QSPR models for Bioconcentration Factor (BCF): are they able to predict data of industrial interest?

Filippo Lunghini1,2, Gilles Marcou1,\*, Philippe Azam2, Remi Patoux2, Marie-Hélène Enrici2, Fanny Bonachera1, Dragos Horvath1, Alexandre Varnek1\*.

1Laboratory of Chemoinformatics, University of Strasbourg, 4 Rue Blaise Pascal, 67081, Strasbourg, France

2Solvay S.A., 20 Rue Marcel Etienne Sembat, 69190, St. Fons, France

\* Corresponding author. email: g.marcou@unistra.fr; [Varnek@unistra.fr](mailto:Varnek@unistra.fr); phone no.: +33-68851304

**SUPPLUMENTARY INFORMATION**

## Section 1 - Collection and curation of BCF data

Databases were pairwise compared for shared compounds, which highlighted a significant overlapping. The correlation coefficients between the experimental BCF values of shared compounds ranged from 0.77 to 0.92, with ECHA having the lowest correlation compared to the other databases, while CEFIC LRI, NITE and most of the data coming from literature reported most of the time the same value. On the other hand, this was not the case of Canadian DSL list and the ECOTOX EPA, with correlation coefficients in the range of 0.2 – 0.3. Moreover, when multiple BCF measures were available for the same molecule, the spread of experimental data was surprisingly high (up to 5 log units). The inspection of several chemicals back to their original publications, revealed that this was caused by wrongly reported information: BCF results were sometimes reported without the exponential notation (omitting thus orders of magnitude), the standard deviation was taken instead of the BCF value or the wrong article was referenced. Some examples are: (i) ECOTOX database, the substance *1,4-Dibromobenzene* (CAS 106-37-6; 2.5 log unit difference) has BCF values reported without exponential expression [1]; (ii) Canadian DSL List, the substance *Fenthion-sulfoxide* (CAS 3761-41-9; 5 log unit difference) has results coming from the wrong articles, which were actually referring to another chemical [34]. Since it was not possible to systematically check and clean these databases, it was decided to exclude them completely, due to the associated high uncertainty. The rest of the databases were merged, ending up with roughly 4500 raw values. Concerning the ECHA database, only results from experiments associated with a Klimisch score [3] of 1 and 2 were considered.

Depending on the type of chemical, the number of raw data values was quite unbalanced, ranging from several tens for “classical contaminants” (e.g. some pesticides, PCBs, PBDE, surfactants) to very few for the “emerging contaminants”, such as per-fluorinated compounds. To provide an example, for the pesticide *Chlorpyrifos* (CAS 2921-88-2) and the PFC *1,1,1,2,2,3,3-Heptafluoro-3-methoxypropane* (CAS 375-03-1), more than 80 BCF values were found for the former while only one BCF value was found for the latter.

### 1.1 – Data curation

The following standardization rules have been applied: removal of salts/solvents, neutralization, removal of explicit hydrogens, aromatic representation for benzene rings, removal of stereo information, transformation of -nitro and -sulfo containing groups. SMILES were checked (or retrieved if not originally available) from the Distributed Structure-Searchable Toxicity (DSSTox) Database [4]. Duplicates were removed: when multiple values were available the median was calculated together with the Median Absolute Deviation (MAD). The MAD was used to select chemicals with highly different experimental values (a threshold of MAD > 0.5 was chosen) and to manually check their reliability/correctness. De-duplication was performed through canonical SMILES matching after the above-mentioned standardization procedure. All these tasks were done with the use of Knime platform [5] and InstantJChem [6].

### 1.2 - Statistics of the curated datasets

Histogram representation of the property distribution (in frequencies) for logBCF is shown in Figure S1. The training set and the Industrial set have the same pattern, with a peak around the logBCF value around 0; the external set on the other hand is quite different, with most of the compounds in the range of logBCF values around 3.

***Figure S1.*** *BCF frequency property distribution for the collected datasets.*

Tanimoto pairwise comparison of the datasets: Industrial/Industrial; Training/Training; Training/Industrial. The comparison has been made using the same DS selected for GTM.

Average similarities:

Industrial/Industrial: 0.402

Training/Training: 0.449

Training/Industrial: 0.411

***Figure S2.*** *Datasets' Tanimoto similarity distributions comparison*

### 1.3 – Model Generation and Validation

Table S1 lists all the employed descriptor spaces. The nomenclature of ISIDA descriptors is here briefly described [7].

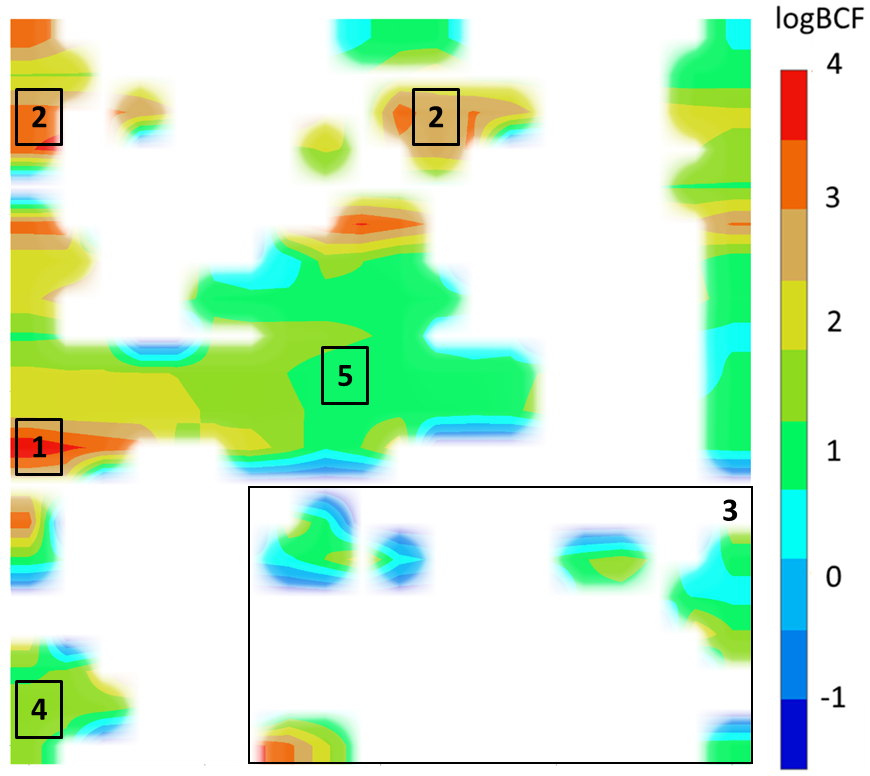
* T = type of fragmentation. Where 1 = sequences of atom, 3 = sequences of atoms and bonds, 4 = atom centred fragments based on sequences of atoms, 6 = atom centred fragments based on sequences of atoms and bonds, 7 = atom centred fragments based on sequences of atoms (fixed length), 9 = atom centred fragments based on sequences of atoms and bonds (fixed length), 10 = triplets
* L and U = minimum and maximum length
* FF = Force Field coloration
* AP = Atom pairs (all constitutional details of a sequence are removed and only the number of constitutive atoms is given)

|  |  |
| --- | --- |
| **# of fragments** | **Descriptor space name** |
| 1475 | t10l2u5 |
| 618 | t1l2u5 |
| 2654 | t3l2u6 |
| 4202 | t3l2u7 |
| 5917 | t3l2u8 |
| 972 | t4l2u2\_FF\_AP |
| 1287 | t4l2u3 |
| 1205 | t4l2u3\_AP |
| 3368 | t4l2u3\_FF\_AP |
| 3093 | t4l3u3\_FF |
| 300 | t6l2u2\_AP |
| 1858 | t6l2u3 |
| 1830 | t6l2u3\_AP |
| 3464 | t6l2u3\_FF |
| 3450 | t6l2u3\_FF\_AP |
| 1065 | t6l3u3\_AP |
| 3157 | t6l3u3\_FF |
| 721 | t7l2u3\_AP |
| 4016 | t7l2u3\_FF |
| 534 | t7l3u3\_AP |
| 3093 | t7l3u3\_FF |
| 300 | t9l2u2 |
| 1651 | t9l2u3 |
| 1346 | t9l2u3\_AP |
| 1370 | t9l3u3 |
| 3157 | t9l3u3\_FF |

***Table S1*** *listing all the employed descriptor spaces with the respective number of fragments.*

## Section 2 - Generative Topographic Mapping: training set visualization

Figure S3 shows the GTM map of the training set. Numbered boxes have been assigned to mark most relevant areas, which are discussed in Table S1 below.



**Figure S3.** GTM map for the training set. Regions of the map are colored according to the property of the molecules contained in the nodes, i.e. logBCF. White areas are empty regions of the map, interpretable as an empty chemical space. Numbered squares mark the most relevant areas of the map: discussion about contained chemicals is reported in Table S1.

|  |  |  |  |
| --- | --- | --- | --- |
| **Region** | **Brief description** | **Most representative members** | |
| **1** | Ca. 80 chlorinated compounds, with one or more benzene rings (e.g. TCDD and PCB families). | 938-22-7 | 32774-16-6 |
| **2** | Ca. 30 perfluorinated compounds are spread along the two #2 regions. On the left, mainly hydrocarbons; on the right carboxylic functional groups and multiple benzene rings are present.  In addition, this area comprises tens of biphenyl-like compounds eventually with halogens as substituents. | 306-98-9 | 2-4099 (MITI no.) |
| **3** | More than 100 compounds are present in this area. The dominant chemotype is the presence of sulfo- group. Molecules are both aromatic and aliphatic.  The only red region of this area (bottom left part) is populated by halogen-containing molecules, like the pesticide Leptophos (2nd image). | 26444-49-5 | 21609-90-5 |
| **4** | Ca. 50 aromatic nitro-containing compounds. Higher BCF segments (e.g. yellow) contain molecules with the same scaffold but substituted with halogen(s). | 99-08-1 | 117-18-0 |
| **5** | Wide and heterogeneous area counting several tens of chemicals which share very low BCF values. Two main families can be identified: (i) aliphatic compounds substituted with multi hydroxylic groups; (ii) higher MW compounds with different functional groups. | 126-58-9 | 66230-04-4 |

**Table S2.** Related to the GTM map of the training set (Figure S1), this table reports the analysis of the marked areas, with a brief description of the chemotypes contained and some examples of molecules.

Compounds (CAS 163702-06-5, 163702-05-4 respectively) have a multimodal responsibility pattern, partially residing into several disparate nodes which are populated by analogous training set compounds. Their (X,Y) position on the map marks the barycenter of their responsibility pattern (SI, section 2). and falls into a zone where there is no actual density accumulation – even though the nodes in which they partially reside benefit from significant populations of training set compounds. This is a well-known artefact arising when compounds with complex responsibility patterns are represented by their (X,Y) coordinates, ignoring the specifics of their multiple partial residences in sometimes disjoined nodes. The apparent discrepancy can arises when a landscape (a property-colored cumulated responsibility pattern) is overlapped to compounds projections.

## Section 3 - Examples of ColorAtom application

### 3.1 – Outliers coloration

The ColorAtom was applied (i) to compare the coloration of the excluded outliers with those of structurally analogues compounds. Top part of Table 7 reports one example of colored outlier (CAS 2528-38-3; with an absolute error of 1.0 logBCF) and its similar compounds. Molecules shown the same coloration pattern, with the phosphate group and the aliphatic residue being correlated to a decrease and increase of the BCF, respectively. Same coloration scheme means that the molecule was predicted using the same learned rules. Additional examples are provided in SI, section 3.

|  |  |  |
| --- | --- | --- |
| **Outliers coloration** | | |
| *Excluded outlier* |  |  |
| 14233-37-5 | 3.85 | 1.5 | 12217-77-5 | 1.2 | 0.2 | 41611-76-1 | 0.54 | 1.5 |
| *Excluded outlier* |  |  |
| 1460-02-2 | 4.14 | 1.32 | 4130-42-1 | 3.51 | 1.31 | 128-37-0 | 3.16 | 0.86 |
|  |  |  |
| 96-76-4 | 2.45 | 0.3 | 120-95-6 | 2.63 | 0.34 | 535-77-3 | 2.73 | 0.61 |
|  |  |  |
| 19715-19-6 | 0.98 | 0.23 | 88-26-6 | 1.52 | 0.21 |  |
| **Key chemotypes** | | |
|  | | |

**Table S3.** Examples of excluded outlier with its most structurally similar compounds (based on Tanimoto score) with the respective experimental and predicted BCF.

### 3.2 – Key chemotypes correlated to BCF

An example of analysis performed through the ColorAtom tool is reported. The first molecule, i.e. “Target molecule”, was considered for the analysis. Three main structural features can be identified, as marked by the black numbered ellipses. Of them, no. 1 and 2 are mainly associated to a reduced BCF value; while no. 3 shows the opposite trend. On such basis, a similarity search has been performed in the training set and structurally related compounds were then analysed through ColorAtom. It can be noticed the presence of a coloration pattern. In particular, when the chemotypes no. 1 and 2 dominate the compound, its associated BCF value is relatively low (from molecule no. 1 to 3, logBCF values ranges from 0.2 to 0.68); while, on the other hand, when the chemotype no. 3 is mainly present logBCF values are much higher (from molecule no. 7 to 10 ranges from 2.0 to 3.44).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **No.** | **Mainly recurring chemotypes** | **Molecule** | **CAS** | **logBCF** |
| Target molecule | 1, 2, 3 |  | 732-11-6 | 0.24 |
| 1 | 1 |  | 55-38-9 | 0.32 |
| 2 | 1 |  | 60-51-5 | 0.2 |
| 3 | 1 |  | 2497-06-5 | 0.68 |
| 4 | 1, 3 |  | 13286-32-3 | 1.24 |
| 5 | 2, 3 |  | 85-41-6 | 1.14 |
| 6 | 2, 3 |  | 26234-46-8 | 1.08 |
| 7 | 3 |  | 84-65-1 | 2.00 |
| 8 | 3 |  | 129-43-1 | 2.32 |
| 9 | 3 |  | 829-26-5 | 3.00 |
| 10 | 3 |  | 40766-31-2 | 3.44 |

**Table S4.** Examples of ColorAtom representations for several molecules. The molecule CAS 732-11-6 has been taken as target example (Discussion part of the manuscript, Figure 5). Its analyzed substructure features are marked by numbered ellipses. These chemotypes were used to mined similar compounds in the training set.

## Section 4 – AD definitions of the employed tools

During benchmarking, compounds inside the model’s training set and predictions outside the AD of the tool were not considered. The AD definition varies depending on the tool.

* For VEGA, the Applicability Domain Index (ADI), an overall assessment that considers different factors, was used. Only compounds with an ADI > 0.85 (i.e. high reliability) were included.
* TEST does not provide a detailed AD assessment. According to the model’s manual, if the consensus prediction is given, the molecule is considered to be inside the AD.
* EPISuite does not provide AD evaluation but, according to the manual documentation, users must manually check if the MW, logP and functional groups are within the ranges of their respective counterparts in the training set. In effect, this is equivalent to perform an AD evaluation with the bounding-box methodology. LogP values were calculated with “RDkit descriptors calculation” node [46] in KNIME. Functional group criteria were not considered, being practically unfeasible when dealing with several compounds.
* OPERA defines the AD based on structure similarity and the leverage approach. The output is automatically given to the user.

[1] Y. Chaisuksant, Q. Yu, and D. W. Connell, “Bioconcentration of bromo- and chlorobenzenes by fish (Gambusia affinis),” *Water Res.*, vol. 31, no. 1, pp. 61–68, 1997.

[2] Karin Fox, Z. Gerd-Peter, and W. Butte, “Kinetics of Bioconcentration and Clearance of 28 Polychlorinated Biphenyl Congeners in Zebrafish (Brachydanio renio).” .

[3] H.-J. Klimisch, M. Andreae, and U. Tillmann, “A Systematic Approach for Evaluating the Quality of Experimental Toxicological and Ecotoxicological Data,” *Regul. Toxicol. Pharmacol.*, vol. 25, no. 1, pp. 1–5, 1997.

[4] US EPA, “Distributed Structure-Searchable Toxicity (DSSTox) Database.” [Online]. Available: https://www.epa.gov/chemical-research/distributed-structure-searchable-toxicity-dsstox-database.

[5] “KNIME: Konstanz Information Miner.” [Online]. Available: https://www.knime.com/.

[6] “Chemaxon.” [Online]. Available: https://chemaxon.com/.

[7] F. Ruggiu, G. Marcou, A. Varnek, and D. Horvath, “ISIDA Property-labelled fragment descriptors,” *Mol. Inform.*, vol. 29, no. 12, pp. 855–868, 2010.