ = 7.24 (t, J = 7.6 Hz, 1H, H4), 7.40 (dt, J = 7.6 and 1.6 Hz, 2H, H3 and H5), 7.55 (dd, J = 8.0 and 0.4 Hz, 2H, H2 and H6), 7.62 (dt, J = 7.2 and 1.2 Hz, 1H, H6'), 7.78 (dt, J = 7.2 and 1.2 Hz, 1H, H7'), 7.99 (dd, J = 8.0 and 0.4 Hz, 1H, H5'), 8.02 (d, J = 8.4 Hz, 1H, H8'), 8.33 (s, 1H, N=CH), 8.38 (d, J = 8.8 Hz, 1H, H4′), 8.59 (d, J = 8.4 Hz, 1H, H3′), 10.38 (s, 1H, NH-C=S), 12.17 (s, 1H, NH-N) ppm. ^{13}C NMR (DMSO-d_6, 100.6 MHz): δ = 118.5 (C3'), 125.7 (C4), 126.3 (C2 and C6), 127.3 (C6'), 127.9 (C4a'), 128.0 (C5'), 128.2 (C3 and C5), 128.8 (C8'), 130.0 (C7'), 136.3 (C4'), 139.0 (C1), 143.1 (N=CH), 147.4

(C8a'), 153.8 (C2'), 176.6 (C=S) ppm. IR (solid): *v* 3316, 3105, 2942, 2176, 1982, 1595, 1537, 1505, 1444, 1427, 1393, 1342, 1313, 1301, 1267, 1251, 1185, 1148, 1140, 1119, 1101, 1071, 1029, 959, 938, 921, 894, 822, 788, 757, 737, 688, 624, 614, 590, 536, 502, 478, 463 cm^{-1}. HRMS: (ESI, positive mode) *m/z* for $C_{17}H_{14}N_4S$, calcd 307.1012; found 307.1002.

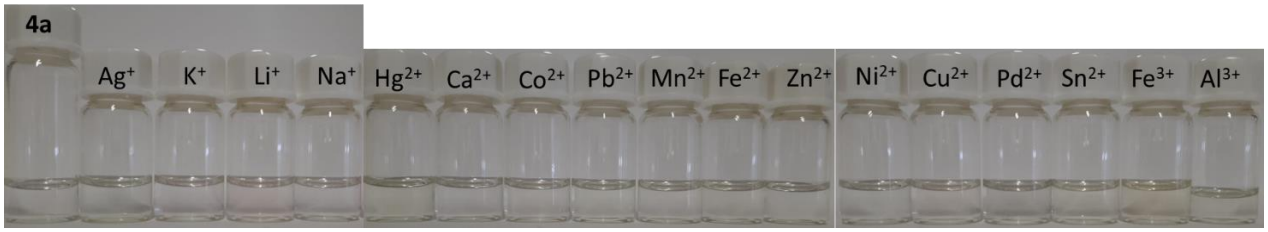
1.7. 2-((8-Hydroxyquinolin-2-yl)methylene)-*N*-phenylhydrazine-1-carbothioamide 5b

Colorless solid (83%). Mp: 214 °C. ^{1}H NMR (DMSO-d_6, 400 MHz): δ = 7.09 (dd, *J* = 1.6 and 7.6 Hz, 1H, H7'), 7.23 (dt, *J* = 0.8 and 7.6 Hz, 1H, H4), 7.41-7.36 (m, 3H, H3 and H5 and H5′), 7.43 (t, *J* = 8.0 Hz, 1H, H6′), 7.53 (dd, *J* = 1.2 and 7.6 Hz, 2H, H2 and H6), 8.28 (d, *J* = 8.8 Hz, 1H, H4′), 8.35 (s, 1H, N=CH), 8.54 (d, *J* = 8.4 Hz, 1H, H3′), 9.91 (s, 1H, OH), 10.35 (s, 1H, NH-C=S), 12.19 (s, 1H, NH-N) ppm. ^{13}C NMR (DMSO-d_6, 100.6 MHz): δ = 112.3 (C7′), 118.0 (C5′), 118.9 (C3′), 125.9 (C4), 126.5 (C2 and C6), 128.4 (C3 and C5), 128.4 (C6′), 129.1 (C4a′), 136.4 (C4′), 138.4 (C8a′), 139.1 (C1), 143.2 (N=CH), 151.8 (C2′), 153.5 (C8'), 176.8 (C=S) ppm. IR (solid): *v* 3611, 3368, 3142, 2988, 2176, 2106, 1872, 1590, 1569, 1534, 1505, 1466, 1446, 1435, 1395, 1328, 1309, 1235, 1199, 1136, 1098, 1024, 945, 929, 915, 901, 872, 843, 771, 740, 719, 690, 641, 609, 563, 548, 516, 492 cm^{-1}. HRMS: (ESI, positive mode) *m/z* for $C_{17}H_{14}N_4OS$, calcd 323.0961; found 323.0964.

Preliminary sensory tests of compound **4a** [10^{-5}M]:

Colorimetric changes in the presence of TBT (50 equiv.):

Colorimetric changes in the presence of other metal ions (50 equiv.):



230703-chemosensor_model

July 13, 2023

1 Model fot TBT chemossensing

[1]:
```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np
from sklearn import linear_model
from sklearn import inspection
import shap
```

```
/home/filipe/.local/lib/python3.10/site-packages/shap/utils/_clustering.py:35:
NumbaDeprecationWarning: The 'nopython' keyword argument was not supplied to
the 'numba.jit' decorator. The implicit default value for this argument is
currently False, but it will be changed to True in Numba 0.59.0. See
https://numba.readthedocs.io/en/stable/reference/deprecation.html#deprecation-
of-object-mode-fall-back-behaviour-when-using-jit for details.
  def _pt_shuffle_rec(i, indexes, index_mask, partition_tree, M, pos):
/home/filipe/.local/lib/python3.10/site-packages/shap/utils/_clustering.py:54:
NumbaDeprecationWarning: The 'nopython' keyword argument was not supplied to
the 'numba.jit' decorator. The implicit default value for this argument is
currently False, but it will be changed to True in Numba 0.59.0. See
https://numba.readthedocs.io/en/stable/reference/deprecation.html#deprecation-
of-object-mode-fall-back-behaviour-when-using-jit for details.
  def delta_minimization_order(all_masks, max_swap_size=100, num_passes=2):
/home/filipe/.local/lib/python3.10/site-packages/shap/utils/_clustering.py:63:
NumbaDeprecationWarning: The 'nopython' keyword argument was not supplied to
the 'numba.jit' decorator. The implicit default value for this argument is
currently False, but it will be changed to True in Numba 0.59.0. See
https://numba.readthedocs.io/en/stable/reference/deprecation.html#deprecation-
of-object-mode-fall-back-behaviour-when-using-jit for details.
  def _reverse_window(order, start, length):
```

```
/home/filipe/.local/lib/python3.10/site-packages/shap/utils/_clustering.py:69:
NumbaDeprecationWarning: The 'nopython' keyword argument was not supplied to
the 'numba.jit' decorator. The implicit default value for this argument is
currently False, but it will be changed to True in Numba 0.59.0. See
https://numba.readthedocs.io/en/stable/reference/deprecation.html#deprecation-
of-object-mode-fall-back-behaviour-when-using-jit for details.
  def _reverse_window_score_gain(masks, order, start, length):
/home/filipe/.local/lib/python3.10/site-packages/shap/utils/_clustering.py:77:
NumbaDeprecationWarning: The 'nopython' keyword argument was not supplied to
the 'numba.jit' decorator. The implicit default value for this argument is
currently False, but it will be changed to True in Numba 0.59.0. See
https://numba.readthedocs.io/en/stable/reference/deprecation.html#deprecation-
of-object-mode-fall-back-behaviour-when-using-jit for details.
  def _mask_delta_score(m1, m2):
/home/filipe/.local/lib/python3.10/site-packages/shap/links.py:5:
NumbaDeprecationWarning: The 'nopython' keyword argument was not supplied to
the 'numba.jit' decorator. The implicit default value for this argument is
currently False, but it will be changed to True in Numba 0.59.0. See
https://numba.readthedocs.io/en/stable/reference/deprecation.html#deprecation-
of-object-mode-fall-back-behaviour-when-using-jit for details.
  def identity(x):
/home/filipe/.local/lib/python3.10/site-packages/shap/links.py:10:
NumbaDeprecationWarning: The 'nopython' keyword argument was not supplied to
the 'numba.jit' decorator. The implicit default value for this argument is
currently False, but it will be changed to True in Numba 0.59.0. See
https://numba.readthedocs.io/en/stable/reference/deprecation.html#deprecation-
of-object-mode-fall-back-behaviour-when-using-jit for details.
  def _identity_inverse(x):
/home/filipe/.local/lib/python3.10/site-packages/shap/links.py:15:
NumbaDeprecationWarning: The 'nopython' keyword argument was not supplied to
the 'numba.jit' decorator. The implicit default value for this argument is
currently False, but it will be changed to True in Numba 0.59.0. See
https://numba.readthedocs.io/en/stable/reference/deprecation.html#deprecation-
of-object-mode-fall-back-behaviour-when-using-jit for details.
  def logit(x):
/home/filipe/.local/lib/python3.10/site-packages/shap/links.py:20:
```

```
NumbaDeprecationWarning: The 'nopython' keyword argument was not supplied to
the 'numba.jit' decorator. The implicit default value for this argument is
currently False, but it will be changed to True in Numba 0.59.0. See
https://numba.readthedocs.io/en/stable/reference/deprecation.html#deprecation-
of-object-mode-fall-back-behaviour-when-using-jit for details.
  def _logit_inverse(x):
/home/filipe/.local/lib/python3.10/site-
packages/shap/utils/_masked_model.py:363: NumbaDeprecationWarning: The
'nopython' keyword argument was not supplied to the 'numba.jit' decorator. The
implicit default value for this argument is currently False, but it will be
changed to True in Numba 0.59.0. See
https://numba.readthedocs.io/en/stable/reference/deprecation.html#deprecation-
of-object-mode-fall-back-behaviour-when-using-jit for details.
  def _build_fixed_single_output(averaged_outs, last_outs, outputs,
batch_positions, varying_rows, num_varying_rows, link, linearizing_weights):
/home/filipe/.local/lib/python3.10/site-
packages/shap/utils/_masked_model.py:385: NumbaDeprecationWarning: The
'nopython' keyword argument was not supplied to the 'numba.jit' decorator. The
implicit default value for this argument is currently False, but it will be
changed to True in Numba 0.59.0. See
https://numba.readthedocs.io/en/stable/reference/deprecation.html#deprecation-
of-object-mode-fall-back-behaviour-when-using-jit for details.
  def _build_fixed_multi_output(averaged_outs, last_outs, outputs,
batch_positions, varying_rows, num_varying_rows, link, linearizing_weights):
/home/filipe/.local/lib/python3.10/site-
packages/shap/utils/_masked_model.py:428: NumbaDeprecationWarning: The
'nopython' keyword argument was not supplied to the 'numba.jit' decorator. The
implicit default value for this argument is currently False, but it will be
changed to True in Numba 0.59.0. See
https://numba.readthedocs.io/en/stable/reference/deprecation.html#deprecation-
of-object-mode-fall-back-behaviour-when-using-jit for details.
  def _init_masks(cluster_matrix, M, indices_row_pos, indptr):
/home/filipe/.local/lib/python3.10/site-
```

```
packages/shap/utils/_masked_model.py:439: NumbaDeprecationWarning: The
'nopython' keyword argument was not supplied to the 'numba.jit' decorator. The
implicit default value for this argument is currently False, but it will be
changed to True in Numba 0.59.0. See
https://numba.readthedocs.io/en/stable/reference/deprecation.html#deprecation-
of-object-mode-fall-back-behaviour-when-using-jit for details.
  def _rec_fill_masks(cluster_matrix, indices_row_pos, indptr, indices, M, ind):
/home/filipe/.local/lib/python3.10/site-packages/shap/maskers/_tabular.py:186:
NumbaDeprecationWarning: The 'nopython' keyword argument was not supplied to
the 'numba.jit' decorator. The implicit default value for this argument is
currently False, but it will be changed to True in Numba 0.59.0. See
https://numba.readthedocs.io/en/stable/reference/deprecation.html#deprecation-
of-object-mode-fall-back-behaviour-when-using-jit for details.
  def _single_delta_mask(dind, masked_inputs, last_mask, data, x, noop_code):
/home/filipe/.local/lib/python3.10/site-packages/shap/maskers/_tabular.py:197:
NumbaDeprecationWarning: The 'nopython' keyword argument was not supplied to
the 'numba.jit' decorator. The implicit default value for this argument is
currently False, but it will be changed to True in Numba 0.59.0. See
https://numba.readthedocs.io/en/stable/reference/deprecation.html#deprecation-
of-object-mode-fall-back-behaviour-when-using-jit for details.
  def _delta_masking(masks, x, curr_delta_inds, varying_rows_out,
/home/filipe/.local/lib/python3.10/site-packages/shap/maskers/_image.py:175:
NumbaDeprecationWarning: The 'nopython' keyword argument was not supplied to
the 'numba.jit' decorator. The implicit default value for this argument is
currently False, but it will be changed to True in Numba 0.59.0. See
https://numba.readthedocs.io/en/stable/reference/deprecation.html#deprecation-
of-object-mode-fall-back-behaviour-when-using-jit for details.
  def _jit_build_partition_tree(xmin, xmax, ymin, ymax, zmin, zmax,
total_ywidth, total_zwidth, M, clustering, q):
/home/filipe/.local/lib/python3.10/site-
packages/shap/explainers/_partition.py:676: NumbaDeprecationWarning: The
'nopython' keyword argument was not supplied to the 'numba.jit' decorator. The
implicit default value for this argument is currently False, but it will be
changed to True in Numba 0.59.0. See
https://numba.readthedocs.io/en/stable/reference/deprecation.html#deprecation-
of-object-mode-fall-back-behaviour-when-using-jit for details.
```

```
  def lower_credit(i, value, M, values, clustering):
```

```
The 'nopython' keyword argument was not supplied to the 'numba.jit'
decorator. The implicit default value for this argument is currently False, but
it will be changed to True in Numba 0.59.0. See
https://numba.readthedocs.io/en/stable/reference/deprecation.html#deprecation-
of-object-mode-fall-back-behaviour-when-using-jit for details.
The 'nopython' keyword argument was not supplied to the 'numba.jit'
decorator. The implicit default value for this argument is currently False, but
it will be changed to True in Numba 0.59.0. See
https://numba.readthedocs.io/en/stable/reference/deprecation.html#deprecation-
of-object-mode-fall-back-behaviour-when-using-jit for details.
```

[2]:
```
data=pd.read_csv('230530-AIM_results.csv')
data
```

[2]:

	Compound	Kind	Signal	C8	lambda_max	log_eps	tbt_lambda_max	tbt_log_eps	\
0	1a	Q	No	H	289	3.9	NaN	NaN	
1	1b	Q	Yes	OH	364	4.1	457.0	3.9	
2	4a	HZ	Yes	H	356	4.6	512.0	4.0	
3	4b	HZ	Yes	OH	353	4.4	389.0	4.1	
4	4c	HZ	No	H	367	4.6	NaN	NaN	
5	4d	HZ	Yes	OH	367	4.6	389.0	4.5	
6	4e	HZ	No	H	368	4.8	NaN	NaN	
7	4f	HZ	No	H	379	4.5	NaN	NaN	
8	4g	HZ	No	OH	385	4.5	NaN	NaN	
9	5a	TSC	Yes	H	339	4.5	412.0	4.4	
10	5b	TSC	Yes	OH	345	4.7	437.0	4.4	

	LOD	E_HOMO_eV	…	QQ_e	p_HOMO_Q	p_LUMO_Q	dHL_Q	Gamma_AV_Q	\
0	NaN	-6.739681	…	-0.358741	97.593	63.424	34.169	1.326027	
1	18.9	-6.173180	…	0.214467	82.378	62.897	19.481	1.306897	
2	4.8	-5.709045	…	-0.403426	22.231	64.393	42.162	1.323987	
3	4.5	-5.652467	…	0.163549	33.711	61.998	28.287	1.303317	
4	NaN	-5.498747	…	-0.441957	18.269	66.699	48.430	1.324291	
5	9.2	-5.472604	…	0.142672	24.297	64.241	39.944	1.304349	
6	NaN	-5.626986	…	-0.431326	18.268	57.840	39.572	1.323891	
7	NaN	-6.232989	…	-0.367311	34.090	5.023	29.067	1.325470	
8	NaN	-5.994008	…	0.197020	66.279	5.525	60.754	1.304354	
9	13.8	-6.101015	…	-0.395870	9.681	57.732	48.051	1.325133	
10	3.1	-5.945862	…	0.194632	61.243	55.498	5.745	1.304147	

	Gamma_SD_Q	Max_Q_Q	Min_Q_Q	AV_Q_Q	SD_Q_Q
0	0.123726	0.522897	-1.126640	-0.035874	0.438848

```
1     0.115439  0.511066 -1.149812  0.021447  0.479385
2     0.126133  0.561283 -1.155190 -0.040343  0.453393
3     0.122781  0.555584 -1.186843  0.016355  0.493021
4     0.126642  0.542098 -1.151266 -0.044196  0.448522
5     0.123422  0.547212 -1.168649  0.014267  0.488326
6     0.126946  0.527573 -1.161340 -0.043133  0.451865
7     0.125149  0.547075 -1.133533 -0.036731  0.445306
8     0.119555  0.538893 -1.170229  0.019702  0.487189
9     0.125147  0.527711 -1.141242 -0.039587  0.444725
10    0.120829  0.553489 -1.175646  0.019463  0.492381

[11 rows x 67 columns]
```

[5]:
```
features = list(data.columns)[9:]
# adapted in order to condense all compounds in one table
qdata = data.loc[data['Kind']!='TSC']
tdata = data.loc[data['Kind']=='TSC']
```

1.1 Preliminary model

[7]:
```
clf=linear_model.RidgeClassifier()
clf.fit(qdata[features],qdata['Signal'])
print(f"{'Compound':^10s} {'Observed':^10s} {'Predicted':^10s}")
for n,v,p in zip(qdata['Compound'],qdata['Signal'],clf.
 ↪predict(qdata[features])):
    print(f"{n:^10s} {v:^10s} {p:^10s}")
```

```
 Compound   Observed  Predicted
    1a         No         No
    1b        Yes        Yes
    4a        Yes        Yes
    4b        Yes        Yes
    4c         No         No
    4d        Yes        Yes
    4e         No         No
    4f         No         No
    4g         No         No
```

[9]:
```
pi=inspection.permutation_importance(clf,qdata[features],qdata['Signal'])
pi=pd.DataFrame({'feature':clf.feature_names_in_ ,
                 'Importance_Av':pi['importances_mean'],
                 'Importance_Sd':pi['importances_std']})

pi.sort_values(by='Importance_Av', ascending=False, inplace=True)
pi.head(20)
```

[9]:

	feature	Importance_Av	Importance_Sd
49	p_HOMO_Q	0.466667	0.108866
31	p_HOMO_8	0.311111	0.108866
30	p_HOMO_7	0.266667	0.150718
28	p_HOMO_5	0.200000	0.083148
34	p_LUMO_1	0.200000	0.083148
29	p_HOMO_6	0.088889	0.044444
39	p_LUMO_5	0.066667	0.054433
33	HOMO_Deloc	0.066667	0.054433
51	dHL_Q	0.066667	0.054433
23	p_HOMO_1	0.066667	0.088889
32	p_HOMO_8a	0.044444	0.054433
35	p_LUMO_2	0.044444	0.054433
37	p_LUMO_4	0.044444	0.054433
50	p_LUMO_Q	0.022222	0.044444
52	Gamma_AV_Q	0.000000	0.000000
56	AV_Q_Q	0.000000	0.000000
55	Min_Q_Q	0.000000	0.000000
36	p_LUMO_3	0.000000	0.000000
38	p_LUMO_4a	0.000000	0.000000
40	p_LUMO_6	0.000000	0.000000

1.2 Final Model

[11]:
```
rf=pi['feature'].head(12).to_list()
clf2=linear_model.RidgeClassifier()
clf2.fit(qdata[rf],qdata['Signal'])
print(f"{'Compound':^10s} {'Observed':^10s} {'Predicted':^10s}")
for n,v,p in zip(qdata['Compound'],qdata['Signal'],clf2.predict(qdata[rf])):
    print(f"{n:^10s} {v:^10s} {p:^10s}")
print('-'*10,'-'*10,'-'*10)
for n,v,p in zip(tdata['Compound'],tdata['Signal'],clf2.predict(tdata[rf])):
    print(f"{n:^10s} {v:^10s} {p:^10s}")
```

```
 Compound   Observed  Predicted
    1a         No         No
    1b        Yes        Yes
    4a        Yes        Yes
    4b        Yes        Yes
    4c         No         No
    4d        Yes        Yes
    4e         No         No
    4f         No         No
    4g         No         No
---------- ---------- ----------
    5a        Yes        Yes
    5b        Yes        Yes
```

[14]:
```
pi=inspection.permutation_importance(clf2,qdata[rf],qdata['Signal'])
pi=pd.DataFrame({'feature':clf2.feature_names_in_ ,
                 'Importance_Av':pi['importances_mean'],
                 'Importance_Sd':pi['importances_std']})

pi.sort_values(by='Importance_Av', ascending=False, inplace=True)
pi
```

[14]:

	feature	Importance_Av	Importance_Sd
0	p_HOMO_Q	0.444444	0.157135
1	p_HOMO_8	0.311111	0.083148
2	p_HOMO_7	0.266667	0.133333
4	p_LUMO_1	0.222222	0.121716
3	p_HOMO_5	0.200000	0.044444
11	p_LUMO_2	0.200000	0.083148
5	p_HOMO_6	0.111111	0.070273
6	p_LUMO_5	0.111111	0.070273
9	p_HOMO_1	0.066667	0.054433
7	HOMO_Deloc	0.044444	0.054433
8	dHL_Q	0.022222	0.044444
10	p_HOMO_8a	0.000000	0.000000

[72]:
```
pi['Importance_Av']/2.0
```

[72]:
```
0     0.222222
1     0.155556
2     0.133333
4     0.111111
3     0.100000
11    0.100000
5     0.055556
6     0.055556
9     0.033333
7     0.022222
8     0.011111
10    0.000000
Name: Importance_Av, dtype: float64
```

[32]:
```
# Edit feature neames for better presentation
prf =␣
  ↪['%HOMO(Q)','%HOMO(C8)','%HOMO(C7)','%HOMO(C5)','%LUMO(N1)','%HOMO(C6)','%LUMO(C5)','DI(HOM
for i,j in zip(rf,prf):
    print(f"{i:10s} {j:10s}")

prettydata=data[rf].rename(columns={a:b for a,b in zip (rf,prf)})
```

```
p_HOMO_Q   %HOMO(Q)
```

```
p_HOMO_8   %HOMO(C8)
p_HOMO_7   %HOMO(C7)
p_HOMO_5   %HOMO(C5)
p_LUMO_1   %LUMO(N1)
p_HOMO_6   %HOMO(C6)
p_LUMO_5   %LUMO(C5)
HOMO_Deloc DI(HOMO)
dHL_Q       %FMO(Q)
p_HOMO_1   %HOMO(N1)
p_HOMO_8a  %HOMO(C8a)
p_LUMO_2   %LUMO(C2)
```

[33]:
```
print(prettydata)
```

```
    %HOMO(Q)  %HOMO(C8)  %HOMO(C7)  %HOMO(C5)  %LUMO(N1)  %HOMO(C6)  \
0     97.593     42.433     38.551      2.431     16.729      4.597
1     82.378     16.382     13.556     23.431     16.821      6.533
2     22.231      2.323      0.522      0.741     14.338      4.223
3     33.711      5.775      1.195      4.824     15.099      5.735
4     18.269      1.800      0.456      0.460     14.314      3.294
5     24.297      3.627      0.523      2.421     15.201      4.255
6     18.268      1.852      0.437      0.545     13.283      3.387
7     34.090      4.636      0.594      2.771      1.457      7.227
8     66.279     13.585      7.122     16.619      1.642      8.062
9      9.681      1.189      0.184      0.596     13.901      2.009
10    61.243     12.383      5.950     14.682     14.361      7.977

    %LUMO(C5)  DI(HOMO)   %FMO(Q)  %HOMO(N1)  %HOMO(C8a)  %LUMO(C2)
0       3.304     33.67    34.169      0.454       6.949     12.638
1       2.116     13.54    19.481      4.788       4.419     11.961
2       4.996     11.04    42.162      5.232       2.096     11.581
3       3.032      8.42    28.287      6.589       4.171     11.996
4       5.526     10.82    48.430      4.646       1.583     11.865
5       3.400      9.24    39.944      5.554       3.014     12.321
6       4.256     10.83    39.572      4.445       1.655     10.675
7       0.182      7.54    29.067      6.270       3.596      0.796
8       0.115      8.74    60.754      7.024       5.863      0.902
9       3.808     38.39    48.051      1.900       1.020     10.721
10      2.224      8.89     5.745      7.191       5.813     10.808
```

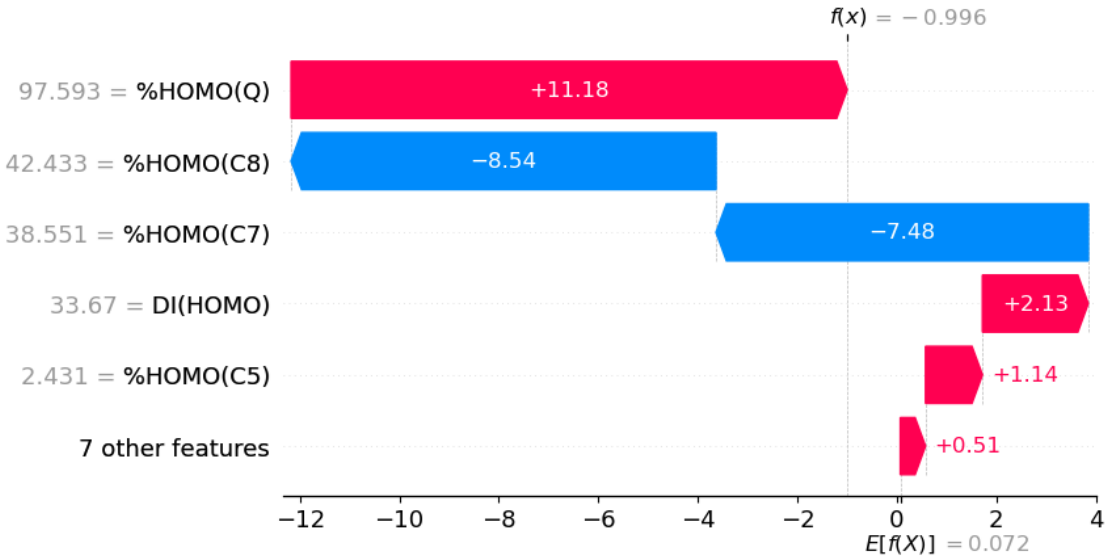
[35]:
```
explainer = shap.Explainer(clf2,prettydata,algorithm='linear',feature_names=prf)
shap_vals = explainer(prettydata)
```

[68]:
```
plt.figure(layout='constrained')
plt.subplot(2,1,1)
shap.plots.waterfall(shap_vals[-2], max_display= 8, show=False)
plt.subplot(2,1,2)
```

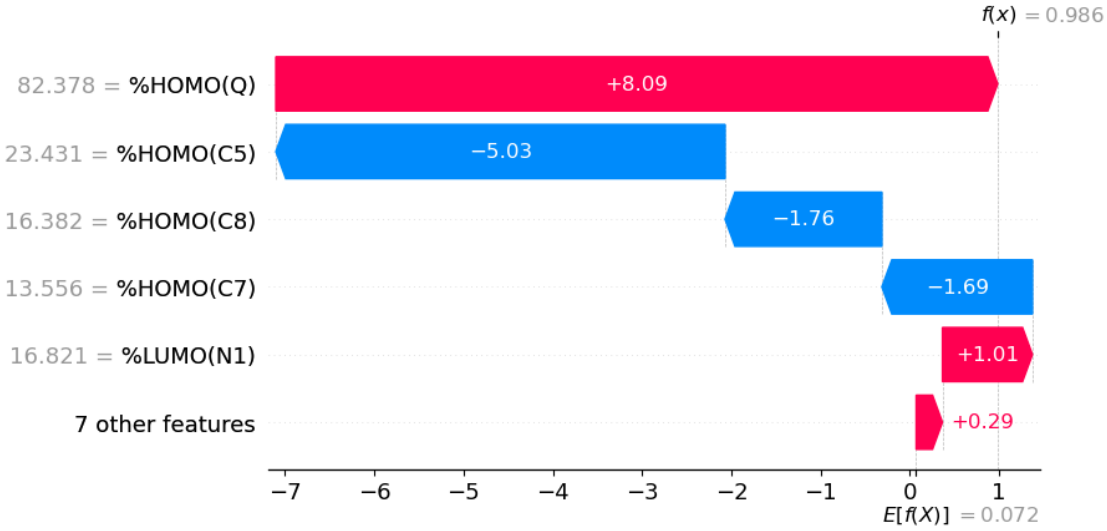
```
shap.plots.waterfall(shap_vals[-1], max_display= 9, show=False)
plt.savefig('fig-SHAP.png')
```

[74]:
```
for i,cn in enumerate(data['Compound']):
    print(cn)
    shap.plots.waterfall(shap_vals[i], max_display=6)
```

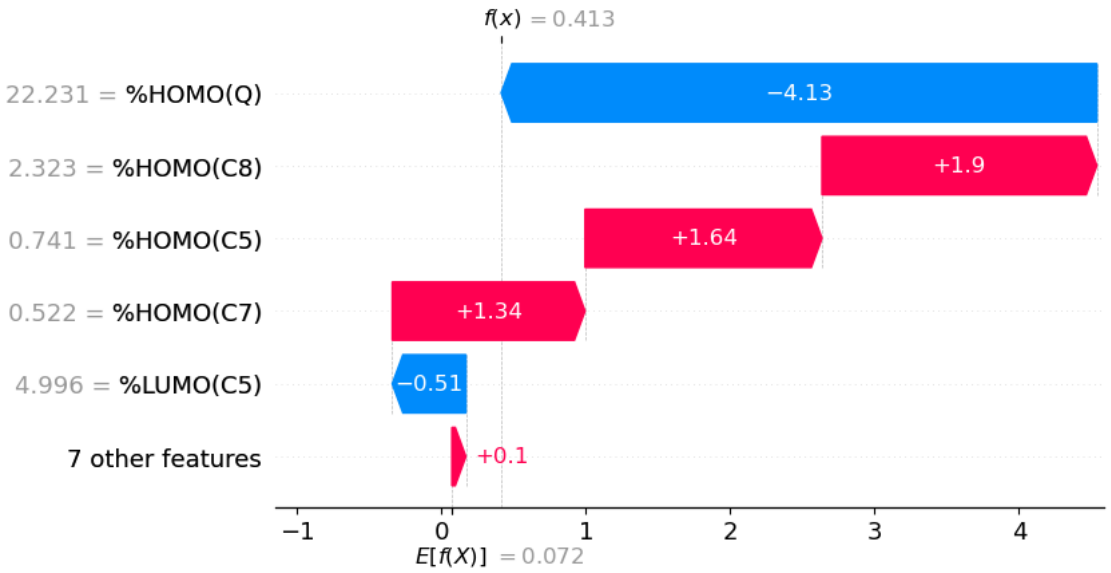
1a



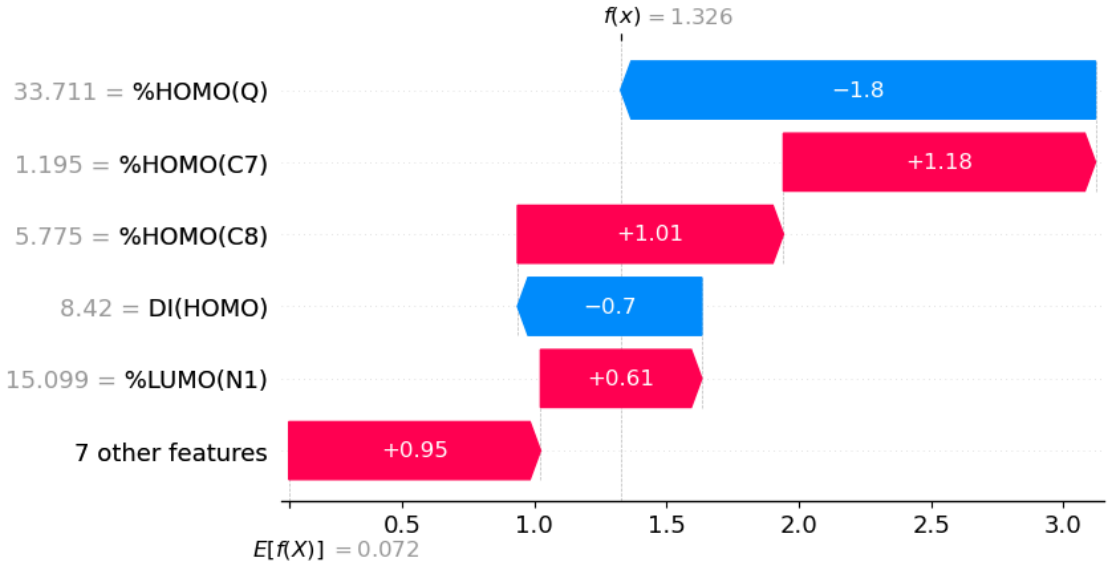
1b



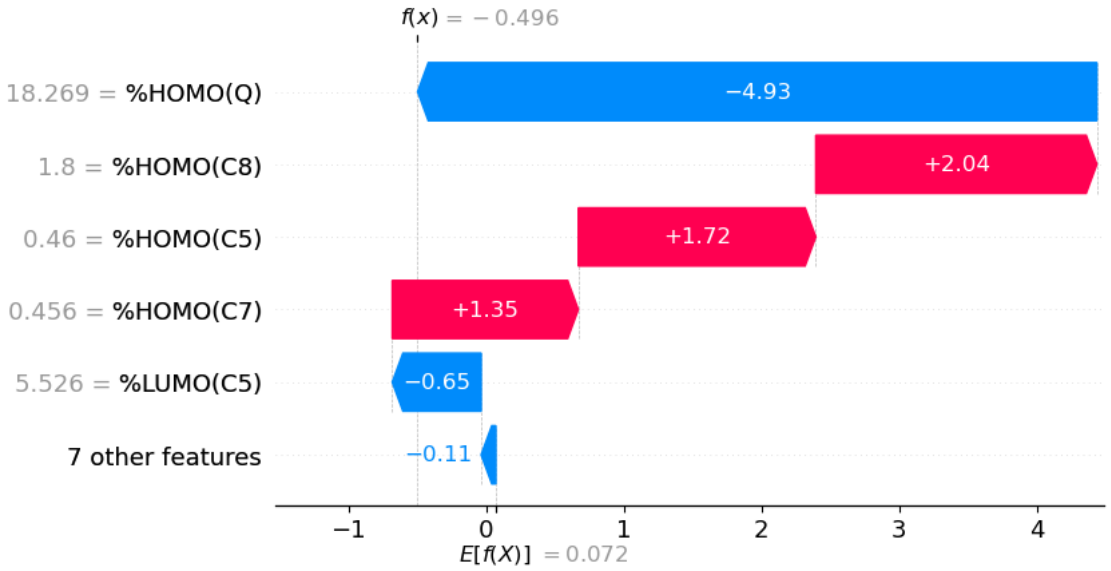
4a


f(x) = 0.413
22.231 = %HOMO(Q)
2.323 = %HOMO(C8)
0.741 = %HOMO(C5)
0.522 = %HOMO(C7)
4.996 = %LUMO(C5)
7 other features
−4.13
+1.9
+1.64
+1.34
−0.51
+0.1
−1
0
1
2
3
4
E[f(X)] = 0.072

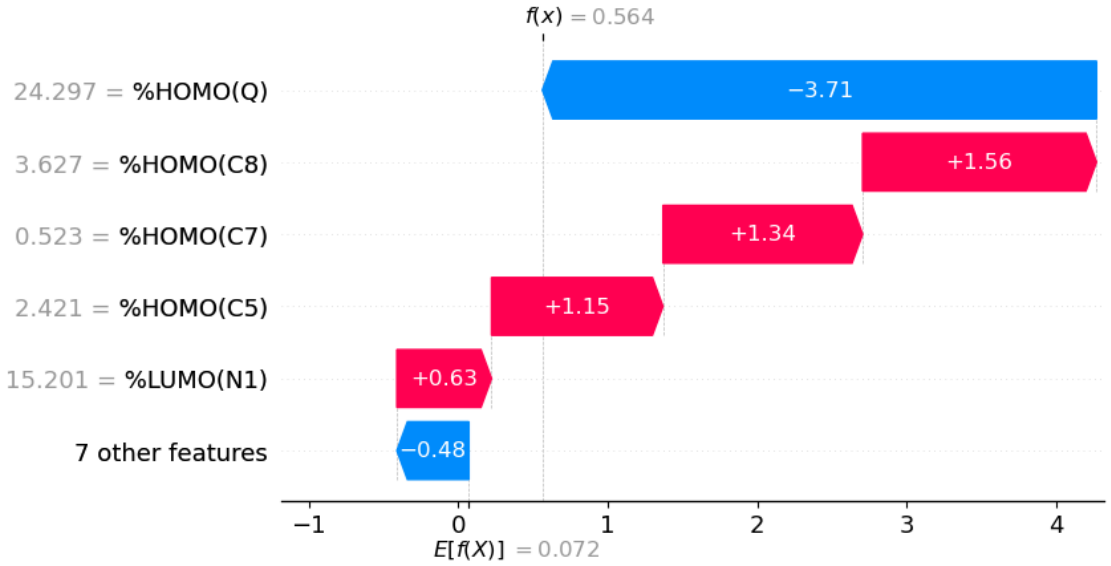
4b


f(x) = 1.326
33.711 = %HOMO(Q)
1.195 = %HOMO(C7)
5.775 = %HOMO(C8)
8.42 = DI(HOMO)
15.099 = %LUMO(N1)
7 other features
−1.8
+1.18
+1.01
−0.7
+0.61
+0.95
0.5
1.0
1.5
2.0
2.5
3.0
E[f(X)] = 0.072

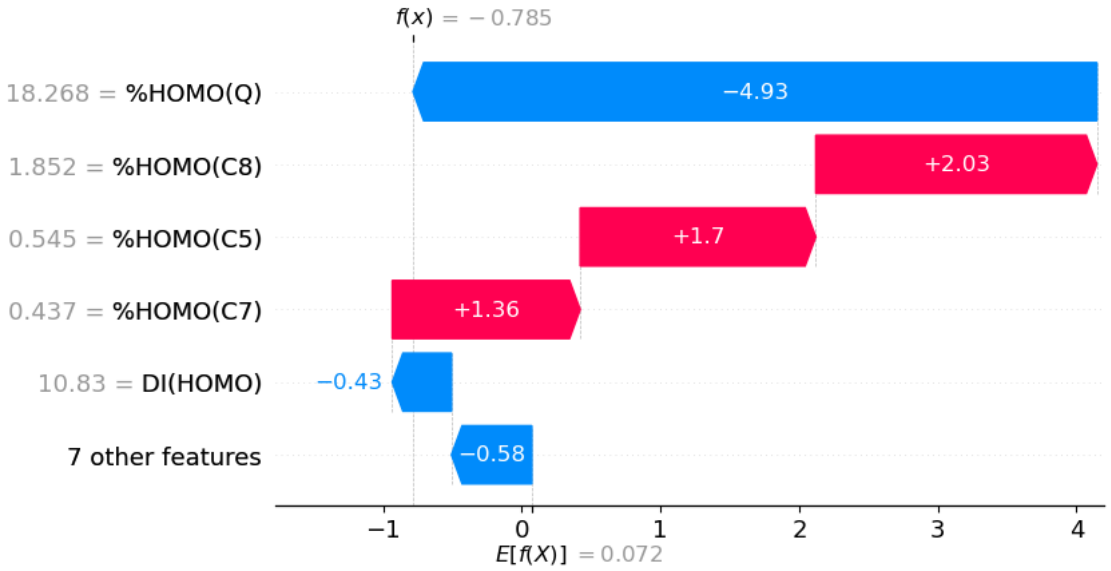
4c


f(x) = − 0.496
18.269 = %HOMO(Q)
1.8 = %HOMO(C8)
0.46 = %HOMO(C5)
0.456 = %HOMO(C7)
5.526 = %LUMO(C5)
7 other features
−4.93
+2.04
+1.72
+1.35
−0.65
−0.11
−1
0
1
2
3
4
E[f(X)] = 0.072

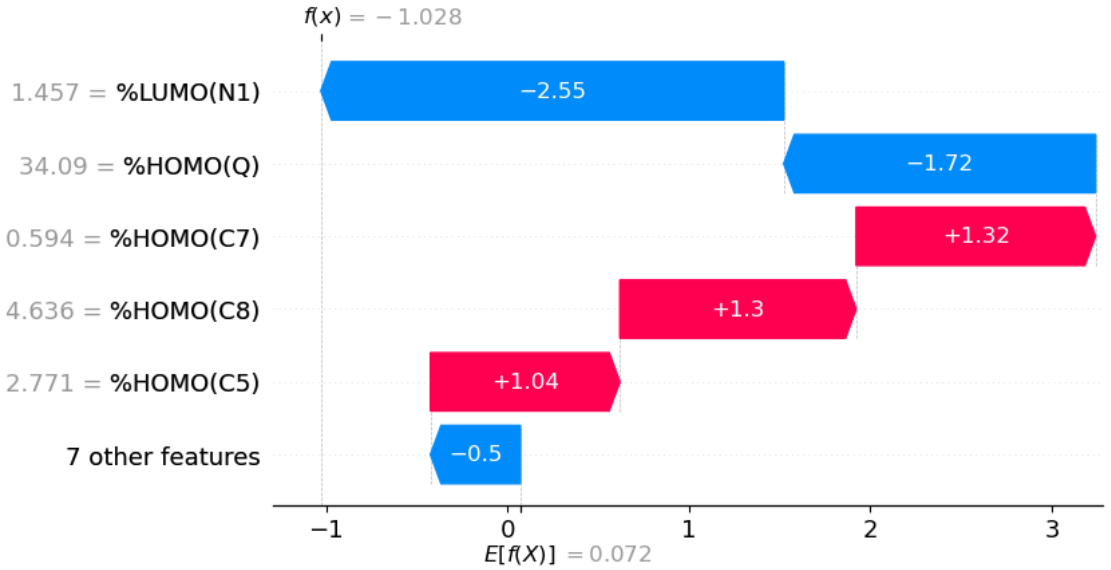
4d


f(x) = 0.564
24.297 = %HOMO(Q)
3.627 = %HOMO(C8)
0.523 = %HOMO(C7)
2.421 = %HOMO(C5)
15.201 = %LUMO(N1)
7 other features
−3.71
+1.56
+1.34
+1.15
+0.63
−0.48
−1
0
1
2
3
4
E[f(X)] = 0.072

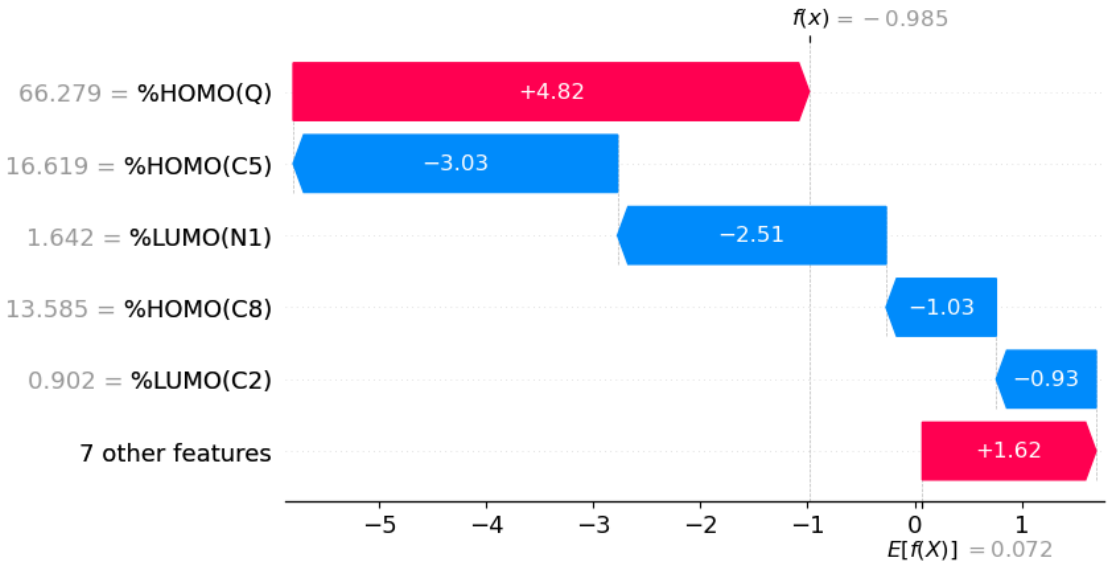
4e


f(x) = − 0.785
18.268 = %HOMO(Q)
1.852 = %HOMO(C8)
0.545 = %HOMO(C5)
0.437 = %HOMO(C7)
10.83 = DI(HOMO)
7 other features
−4.93
+2.03
+1.7
+1.36
−0.43
−0.58
−1
0
1
2
3
4
E[f(X)] = 0.072

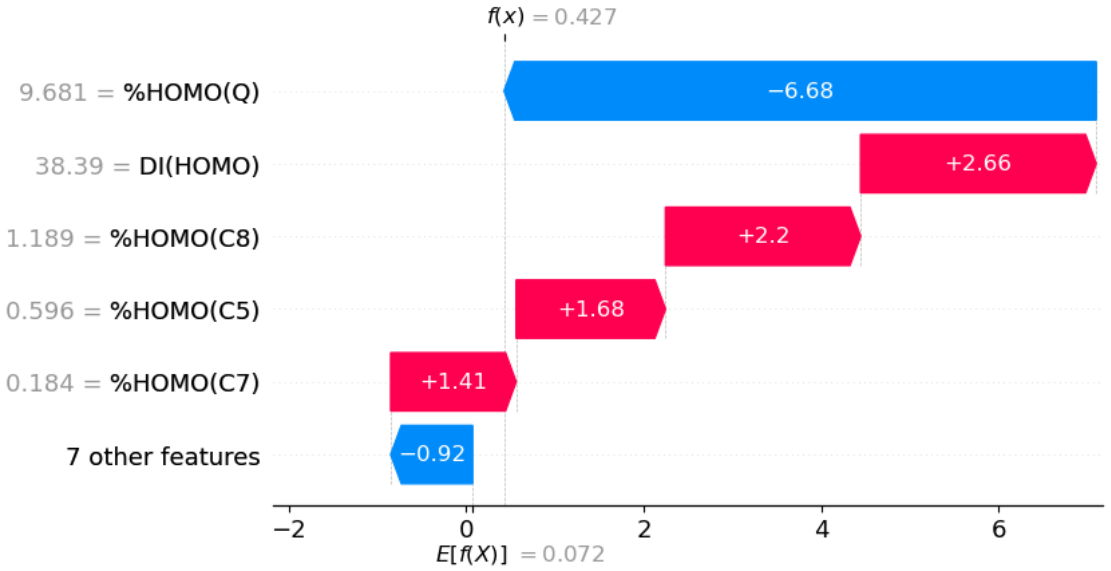
4f


f(x) = − 1.028
1.457 = %LUMO(N1)
34.09 = %HOMO(Q)
0.594 = %HOMO(C7)
4.636 = %HOMO(C8)
2.771 = %HOMO(C5)
7 other features
−2.55
−1.72
+1.32
+1.3
+1.04
−0.5
−1
0
1
2
3
E[f(X)] = 0.072

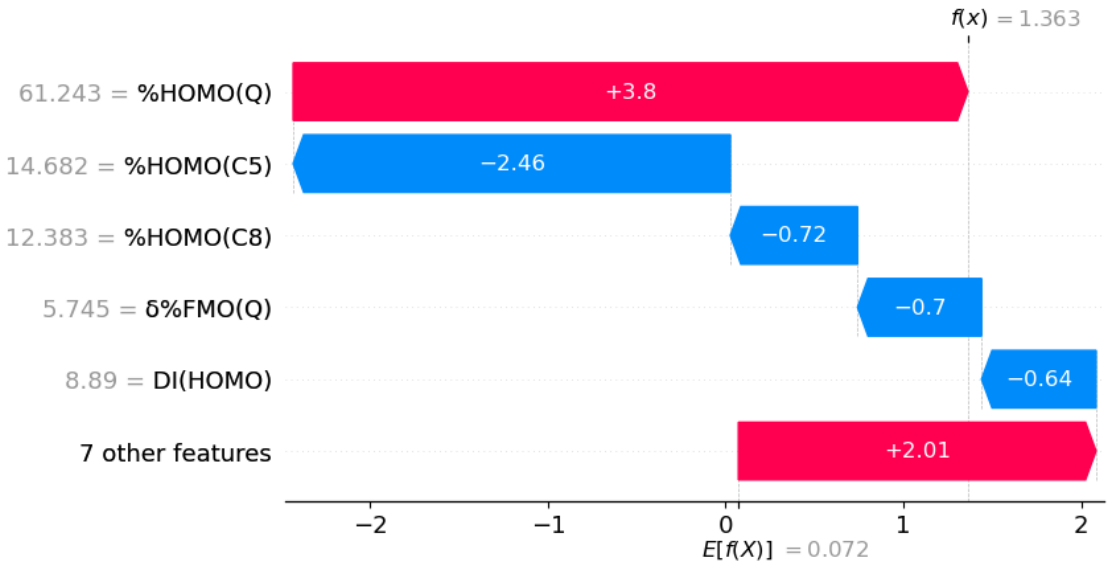
4g


f(x) = −0.985
66.279 = %HOMO(Q)
16.619 = %HOMO(C5)
1.642 = %LUMO(N1)
13.585 = %HOMO(C8)
0.902 = %LUMO(C2)
7 other features
+4.82
−3.03
−2.51
−1.03
−0.93
+1.62
−5
−4
−3
−2
−1
0
1
E[f(X)] = 0.072

5a


f(x) = 0.427
9.681 = %HOMO(Q)
38.39 = DI(HOMO)
1.189 = %HOMO(C8)
0.596 = %HOMO(C5)
0.184 = %HOMO(C7)
7 other features
−6.68
+2.66
+2.2
+1.68
+1.41
−0.92
−2
0
2
4
6
E[f(X)] = 0.072

5b


f(x) = 1.363
61.243 = %HOMO(Q)
14.682 = %HOMO(C5)
12.383 = %HOMO(C8)
5.745 = δ%FMO(Q)
8.89 = DI(HOMO)
7 other features
+3.8
−2.46
−0.72
−0.7
−0.64
+2.01
−2
−1
0
1
2
E[f(X)] = 0.072

[ ]: