Prediction of Physicochemical Properties of Organic Compounds from 2D Molecular Structure – Fragment methods vs. LFER models

Gerrit Schüürmann [a,b] *, Ralf-Uwe Ebert [a], Ralph Kühne [a]

[a] Department of Ecological Chemistry, UFZ Centre for Environmental Research, Permoserstrasse 15, 04318 Leipzig, Germany

[b] Institute for Organic Chemistry, Technical University Bergakademie Freiberg, Leipziger Strasse 29, 09596 Freiberg, Germany.

Supplemental Material:

*Corresponding author: phone +49-341-235-2309, fax +49-341-235-2401, e-mail gerrit.schuurmann@ufz.de

Table A1. ChemProp [1] implementation of the log K_{oc} model by Tao et al. [2]. Substructures are coded in SMARTS [4]. Comments are added if necessary. The increment values are taken from [2] and repeated here for clarity.

---

1. General rules.

Search and counting procedure. Atoms and bonds have strictly to be looked for in the order listed here (from top to bottom). Each atom of the molecule has to occur in only one fragment (corresponding to the first hit), and each atom has to be assigned to a fragment, otherwise the model result would be invalid. According to SMARTS, atoms marked by $ in the search string are not assigned to a fragment. In addition to SMARTS, § is used here as a syntax extension of the $ symbol: Several atoms marked by § in a search structure must be different real atoms in the molecule, while different atoms marked by $ in the search string may be the same actual atom in the molecule according to SMARTS. Atoms marked by $ or § must be found elsewhere to make the overall decomposition of the molecule into fragments complete. The same applies to bonds within fragments, but not to bonds to atoms marked by $ or §.

Furthermore, the model application is restricted to structures with an uninterrupted bonding system (e.g. no salts or otherwise disconnected structures).

Aromaticity. Rings fulfilling the Hückel rule are considered to be aromatic. This includes respective 5-member rings with, e.g., NH, S or O. In general, 6-member rings with O= or S= directly attached to a ring member are excluded. However, in some marked cases particular substructures of the latter type listed below are considered to be aromatic.

---

2. Abbreviations.

In the following text, a number of substructures will be denoted by abbreviations in analogy to simple atom symbols. They define atoms in particular structural environments and will be used in the same manner as normal atom symbols, meaning any atom fulfilling the respective conditions.

Symbol **Rb** for
(-,:[$(*@&!:*)]),~[$(@&!:*)],
*(-,:[$(*@&!:*)]),~[$(*)[1)]]

Halogen and non-halogen substructures:

Symbol **Xx** for [F,Cl,Br,I]~[$(C)]
Symbol **Xc** for [#6,F,Cl,Br,I]|
Symbol **Xy** for [Xx]-[$(C-[Xx])]

Particular branches (short chain or hydrophilic group):

Symbol **Cv** for any of
CH3,
CH2(-[$(CH3)]),
CH1(=[$(CH2)]),
CH0(#[$(CH)]),
[!#6],

[#6](-[$(Xc)]),
[#6](-[$(#6)]$^{2)}$),
[#6](-[$(#6R1)]),
[#6]((-[$(#6)]$^{3)}$)
attached by a single bond to [$(AR0)]

Cv will be used in a more complex context:
Cx for [AR0]-[$(Cv)]
Cy for [Cx]-[$(Cv)]
Cb for [Cy]-[$(Cv)]
Cc for [Cb]-[$(Cv)]

Particular aromaticity:

Ci for any of
c$^{4)}$ in a ring together with 2 of c$^{4)}$=O,
c$^{4)}$(=[$(O)] in a ring together with c$^{4)}$=O,
c$^{4)}$(=[$(O)]n$^{4)}$,
c$^{4)}$ in a ring together with n$^{4)}$c$^{4)}$=O,
c$^{4)}$([§(c$^{4)}$=O)])[§(c$^{4)}$n$^{4)}$)],
c$^{4)}$([$(n$^{4)}$)])[$(c$^{4)}$c$^{4)}$c^{4}=O)],
c$^{4)}$(=[$(O)])[$(c$^{4)}$c$^{4)}$n$^{4)}$)],
n$^{4)}$[$(c$^{4)}c^{4)}c^{4)}$=O)],
C$^{4,5)}$(=[$(O)])-[$(N-C$^{4)}$(=O)cc$^{5)}$)]

Ci will be used in a more complex context:
Ai for any of
[OR1,SR1]([§(c,[Ci])])[§(c,[Ci])],
[a;!c] $^{4)}$ in a ring together with [Ci],
N(-[§(C$^{4,5)}$=O])-[§(C$^{4)}$(=O)cc$^{5)}$)],
Cj for [Ci] or [Ai]
Cm for a (aromatic atom corresponding to the general definition) or [Cj]

Respective non-aromaticity:

Before looking for the following substructures, mark all [Cm] atoms
Cq for [CH1$^{4)}$,cH1$^{4,6)}$]
Cp for any of
[CH2,CH3,CH4],
[CH0$^{4)}$,cH0$^{4,6)}$]
Cn for either [Cq] or [Cp]
Now, delete all marks for [Cm].

H-Polar factors:

Before looking for the following substructures, mark all of
[NH1](-[§([SX4](=O)(=O)-[Cm])])-[§([Cn](=O)-[NH1]-[Cm])]
Symbol **Hp** for any of
[SX4](=O)(=O)-[$([#6],-[NH1]-[#6])],
C(=O)(-[$(C)])-[NH2],
C(=O)(-[§(C)])-[NH1]-[§(C)],

[O,S](-[§(#6X4)]-[§([#6X4],[#6]=[#6],C#C,[!#6])],
[O,S](-[§([#6]=[!0&!S])]-[§([#6X4],[#6]=[#6],C#C,[!#6])],
[OH1]-[$(CX4,c)],
[OH1]-[$([#6]=[!#8&!#16])],
[CH1](=O)(-[$(!#7&!#8)]),
C(=O)(-[§(#6)])(-[§(!#7&!#8)]),
C(=O)(-[§(!N)])-[OH1,O~[§(#6)]],
[NH2]-[$(CX4,c)],
[NH2]-[$([#6]=[!#8&!#16])],
[NH1](-[§(CX4,c)])-[§(CX4,c)],
[NH1](-[§(CX4,c)])-[§([#6]=[!#8&!#16])],
[NH1](-[§([#6]=[!#8&!#16])])-[§([#6]=[!#8&!#16])],
[NH0](-[§(CX4,c)])(-[§(CX4,c)])-[§(CX4,c)],
[NH0](-[§(CX4,c)])(-[§(CX4,c)])-[§([#6]=[!#8&!#16])],
[NH0](-[§(CX4,c)])(-[§([#6]=[!#8&!#16])])-[§([#6]=[!#8&!#16])],
[NH0](-[§([#6]=[!#8&!#16])]) (-[§([#6]=[!#8&!#16])])-[§([#6]=[!#8&!#16])],
[NX3][6)]-[$([SX4](=O)=O)],

Now, delete all marks for H-polar factors.

Extended H-polar factors for particular cases:

Hq for [Hp] or any of

[NH1](-[§(c)])-[§([#6]=[#8,#16])],
[NH0](-[§(CX4)])(-[§(c)])-[§([#6]=[#8,#16])],
[NH0](-[§(c])(-[§(c)])-[§([#6]=[#8,#16])]

Isolating carbon definition:

Ic for [Cn] or [Cm]

Hc for any of

[n,o,s]~[$(c)],
[Ai]~[$(Ci)],
[Ci]=[$(O)]

---

3. Isolating carbon handling.

In the original method description, the isolating carbon concept is not strictly applied. For that reason, a complex workaround is required.

Mark all C that belong to any of the following fragments (i.e., to be non-isolated):

C(-[§(Ic)]=[§(N-O-C(=O)-[NH1,N(-[Ic]),N[SX2]-[Ic])],
[CH1](-[§(Ic)]=[§(O)],
C(-[§(Ic)](-[§([OH1])])=[§(O)],
C(-[§(Ic)](-[§([NH1]-Ic)])=[§(O)],
[CH1](-[§(Ic)]=[§([NH1]-[Ic])],
C(-[§(Ic)](=[§(O)])-[§([NH2])],
C(-[§(O-[Ic])])(=[§(O)])-[§([NH1]-[Ic])],
C(-[§([OH1])])(=[§(O)])-[§([NH1]-[Ic])],
C(-[§(O-[Ic][5)])])(=[§(O)])-[§([NH1][5)])],
C(-[§(O-[Ic])])(=[§(O)])-[§([NH2])],
[CJ1](-[§(Ic)])=[§(N-[Ic])],
C(-[§([NH1]-[Ic])])(=[§(O)])-[§([NH1]-[Ic])],
C(-[§([NJ1]-[Ic][5)])])(=[§(O)])-[§([NH1][5)])],
C(-[§([NH1]-[Ic])])(=[§(O)])-[§([NH2])],

C(-[§([NH1]-[Ic])])(=[§(O)])-[§([NH1]-[Ic])],
C(-[§([NJ1]-[Ic][5)])])(=[§(O)])-[§([NH1][5)])],
C(-[§([NJ1]-[Ic][5)])])(=[§(O)])-[§([NH0][5)]-[Ic])],
C(-[§([NH1]-[Ic])])(=[§(O)])-[§([NH1]-O-[Ic])],
C(-[§([NH1]-[Ic])])(=[§(O)])-[§(N(-[Ic])-O-[Ic])],
C(-[§([NH1,O]-[Ic])]) (=[§(O)])-[§([NH1]-[SX4](=O)(=O)-[Ic])],
C(-[§(Ic)])(-[§(F)]) (-[§(F)]) -[§(F)],
C(-[§(Ic)])#[§(N)],
C(-[§(Ic)]) (=[§(O)])-[§(NH2)],
C(-[§(Ic)]) (=[§(O)])-[§(NH1(-[Ic])-[Ic])],
C(-[§(Ic)]) (=[§(O)])-[§(NH0(-[Ic])(-[Ic])-[Ic])],
C(-[§(Ic)]) (=[§(O)])-[§(Ic)],
C(-[§(Ic)]) (=[§(O)])-[§(O-[Ic])],
C(-[§(Ic)]) =[§(N-[Ic])],
C(-[§(O-[Ic])]) (=[§(O)])-[§(NH2, NH1(-[Ic])-[Ic], NH0(-[Ic])-[Ic])],
C(-[§([SX2]-[Ic])]) (=[§(O)])-[§(NH2, NH1(-[Ic])-[Ic], NH0(-[Ic])-[Ic])],
C(-[§(Ic)]) (=[§(O)])-[§(O-N(-[Ic])-[Ic])],
C[5)](=[§(O)])-[§(O-N(-[Ic])-[Ic][5)])],
C(-[§([NH1]-[Ic])]) (=[§(O)])-[§([NH0](-[Ic])-[Ic])]

By definition, all remaining C atoms are isolating carbons, and as such form simple fragments.

---

4. Fragments, Part 1.

Fragments can only be attached to isolating C atoms.

While Part 1 fragments affect the consideration of correction factors (as defined subsequently), this is not the case for Part 2 fragments. Accordingly, Part 2 fragments are defined below the section defining correction factors.

Note:
Groups with alternative aromaticity[4)] belonging to the subgroup "*Fused in Aromatic Ring*" will be looked for before searching for any other group:

-0.898 O=[$(Ci)]
0.413 [nH1][4)] in ring with [Ci]
or: [NH1](-[§(C[4,5)]=0)])-([§(C[4)](=O)-cc[5)])]) (mark all bonds within this substructure)
-0.739 [nH0X3][4)] in ring with [Ci]
or: [NH0](-[§(C[4,5)]=0)])-([§(C[4)](=O)-cc[5)])]) (mark all bonds within this substructure)
-0.308 [nH0X2][4)] in ring with [Ci]
0.251 [cH0][4)] in ring with [Ci]
0.305 [cH1][4)] in ring with [Ci]
0.533 [o][4)] in ring with [Ci]
or: [OR1](-[§(c,[Ci])]) -[§(c,[Ci])]
0.748 [SR1](-[§(c,[Ci])]) -[§(c,[Ci])]

Next, look for
O=[7)]C[4,5)]-cc-C[4)](=O)-N[5)]
to mark the indicated bond.

In the following, [Cn] can only be found if not marked already by the rules above. However, it will not be marked from now on. [Cm] may always be found (i.e. there is no restriction due to possible mark status).

Without C or H

1.149 F-[$(Cn)]
0.087 F-[$(Cm)]
0.523 Cl-[$(Cn)]
0.439 Cl-[$(Cm)]
0.558 Br-[$(Cn)]
0.404 Br-[$(Cm)]
-0.584 [§(Cn)]-O-[§(Cn)]
-0.723 [§(Cm)]-O-[§(Cn)]
-0.719 [§(Cm)]-O-[§(Cm)]
-0.074 [§(Cn)]-[SX2]-[§(Cn)],
[§(N)]-[SX2]-[§(N)], [§(N=C)]-[SX2]-[§(Cn)][8)]
-0.307 [§(Cm)]-[SX2]-[§([Cn],[Cm])]
0.168 [$(Cm)]-[NX3](=O[7)])~[7)][OH0X1]
-0.970 [§(Cn)]-[SX4](=[7)]O)(=[7)]O)-[§(Cn)]
-1.153 [§(Cm)]-[SX4](=[7)]O)(=[7)]O)-[§(Cn)]
-1.737 [§(Cm)]-[SX4](=[7)]O)(=[7)]O)-O-[§(Cn)] (= -1.153 + (-0.584))
-1.693 [§(Cn)]-[SX4](=[7)]O)(=[7)]O)-O-[§(Cm)] (= -0.970 + (-0.723))
-0.709 [§(Cn)]-[SX3](=[7)]O)-[§(Cn)]
-1.133 [§(Cm)]-[SX3](=[7)]O)-[§(Cn)]
0.002 [$(Cn)]-[SX2]-[7)][PX4](=[7][SX1])(-[7)]O-[§(Cn)])-[7)]O-[$([Cn],[Cm])]
-0.964 [$(Cn)]-O-[7)][PX4](=[7)]O)(-[7)]O-[§(Cn)])-[7)]O-[$(Cn)]
-0.456 [§(Cm)]-O-[7)][PX4](=[7)][SX1])(-[7)]O-[$(Cn)])-[7)]O-[$(Cn)]
-1.455 [§(Cn)]-O-[7)][SX3](=[7)]O)-[7)]O-[§(Cn)]
-0.934 [$(Cn)]-[PX4](=[7)][SX1])(-[7)]O-[§(Cn)])-[7)]O-[$(Cn)],
-1.309 [$(Cn)]-[SX2]-[7)][PX4](=[7)]O)(-[7)]O-[§(Cn)])-[7)]O-[$(Cn)]
-1.204 [§(Cm)]-[SX2]-[7)][PX4](=[7)]O)(-[7)]O-[$(Cn)])-[7)]O-[$(Cn)],
[§(Cm)]-[PX4](=[7)]O)(-[7)]O-[§(Cn)])-[7)]O-[§(Cn)] [8)]
-1.906 [$(Cn)]-[PX3](=[7)]O)-[H,A]
-1.160 [$(Cn)]-[PX3](=[7)][SX1])-[H,A]
-0.934 [§(Cm)]-[SX2] -[7)][PX4](-[§(Cn)])-O-[§(Cn)]

With C, without H

0.521 [§(Cm)]-[§(C)](-[7)]F)(-[7)]F)-[7)]F
0.075 [§(Cm)]-[§(C)]#[7)]N
-1.767 [$(Cn)]-Cn(=[7)]O)-[7)][NH0](-[§(Cn)])-[$(Cn)]

-1.833 [$(Cn)]-Cn(=[7)]O)-[7)][NH0](-[$(Cn)])-[§(Cm)],
[§(Cm)]-[Cn](=[7)]O)-[7)][NH0](-[§(Cn)])-[§(Cn)]
-0.839 [§(Cm)]-Cn(=[7)]O)-[§([Cn],[Cm])]
-0.434 [§(Cn)]-Cn(=[7)]O)-O-[§(Cn)]
-0.427 [§(Cm)]-Cn(=[7)]O)-O-[§([Cn],[Cm])]
-1.355 [§(Cn)]-Cn(=[7)]O)-[§(Cn)]
-0.272 [$(Cn)]-[Cp]=[7)]N-[$(Cn)]
-1.831 [$(Cn)]-O-[7)]C(=[7)]O)-[7)][NH2,[NH1]-[§(Cn)],[NH0](-[§(Cn)])-[$(Cn)]],
[§(Cm)]-O-[7)]C(=[7)]O)-[7)][NH0(-[§(Cn)])-[§(Cn)]

-0.365 [$(Cn)]-S-$^{7)}$C(=$^{7)}$O)-$^{7)}$ [NH2,[NH1]-[§(Cn)],[NH0](-[§(Cn)])-[$(Cn)]]

With H, without C

1.487 -H (only in connection with some other fragments, as indicated respectively)
-0.300 [OH]-[$(Cn)]
-0.176 [OH]-[$(Cm)]
-1.715 [§(Cm)]-O-$^{7)}$[PX4](=$^{7)}$O)(-$^{7)}$[NH1]-[$(Cn)])-$^{7)}$O-[$(Cn)]

-0.963 [$(Cm)]-[SX4](=$^{7)}$O)(=$^{7)}$O)-[NH2] (= -1.153 + (-0.545) + 0.735)
-1.268 [§(Cm)]-[SX4](=$^{7)}$O)(=$^{7)}$O)-[NH1]-[§(Cn)] (= -1.153 + (-0.545) + 0.430)

With C and H

-1.109 [$(Cn)]-[Cq]=$^{7)}$O,[CH2]=$^{7)}$O
-0.678 [$(Cn)]-[Cn](=$^{7)}$O)-$^{7)}$[OH1]
-0.425 [$(Cm)]-[Cn](=$^{7)}$O)-$^{7)}$[OH1]
-1.406 [$(Cn)]-[Cn](=^{7}O)-$^{7)}$[NH1]-[$(Cn)]
0.081 [§(Cq)](=$^{7)}$O)-$^{7)}$[NH1]-[§(Cn)] (= -1.406 + 1.487)
-0.534 [§(Cm)]-[Cn](=$^{7)}$O)-$^{7)}$[NH1]-[§(Cn)]
-0.875 [§(Cn)]-[Cn](=$^{7)}$O)-$^{7)}$[NH1]-[§(Cm)]
0.612 [§(Cq)](=$^{7)}$O)-$^{7)}$[NH1]-[§(Cm)] (= 0.875 + 1.487)
-2.515 [$(Cm)]-[Cn](=$^{7)}$O)-$^{7)}$[NH1]-[$(Cm)]
-0.315 [$(Cm)]-[Cn](=$^{7)}$O)-$^{7)}$[NH2]
-0.837 [§(Cm)]-O-$^{7)}$[Cn](=$^{7)}$O)-$^{7)}$[NH1]-[§(Cn)]
-0.600 [§(Cn)]-O-$^{7)}$[Cn](=$^{7)}$O)-$^{7)}$[NH1]-[§(Cm)]
0.887 [OH1]-$^{7)}$[Cn](=$^{7)}$O)-$^{7)}$[NH1]-[$(Cm)] (= -0.600 + 1.487)
-0.522 [$(Cm)]-O-$^{7)}$[Cn](=$^{7)}$O)-$^{7)}$[NH1]-[$(Cm)]
-0.479 [$(Cm)]-O-$^{7)}$[Cn](=$^{7)}$O)-$^{7)}$[NH2]
-1.738 [§(Cm)]-[§(Cq)]=$^{7)}$[NH1]
-1.158 [§(Cm)]-[NH1]-$^{7)}$[Cn](=$^{7)}$O)-$^{7)}$[NH1]-[§(Cn)]
-0.219 [$([Cn],[Cm])]-[NH1]-$^{7)}$[Cn](=$^{7)}$O)-$^{7)}$[NH2]
-0.533 [$(Cn)]-[Cn]=$^{7)}$N-$^{7)}$O-$^{7)}$[Cn](=$^{7)}$O)-$^{7)}$[NH1,N(-[$(Cn)]),N(-[§(SX2)]]-$(Cn)]
-1.476 [§(Cm)]-[NH1]-$^{7)}$[Cn](=$^{7)}$O)-$^{7)}$[NH0](-[§(Cn)])-[§(Cn)]
-1.204 [$(Cn)]-[NH1]-$^{7)}$[Cn](=$^{7)}$O)-$^{7)}$N(-[$(Cn)]-[§(Cm)]
-1.026 [§(Cm)]-[NH1]-$^{7)}$[Cn](=O)-$^{7)}$[NH1X3]-$^{7)}$O-[§(Cn)],
[§(Cm)]-[NH1]-$^{7)}$[Cn](=O)-$^{7)}$[NX3(-[$(Cn)])]-$^{7)}$O-[$(Cn)]
-1.571 [$(Cm)]-[SX4](=$^{7)}$O)(=$^{7)}$O)-$^{7)}$[NH1]-$^{7)}$[Cn](=$^{7)}$O)-$^{7)}$[NH1]-[$(Cm)]
-2.050 [§(Cm)]-[SX4](=$^{7)}$O)(=$^{7)}$O)-$^{7)}$[NH1]-$^{7)}$[Cn](=$^{7)}$O)-$^{7)}$O-[§(Cn)]
(= -0.219 + (-1.831))
0.378 [CH2]=$^{7)}$O (= -1.109 + 1.487)

Fused in Aromatic Ring

0.305 [cH1]
0.251 [cH0]
-0.308 [nH0X2]
0.748 [s]
-0.739 [nH0X3]
0.413 [nH1X3]
0.533 [o]

---

5. Correction factors.

Correction factors are applied according to particular rules as specified below, and in addition to the fragments.
As noted above, the consideration of correction factors is affected by Part 1 fragments.

Before starting to apply any correction factor, save all atom and bond marks. But do not un-mark them.

Aliphatic chain bond factor

Look for one occurrence of [§(*)]-&!@[§([AR0])]-&!@ [§(*)]
- If found, apply 0.088 to each of [§(*)]-&!@[§(*)]
- Mark bond as counted
- Look once for [$(*)]
  - If found, reduce by 0.088 and apply the following rules:
  - Look once for [$(*R1)]
    - If found, add 0.088 and apply the following rules to account for ring-chain-interruptions:
      - Delete all atom and bond marks
      - Look for [§(*R1)]-&!@ [§(*R1)]
      - and add 2*0.088 for each occurrence, mark the bond
      - Look for [§(*R1)]-&!@ [§(*R0)]
      - and add 0.088 for each occurrence, mark the bond
      - Delete all atom and bond marks again
      - Look for [§(*)]~[§(*)]
      - and add 0.088 for each occurrence, mark the bond
      - Look for any of
        - [$(*)]
        - and [*R3]
        - and twice[*R2] (not necessarily bonded together)
      - and reduce by 0.088 for each occurrence, mark bond
      - Restore the atom and bond counts saved above, and invert them
      - Look for [§(*)]*-&!@[§(*R1)] and mark the bond
      - Then, look for *-&!@[$(*R1)], mark the bond
        - and add 0.088 for each

Restore the atom and bond counts saved above again.

Aliphatic ring bond factor

0.256 [§(*)]@[§(*)] (mark the bond)
If any occurrence, the following procedure applies:
- Un-mark all chain bonds and aromatic bonds by:
  - Invert all marks
  - [§(*)]-&!@[§(*)] mark the bond
  - [§(*)]:[§(*)] mark the bond
  - Invert all marks again
- Look for [§(*)]@[§(*)]-&!@[§(Rb)].
  - For each occurrence, reduce by 0.256
  - If any occurrence, look for the following substructures:
    - [§(a)]@[§(a)]:[§(a)] For each occurrence,
      - reduce by 0.256 and mark the bond

[§(*)]@[§(a)]:[§(a)] For each occurrence,
reduce by 0.5*0.256 and mark the bond
Else, if no occurrence:
Look for [§(a)]@[§(a)]:[§(a)]
For each occurrence, reduce by 0.256 and mark the bond
If any occurrence, look for [§(*)]@[§(a)]:[§(a)]
For each occurrence, reduce by 0.5*0.256 and mark the bond
Else, if no occurrence:
Look for [§(*)]@[§(a)]:[§(a)]
For each occurrence, reduce by 0.5*0.256 and mark the bond
If no occurrence, look once for [§(*)]@[§(*)]
Reduce by 0.256 and mark the bond
Delete all atom and bond marks
Look for [§(Cj)]@ [§(Cj)]@ [§(Cj)]@ [§(Cj)]@ [§(Cj)]@ [§(Cj)]
If found, reduce by 5*0.256 for each, mark the bond, save all marks,
delete all marks, and
Look for [§(Cj)]@[7)] [§(Cj)]@ [§(Cj)]@ [§(Cj)]@ [§(Cj)]
Invert all marks
[§(a)]=[§(a)] both atoms fused aromatic[4)]
Add 0.256 for each occurence
Restore the last saved marks
Look for[§(Cj)]@ [§(Cj)]@ [§(Cj)]@ [§(Cj)]@ [§(Cj)]
If found, reduce by 4*0.256 for each, mark the bond, save all marks,
delete all marks, and
Look for [§(Cj)]@[7)] [§(Cj)]@ [§(Cj)]@ [§(Cj)]
Invert all marks
[§(a)]=[§(a)] both atoms fused aromatic[4)]
Add 0.256 for each occurence
Restore the last saved marks
[OX2R1],[SX2R1](-[§(c,[Ci])]) -[§(c,[Ci])]
Reduce by 0.256 for each occurence

Multiple halogenation factors

Delete all atom and bond marks
0.009 [Xx][§(#6)]-[§(Xy)],
[§(Xx)][#6]-[§([#6]-[Xy])]
Restore the atom and bond counts firstly saved
4*-0.035 [Xy]-[#6](-[§(Xx)]-[§(Xx)]
3*0.131 [Xy]-[#6]-[$(Xx)]
2*0.045 [Xy]-[#6]
Invert all marks
Mark all C atoms
Invert all marks
Mark any of
[Xy]-[#6](-[§(Xx)])-[§(Xx)],
[Xy]-[#6]-[$(Xx)],
[Xy]-[#6],
[§(Xx)]#6-[§([#6][Xy])],
[!C]
2*-0.102 [Cc]
-0.102 [Cb]

Delete all atom and bond marks
Mark any of
[NH0X3R0](-[§(O)])-[§(C=O)],
[NH0X3R0]-[$(C(-O)=O)]
-0.102 [NH0X3R0]-[$(C=O)]

Aliphatic bond factors

Restore the atom and bond counts firstly saved again
0.114 [§(*)]=&!@[§(*)]
Look for [§(a)]:[§([!n]&a)] and mark the bond
0.126 [§(*)]=&@[§(*)] <F=0> "aliphatic ring double bond factor"
-1.028 [§(*)]#[§(*)]
Restore the atom and bond counts firstly saved again
[§(a)]=[§(a)] both atoms fused aromatic[4], mark the bond
-0.126 [§(Cj)]@[7] [§(Cj)]@ [§(Cj)]@ [§(Cj)]@ [§(Cj)]
-0.126 [§(Cj)]@[7] [§(Cj)]@ [§(Cj)]@ [§(Cj)]

H-polar factors

Delete all atom and bond marks
0.373 [§(Hp)]-[CR0]-[§(Hp)]
Delete all atom and bond marks
0.429 [§(Hp)]-C!:&!@C-[§(Hp)]
Delete all atom and bond marks
0.636 [§(Hp)]-[CR1]-[§(Hp)]
Delete all atom and bond marks
Now mark any [§(O=c)]cc[§(Hp)] (all C atoms alternative aromaticity[4])
-0.085 [§(Hp)]-C!:&@C-[§(Hp)]
Delete all atom and bond marks
2*0.075 [§(Hc)]:c(:[§(Hc)])-[§([Hq],C(=O)-O,O-C=O)]
2*0.075 [§(Hc)]:c(:[§(Hc)])-[§(Hq)]
0.075 [§(Hc)]:c-[§(Hq,C(=O)-O,O-C=O)]
Delete all atom and bond marks
2*0.166 [§(Hc)]:c(:[§(Hc)]) :[§(Hc)]
2*0.166 [§(Hc)]:c(:[§(Hc)])-&@[§(([OX2R1],[SX2R1])-c)]
0.166 [§(Hc)]:c:[§(Hc)]
Delete all atom and bond marks
0.560 [§([Hp],[Hc])]- [§([Hp],[Hc])] (mark the bond)
Delete all atom and bond marks
-0.107 [§(F,Cl,Br,I)]-[§(c)]: [7] [§(Hc)]
Restore the atom and bond counts firstly saved again

Branched carbon and nitrogen factors

-1.023 [CH0X4]
-0.731 [CH1X4]
-0.376 any of
[§(#6)]~[CH0X2]~ [§(#6)],
[CH2X4],
[CH1X3],
[CH0X3]=[$(C)],

0.430 [NH1],[NH0X2]
0.735 [NH2]

---

6. Fragments, Part 2.

Part 2 fragments include isolating carbons and other basic fragments as defined below.

In the following substructures, look for C atoms not marked yet, but do not mark them
-0.545 [§(C)]-[NH2,[NH1](-[§(C)]),[NH0](-[§(C)])-[§(C)]]
-0.626 [§(Cm)] -[NH2,[NH1](-[§(C)]),[NH0](-[§(C)])-[§(C)]]
-0.562 [NH1,[NH0](-[§(#6)])]([§(Cm)])[§(Cm)]

0.010 [§(c)]-C-[§(c)]
0.423 C-[$(c)]
0.519 C

[1)] one or chain of any atoms connected with *@&!:*
[2)] one or chain of respective atoms connected with [§(Xc)]
[3)] one or chain of respective atoms connected with [§(#6R1)]
[4)] any atom in rings fulfilling the Hückel rule considered to be aromatic. This includes 6-member rings with O= or S= directly attached to a ring member
[5)] atoms connected via ring closure
[6)] do not consider atoms marked before
[7)] mark this bond
[ü)] both atoms may be identical (i.e. the same real atom) in a ring structure
[8)] additionally in contrast to the original publication

Table A2. ChemProp [1] implementation of the model for B^H, B^O,E, S and L of Platts et al. [3]. Substructures are coded in SMARTS [4]. The fragment numbers refer to the original paper [3], from where the respective increment values (Table A4) have been taken.

1. Fragments.

Atoms have strictly to be looked for in the order listed here (from top to bottom). Each atom of the molecule has to occur in only one fragment (corresponding to the first hit), and each atom has to be assigned to a fragment, otherwise the model result would not be valid. For the explanation of the symbols $ and § please see Table A1. In case of deviations from the originally published model, the respective numbers are marked by *.

#1	[CH3]
#2	[CH2X4]
#3	[CH1X4]
#4	[CH0X4]
#5	[CH2X3]
#6	[#6H1X3]
#7	[#6H0X3]-[$(*)]
#8	[§(a)]:c(:[§(a)]):[§(a)]
#21*	N#[§(C)]-[§(A)]
#22*	N#[§(C)]-[§(a)]
#9	C#[$(*)]
#10	[NH2X3]-[$(A)]
#11	[NH2X3]-[$(a)]
#12	[§(A)]-[NH1X3]-[§(A)]
#13	[§(a)]-[NH1X3]-[§(*)]
#14	[nH1]
#15	[§(A)]-[NH0X3](-[§(A)])-[§(A)]
#16	[§(a)]-[NH0X3](-[§(*)])-[§(*)]
#17	[nH0X3]
#18	[NX2R0]
#19	[NX2R1]
#20	[nX2]
#23	[$(A)]-[NX3](=O)~[OH0X1]
#24	[$(a)]-[NX3](=O)~[OH0X1]
#25*	[$(OH0X2)]-[NX3](=O)~[OH0X1]
#26	[OH1]
#27	[OH0X2R0]
#28	[OH0X2R1]
#29	o
#30	[OH0X1]
#31	[SH0X2]
#32	s
#33	[SH0X1]
#34	[SX3]
#35*	[$(OH0X2)]-S(=O)=O
#36	S
#37	P
#38	F-[$(A)]

#39	F-[$(a)]
#40	Cl
#41	Br
#42	I

---

2. Corrections.

To be applied independently from the fragments, and independently from each other. Each atom of the molecule may, but does not need to occur in any correction. Again, deviations from the published model are marked by * at the respective numbers.

#43	[OH0R0X2]-C=O
#44	[OH0R1X2]-C=O
#45	[OX1,SX1]=P-[OX2,SX2]
#46	[OH0X2]-C(=O)-[OH0X2]
#47	[OH1]-C=O
#48	[$(a)]-C(=O)-[NH1X3,NH0X3]
#49*	[[§(A)]-[CR0](=O)-[NR0X3], [§(a)]-C(=O)-[NH2]]
#50	[NR1X3]-[CR1]=O
#51	O=S(=O)-[NX3]
#52	[NX3]-C(=O)-[NX3]
#53	[NX3]-C(=O)-[OH0X2]
#54	O=CNC=O
#55	O=C~#6=#6~C=O
#56	[§(F,Cl,Br,I,[NX3](=O)~[OH0X1],C#N,FC(F)F)]-[CX3]-[§(F,Cl,Br,I,[NX3](=O)~[OH0X1]
#57	[§(F,Cl,Br,I,[NX3](=O)~[OH0X1],C#N,FC(F)F)]-C-C-[§(F,Cl,Br,I,[NX3](=O)~[OH0X1]
#59-#67	H-bond 1-9 as explained below
#68	[OH1]-[CX3]-[CX3]-[OH1]
#69	n:n
#70	[n,o,s]:[o,s]
#71	n:c:n
#72	[n,o,s]:c:[o,s]
#73	[n,o,s]:c:c:[n,o,s]
#74*	[!C]-c:c-[!C] except [OH1]-c:c-[Cl,Br,I] (cf. text)
#75	[!C]-c:c:c-[!C]
#76	[!C]-c:c:c:c-[!C]
#77	[O,S]=P-[NX3]
#78	N-c:n
#79	[OH1]-C-c (except [OH1]-C(=O)-c)
#80	O=NX2
#81	[OH0X2]-c:c-[OH0X2]

---

3. Indicator.

To be applied once, if appropriate. The implementation is the best guess with regard to the meaning “steroid” as mentioned in the original publication.

#58 [CR2,CR3]

4. H-bond corrections.

Independent from fragments and other corrections, but ***not*** independent from each other. Atoms have strictly to be looked for in the order listed here (from top to bottom). Each donor atom of the molecule may, but does not need to occur in only one H-bond correction.

H-bond 1	[OH1]-C(-[$(*)])-C-[O,N], [OH1]-[CX4]-C(-[$(*)])-[O,N], [OH1]-C(=[C,N])-C(-[$(*)])-[O,N]
H-bond 2	[OH1]-C(-[$(*)])-[C,N]=[O,S], [OH1]-C-[C,N](-[$(*)])=[O,S]
H-bond 5	[OH1]-c:c-[NX3](=O)~[OH0X1]
H-bond 3	[OH1]-c:c-[O,N]
H-bond 4	[OH1]-c:c-C=[O,S]
H-bond 6	[NH1,NH2]-C(-[$(*)])-C-[O,N], [NH1,NH2]-C-C(-[$(*)])-[O,N]
H-bond 7	[NH1,NH2]-c:c-[O,N]
H-bond 8	[NH1,NH2]-c:c-[C,N]=[O,S]
H-bond 9	[OH1]-c:c-[Cl,Br,I]
H-bond 10	[OH1]-[CX4]-C(-[$(*)])-C=O

5. Aromaticity.

Rings fulfilling the Hückel rule are considered to be aromatic. This includes respective 5-member rings with, e.g., NH, S or O. However, 6-member rings with O= or S= directly attached to a ring member are excluded.

Table A3. ChemProp [1] implementation of the model for *A* of Platts et al. [3]. Substructures are coded in SMARTS [4]. The fragment numbers refer to the original paper [3], from where the respective increment values (Table A4) have been taken.

1. Corrections.

To be applied independently from the fragments, and independently from each other. Each atom of the molecule may, but does not need to occur in any correction. #47 differs from the published model. For the explanation of the symbols $ and § please see Table A1.

#1	[OH1]-[$(A)]
#2	[OH1]-[$(c)]
#3	[NH2]-[$(A)]
#4	[NH2]-[$(c)]
#5	[§(A)]-[NH1R0]-[§(A)]
#6	[§(A)]-[NH1R1]-[§(A)]
#7	[NH1]-[$(c)]
#8	[nH1]
#9	[OH1]-C=O
#10	[NH2]-C=O
#11	[NH1]-C(=O)-[$(A)]
#12	[NH1]-C(=O)-[$(a)]
#13	O=S(=O)-[NH1,NH2]
#14	[NH1]-C(=O)-[NH1]
#15	[NH1]-C(=O)-[NH0X3]
#16	[NH1]-C(=O)-[OJ0X2]
#17	[NH1]-C(=N)-[NH0X3]
#18	[CH1]#*
#19	P~[OH1,SH1]
#21	[§(F,Cl,Br,I,[NX3](=O)~[OH0X1],C#N,FC(F)F)]~CH1~[§(F,Cl,Br,I,[NX3](=O)~[OH0X1]
#20	CH1~[$(F,Cl,Br,I,[NX3](=O)~[OH0X1],C#N,FC(F)F)] but not when #21applies
#22	[OH1]C(=O)-CX4-CX4-C(=O)[OH1]
#23	[OH1]C(=O)-CX4-[$(F,Cl,Br,I,[NX3](=O)~[OH0X1],C#N,FC(F)F)]
#24	[OH1]-CX4-[$(F,Cl,Br,I,[NX3](=O)~[OH0X1],C#N,FC(F)F)]
#25	[OH1]-CX4-CX4-[$(F,Cl,Br,I,[NX3](=O)~[OH0X1],C#N,FC(F)F)]
#26	[nH1]:[!c]
#27	[nH1]:c:[!c]
#28...#37	H-bond corrections as explained in Table A2. Please note, these corrections must not be applied independently from each other!
#38	[nX2]:c(:[$a]):c-[OH1]
#39	[OH1]-c:c:c-[$(F,Cl,Br,I,[NX3](=O)~[OH0X1],C#N,FC(F)F)]
#40	[OH1]-c:c:c:c-[$(F,Cl,Br,I,[NX3](=O)~[OH0X1],C#N,FC(F)F)]
#41	[NH2]-c:c:c-[$(F,Cl,Br,I,[NX3](=O)~[OH0X1],C#N,FC(F)F)]
#42	[NH2]-c:c:c:c-[$(F,Cl,Br,I,[NX3](=O)~[OH0X1],C#N,FC(F)F)]
#43	[OH1]-C(=O)-c:c:c-[$(F,Cl,Br,I,[NX3](=O)~[OH0X1],C#N,FC(F)F)]
#44	[OH1]-C(=O)-c:c:c:c-[$(F,Cl,Br,I,[NX3](=O)~[OH0X1],C#N,FC(F)F)]
#45	CX4-c:c(-[OH1]):c-CX4
#46	CX4-c:c(-[NH2]):c-CX4
#47	[OH]-c:c-C-[$(F,Cl,Br,I,[NX3](=O)~[OH0X1],C#N,FC(F)F)] or [OH]-c:c-C#N

#48	[OH]-c:c:c-C(=O)-[OH1]
#49	[OH]-c:c:c:c-C(=O)-[OH1]
#50	[OH]-c:c:c-C(=O)-[H,C]
#51	[OH]-c:c:c:c-C(=O)-[H,C]

2. Aromaticity.

See Table A2.

Table A4. Increment values for the implementation of the model for B^H, B^O, E, S and L (first part of the table) and A (second part) of Platts et al. [3], taken from the original paper.

1. Model for B^H, B^O, E, S and L

Fragment #	E	S	B^H	B^O	L
1	-0.104	-0.075	0.007	0.000	0.321
2	0.000	0.000	0.000	0.000	0.499
3	0.089	0.036	0.011	0.020	0.449
4	0.187	0.071	0.037	0.047	0.443
5	-0.045	-0.085	0.019	0.024	0.244
6	0.068	0.050	0.011	0.012	0.469
7	0.180	0.101	0.000	0.000	0.624
8	0.300	0.121	0.019	0.018	0.744
9	0.040	0.034	0.028	0.032	0.332
10	0.085	0.175	0.481	0.486	0.781
11	0.163	0.383	0.275	0.326	0.949
12	0.138	0.265	0.541	0.543	0.568
13	0.192	0.311	0.415	0.426	0.912
14	-0.030	0.221	0.316	0.267	1.250
15	0.220	0.323	0.653	0.655	0.400
16	0.346	0.295	0.321	0.338	0.869
17	0.083	0.265	0.392	0.338	0.794
18	0.117	0.125	0.200	0.202	-0.235
19	0.121	0.254	0.596	0.589	-0.240
20	0.046	0.223	0.321	0.300	0.574
21	0.000	0.694	0.242	0.245	0.757
22	0.000	0.390	0.103	0.093	0.732
23	0.200	0.000	-0.476	-0.595	0.278
24	0.210	-0.231	-0.525	-0.533	0.347
25	0.000	-0.476	-0.204	-0.202	0.000
26	0.061	0.247	0.307	0.311	0.672
27	0.014	0.185	0.211	0.226	0.360
28	0.013	0.185	0.331	0.330	0.359
29	-0.125	0.000	0.047	0.060	0.057
30	-0.041	0.370	0.334	0.339	0.495
31	0.330	0.189	0.168	0.175	1.258
32	0.116	0.000	0.043	0.083	0.848
33	0.364	0.618	0.071	0.069	0.954
34	0.413	1.065	0.448	0.319	2.196
35	0.000	-0.505	-0.188	-0.190	0.000
36	0.465	0.643	0.000	0.000	0.554
37	0.295	0.703	1.183	1.189	2.051
38	-0.180	-0.042	-0.036	-0.033	-0.143
39	-0.230	0.000	0.000	0.000	-0.147
40	0.023	0.082	0.000	0.000	0.669
41	0.196	0.161	-0.011	0.000	1.097
42	0.533	0.198	0.000	0.000	1.590
43	-0.113	-0.225	-0.206	-0.223	-0.390
44	0.000	0.360	-0.214	-0.169	0.406
45	-0.100	-0.240	-0.394	-0.408	-0.483
46	0.000	-0.190	-0.267	-0.298	0.000
47	-0.192	-0.412	-0.308	-0.312	-0.369
48	0.221	-0.076	-0.095	-0.038	0.000
49	0.000	0.175	-0.287	-0.292	0.603
50	0.061	-0.100	-0.231	-0.242	0.583
51	-0.111	-0.569	-0.446	-0.443	0.000

52	-0.110	-0.553	-0.076	-0.054	0.000
53	0.000	-0.588	-0.252	-0.251	0.000
54	0.000	-0.510	-0.148	-0.149	0.000
55	0.000	-0.411	-0.051	-0.050	0.000
56	-0.017	-0.050	-0.014	-0.016	-0.111
57	0.012	0.000	0.013	0.010	0.054
58	0.285	1.029	0.267	0.218	0.488
59	0.029	-0.067	0.000	0.000	-0.072
60	0.000	-0.095	-0.068	-0.090	-0.337
61	-0.069	-0.237	-0.079	-0.122	0.000
62	0.000	-0.344	-0.387	-0.403	-0.303
63	0.000	-0.276	-0.126	-0.120	-0.364
64	0.000	-0.102	0.000	0.000	0.062
65	0.000	0.000	-0.059	-0.027	0.000
66	0.000	-0.140	-0.045	-0.069	0.169
67	-0.100	-0.120	-0.130	-0.130	-0.400
68	-0.043	0.052	0.000	-0.018	0.100
69	0.092	0.024	-0.132	-0.094	-0.179
70	-0.113	0.047	-0.157	-0.141	0.000
71	0.000	-0.040	-0.098	-0.113	0.042
72	0.052	0.087	-0.170	-0.184	0.209
73	0.000	-0.051	-0.089	-0.073	-0.058
74	0.000	-0.043	0.031	0.025	-0.081
75	0.000	-0.038	-0.035	-0.033	-0.026
76	0.000	0.000	-0.023	-0.025	0.000
77	-0.080	-0.452	-0.668	-0.668	0.000
78	0.185	0.098	-0.042	-0.057	0.149
79	0.000	0.000	0.131	0.129	-0.145
80	0.000	0.434	-0.408	-0.405	0.000
81	0.000	0.380	-0.216	-0.218	0.000
intercept	0.248	0.277	0.071	0.064	0.130

2. Model for *A*

Fragment #	***A***
1	0.345
2	0.543
3	0.177
4	0.247
5	0.087
6	0.321
7	0.194
8	0.371
9	0.243
10	0.275
11	0.281
12	-0.091
13	0.356
14	-0.165
15	-0.119
16	-0.105
17	0.170
18	0.082
19	0.493
20	0.019
21	0.050
22	-0.362
23	0.118
24	0.100

25	0.051
26	0.194
27	0.042
28	-0.089
29	-0.161
30	-0.251
31	-0.418
32	-0.450
33	-0.155
34	0.0
35	-0.093
36	-0.110
37	-0.601
38	-0.475
39	0.119
40	0.176
41	0.080
42	0.084
43	0.085
44	0.055
45	-0.162
46	-0.181
47	0.195
48	-0.203
49	0.096
50	0.185
51	0.203
intercept	0.003

References:

[1] G. Schüürmann, R. Kühne, F. Kleint, R.-U. Ebert, C. Rothenbacher, P. Herth, in '*Quantitative Structure-Activity Relationships in Environmental Sciences - VII*', Eds. F. Chen, G. Schüürmann, SETAC Press, Pensacola, **1997**, pp. 93-114.

[2] S. Tao, H. Piao, R. Dawson, X. Lu, H. Hu, *Environ. Sci. Technol.* **1999**, *33*, 2719-2725.

[3] J.A. Platts, D. Butina, M.H. Abraham, A. Hersey, *J. Chem. Inform. Comput. Sci.* **1999**, *39*, 835-845.

[4] Daylight Chemical Information Systems,Inc. http://www.daylight.com/dayhtml_tutorials/languages/smarts.