CONCERTREACH
CONCERTING EXPERIMENTAL DATA
AND IN SILICO MODELS FOR REACH

LIFE17 GIE/IT/000461

In silico models: theory, guidance and applications within VEGAHUB

CONCERTREACH
CONCERTING EXPERIMENTAL DATA
AND IN SILICO MODELS FOR REACH

LIFE17 GIE/IT/000461

E-book

Editor
Emilio Benfenati
Istituto di Ricerche Farmacologiche “Mario Negri”
Via Mario Negri 2, 20156, Milano (Italy)

First edition 2023

Foreword

This eBook updates the previous eBook Theory, guidance and applications on QSAR and REACH, published in 2012.

R. L. Stevenson, in his masterpiece Strange Case of Dr Jekyll and Mr Hyde, imagined that it was possible to separate the good and evil in man. A chemical potion could reveal the evil side of humankind. At the basis of this work of fantasy there is an ethical issue, merged with the scientific fiction of the revelatory technical device.

Modern science is currently working towards the more modest goal of identifying the good and bad nature of chemical substances, with the aim of eventually pinpointing the components which make the chemical toxic. This can be done through so-called QSAR models.

The basic hypothesis of an *in silico* model is that a given property or effect can be put into relationship with the structure of a chemical, which is described by reference to certain parameters. To achieve this, we need a good mathematical algorithm as well as suitable ways to describe the chemical.

With this eBook you will learn about the state of the *in silico* models. This book composed of two parts: The first will address the scientific aspect of modelling. Then we will examine some practical cases, discussing the possible applications of *in silico* models.

We will refer to the debate on the suitable and appropriate use of these models for specific applications, and thus we will also deal with regulatory and cultural matters. The acceptance and broad application of *in silico* models are not solely related to the statistical power of a model. This eBook aims to give you tools to evaluate when QSAR models can help further the goal of protecting health and the environment.

The EC project

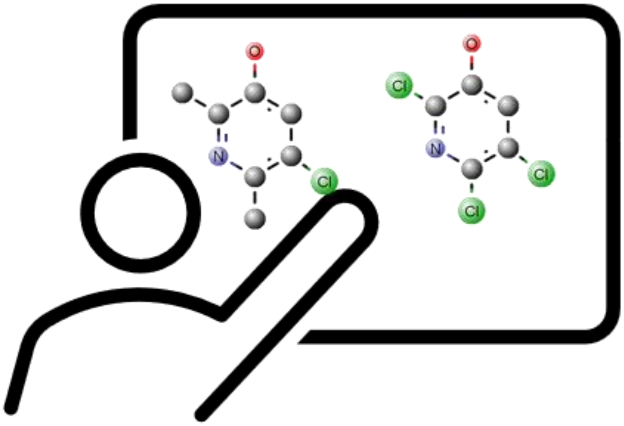
We acknowledge the EC project CONCERT REACH (https://www.life-concertreach.eu/). The LIFE funded CONCERT REACH project (LIFE17 GIE/IT/000461. 2018-2022.) promoted a wider understanding, awareness and appropriate use of *in silico* methods, establishing a network of *in silico* platforms available in Europe: VEGAHUB, the Danish (Q)SAR Database, OCHEM and AMBIT. A gateway (https://www.life-concertreach.eu/results/results-gateway/) to find hundreds of models has been created.

The gateway is a clearinghouse of knowledge and experience for professionals who are developing and using *in silico* models. For industry users, regulators and academics, it provides downloadable software for the *in silico* models.

CONCERTREACH
CONCERTING EXPERIMENTAL DATA
AND IN SILICO MODELS FOR REACH

LIFE17 GIE/IT/000461

Table of content



Emilio Benfentati[a], Anna Lombardo[a], Giuseppina Gini[b]

[a]Istituto di Ricerche Farmacologiche Mario Negri IRCCS, Milano Italy

[b]Politecnico di Milano, Milano Italy

Part A

Theory

Introduction to in vivo, in vitro and in silico methods

In our everyday lives we have to deal with an exponentially increasing number of different chemical compounds. More than 200 million substances have been registered so far, including food colouring and preservatives, drugs, varnishes, paints, pesticides, and many others. It is well recognised that chemicals may pose a high risk to the environment and to humans, and therefore their toxic activity has to be assessed.

Biological active substances interact with biomolecules, triggering specific mechanisms, like the activation of an enzyme cascade or the opening of an ion channel, which finally lead to a biological response. These mechanisms, determined by the chemical composition of the relevant substances, are unfortunately largely unknown; thus, toxicity must be studied experimentally.

It is possible to use three types of approaches to assess the biological activity of a molecule (*Figure 1*): *in vivo* experiments, i.e. animal testing; *in vitro* experiments, which involves tissue culture cells; and *in silico* simulations, which refers to computer-based predictions.

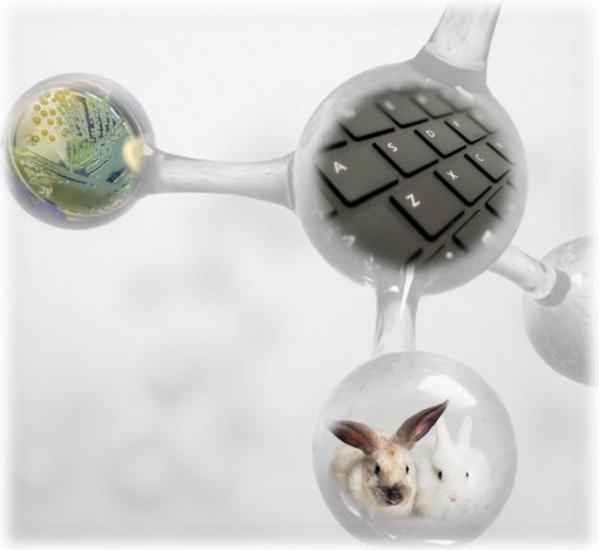


Figure 1: The three experimental ways to evaluate chemical substances: in vivo, in vitro and in silico.

Both animal testing and *in vitro* experiments are time consuming and expensive. Additionally, animal testing is now considered ethically unacceptable by a growing majority of people and more and more regulations are moving towards replacement. In Europe, *in vivo* studies cannot be done for cosmetics. For these reasons, and also thanks to improvements in computational power, the scientific community and the industrial world have started to use *in silico* approaches, or at least view them as a possible viable alternative; they have developed thousands of models and strategies able to predict the properties of the compounds.

Computational chemistry has changed the classical way to engage in experimental science. We are increasingly moving from experiments to simulations. We can model molecules according to different views, from the basic valence model to graph representation, from electronic clouds to 3D structure. Algorithms are available to compute the so-called “molecular descriptors”, which are variables (either continuous or discrete) representing structural properties. These descriptors range from simple properties to complex molecular fingerprints. Descriptors can help in transforming the study of interactions between molecules and living organisms in a data mining problem. Data mining could reveal relevant correlations between the descriptors and the response of interest.

The algorithms look for correlations between the properties of the chemical structure of a compound and a measure of its activity/toxicity in a specific area, such as mutagenicity, carcinogenicity, or skin sensitization. This is called the "endpoint of interest".

In other words, once the structure of a compound is quantified in a set of molecular descriptors, these algorithms may be able to establish a mathematical relationship between the compound and, for example, its toxicity. To obtain the most reliable relation possible, a sufficiently large dataset of compounds with known structure and experimentally determined property of interest is necessary to "train" the model.

The underlying idea of these models is that chemicals with similar structures, i.e. with similar values for the considered descriptors, must behave in a similar way. Thus, once the model is built, it can be used as a predictive tool in drug design, environmental protection and hazard analysis for all those compounds whose structure is similar to the structure of the ones used to tune the model.

The use of these models is growing, since they aim to provide fast, reliable and quite accurate estimates of the chemicals' activity. These features also make them suitable for legislative purposes, and that is why they have been included as an alternative tool for risk assessment in the European legislation on chemical production, called REACH (Registration, Evaluation, Authorisation and Restriction of Chemicals). This legislation sets the rules for chemical production in the E.U., and one of its key points is the requirement of a risk analysis for each chemical placed in the European market in an amount greater than 1 tonne/year. To further underline the breadth of this law, suffice it to say that the document is 849 pages long, and international mass media defined REACH as "the most important legislation in European Union in 20 years" (BBC News, 28 November 2005) and "the strictest law to date regulating chemical substances" (San Francisco Chronicle, 14 December 2006). REACH prompted regulations in other parts of the world, which require sufficient data to evaluate the impact of a substance towards the environment and humans.

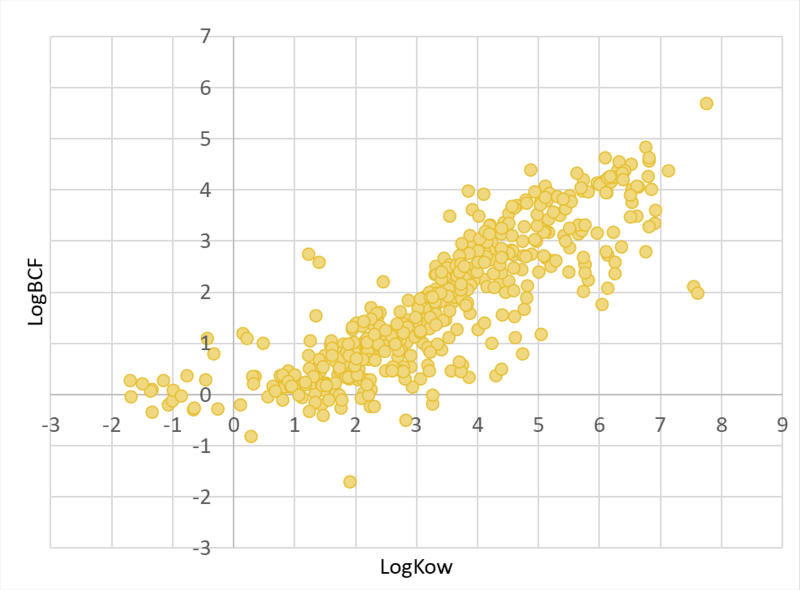


Figure 2: An example of QSAR analysis. The descriptor is LogKow; the endpoint is the bioconcentration factor (BCF) in log units.

In silico for predicting properties: the QSAR approaches

Quantitative structure-activity relationship (QSAR) models are models linking a property or effect, such as boiling point or toxicity, to parameters associated with chemical structure, such as certain molecular descriptors. They can be used to assess chemical substances within the so-called *in silico* approach (as an analogy to the *in vitro* and *in vivo* approach). Thus, the three main components of the QSAR models are (*Figure 3*):

1. The property to be modelled;
2. The chemical information;
3. The algorithm linking the property and the chemical.

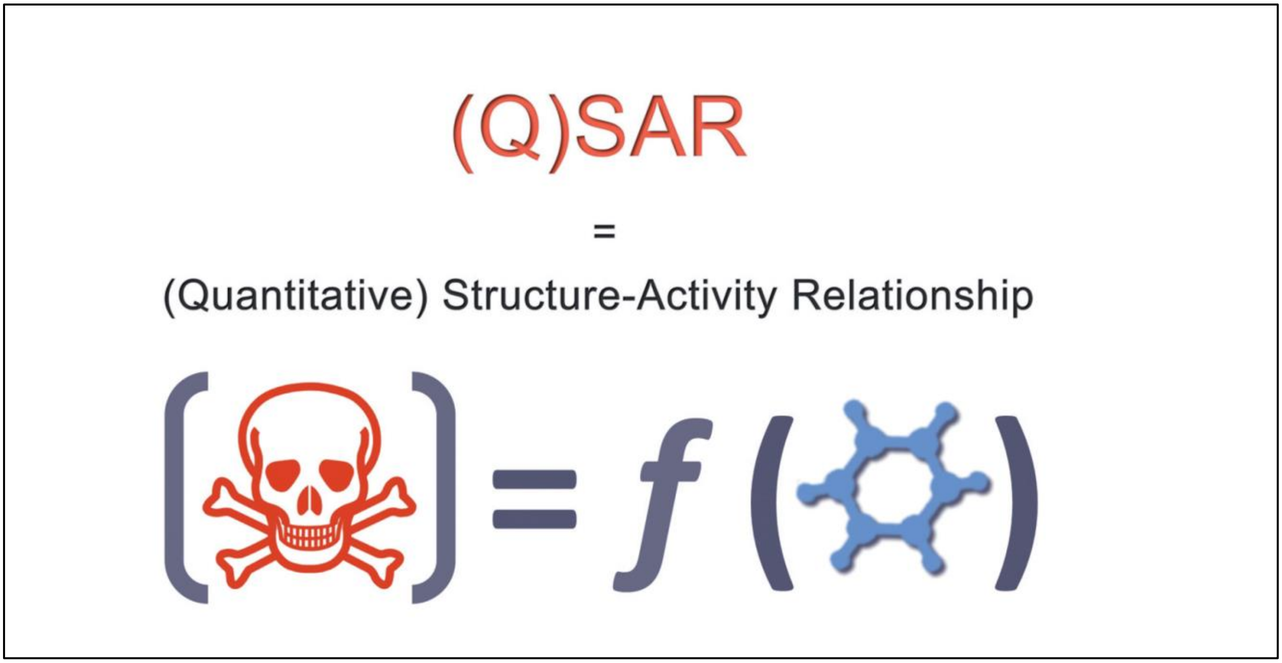


Figure 3: QSAR methods aims at finding the correlation between structural properties of chemicals and their activity.

The QSAR world is very complex, and it would be wrong to think that QSAR is one single method. There are thousands of chemical descriptors and thousands of chemical fragments, many diverse algorithms, studies addressing different endpoints, and for the same endpoint different sets of substances that may interest us. Thus, the number of possible models is stupendously high, and indeed, thousands of models have been developed.

The communities working with the QSAR are also diverse, and there are typically conferences dedicated to applications to toxicological or environmental matters or to the development of new chemicals, in particular pharmaceuticals. Recently there has been an exchange of opinions and methods between these two major communities. Historically, some studies started from the interest in the identification of the physicochemical properties associated with a certain effect. For instance, in the 60’s Corwin Hansch studied ecotoxicity, and put it in relationship with LogKow the partition coefficient between octanol and water, expressed as a logarithm. The idea behind this was that the partitioning between water and the organic solvent represents a model for the partitioning between water and the fish body, and the uptake of the toxic compound into the fish body is a good indicator of the toxic effect. The driving force to conduct these studies was the identification of the key physico-chemical phenomena, which could underpin the observed toxicity phenomena.

Under another approach, some studies sought to identify if chemical toxicity was related to the occurrence of a certain chemical moiety, such as a specific fragment for mutagenicity. In both cases there was an attempt to identify the cause of the toxicity, as explained by a descriptor or a fragment. The idea behind it all was that once the cause is known, we can govern the phenomenon, and thus predict the effect. Unfortunately, the situation is more complex, and these approaches were only partially successful in prediction. As we will see, other approaches have been used to analyse equally complex situations using probabilistic tools. Indeed, it is possible, even if the phenomenon is complex, to summarize the general behaviour within a certain probability.

There are some differences between the ecotoxicity and mutagenicity studies. The fish toxicity case introduced above refers to a toxic effect which has a modulation of the value, and is measured with a continuous value, such as fish acute toxicity. Another common endpoint, LD50, which is the dose that kills 50% of the animals (*rat or mouse, most typically*), is conceptually addressed in a similar way to fish toxicity, with toxicity levels increasing depending on the chemicals.

Conversely, toxic effects such as carcinogenicity or mutagenicity are often expressed as a binary conceptual system: toxic or not. The idea behind this is that a chemical can be carcinogenic even in a minute amount, starting a process which magnifies its effect with time. In such a case the idea is to identify the molecular component which provokes the phenomenon. Therefore, a structure-activity relationship (SAR) model, without the quantitative assessment, is sufficient.

In the first case, quantitative tools are used involving quantitative descriptors and algorithms. In the second case, often fragments are used, and the algorithms are classifiers.

In practice, there are many instances of QSAR models which use combinations of the different chemical descriptors and fragments, and algorithms which can address quantitative or qualitative outputs.

The components of QSAR models

As previously mentioned, the basic hypothesis of a QSAR model is that the activity (*or effect or property*) can be put in relationship with the chemical, using some parameters to describe the chemical compound. Below we will analyse in more detail these three components of QSAR models.

Experimental values, their quality and uncertainty

Even before considering QSAR models, experimental data are necessary for many purposes such as chemical risk assessment. Animal models are used to assess the effects on more complex targets. Obviously, the laboratory model is much simpler than any ecosystem. For instance, when evaluating the effects on the ecosystem, models using fish can be used. In this case, a certain number of fish are put in a tank. This situation is much simpler than any ecosystem, where many organisms are present, in conditions which are more complex than those adopted within the experimental model.

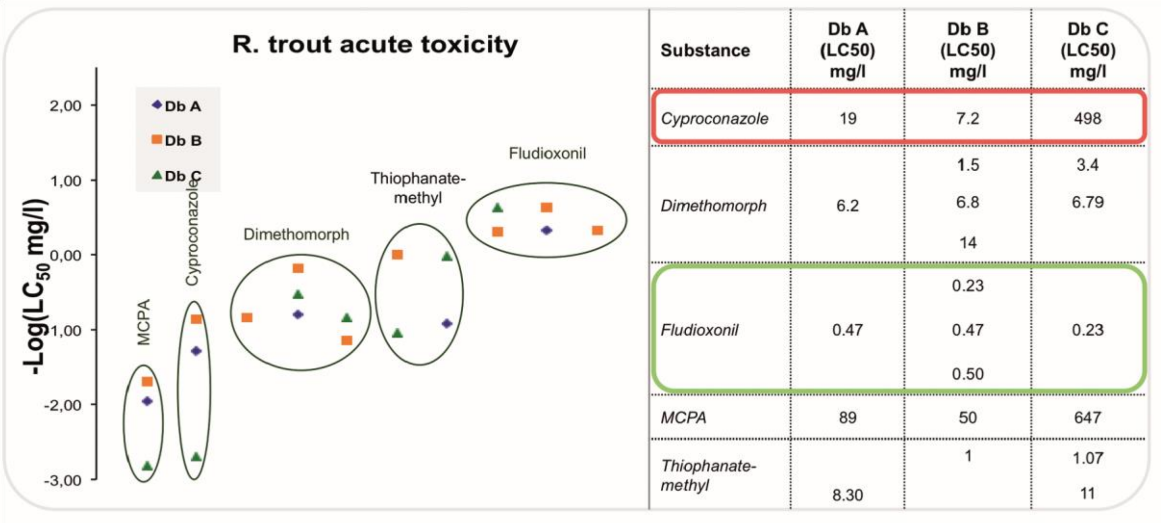
Similarly, animal models are frequently used for human toxicity. In this case there is the issue of extrapolation from one species (a rat, for instance) to humans. Furthermore, the issue of the effects on the whole human population includes many varying situations, because children, pregnant women, sick people, and other sensitive parts of the population have to be protected.

For practical reasons, experiments on animals are done using a limited number of species and situations. Furthermore, it is necessary to get comparable results when using the same test, and thus the experimental parameters are typically fixed within defined protocols.

A fundamental concept is that any experimental value is associated with an uncertainty value. This is true for physical measurements, and even more so for biological data. Toxicity values are always affected by high levels of uncertainty. For instance, when one biotest is compared to a second for equivalence, it is accepted that the value changes by a factor of five [1]. In the case of BCF the reported experimental uncertainty can be up to 0.75 in Log unit [2].

As another example, the reproducibility of the Ames mutagenicity test, which is a quite simple model using bacteria, rather than complex organisms, is about 85%. In other words, if we give 6 substances to two laboratories, we can expect a contradictory output for one of these 6 chemicals. *Figure 4* shows the reproducibility of experimental toxicity data within three different high-quality databases.

Reproducibility of experimental ecotoxicity data LC50 (Lethal concentration) values from different databases



Substance	Db A (LC50) mg/l	Db B (LC50) mg/l	Db C (LC50) mg/l
Cyproconazole	19	7.2	498
Dimethomorph	6.2	1.5 6.8 14	3.4 6.79
Fludioxonil	0.47	0.23 0.47 0.50	0.23
MCPA	89	50	647
Thiophanate-methyl	8.30	1	1.07 11

Figure 4: Reproducibility of the experimental results for toxicity studies. Multiple experimentally-determined values for acute toxicity on Rainbow trout (LC50) have been taken for the same chemicals from different databases (in some cases even the same database contains several different values). The graphic shows that differences exist in these values even though they have been obtained using well known and accepted protocols.

Unfortunately, the information regarding the uncertainty of the experimental models is not always available, and often users ignore the fact that this assessment is fundamental. People may think that the lethal dose that kills rats is sharply defined, an absolute threshold which separates a bad or good effect for a chemical, but there is always an uncertainty factor with *in vivo* values.

Furthermore, these values have a probabilistic meaning. Let's consider for instance the lethal dose, which is in fact defined as the dose that kills 50% of the animals. The obvious meaning is that half of the animals die, and 50% don't. Why this happens and what mechanism it is that impacts half but not all of them are unknown. The meaning of the toxicological test is probabilistic for this reason. The content of this experiment, in its statistical nature, is perfectly useful within the risk assessment framework we mentioned above. We know that a certain effect is expected, and the risk assessment procedure will elaborate the value of the possible risk for other situations. The nature of the input is probabilistic, as is the nature of the output.

Of course, the uncertainty typical of a given endpoint, when assessed using a single protocol, should be characterised because this affects the uncertainty of the *in silico* model. The uncertainty of the final model cannot be inferior to the uncertainty of the input data, and it is suspicious to see values predicted with a precision superior to that of the experimental laboratory model.

The input values should be checked to avoid noise. Indeed, it is well known for any model that we cannot extract correct information if we feed the model with poor value: garbage in, garbage out, as it is said. The availability of data from different sources can provide a way to compare and integrate data. It is important to have access to multiple values for the same chemical, and also to know the uncertainty related to a given endpoint.

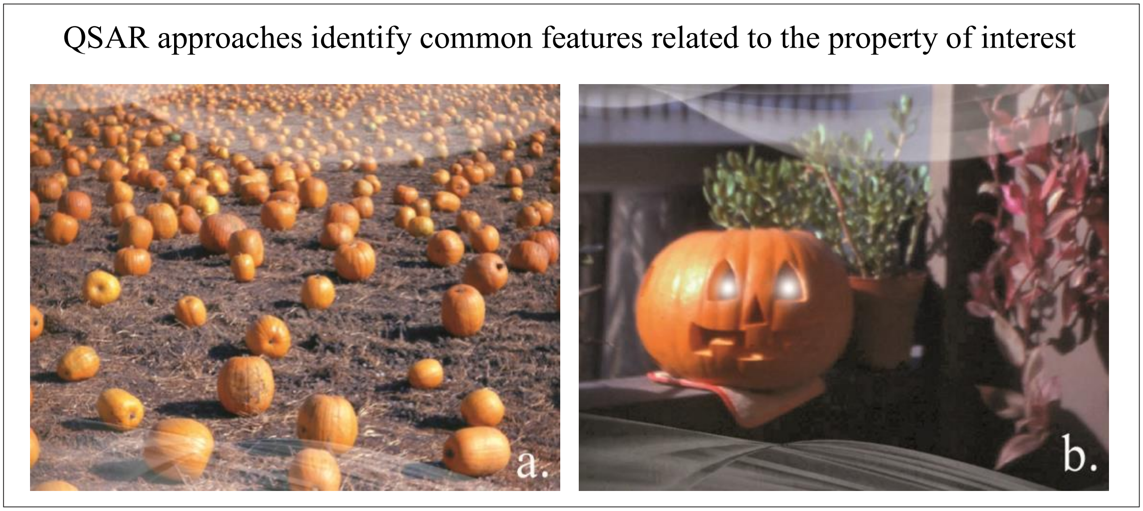
In the case of toxicity values, some databases are good, other less so. We compared different official databases on pesticides and found differences among reported values [3]. Worse may be the case of data taken from the literature. But this matter is not limited to the property values. We also found many mistakes in the chemical structures reported in journals [4]. All these checks require time and effort, therefore are not typically done when a QSAR model is applied for academic research. However, in the case of a model to be proposed for regulatory purposes, efforts should be made to quality check the data. For instance, within the CAESAR [5] and DEMETRA [6] projects we spent about one year checking the data, before starting the modelling activities. Some researchers have clearly identified this problem and dedicated efforts to increasing the quality of the toxicity data available. One such researcher is Ann Richard with the DSSTox database [7]. Today, other colleagues within US EPA proceeded in her work, and the results are within the DashBoard (https://www.epa.gov/chemical-research/comptox-chemicals-dashboard).

Property data are crucial for further development of QSAR. DSSTox, ECOTOX, and AMBIT are examples of databases. The OECD Toolbox has gathered data on many properties.

An interesting feature of these databases is the availability of toxicity and chemical data/structures together. There are several examples such as the databases we mentioned: the US EPA CompTox Chemicals DashBoard [8], the OECD QSAR Toolbox [9], and those related to the CONCERT REACH project [10].

The chemical information: descriptors and fragments

There are two main ways to describe a chemical compound: using global descriptors or using specific fragments.



QSAR approaches identify common features related to the property of interest

Figure 5: Identification of important structural properties and/or fragments. a) QSAR approaches are able to extract, from a training set, the most significant structural properties linked to the specific property to model, which are global ones (in this case: the size, shape, colour of the pumpkin). b) In other cases, it is useful to analyse peculiar features. The pumpkin in figure b) is characterized by very particular features.

If we imagine that the pumpkins in *figure 5a* are chemical compounds, we can distinguish them on the basis of their size, shape, colour, etc. However, looking at the pumpkin in *figure 5b* we immediately see that there are peculiar features which make this pumpkin different from any other pumpkin.

In case of the chemical compounds, some descriptors are global, or general, such as molecular weight or molecular size; however, in some cases it may be preferable to evaluate if there is a specific molecular moiety in the chemical compound. There are many QSAR models using global descriptors, but also a certain number of them using fragments.

The molecular descriptors can be classified as:

- Constitutional descriptors are quite simple; they include molecular weight, number of atoms present in a molecule (*for instance number of chlorine atoms*), number of double bonds, etc.
- Topological descriptors indicate the bonds between atoms and can be used to represent the ramification of the molecule. Indeed, a molecule can be represented as a graph.
- Certain descriptors take into consideration the electronic charge of a certain atom, or its polarity.

The procedure to calculate molecular descriptors

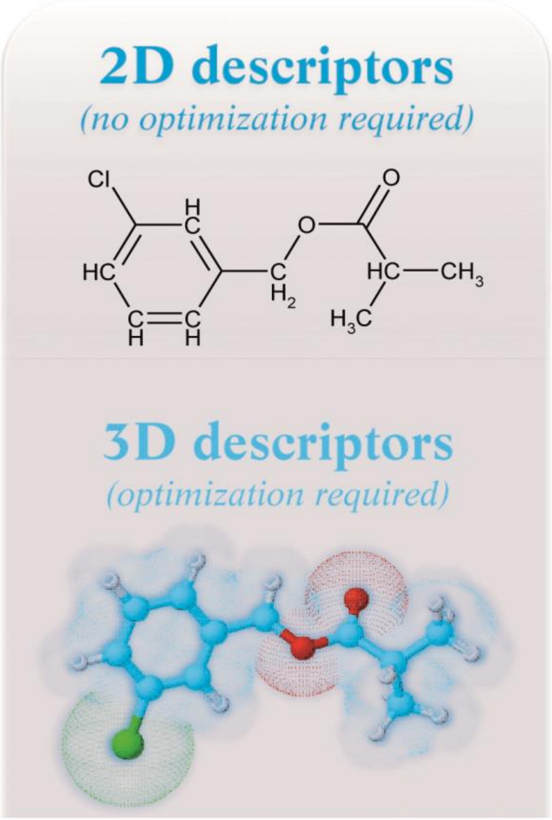


Figure 6: Molecular descriptors can be of different levels of complexity. The so-called 2D descriptors (e.g. topological information) do not need any conformational information regarding the molecule. The 3D descriptors (e.g. charge distribution) need the molecule to be optimized, creating a series of problems, from increased time for calculation, to more difficult reproducibility.

- Some descriptors refer to the molecular orbitals of the molecule. Some descriptors calculate the energy of the molecular orbitals, for instance HOMO refers to the energy of the highest occupied molecular orbital, and LUMO refers to the energy of the lowest unoccupied molecular orbital.
- Another kind of descriptor is the so called physico-chemical. They include LogKow, lipophilicity, etc. LogKow (also called logKow) is the logarithm of the partition coefficient between octanol and water. This descriptor has been used since the first QSAR models, and originally it was measured. Nowadays it is much more common to calculate it.
- There are programs that have a list of pre-codified fragments, even thousands of them, and the software checks whether they are present in the molecule of interest or not. There are programs which check for the presence of fragments in a molecule with reference to a list of fragments; these programs are quite fast, and are often used to process huge databases, for similarity purposes. The pharmaceutical industry uses models based on fragments quite often.

A few decades ago, the use of chemical descriptors was very limited. For instance, Corwin Hansch studied ecotoxicity, and put it in relationship with LogKow. The idea behind this was that the partitioning between water and the organic solvent represents a model for the partitioning between water and the fish body, and the uptake of the toxic compound into the fish body is a good indicator of the toxic effect.

This physicochemical parameter has been used in most of the QSAR models of aquatic toxicity.

Slowly, other descriptors have been investigated, in an attempt to better explain certain effects. In particular, further descriptors were introduced to better explain chemical reactivity, molecular size, etc. Nowadays thousands of chemical descriptors can be calculated. Quite often the molecular descriptors are combined. Therefore, dividing one molecular descriptor with a second one, for example, can easily produce a new one.

2D and 3D descriptors

We can also distinguish the descriptors on the basis of the kind of detail needed to represent the molecule. A major difference is between descriptors which need a tri-dimensional (*3D*) representation and other descriptors. Indeed, some descriptors, such as the number of certain atoms, or topological descriptors, do not need a 3D representation of the molecule. Conversely, descriptors like molecular volume of quantum-mechanical molecules require a 3D representation of the molecule.

In the case of 2D descriptors, the molecule can be represented flat. In case of 3D descriptors we need a 3D representation (*figure 6*). Most typically, the 3D representation has to be optimized, and this is done manually. Indeed, there are many different conformations that the molecule may exhibit, depending on the rotations of the bond and the angle between bonds. On the basis of the different conformation, many 3D descriptors can vary, such as molecular volume, length, etc. Thus, the values of the 3D descriptors typically change depending on the user, the software, the approximations, etc. Furthermore, the calculation of the 3D descriptors takes more computer time for the necessary optimization of the values.

For these reasons, the reproducibility of the 3D descriptors is lower, compared to that of the 2D descriptors. Thus, if we want fast and reproducible QSAR models, 3D descriptors may be counterproductive.

However, 3D descriptors have some advantages, and can provide better results for specific cases, such as models on restricted chemical classes.

Most typically, the variability and the uncertainty of the experimental values of the property to be predicted is so large that improvements related to the use of 3D descriptors are negligible. For instance, in the case of global models for toxicity developed within CAESAR, we found that the statistical performance of the models based on 2D descriptors was the same as the models based on 3D alone, or 2D and 3D descriptors combined.

We notice that also 2D descriptors may suffer from poor reproducibility. Certain parameters, which are related to the representation of the molecular bonds and the tautomers, may vary the output of the 2D descriptor. This is the case for the number of double bonds. Different numbers may be obtained if we calculate certain bonds as aromatic bonds, or as double bonds. For instance, the benzene ring may be represented as a ring with three double bonds and three single ones, or as a ring with six aromatic rings. While in the case of benzene rings this can be easily solved, it may be more critical for heteroaromatic rings, because it may be more difficult to distinguish if a bond is aromatic or not. This often depends on the formalism and the conventional approach adopted.

Even in the case of tautomers, certain 2D descriptors may change depending on which tautomer is used. Thus, if we want to have reproducible results in QSAR models it is necessary to use the same software to calculate the chemical descriptors, and to use the same format to represent the chemical structure. For instance, an error which may occur if different formalisms are used is related to the representation of the nitro group.

Beyond 3D descriptors, more complex descriptors exist, including 4D, 5D, etc. In this simple introduction to QSAR do not cover them, since for the common publicly available models they are not used.

The way to represent the chemical structure

Before calculating the chemical descriptors or fragments, the chemical formula has to be represented in a suitable way.

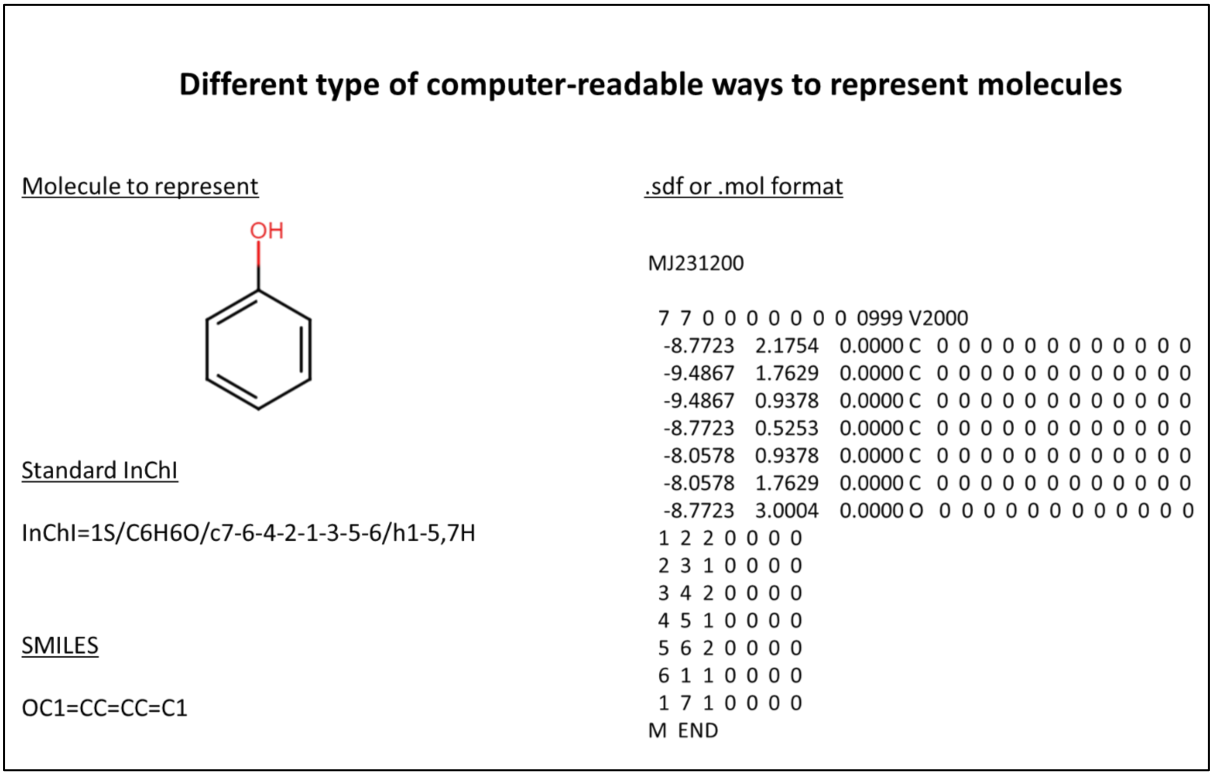
There are several ways to do this. Typical ways are InChI (International Chemical Identifier) [11], SMILES (Simplified Molecular Input Line Entry System) [12], or sdf1 format.

InChIs are flexible layered alphanumeric codes to represent the structure of a molecule. Layers represent the different details that can be represented; therefore, InChIs is not univocal. To solve this problem, the Standard InChIs were created, less flexible but unique. So far, InChIs (Standard or not) are less used than SMILES, simpler alphanumeric strings in which atoms, bonds and stereochemistry are codified. However, care should be taken with SMILES because there may be more than one correct SMILES for the same chemical. There are different formalisms to write the structure with SMILES; for instance for the nitro group and the kind of bonds between the oxygen and nitrogen. Indeed, the bond between N and O can be written as a double bond or with a separation of charges: N=O or N^+O^-. The chemical may be read in different ways, generating different results. Thus, the user should not mix SMILES taken from different sources, unless they are standardized. In this way, the same structure is written in the same *figure7*.

The choice of the suitable complexity in the chemical descriptors and chemical

The choice of how to represent the chemical through an appropriate format and structure has to be related to the purpose.

In most models useful for REACH, related to typical industrial chemicals, the substances are not pure enantiomers. It is quite difficult to find in the literature two experimental values specific for the two enantiomers. Thus, to verify the appropriateness of the substance for the specific case, and whether the predictive model has been developed on the basis of sufficiently specific data, the detail of the chirality of the substance should be checked.



Different type of computer-readable ways to represent molecules

Molecule to represent

Standard InChI

InChI=1S/C6H6O/c7-6-4-2-1-3-5-6/h1-5,7H

SMILES

OC1=CC=CC=C1

.sdf or .mol format

```
MJ231200

 7  7  0  0  0  0  0  0  0  0999 V2000
  -8.7723   2.1754   0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
  -9.4867   1.7629   0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
  -9.4867   0.9378   0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
  -8.7723   0.5253   0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
  -8.0578   0.9378   0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
  -8.0578   1.7629   0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
  -8.7723   3.0004   0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
 1  2  2  0  0  0  0
 2  3  1  0  0  0  0
 3  4  2  0  0  0  0
 4  5  1  0  0  0  0
 5  6  2  0  0  0  0
 6  1  1  0  0  0  0
 1  7  1  0  0  0  0
M  END
```

Figure 7: Examples of representation of molecules for software input. Different methods have been developed throughout the years to represent and archive structural and conformational information of molecules for computer utilisation purposes. The .mol (for one chemical) or .sdf (for a list of chemicals) format save the type and position of each atom as well as the topological information explicitly; these format are complex but define the molecule clearly. The software has only to place the atoms in the defined position and link them with the defined bond. The simplest way to represent the molecules is the Simplified Molecular Input Line Entry Specification (SMILES); SMILES are text strings formatted in a way that allows the software to know how the atoms are linked to each other; the position is then calculated on the basis of the topology. The International Chemical Identifier (InChI) is an alphanumeric string more flexible but also more complex than SMILES.

The user may wish to get the maximum detail and information on the chemical structure, chirality, etc. However, the real information should refer to the chemical which has been used for the experiment, and thus if it is a mixture of enantiomers, this should be used. Furthermore, the analysis of the results may be limited by the lack of other, related chemicals with the same detail of information. Indeed, if we want to build up a model for enantiomers, we need a series of cases, not just a few. However, as we have explained this may be very difficult.

Another important aspect related to the chemical structure, is that most typically the salt is not used for QSAR modelling. Thus, the salt is transformed into the neutral form, loosing, for instance, the sodium, or the chloride ion. This has to be considered, when using the QSAR model.

Software for chemical descriptors calculation

There are several commercial and free programs for chemical descriptors; below we provide some examples of free programs.

- OCHEM (*http://ochem.eu/*)
- CDK (ref to annexes: Annex 4. A free and open source informatics library for chemistry: Chemistry Development Kit (CDK))
- PaDEL (Yap, C.W. (2011), PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. J. Comput. Chem., 32: 1466-1474. *https://doi.org/10.1002/jcc.21707*)

More on chemical descriptors

See these for more info on descriptors:

- http://qsar.sourceforge.net/dicts/qsar-descriptors/index.xhtml for a list and classes of molecular descriptors;
- https://www.sciencedirect.com/topics/medicine-and-dentistry/molecular-descriptor#:~:text=A%20molecular%20descriptor%20is%20a,endpoint%20%5B9%2C44%5
- https://bigchem.eu/sites/default/files/School1_Horvat.pdf
- https://www.frontiersin.org/articles/10.3389/fchem.2022.852893/full
- https://pubmed.ncbi.nlm.nih.gov/29934886/

Moreover, there are some books or chapters on chemical descriptors [13-14].

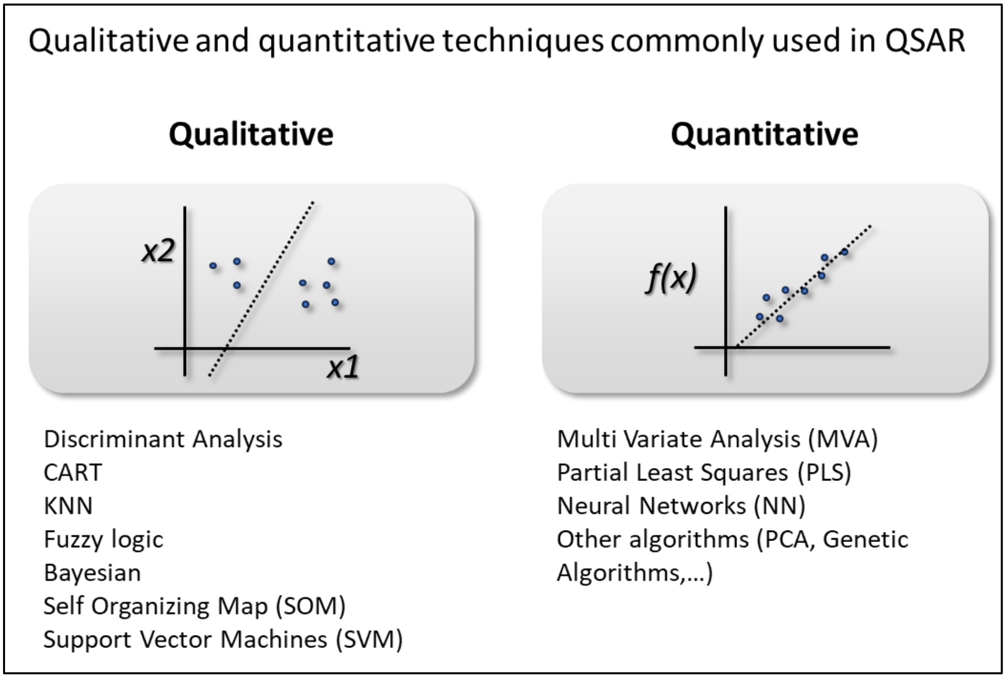
The modelling algorithms: qualitative and quantitative models

In the last decades in addition to the thousands of chemicals descriptors that have been made available, many advanced, powerful modelling algorithms have also been developed. The older QSAR models were linear equations with a few parameters. Then, other tools were introduced, such as artificial neural network, fuzzy logic, and data mining algorithms, making possible nonlinear models and automatic generation of mathematical solutions [15-16].

We can distinguish two kinds of algorithms: qualitative and quantitative. Quantitative methods get a continuous value. Qualitative algorithms find the category, e.g. the toxicity class. *Figure 8* lists some common qualitative and quantitative models used in QSAR.

In fact, in the case of qualitative models the appropriate definition would be Structure-Activity Relationship (SAR), since the purpose of qualitative models is not to obtain a quantitative evaluation, but a category. In some cases, the categories refer to different thresholds, such as the toxicity classes for acute toxicity for mammals.

The definitions of qualitative models as SAR and quantitative models as QSAR, and the distinction between qualitative or quantitative algorithms are useful as general rules, to describe the general boundaries of the methodology. However, there are cases where the boundaries are indistinct. With the fuzzy logic approach, it is possible to go from one paradigm to another quite easily. Even with other mathematical methods it is possible to talk about the probability that a certain compound is toxic or not, and thus to think about a classification problem (*difference between toxicity 0 = not toxic, and 1 = toxic*) using a continuous scale: for example, the probability that a certain chemical is toxic (*100% probability*) or not toxic (*0% probability*).


Qualitative and quantitative techniques commonly used in QSAR

Qualitative

Quantitative

Discriminant Analysis
CART
KNN
Fuzzy logic
Bayesian
Self Organizing Map (SOM)
Support Vector Machines (SVM)

Multi Variate Analysis (MVA)
Partial Least Squares (PLS)
Neural Networks (NN)
Other algorithms (PCA, Genetic Algorithms,…)

Figure 8: Mathematical and statistical techniques commonly used to build QSAR models.

Different algorithms

There is a variety of methods for building QSAR models. They can be classified as supervised (for example, Multiple Linear Regression, Discriminant Analysis, Partial Least Squares, Classification and Regression Trees, Neural Networks, etc.) or unsupervised (for example, Principal Component Analysis, Cluster Analysis, k-Nearest Neighbours, Nonlinear Mapping, etc.), where supervision refers to the use of the response data that is being modelled, or unsupervised. Unsupervised learning makes no use of the response, meaning that the algorithms seek to recognize patterns in the descriptor data only. The advantage of unsupervised learning is the lower likelihood of chance effects, due to the fact that the algorithm is not trying to fit a model. On the other hand, supervised learning does use the response data and care needs to be taken to avoid chance effects. Another significant difference between supervised and unsupervised learning methods is the ratio of compounds (p) to variables (n) in a data set. When n ≥ p, some supervised learning techniques may not work due to the failure to invert a matrix, while others may give a false, but apparently correct, classification. Even though this is not a problem for unsupervised methods, the presence of extra variables that have no useful information may obscure meaningful patterns.

The nature of the response data that are capable of handling is another important feature of modelling methods. In this context, there are two types of methods: those that deal with classified responses (*for example, mutagen / not mutagen, toxic / slightly toxic / non-toxic*) and those that handle continuous data (*the response is a potency of an endpoint*). For the modelling of categories, a wide range of classification methods exists, including: Discriminant Analysis, k-Nearest Neighbours (*KNN*), Classification and Regression Trees (*CART*), Support Vector Machine, etc. For the modelling of continuous data, the most widely used method is Multiple Regression Analysis (*MRA*), a simple approach that leads to an easily understandable result. MRA is a powerful means of establishing a correlation between independent variables (*molecular descriptors*) and a dependent variable (*biological activity*). The main characteristics of MRA are:

- Linear relationship between Y and several descriptors (Xi);

$$Y = aX1 + bX2 + cXn + \ldots + \text{const.}$$

- Errors are minimized by least square;
- Polynomial terms may be included.

In addition, Artificial Neural Networks (*figure 9*) can be used for modelling both classified and continuous data. The main characteristics of ANN are:

- The structure is inspired by biology;
- ANNs are a set of connected nonlinear elements making transformation of input.

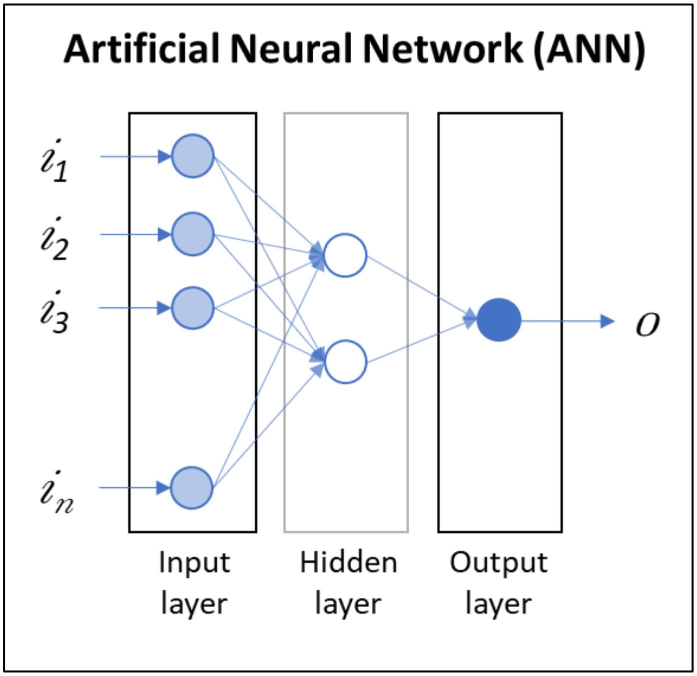


Figure 9: The organization of Artificial Neural Networks is similar to that of biological systems. The elements of the network (called neurons) are organized in layers and interconnected in a way that the output of a neuron is the input of the subsequent one.

More recently, several machine-learning tools have been proposed, obtaining better results compared with classical methods. Deep learning emerged as an interesting approach. We will present in detail this at the end of Part A of this eBook.

Free tools & algorithms for QSAR modelling

Free and open sources tools and algorithm have been developed and currently maintained to build QSAR models. Some examples of tools developed specifically for building QSAR models are:

- OCHEM (http://ochem.eu/)
- CORAL (http://www.insilico.eu/coral/SOFTWARECORAL.html /)
- SARpy [17]
- and those downloadable from VEGAHUB in the download page (https://www.vegahub.eu/download/).

Other more general free resources, including mathematical and statistical approaches which can be used for QSAR modelling are:

- WEKA (http://www.cs.waikato.ac.nz/ml/weka/)
- R (http://www.r-project.org/)
- KNIME (https://www.knime.com/)

The way that a QSAR model is built up

In some cases, especially in case of genotoxicity models, the human expert identified fragments which can be related to the genotoxic effect. For instance, it is known that nitrosoamines are genotoxic. The visual examination of a series of chemicals sharing the same fragment may be used for this purpose. In this case the effect is simply the toxic effect (*genotoxic or not, for instance*), and the chemical information is simply the fragment. The algorithm is, in this case, the rule. Expert systems have been built up in this way. Examples of this kind of model include the models developed by the group of Romualdo Benigni, available within Toxtree (https://toxtree.sourceforge.net/).

Most typically a QSAR model is built up starting with a set of chemicals with known property values.

The very first step in QSAR modelling is the translation of the structural information in some numerical values (*the molecular descriptors*).

Therefore, conceptually, a table of all the chemical compounds is built up. For each chemical, one needs the descriptors and the property values of that set of compounds. The compounds are typically arranged in a column, and there are several columns for the chemical descriptors, as well as one column for the property value. The chemical descriptors are the x (*the input*) of the model, and property is the y (*the output of the model*).

However, a major problem may result from the uncritical use of powerful mathematical tools; the risk is that the model does not work when applied to new compounds, because it is only capable of replicating the toxicity of the chemicals used to train the model. In model development the procedure is to use some chemical compounds with known toxicity as training set. Then, using chemical parameters and a suitable algorithm, the model is developed.

However, to check if the model is really a predictive one, an assessment has to be done. This obvious consideration applies to all kinds of models, the simple, with a single parameter, and the more complex ones. The risk of chance correlation is higher when a high number of descriptors or parameters is used and when few examples or molecules are used. This may also lead to over-fitting, a phenomenon in which the model gives high performance on the training set, but then results decrease dramatically when the model is applied to new chemicals, such as in an external validation set (as shown in *figure 10*).

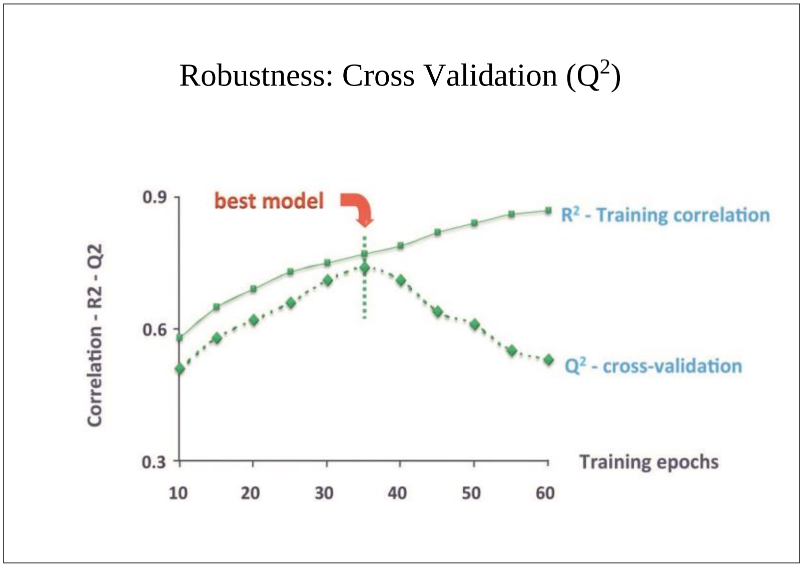


Figure 10: QSAR models have to be tested on molecules not used to build them in order to test their real predictive power. Cross validation consists of leaving out N compounds from the training set and rebuilding the model. The N compounds are then used as test set. This procedure is iterated many times. The parameters R2 and Q2 are both measure of correlation between experimental and predicted values; the difference is that Q2 refers only to the compounds outside the training set of the model.

This leads to the need for dimension reduction, variable elimination and selection, which are different techniques for reducing the complexity of a problem to recognize useful and informative patterns in the data. Dimension reduction is the process of reducing the number of random variables under consideration and is usually performed by a mathematical procedure called Principal Component Analysis (*PCA*) in which new variables, called principal components, are created from linear combinations of the original variables. Variable elimination is the process by which unhelpful or unnecessary variables are removed from a data set. Common procedures for variable elimination are Corchop [18] and unsupervised forward selection. Even after eliminating unnecessary variables from a data set, there may still be many variables to choose from when building a model. In this case variable selection is used, with the aim to choose descriptors useful in mathematical models; this will lead to a model that will generalize to other unseen compounds. There are many diverse procedures for variable selection, and some are built in to the process of model building, such as forward stepping multiple regression.

The validation of the model

The models have to be validated. This simple statement may appear obvious today. QSAR models developed decades ago mainly addressed the discovery of certain relationships between a given parameter and the effect of interest. For instance, it was satisfactory to find that there was a linear relationship between LogKow and aquatic toxicity. Conversely, models for regulatory purposes require stringent validation procedures on the prediction of the property/effect.

For regulatory purposes there must be proof that this relationship applies to the prediction of the properties of *new* chemicals; thus, this has to be specifically addressed. This statement puts emphasis on the statistical validation of the model. This is clearly addressed within the five OECD principles for QSAR [19]. New statistical tools and evaluation procedures have been introduced, compared to the simple fitting measurement based on the training set. The importance of an external test set has been stressed in many cases. If the total number of compounds is low, this imposes limitations on the external validation. Internal validation is in any case recommended, and for this purpose a number of tools has been developed, such as leave-one-out, y-scrambling, etc. Leave-One-Out Cross Validation (*LOO or LOO CV*) involves leaving out one compound, fitting the model to the remainder of the set, making a prediction for the left-out compound and repeating the process for each of the compounds in the set. A variety of statistics can be generated using this procedure, for example LOO R^2 (*called* Q^2) and a predictive residual sum of squares (*PRESS*).

The disadvantage of LOO is that only a small part of the data set is omitted and if outliers occur in pairs or groups they will not be identified. A better approach is to leave out some larger portion of the set (*10 or 20%*) and to repeat this a number of times (Leave-Many-Out, *LMO*). This allows the generation of a set of predicted values for the compounds so that estimates may be made of the likely errors in prediction. The disadvantage of this approach is that it is computationally intensive and suffers from a combinatorial explosion as the sample size is increased. Some examples of techniques for the model validation are summarized below (*Table 1*).

Evaluation of a classifier

Typically, qualitative models are evaluated using the Cooper statistic. In the simple case of a binary classification, there are two classes, such as toxic (*positive*) or not (*negative*). The results of a classifier could be therefore grouped in four categories: toxic compounds predicted as toxic (*True*

Positive or TP) or as non-toxic (*False Negative or FN*) as well as non-toxic compounds predicted as non-toxic (*true negative or TN*) or as toxic (*False Positive or FP*). These four classes are usually represented in the so-called *confusion matrix* (as shown in *Figure 11*).

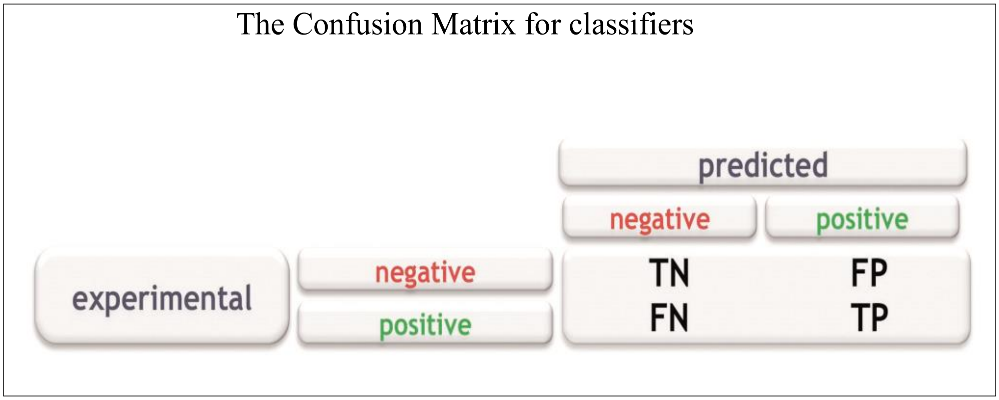
Table 1: Common approaches used to validate QSAR models.

Cross validation	Bootstrapping	Y-scrambling
Leave-One-Out • All the data are used for fitting except for one compound • Predict the excluded sample • Repeat it for all samples • Calculate Q^2 or R^2cv similarly to R^2 on the basis of these predictions. Problem: this approach may be too optimistic if there are many samples Leave-Many-Out • Use larger groups to obtain a more realistic outcome	• Bootstrapping simulates what happens by randomly re-sampling the data set with n objects • K n-dimensional groups are generated by a randomly repeated process which eliminates some objects • The model obtained on the different sets is used to predict the values for the excluded sample • From each bootstrap sample the statistical parameter of interest is calculated • The estimation of accuracy is obtained by the average of all calculated statistics	• Randomly permute Y responses while X variables are kept in the same order for several times

Three main statistical parameters can be derived from the combination of these four cases for model evaluation:

Accuracy (*A*), also referred as concordance, is the measure of the correctness of prediction. This parameter gives a general evaluation of the errors made and is defined as the ratio between the compounds correctly predicted and the total number of compounds. A good model has high accuracy value.

$$A = (TP + TN) / TOTAL$$



The Confusion Matrix for classifiers

		predicted	
		negative	positive
experimental	negative	TN	FP
	positive	FN	TP

Figure 11: The confusion matrix of a binary classification. Positive classification is usually associated with active compounds (e.g. toxic) whereas negative is associated to inactive ones (non-toxic). Compounds correctly predicted are called True Positives (TP) or True Negatives (TN) depending on whether they are active or inactive. Active compounds predicted inactive are referred as False Negatives (FN), whereas inactive compounds incorrectly predicted are False Positives (FP).

Sensitivity (*S*) is the measure of the positive compounds correctly predicted. Especially for regulatory purposes, it is important not to declare safe a chemical which is actually toxic (*FN*). Sensitivity is defined as the ratio of the TP tests to the total number of positives. A good model has high sensitivity.

$$S = TP / P$$

Specificity (*SP*) is the measure of the negative compounds correctly predicted. Specificity is defined as the ratio of the TN tests to the total number of negative compounds. Sometimes the 1 - SP parameter is reported.

$$SP = TN / N$$

It is our opinion that for regulatory purposes it is important to verify that the qualitative model has a high sensitivity, to reduce the number of false negatives.

When the dataset used to build up a model is not balanced, it is preferable to take this into consideration, also in the evaluation of the results. In the case of unbalanced datasets, it is preferable to use the Matthews correlation coefficient (MCC) or the balanced accuracy (BA).

$$MCC = \frac{TN * TP - FN * FP}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

$$BA = (S + SP)/2$$

In some cases, more than two classes are defined. For instance, a chemical can be not bioaccumulative, bioaccumulative or very bioaccumulative (three classes).

Evaluation of a quantitative model

Quantitative models are most typically evaluated using statistical parameters which take into account the errors of the model. These errors are measured on the basis of the training set, and this gives an idea of the model robustness. However, this is not sufficient since the main interest of REACH is to understand if a certain model can be used for prediction purposes. Thus, for regulatory purposes, additional statistical measurements are used for prediction. Some measurements use internal validation whereas other tools refer to an external test set.

The values predicted by the model (*on training, test and/or external validation set*) are put in correlation with the experimental values using a graph and then the coefficient of determination (*R^2*) is calculated giving an estimation of the model goodness. In the following formula, obs represents the observed (experimental), $pred$ the predicted, and $\overline{obs}$ the mean of the observed values.

$$R^2 = \frac{1 - (obs - pred)^2}{(obs - \overline{obs})}$$

In the case of qualitative models the emphasis is on false positives and negatives. However, we underlined the importance of also paying attention to false negatives for quantitative models [3]. Indeed, regulators pay much more attention to false negatives. With regard to models for regulatory purposes, attention can be paid to false positives within a wider strategy in which intelligent testing methods are used, taking into account the sensitivity and selectivity properties of each individual element in the combined strategy. The usual approach to evaluating quantitative methods, using R^2, clearly shows that we do not take into account whether or not the error has a sign, because we use the square. However, we should evaluate this. This was addressed in the EC projects DEMETRA and later on in the project CAESAR.

The validation of the QSAR models and non-testing methods and regulatory needs

In the classical QSAR models it was assumed that a convincing explanation of the phenomenon was sufficient. The interest in the possible use of such a model to predict the properties of related compounds was not the declared target of the study, and the statistical proof of the predictive power of the model was usually not sufficiently checked. Indeed, the older QSAR models were based only on the fitting description of the mathematical equation.

We emphasise that the target for these studies was not the prediction of the property of the chemical compound, but the understanding and modelling of the mechanism at the basis of the phenomenon. In these kinds of studies, all the property values of the set of compounds were known. What was unknown was when a certain phenomenon occurred. Even today there are studies conducted to explore the possible reasons for the occurrence of a certain phenomenon and the emphasis may be placed on exploring the mechanism. Additionally, if we want to shift the model to the predictive field, we need appropriate ways to validate our model.

Another reason to develop or use QSAR models is to predict the property values of a certain chemical. In this case, the model is built on the basis of the property values of other chemicals, with known values, and the model calculates or derives the property value of the chemical on the basis of certain rules. In order to use a model for this purpose, a suitable check of the predictive performance of the model has to be done, as explained in the paragraph “The validation of the model”.

If the predictive performance of models is not statistically checked, there is a serious risk in using the model to predict a property value. The careful check of the predictive power is one of the principles for the correct use of the QSAR models defined by the OECD [19].

It is possible that a model aims both to predict the mechanism at the basis of the phenomenon and also to predict the property of a target chemical. Indeed, in a certain way all models would aim to achieve both goals. The main difference is the emphasis placed on one aspect over another; is the plausible explanation of the phenomenon, or the predictive power of the model more important? Indeed, a model can be optimized and carefully checked towards the suitable, high predictive power. However, the components of this model may be not fully understandable. This may be the case with integrated models, which combine different models (see the section Hybrid models). It has been theoretically described and experimentally proved that these models are more robust and predictive. However, since they refer to different parameters which are statistically optimized, their interpretation is more difficult than simpler models based on a few rules.

On the other hand, models which are based on plausible mechanisms may be more easily accepted.

However, we have to remember that these models also have a statistical basis: according to what is observed from related compounds, a certain mechanism can be *hypothesized*. This does not mean that the specific mechanism will occur for the target compound.

Unfortunately, phenomena occurring in nature and life are very complex. Currently we are capable only of gathering limited information on them.

In information technology, there is a concept of explicit and implicit knowledge. Explicit knowledge is knowledge which has been already codified into explicit rules. This is the case of QSAR models where, for instance, referring to expert system, some fragments associated to carcinogenicity have been identified. However, these lists are not univocal; several of them exist, but not all fragments have been identified and thus there are chemical which are carcinogenic which are not recognized as toxic (*false negatives*). Furthermore, there are chemicals showing a carcinogenic fragment which are not toxic (*false positives*).

On the other hand, there are models based on the data, which extract the knowledge directly through a process of data mining and knowledge engineering. In some cases, these models showed higher performance than the models based on explicit knowledge, due to the fact that they can incorporate information which has not yet been codified by human experts [17]. Considering that each approach has limits, it is preferable to use both approaches and integrate them, if possible.

Mechanistic or probabilistic models?

The general theoretical context

Quite often there is a debate between supporters of the mechanistic models versus other in favour of probabilistic ones. We already mentioned above that in information technology the matter is addressed distinguishing between *explicit* versus *implicit* knowledge. In both cases the source of the information contained within the experimental data: in the case of the explicit knowledge, the human experts have codified a series of rules, which are used within the expert systems, for instance.

The matter has a long history: thousands of years! When Rafael painted the School of Athens for the Pope Julius the second, at the centre of the scene there are two great philosophers: Plato and Aristotle. Indeed, at the time of Rafael there was still the debate about the two opposite visions: the supremacy of the theory, as sustained by Plato, or the supremacy of the real world – the observation.

This dilemma has been solved by Galileo. The observation generates the theory, which must be verified by further observations. Later, this concept has been further elaborated by Hume, through the process of induction and deduction, moving from the data to the theory, and vice versa. Finally, in the last century, the Nobel price was assigned to Pauli for his studies on probability, introducing a further level of complexity to the matter.

The pros and cons of each approach

Coming back to the mechanistic or probabilistic models, there are close similarities between what we discussed above. The mechanistic approach emphasizes the theoretical aspects, while the probabilistic one underlines the fact that it is possible to extract the correct knowledge from the data using good algorithms. By the way, the more recent algorithms go beyond the statistical ones, as we shortly discussed introducing machine learning.

The great advantage of the mechanistic models is their reliance on sound theoretical basis. This increases the confidence in their use; thus, they are more convincing. The interpretation of the results helps assessors, using a language close to their practice.

However, the matter is more complex. For instance, there are multiple lists of fragments associated to mutagenicity (structural alerts) proposed within different mechanistic models, and their overlap is only partial. Indeed, the list of reasons why a substance is mutagenic or not is somehow subjective. Furthermore, there is not a complete knowledge of the causes of mutagenicity, thus this method will have false negatives. Another issue is that this approach may be simplistic. Its basis refers to a single event determining the complete process; however, we know that there are modulations in the process, and multiple concurrent causes are likely responsible for the final outcome. For instance, aldehydes are associated with mutagenicity, and indeed the short ones, with few carbon atoms, are mutagenic. However, the aldehydes with a longer chain are not mutagenic. Indeed, most, if not all, the structural alerts are associated with a prevalence of mutagenic substances which is not 100%, and in some cases the prevalence is less than 50%. This means that there are false positives. To partially mitigate this problem, for some structural alerts there are conditions there which stop the adverse effect. However, while for the toxic effects the mechanism is explicit, for the mitigation the mechanism is often not explained. Thus, this approach introduces a single event at the basis of the toxic phenomenon, typically associated to a fragment, disregarding the whole molecule. In some cases, some other conditions are introduced, but these cannot replicate the complex phenomena occurring in a living organism, which usually are not linear and imply feed-back and circular processes, within a complex network of events. The group of Daniel Boley studied the network of biological processes occurring in simple monocellular systems. They calculated tens of millions of steps in complex networks within the process of conversion of sugar to ethanol done by yeast, for example [20]. This points out the relevance of Pauli’s lesson. Finally, it is sometimes assumed that the result of a mechanistic model indicates the real mechanism. Actually, it indicates a possible mechanism, but a practical experiment should be done to verify if the assumption is correct. This, according to Galileo’s lesson.

As a conclusion, the mechanistic approach is:

- 👍 Supported to theoretical considerations
- 👍 Clear
- 👍 Close to the common practice used by assessors
- 👍 Convincing
- 👎 Not unique
- 👎 Simplistic
- 👎 Partial
- 👎 Generates false negatives
- 👎 Generates false positives
- 👎 Explanation is not complete
- 👎 It produced a possible explanation, not a demonstrated one

The machine learning and statistical approaches (we will call them the computer approach) have the advantage of relying on observations, preferably a large set of data. In the case of large datasets, it would be very difficult for the human expert to manually screen tens of thousands of data. Indeed, even the mechanistic approach often applies computer methods to screen and prune data. Thus, there is a mutual advantage of the two approaches (mechanistic and computer approaches), which may interact. Indeed, in the case of the computer approach, it may happen that the developer uses molecular descriptors which have been proven to be theoretically associated with the effect. Thus, in the real cases, there is not a dualism between the two practices.

The computer approach can deal with complexity, addressing the case of the multiple causes associated with a single effect. Even more, through multitask methods, multiple inputs can be associated with multiple outcomes, and this increases the possibility of exploitation of the data (not from a single endpoint, but from many endpoints) simplifying the process and identifying links between different endpoints. The computer approach can deal with non-linear processes. It can cope with quantitative outcomes, identifying threshold values where the change of behaviour occurs. It can associate the probability that a certain event occurs. The computer approach can easily associate the outcome with the set of substances with their effects, following a strategy close to read-across. Indeed, several predictive tools apply this strategy, which is non-parametric.

Without a computer, these aspects would be impossible to handle. Probably the most fascinating advantage of the machine learning tools is that they are heuristic: they can learn from the data. Conversely, the mechanistic tools based on human knowledge require a previous process of knowledge extraction, which can be quite long. For instance, several of the collections of rules for mutagenicity, such as the Benigni-Bossa rules, are based on previous manual work, quite has been time consuming. The computer models can extract collections or rules in a matter of seconds and can identify rules which were not identified manually.

The computer approach has disadvantages, too. It uses parameters, such as molecular descriptors, which often are obscure in their meaning. Some descriptors have a clear meaning, and this is an advantage. However, also in this case, there is a difference, compared to the mechanistic models. In the case of the mechanistic models the role of the parameter is related to a mechanism (as we said, a likely mechanism, but anyhow there is a causal relationship between the parameter and the effect). Conversely, in the case of the computer approach, there is an association, which does not mean a causal one. Thus, after a model is built, we can explore the parameters used, and investigate individually their role. However, this is another difference between the two approaches. The parameters used in the computer approach in most cases play a joint role. Thus, it may be difficult to analyse them individually, and surely it is not correct. This complexity of the computer approach, which is an advantage, is at the same time a disadvantage if we look for a simple explanation.

The difficulty for the explanation is not only for the use of a large set of parameters. The use of complex machine learning tools, such as deep learning, also poses a great challenge to explanation, because the link between the inputs (such as the molecular parameters) and the output (such as the toxicity value) is mediated by multiple steps. Furthermore, the complex algorithms represent a barrier for the users.

Furthermore, the scheme of a typical model developed with machine learning or statistical tools depends on the training set, on the parameters, and on the algorithm. Thus, a huge amount of models can be developed, and several of them will have quite similar performance. Thus, there is a potential redundancy of models, and this fact is quite different from the mechanistic vision of a well--defined sequence of factors provoking the final effect.

As a conclusion, the computer models

- 👍 Can replicate the complex processes associated with the toxic effects
- 👍 Can model non-linear processes
- 👍 Can handle the interaction among multiple parameters concurring to the final outcome
- 👍 Can model quantitative values
- 👍 Can provide the uncertainty associated to the prediction
- 👍 Can be multitask
- 👍 Are heuristic
- 👎 Are often opaque
- 👎 The mathematical basis of the model is not familiar to most users
- 👎 Identify association, but not causality

Multiple models are similar, complicating the choice of the best model to be used.

Concluding remarks on mechanism and statistics

We have seen the differences between the two approaches. However, the borders between these categories are not as strict as it may appear. For instance, we may consider mechanistic models those on bioconcentration factor (BCF) derived from EPISuite. They are based on logKow (which has a mechanistic meaning) plus other factors, used to correct the predictions. The value for logKow is obtained using molecular descriptor and fragments with empirical coefficients – not mechanistic reasoning. Thus, even the mechanistic model is “contaminated” by empirical and statistical components. The CAESAR model (see below Part B), which is a “statistical” one, uses logKow plus few other descriptors. Thus, in practice, the difference between these models is quite low. Similarly, most of the structural alerts within the Benigni-Bossa rule set (see below, part B) are also present in the list of the fragments generated by SARpy. Thus, in practice, also in this case, the two systems are not so different, and they provide the same info. But even the statistical model may benefit from the knowledge on the mechanism, since in many cases the modeller uses descriptors which have been already reported to be associated with the effect, and thus the approach is not completely blind. But this cross-fertilization is useful, and it is not a matter of competition, but of collaboration, to achieve a higher goal.

Thus, this representation may be partial and misleading if not properly contextualized. The real challenge is to assess if a substance represents a risk or not. This requires exploiting all possible sources of evidence, since our knowledge is limited. It would be a pity to renounce one or the other approach, for antagonism between supporters of one approach. Indeed, authorities require applying both approaches, as in the case of the evaluation of impurities in pharmaceuticals [21]. From a theoretical point of view the integration of both visions, based on the theory or on the observations, is consistent with Galileo’s lesson, and successive elaboration. However, there is an important difference. When we use *in silico* models, we are dealing with virtual data, obtained by mechanistic or statistical models. In both cases, the prediction is based on assumptions. We assume that a certain mechanism applies to our substance (which is not proven) or we assume that our substance is similar to the other substances used to build up the machine learning model.

Thus, the virtuous circle from the observations to the theory, and then back from the theory to observations, in the case of *in silico* models is replicated by a parallel, virtual, circle:

- We have observations, i.e. data on similar substances in the case of read-across, which are not the real data on the target substance, because mediated by the similarity between the target and the similar substances.
- We have the theory, the mechanism. This is the same as in the classical scheme, but we have to remember that we assume that this mechanism is correct for the target substance, but this has not been proven: thus, the circle is not closed, because we do not have the validation of the theory.
- We have the predicted value. This value is supported by the observations on the substances in the training set, and by the theoretical aspects associated to the algorithm: descriptors,

relationships, threshold values, etc. depending on the model. The prediction combines data and theory in the same system.

Furthermore, the categorization in two kind of models, expert-based or statistical-models, does not consider all the kinds of models and their complexity. For instance, quite often there are statistical models which take advantage of some explicit knowledge, for instance because in the development of the model some specific models have been "suggested" to the model. Similarly, also the models classified as expert-based~~,~~ quite often benefit from empirical knowledge, and are based on some observation (see VEGA models in Part B for more discussion). Furthermore, many modern *in silico* models are obtained with machine learning, which is not a statistical approach. Finally, the KNN models are not parametric, thus not statistical ones. They are a kind of automatic read-across, based on similarity, without using descriptors.

Read-across or in silico models?

Read-across is used for the REACH registrations much more frequently than *in silico* models. Again, we recommend using all possibilities to evaluate a substance, not only a single approach. In Part B we will provide examples where we demonstrate that their combination is very useful. Anyhow, it is important to see which are the differences between the two approaches, and their limitations and advantages.

In principle, *in silico* models include the data used for read-across, and the algorithm which extracted the parameters explaining the behaviour of the substances. Thus, a good *in silico* model should contain all the elements used for read-across. However, this is the ideal case, but in practice within a population of many substances we have different situations and behaviours, and the overall model may fail to capture all the nuances associated with the different substances.

Let's imagine that we have a model based on a training set, and that within this training set there is only one substance containing germanium in its structure. If we want to evaluate a new substance containing germanium, the prediction will be based on the general behaviour of the whole population. If the substance with germanium in the training set has an effect not influenced by germanium, the use of the model to predict the unknown substance may be correct. However, if the germanium in the structure is responsible for the adverse effect, for instance,~~ ~~it is very likely that the model fails to learn this peculiar effect. In this case, the read-across approach may instead provide the useful information regarding the adverse effect associated with germanium. Note that here we simply rely on observations, not on mechanism, which has not been demonstrated.

The extreme case of germanium is clear, but more frequently we may have rare structural features (peroxide, particular heterocyclic rings, etc.) which may represent an issue, for the lack of information. How much we deviate from the well-known situation may be difficult to assess. Software like VEGA can immediately recognize this condition, through the tool for the applicability domain index (ADI) (see part B). Indeed, the ADI tool investigates exactly this: how "dense" and consistent is the information in the local area around the target substances, compared with the general population used by the model.

Hybrid models

As previously mentioned, it is likely that more than one model exists for the same endpoint. Thus, we may face the issue of how to compare the results. We may adopt simple or sophisticated strategies.

A simple approach is to take the worst-case scenario. If we have two models, even if only one of them predicts the chemical as toxic, we can follow the precautionary principle and assume the toxicity of the chemical. As an alternative, we can adopt the approach of the majority vote, in case of several results.

If we are dealing with continuous values, we could process them from a statistical point of view and take the average, or another statistical parameter. Alternatively, we could use the results only when they match, and disregard results where there is a conflict.

None of the above approaches takes into consideration the reliability of the individual models, which may be different. Thus, a more sophisticated way is to take into account the model reliability. A possible way to handle this is through the Bayesian system. Such an approach has been adopted for instance within the OSIRIS project, to integrate results from different models. An even more sophisticated approach also takes into account the individual results on the basis of the chemical, for the individual models.

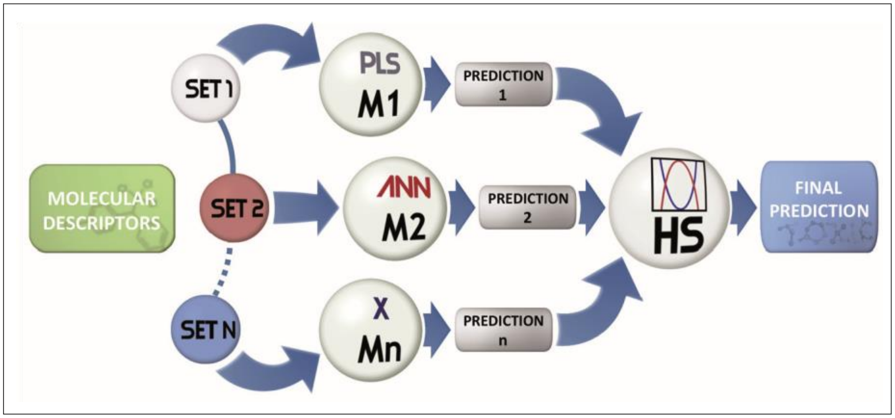


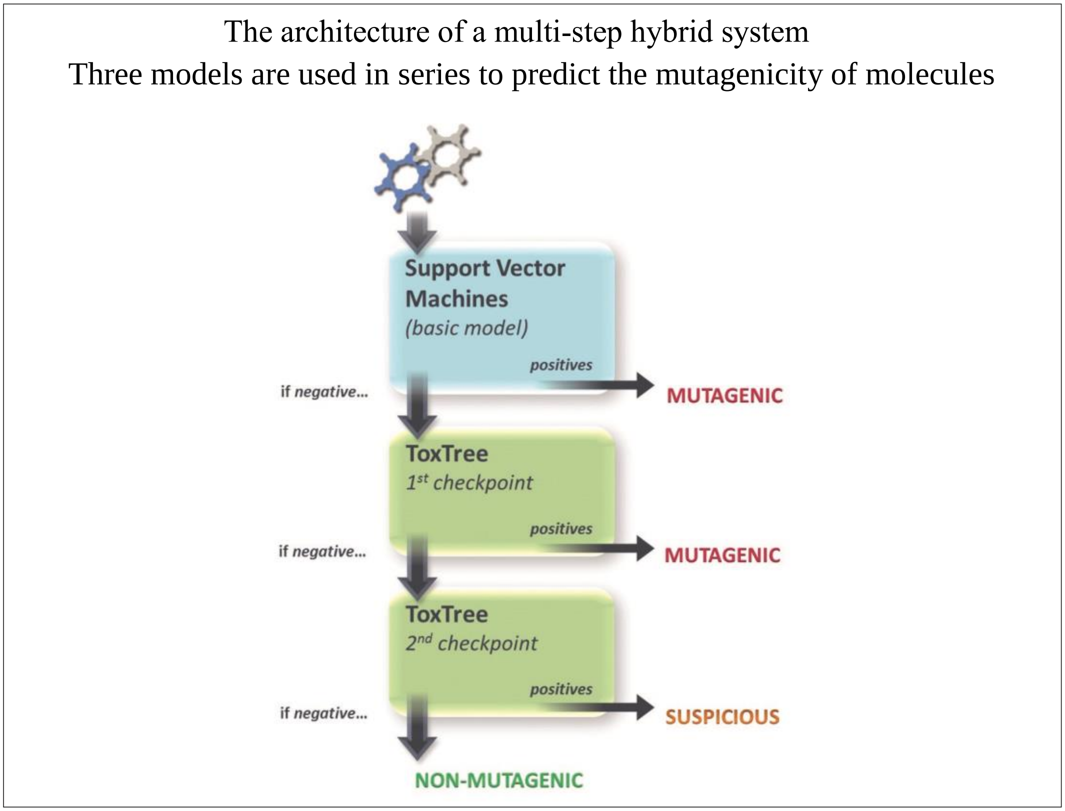
Figure 12: Multiple models can be utilised to obtain evaluation which will be then used as input by another model in a hybrid system approach.

This strategy has been adopted in the DEMETRA project, which developed and integrated a series of models [1]. The strategy of using the results of several QSAR models as input for a final model is depicted in *Figure 12*.

Another example is the T.E.S.T. model, which contains integrated models, called consensus models. The consensus model takes into account the results of the reliable models only, and this depends on the chemical compound.

CAESAR also developed hybrid models, which integrated the results of different models, depending on the chemical. Here we can discuss two different examples, adopting two strategies.

The CAESAR model for bioconcentration (*BCF*) is based on two separate models, which are combined. These two models use different descriptors and algorithms, providing different results. A third model, the hybrid one, uses the output of the two models as input instead of the chemical descriptors [4,5]. The strategy is similar to what represented in *Figure 13*.


The architecture of a multi-step hybrid system

Three models are used in series to predict the mutagenicity of molecules

Figure 13. A three-step hybrid system for mutagenicity prediction.

The CAESAR model for mutagenicity uses two models in cascade (*Figure 13*). The first model makes the prediction. If the output is “non-mutagenic”, at this point the chemical is processed by a second model. This is repeated once again, with a third model. At this point the output of the third model is: “non-mutagenic” or “suspicious”. Thus, in this case we have a sequence of models, which are switched depending on the output of the previous model.

These two strategies are quite well suited for two different purposes: a continuous output, or a category.

As a general comment, we notice that it is preferable to have combinations of models based on different approaches. This maximizes the exploitation of the approach. *Figure 14* shows the improvement of the results, using the hybrid models.

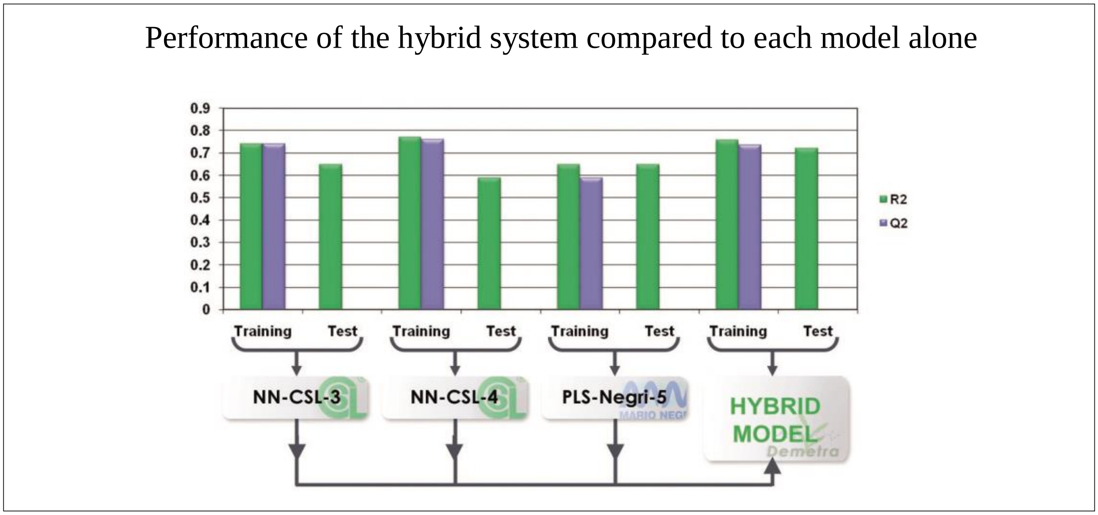

Performance of the hybrid system compared to each model alone

Figure 14: Hybrid system predictive performances compared to the single models.

For more discussion on the integration between different models, see [22] and the JANUS approach, in Part B.

Deep learning and further perspectives on in silico models

A broader vision, beyond QSAR

Models to explore biochemical mechanisms may follow different methods from those mentioned above, which address the prediction of the activity, and not necessarily of the mechanism. Expert modellers can explore complex situations using all their experience and subjective processes, even though the possibility of disseminating the procedure will be limited. Industry can of course use its own confidential data for internal purposes and the model will not suffer because of these conditions.

The scenario of QSAR models is very broad. Many techniques exist, which are not, strictly speaking, classical QSAR models. Methods such as docking offer the possibility of studying the interactions between a ligand and the receptor. Methods such as COMFA can investigate parts of the molecule involved in the toxicity process [23]. While QSAR models explore the hypothesis of a relationship between a certain chemical descriptor and the property without specific a priori mechanistic information, docking allows the introduction of specific knowledge to do with the biochemical environment in which the chemical should be active. Forces affecting the binding are used for modelling. The model is suitable when the property of interest is mediated by the binding and is very appealing in its capability to show the direct biochemical interaction. However, in the event that the process is more complex, and several steps are involved, binding alone may overlook important parts of the phenomenon to be studied.

Models such as COMFA are useful to identify the steric and electrostatic factors of the molecule affecting the process, showing the specific parts of the molecule where this occurs. There are examples of the useful integration of different methods, to better explore the toxicity phenomena and the factors involved in the phenomenon.

Deep learning

In the last decade deep learning (DL), which means learning from a deeply layered neural network (NN) structure [24], has been applied to various chemical problems. All those systems could access a large body of quality data for training the models, which makes DNNs tools of election for domains where large datasets are available, as for in vitro tests, genomics, proteomics and genetic data. One of the first of such models was for high-throughput screening in drug discovery, and used the Tox21 data, which included 12,000 chemicals for 12 different toxic effects [25].

DL methods are by nature apt to extract from raw data their implicit representation in terms of features. Today it is possible building QSARs without pre-computing chemical features, and instead using DL to learn directly from the chemical graph. Analysing the net it is possible to extract information that in some sense can be compared with Structural Alerts (SAs).

DL methods can offer the advantages of SAR and QSAR together. They apply self-extracted SAs as in SAR and make statistical predictions as in QSAR, so considering the whole molecule and also reasoning on its subparts. They better avoid the similarity paradox, which is the problem that similar molecules sometime do not have similar properties [26].

From neural networks to deep neural networks

The basic behaviour of NNs is to induce an implicit function from labelled data by optimizing the weights assigned to many individual simple computation units, called neurons, organized in a network. After a random initialization of the weights, training data are passed through the net, the output computed and compared to the true output to generate a value of the loss (error) function. The gradient of the error is used in backward direction to update the weights, and the training continues with new runs until a stopping criterion is applied.

Training a neural network is empirically done with many trials of various meta-parameters. Training requires to pass many times through the network the input training data, and this number is called epoch. Big data sets are divided into batches, whose size is also a parameter to set; iteration is the number of batches needed to complete one epoch. Another important parameter is the learning rate (a value in (0, 1)), which determines the step size to change the weights at each iteration.

Different regularization methods are used to avoid overfitting. For big networks and large data sets the most used is dropout during training. It means that nodes of the networks, with a predefined probability, are randomly deleted (in practice set to zero) so to avoid that the net learns too much of the data and eventually becomes unable to generalize.

DNNs are NNs with many layers; they can have hundreds of neurons and the number of weights to set can easily be in tens of thousands. For this reason various DNN architectures make the choice to use the same weights for more neurons, and pay much attention in reducing the computation time. The availability of GPUs, apt to speed up matrix operations, makes the training of such big networks easily affordable.

The architectural choices needed to fully define a DNN are the number of neurons, the type of computation each neuron does, the learnable parameters, the number of layers, the parameters shared across neurons, and the transfer functions.

The basic DNN architectures are Convolutional Neural Network (CNN), proposed for image understanding, Recurrent Neural Network (RNN), proposed for text analysis, and Graph Convolutional Network (GCN) used to compute properties on graphs.

Convolutional NN (CNN)

A CNN is a network that breaks down an input, typically an image, into smaller pieces and extracts the feature to be used to make a classification decision [24].

CNNs combine convolutional layers, pooling layers, fully connected layers and activation layers (*Figure 15*).

- Convolutional layers use a set of fixed-sized weight matrices, called filters, which perform element-wise multiplication on the image pixels. Weights are learned.
- Pooling layers usually come after or in between convolutional layers to reduce the dimension of the original input. Average pooling smoothens the features taking the average, while max pooling picks the largest value to extract distinctive features.
- Fully connected layers flatten the previous layer and connect to all nodes of the previous layer to each of its nodes. Then output is fed into a non-linear layer.

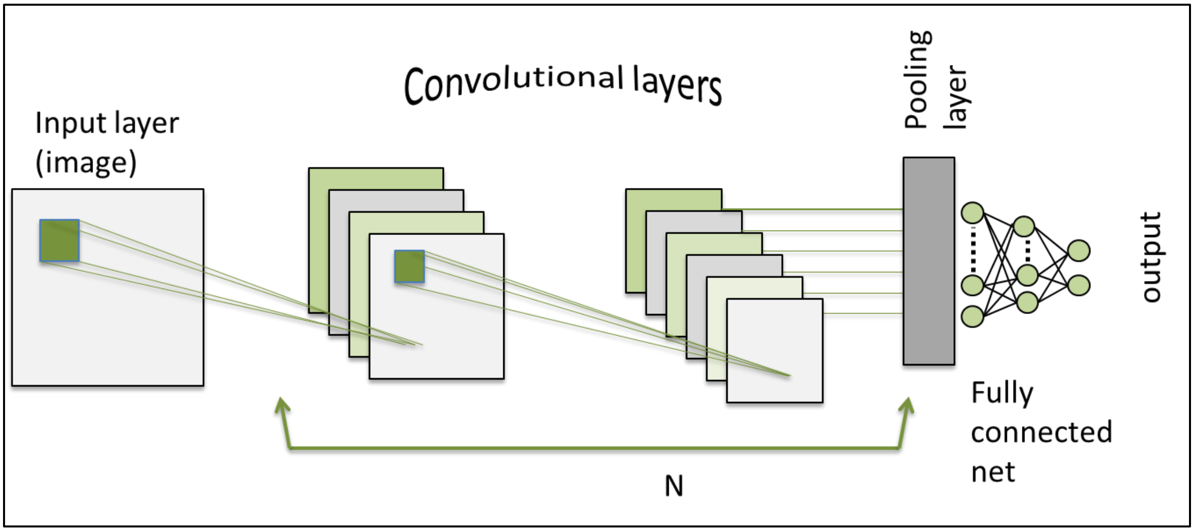


Figure 15. The CNN, with N convolutional layers.

Recurrent NN (RNN)

Feedforward networks can receive a fixed sized number of data to the hidden layers. The inability to handle variable length input, as in case of texts, and the necessity to consider long-term dependencies give rise to using recurrent units with feed forward neural network.

RNNs [27], in *Figure 16*, use the same function and the same set of parameters for every time step: at each time step, the previous hidden state and the current input is fed through the function to update the hidden state. The loss is defined as the sum of the loss from each time step. RNNs are effective in using the temporal interconnections present in the data.

Backpropagation through time (BPTT) is used for training.

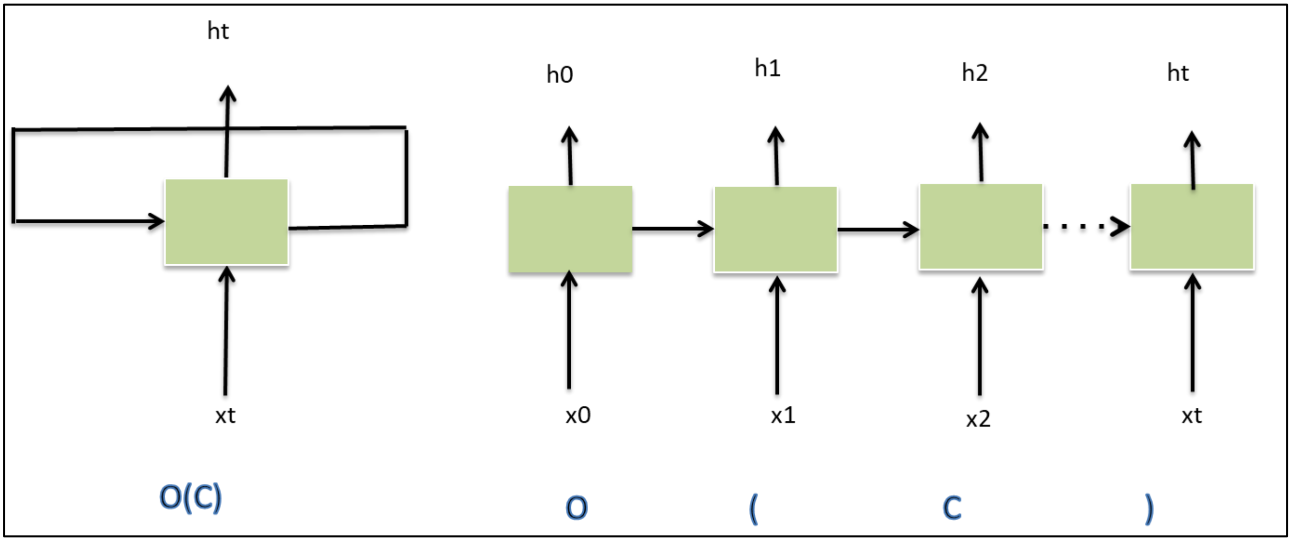


Figure 16. RNN as a cycle and unfolded at each time step.

Graph convolutional neural networks (GCN)

A fully connected feedforward NN can input a graph of N nodes to N neurons, losing the information about the edges. Representing graphs as adjacency matrices instead allows considering the connections, but at the cost of using N^2 values instead of N. To make the computation affordable even for large graphs it is necessary to share parameters and weights in the NN.

GCN [28] receive in input graphs adjacency matrices and apply the same winning principles of the CNN: shared weights, and multi-layer refinement. Images are represented as matrices, the products of pixels and weights. Graphs too are represented as matrices, the product of adjacency matrix and weights. A GCN in general considers an adjacency matrix concatenated with a matrix of node features, and instead of a window to select the neighbouring pixels it uses an aggregation function.

The GCN, as in *Figure 17*, contains convolutional blocks, readout layer, and fully connected layers. Convolutions, followed by filtering and pooling, reduce the input to extract the features using the same weights or parameters for different neurons attached to different parts of the input or to neurons in previous layers. The GCN applies a filter over the graph to look for essential vertices and edges; pooling (an operation as max, mean or sum) generates a smaller graph, where higher-level features emerge.

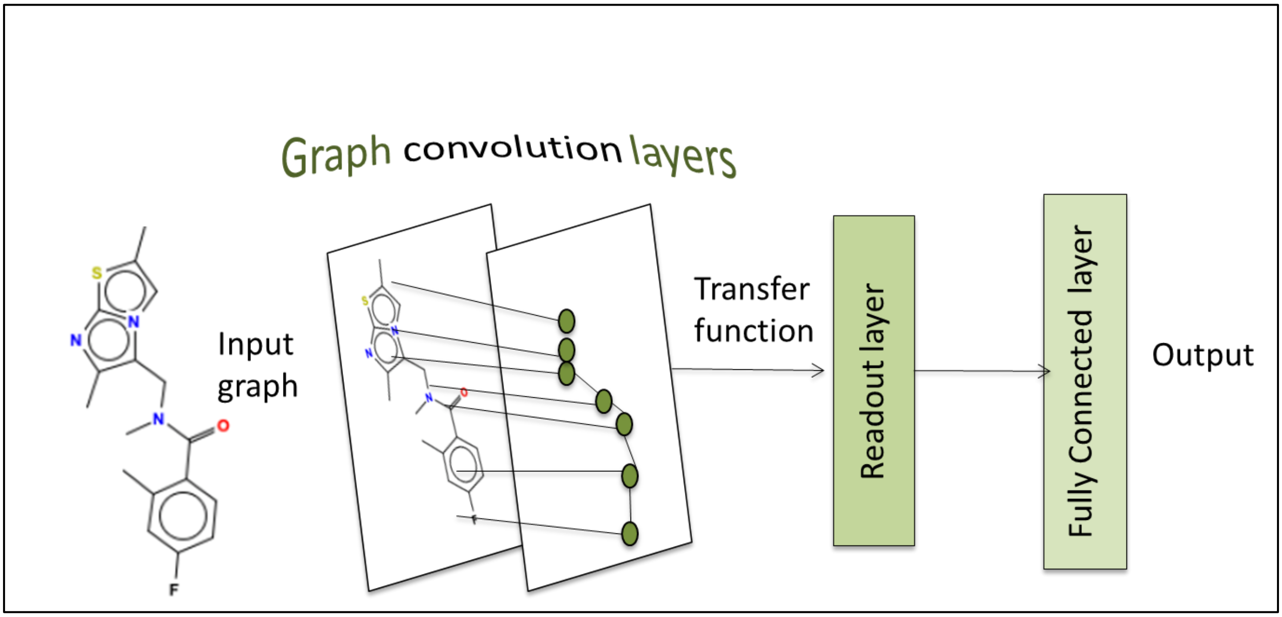


Figure 17. The structure of a GCN.

A way to make more understandable the output of a DNN is to add the attention mechanism. In animals it means focusing on a specific part of the sensorial data to help interpreting the scene. In the same way, a neural network can define the important parts of the input it receives in order to make a correct prediction. This can be obtained adding a layer connected to the network that receives the context vector and calculates the weight this has on the final prediction. Examples of this use are in the various models developed with NN, RNN, and GCN for Ames mutagenicity [29, 30].

Chapter references

1. Benfenati E., Clook M., Fryday S. and Hart A., QSARs for regulatory purposes: the case for pesticide authorization, in: Benfenati E. (Ed.), Quantitative Structure-Activity Relationships (QSAR) for Pesticide Regulatory Purposes, Elsevier Science Ltd, Amsterdam, The Netherlands (2007), 1-57
2. Dimitrov A., Dimitrova N., Parkerton T., Comber M., Bonnell M. & Mekenyan O., Base-line model for identifying the bioaccumulation potential of chemicals, SAR QSAR Environ Res. 2005 Dec;16(6):531-54.
3. Benfenati E., Boriani E., Craciun M., Malazizi L., Neagu D., Roncaglioni A., Databases for pesticide ecotoxicity, in: Benfenati E. (Ed.), Quantitative Structure-Activity Relationships (QSAR) for Pesticide Regulatory Purposes, Elsevier Science Ltd, Amsterdam, The Netherlands (2007), 59-81
4. Zhao C., Boriani E., Chana A., Roncaglioni A. and Benfenati E., A New Hybrid QSAR Model for Predicting Bioconcentration Factor (BCF), Chemosphere 2008, 73:1701-1707.
5. EC funded project CAESAR (Computer Assisted Evaluation of industrial chemical Substances According to Regulation) - http://www.caesar-project.eu/
6. EC funded project DEMETRA (Development of Environmental Modules for Evaluation of Toxicity of pesticides Residues in Agriculture)
7. US EPA DSSTox (Distributed Structure-Searchable Toxicity) Database Network - https://www.epa.gov/chemical-research/distributed-structure-searchable-toxicity-dsstox-database
8. US EPA CompTox Chemicals Dashboard. https://www.epa.gov/chemical-research/comptox-chemicals-dashboard
9. OECD QSAR Toolbox - https://www.oecd.org/chemicalsafety/risk-assessment/oecd-qsar-toolbox.htm
10. The Results Gateway of the CONCERT REACH LIFE Project - https://www.life-concertreach.eu/results/results-gateway/
11. The IUPAC International Chemical Identifier (InChI) - https://publications.iupac.org/PAC2/PAC%20-%20IUPAC/old2015.iupac.org/home/publications/e-resources/inchi.html
12. Daylight SMILES (Simplified Molecular Input Line Entry System) - https://www.daylight.com/dayhtml/doc/theory/theory.smiles.html
13. Todeschini R. & Consonni V., Handbook of Molecular Descriptors, J. Wiley & Sons, New York, 2008, 688 pp.
14. Benfenati E., Casalegno M., Cotterill J., Price N., Spreafico M., and Toropov A., Characterization of chemical structures, in: Benfenati E. (Ed.), Quantitative Structure-Activity Relationships

(QSAR) for Pesticide Regulatory Purposes, Elsevier Science Ltd, Amsterdam, The Netherlands (2007), 83-109.

15. Devillers J., Genetic Algorithms in Molecular Modeling (Principles of QSAR and Drug Design), Elsevier Science Ltd, Amsterdam, The Netherlands, 1996, 327 pp.
16. Devillers J., Neural Networks in QSAR and Drug Design (Principles of QSAR and Drug Design), Academic Press, London, UK, 1996, 284 pp.
17. Ferrari T., Gini G., Bakhtyari N.G. & Benfenati E., Mining toxicity structural alerts from SMILES: A new way to derive Structure Activity Relationships, in: Computational Intelligence and Data Mining (CIDM), 2011 IEEE Symposium on, 2011, pp. 120-127.
18. Livingstone D.J., Rahr E., Corchop – an Interactive Routine for the Dimension Reduction of Large QSAR Data Sets, Quant. Struct.-Act. Relat. 2006, 8(2):103-108.
19. OECD Principles for the Validation, for Regulatory Purposes, of (Quantitative) Structure-Activity Relationship Models. Paris, France. https://www.oecd.org/chemicalsafety/risk-assessment/37849783.pdf
20. Jevremović D, Trinh CT, Srienc F, Sosa CP, Boley D. Parallelization of Nullspace Algorithm for the computation of metabolic pathways. Parallel Comput. 2011 Jun;37(6-7):261-278. doi: 10.1016/j.parco.2011.04.002.
21. ICH Harmonised Tripartite Guideline (2017) Assessment and control of DNA reactive (mutagenic) impurities in pharmaceuticals to limit potential carcinogenic risk - M7. Current Step 4 version.
22. Benfenati E., Chaudhry Q., Gini G., Dorne J.L., Integrating in silico models and read-across methods for predicting toxicity of chemicals: A step-wise strategy. Environ Int. 2019 Oct;131:105060. doi: 10.1016/j.envint.2019.105060.
23. Roncaglioni A., Benfenati E., Computer-aided methodologies to predict endocrine-disrupting potency of chemicals in Shaw I (Ed.), CRC Press, Boca Raton (2009), 306-321.
24. LeCun, Y. & Bengio, Y. (1995). Convolutional networks for images, speech, and time series, in M.A. Arbib (Ed) The handbook of brain theory and neural networks, vol. 3361, no. 10.
25. Mayr, A., Klambauer, G., Unterthiner , T.& Hochreiter, S. (2016). DeepTox:Toxicity Prediction using Deep Learning. Frontiers in Environmental Science, 3: 80.
26. Gini, G. (2020). The QSAR similarity principle in the deep learning era: Confirmation or revision? Foundations of chemistry, 22: 383–402.
27. Williams, R. J., Hinton, G. E.; Rumelhart, D. E., 1986. Learning representations by back-propagating errors. Nature. 323: 533–536.
28. Wu, Z.; Pan, S.; Chen, F.; Long, G.; Zhang, C.; Yu, P. S. (2020). A Comprehensive Survey on Graph Neural Networks. IEEE Transactions on Neural Networks and Learning Systems, 1-21.

29. Gini, G., Zanoli, F., Gamba, A., Raitano, G. & Benfenati, E. (2019). Could deep learning in neural networks improve the QSAR models? SAR&QSAR in Environmental Research 30(9), 617-642.
30. Gini, G., Hung, C., Benfenati, E. (2022) Big data and deep learning: extracting and revising chemical knowledge from data. In S. Basak and M. Vracko (Eds) Big Data Analytics in Chemoinformatics and Bioinformatics", Elsevier, 2022, p 115-150, *doi: 10.1016/B978-0-323-85713-0.00030-X*

Emilio Benfenati, Anna Lombardo, Erika Colombo, Gianluca Selvestrel, Edoardo Viganò

Istituto di Ricerche Farmacologiche

Mario Negri IRCCS, Milano Italy

PART B

Using VEGAHUB

B1. What is available within VEGAHUB?

VEGAHUB is a platform of tools for the evaluation of chemical substances using *in silico* tools. It includes tools suitable for specific endpoints or programs to develop your own model.

Pre-built models

VEGAHUB offers tools for:

- QSAR
- Read-across

- Weight-of-evidence wrapping results from QSAR and read-across
- Prioritization
- Risk assessment
- Substitution of substances of concern

Among these tools, some have been developed to be used without a particular deep experience. We will focus on them in the following chapters.

Furthermore, there are tools to build-up your new models.

Programs to build-up new models

- SARpy
- QSARpy
- CORAL
- aiQSAR
- and several links to other systems.

We will not discuss these tools to build-up new models here.

B2. The VEGA models

VEGA provides *in silico* models for the evaluation of the properties of chemical substances. There are more than 100 models, continuously updated. They can be split into five categories considering five kinds of properties:

- Toxicity
- Ecotoxicity
- Environmental properties
- Physico-chemical properties
- Toxicokinetics

For several endpoints, there is more than one model available. In this case, we recommend using all of them.

The **VEGA** models derive from several sources, and in many cases have been developed within EC projects, listed at the VEGAHUB website (https://www.vegahub.eu/community/). There are also models obtained from other systems, such as T.E.S.T., EPISuite, OPERA, QSARINS, and Toxtree. Indeed, we promote exchange of tools within the broad community of model developers, and the VEGA models are available not only within VEGAHUB, but also within other systems, such as AMBIT, the Danish QSAR Database, MERLIN-Expo, SILIFOOD, CLICC (https://clicc.ucsb.edu/clicc-tool)and the OECD QSAR Toolbox.

The VEGA models have been built up using different collections of substances, different molecular descriptors, and different algorithms. For instance, some models have been built up using the SARpy (see Part A) software, others CORAL (see Part A), or KNN. Some models are classifiers, others are regression models. Within VEGAHUB, the user can get the complete description of the model according to the QSAR Model Reporting Format (QMRF; https://www.vegahub.eu/portfolio-item/vega-qsar-models-qrmf/).

Few are mechanistic models:

- Mutagenicity (Ames test) model (ISS) (version 1.0.2)
- Carcinogenicity model (ISS) (version 1.0.2)
- Skin Sensitization model (TOXTREE) - ver. 1.0.0
- Cramer classification (TOXTREE) (version 1.0.0)
- Verhaar classification (TOXTREE) (version 1.0.0)
- Skin Permeation (LogKp) model (Potts and Guy)
- BCF model (Arnot-Gobas) (version 1.0.0)
- BCF model (Meylan) (version 1.0.3) (it includes empirical parameters)
- kM/Half-Life model (Arnot/EpiSuite) (version 1.0.0) (it includes empirical parameters)

Few are a combination of mechanistic and statistical/machine learning models:

- Mutagenicity (Ames test) model (CAESAR)
- Mutagenicity (Ames test) CONSENSUS model (version 1.0.3)
- Skin Sensitization model (NCSTOX)

We have discussed in Part A that the borders between these two kinds of models are not always well defined.

The VEGA models can be used individually. They are also included in other programs within VEGAHUB, such as JANUS and VERMEER. Using these systems, the user can automatically run tens of models for specific purposes. This can be useful, also because in this way the results of multiple models for the same endpoint are already wrapped into a single property value, considering the uncertainty associated with each specific prediction. Thus, we recommend using JANUS also for this purpose.

For each model, VEGA includes an independent software helping the user in the evaluation of the result, thus VEGA offers both the prediction and the measurement of the reliability of the predictions, through the Applicability Domain Index.

B2.1. The Applicability Domain Index

Any **QSAR** model is based on three pillars: the property to be studied (for instance mutagenicity), the chemical information, and the function linking the property and the chemical (see Part A). Thus, to evaluate the prediction, we developed parameters which refer to each of these three components of the conceptual **QSAR** models: the property, the chemical information, and the algorithm. These parameters are then merged into a single value, called applicability domain index (ADI).

Other programs address the applicability domain in a qualitative way (in or out), some also in a quantitative way with a numerical value, most refers to the chemical similarity only, but the tool developed within **VEGA** is unique since it conceptually refers to all the pillars of the **QSAR** model: the chemical information, the property, and the algorithm linking the chemical and the property.

The ADI is a guidance for the user. As we said, the ADI tool investigates how “dense” and consistent is the information in the local area around the target substances, compared with the general population of substances used by the model.

Within a strategy of prioritization and screening of large series of substances, the ADI can be applied as a first filter, whereas the manual assessment described below cannot be applied for large numbers of substances. Indeed, **QSAR** models can be used for screening, or for the assessment of a single substance, and the role of the user is much more important in the second case.

We have noticed that the use of ADI is useful to improve the reliability of the predictions, working on collections of substances, for different endpoints. Indeed, we demonstrated that when the ADI value is high, the predictions are better, when it is low, there are more errors, and when it is medium, there are less errors than when the ADI is low. See full details in this open article: https://doi.org/10.3390/ijms24129894.

However, the user may disagree on the prediction, or on the relevance of certain score indicated within the ADI. Below we provide some examples, to better explain the use of all the pieces of information provided by **VEGA**, and what the user should do.

The elements to be evaluated for the ADI

The similarity of the related compounds. This relates to the chemical information of the model. Similarity depends on many factors, and there is no absolute measurement for it. We optimized the algorithm of similarity used in **VEGA** on the basis of a check with 4 million compounds. This is an advantage of **VEGA** compared with other programs. The similarity is calculated as described (http://jcheminf.springeropen.com/articles/10.1186/s13321-014-0039-1). The software calculates how similar the similar compound is providing a score between 1 (in case of identity) and 0. Values of 0.9 for similarity indicate a good similarity. Usually values lower than 0.75 indicate that the similar compound has important differences compared to the target.

The presence of unusual fragments. This relates to the chemical information of the model. **VEGA** identifies the presence of rare fragments, not common in the set of compounds at the basis of the specific model. Thus, this factor identifies a lack of knowledge on a component present in the molecule.

The check of the descriptor range. This relates to the information on the algorithm of the model. **VEGA** evaluates if the descriptors of the target compound have values in the range of those related to the substances in the training set. The molecular weight is also checked.

The sensitivity analysis of the descriptors. This check is used only in very few models. It relates to the information on the algorithm of the model. **VEGA** evaluates if a change of 10% of the descriptor values of the target compound provokes a large variation of the predicted value. This indicates an area of higher uncertainty of the model, associated with possible activity cliffs. This algorithm included in **VEGA** is quite unique.

The concordance between the predicted value of the target compound, and the experimental values of the similar compounds. This relates to the information on the toxicity/property value. The experimental values can be used alone, without the predicted values, and thus, in practice, it can be used for read-across. If there is agreement between the predicted value and the experimental values of the similar compounds, this is the ideal situation. If there is disagreement, the user should decide if there are sufficient elements to take a decision. If part of the information is not useful (see the example below), the user may disregard part of the information and work with the remaining information, if sufficient and convincing. In this process, the user should carefully evaluate the reasons for the effect, also considering the eventual presence of structural alerts, as below explained.

The accuracy of the prediction. This relates to the information on the toxicity/property value. This parameter indicates if the specific **VEGA** model face problems in the prediction of similar compounds, and thus if in the local area of the target compound there may be issues. This parameter identifies activity cliffs.

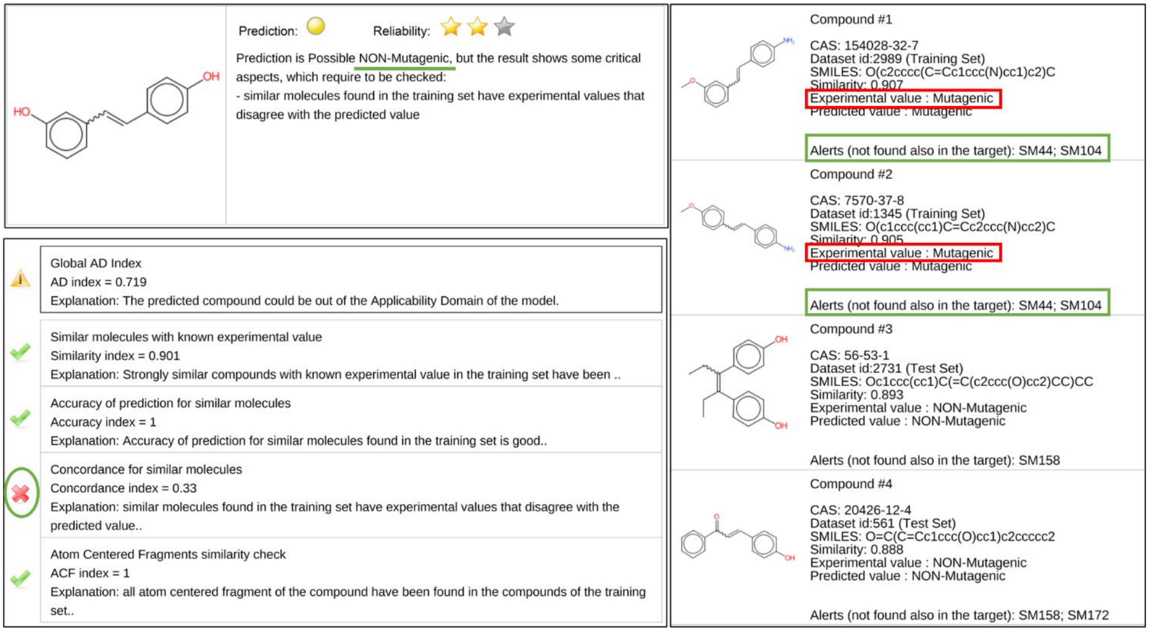
The maximum error in prediction. This relates to the information on the toxicity/property value. In case of models for continuous values, **VEGA** provides this parameter, which indicates if uncertainty is very high.

The presence of structural alerts. This relates to the information on the toxicity/property value. There are structural alerts (SA) indicated by **VEGA.** Each model has its own series of SA (if you want to use the complete list of SA you should use ToxRead). In the report, VEGA will show the three most similar substances with each of the SA found in the target substance. In this way, you can evaluate the reliability of each SA for the specific case. The presence of SA does not modify the ADI value. They are provided for the manual evaluation.

The similar compounds should be used also within a read-across perspective. Indeed, **VEGA** combines **QSAR** and read-across, but the user may decide to use only the **QSAR** result or the read-across approach. As much as possible we recommend using both approaches, within a weight-of-evidence strategy. This reinforces the assessment. If the similar compounds are not so similar, it is inappropriate to use read-across.

BOX 1

Concordance for similar molecules



Prediction: ● Reliability: ★★☆

Prediction is Possible NON-Mutagenic, but the result shows some critical aspects, which require to be checked:
- similar molecules found in the training set have experimental values that disagree with the predicted value

⚠ Global AD Index
AD index = 0.719
Explanation: The predicted compound could be out of the Applicability Domain of the model.

✓ Similar molecules with known experimental value
Similarity index = 0.901
Explanation: Strongly similar compounds with known experimental value in the training set have been ..

✓ Accuracy of prediction for similar molecules
Accuracy index = 1
Explanation: Accuracy of prediction for similar molecules found in the training set is good..

✗ Concordance for similar molecules
Concordance index = 0.33
Explanation: similar molecules found in the training set have experimental values that disagree with the predicted value..

✓ Atom Centered Fragments similarity check
ACF index = 1
Explanation: all atom centered fragment of the compound have been found in the compounds of the training set..

Compound #1

CAS: 154028-32-7
Dataset id:2989 (Training Set)
SMILES: O(c2cccc(C=Cc1ccc(N)cc1)c2)C
Similarity: 0.907
Experimental value : Mutagenic
Predicted value : Mutagenic

Alerts (not found also in the target): SM44; SM104

Compound #2

CAS: 7570-37-8
Dataset id:1345 (Training Set)
SMILES: O(c1ccc(cc1)C=Cc2ccc(N)cc2)C
Similarity: 0.905
Experimental value : Mutagenic
Predicted value : Mutagenic

Alerts (not found also in the target): SM44; SM104

Compound #3

CAS: 56-53-1
Dataset id:2731 (Test Set)
SMILES: Oc1ccc(cc1)C(=C(c2ccc(O)cc2)CC)CC
Similarity: 0.893
Experimental value : NON-Mutagenic
Predicted value : NON-Mutagenic

Alerts (not found also in the target): SM158

Compound #4

CAS: 20426-12-4
Dataset id:561 (Test Set)
SMILES: O=C(C=Cc1ccc(O)cc1)c2ccccc2
Similarity: 0.888
Experimental value : NON-Mutagenic
Predicted value : NON-Mutagenic

Alerts (not found also in the target): SM158; SM172

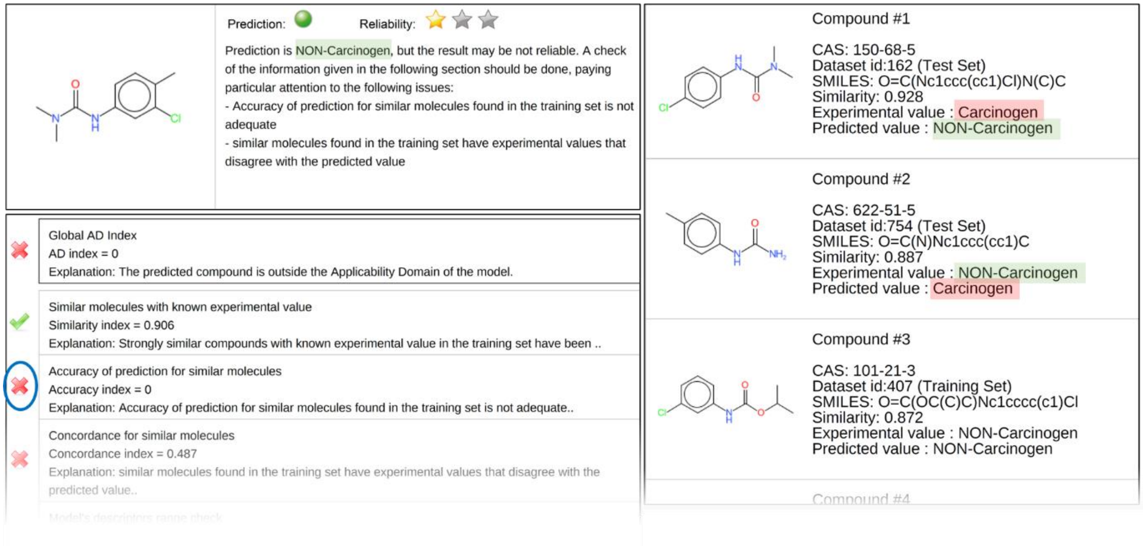
Figure B1. The concordance for similar molecules

In the example of *Figure B1*, the target substance is on upper left side of the figure, while on the right there are four similar substances. The model predicts the target substance as Possible NON-Mutagenic, while the ADI tool in VEGA shows that the first two similar substances are mutagenicity. As a consequence, the Concordance value is only 0.33 (the value uses the first three substances and two substances out of three are mutagenic). However, VEGA indicates that the first two substances contain alerts which are not present in the target. These two alerts are related to the aromatic amine (VEGA indicates that there are alerts not present in the target substance; in the report of the VEGA pdf report of the target substance VEGA shows only the alerts present in the target; to see the alerts present only in the similar substance, you can copy and paste the SMILES of the similar substance and run VEGA, the same model, the SARpy model for mutagenicity,

BOX 1

in this case). Since the target substance is not an aromatic amine, you can disregard the first two similar substance in your evaluation and do not consider the potential issue indicated by the Concordance parameter. As a consequence, the overall ADI will be high.

Accuracy of prediction for similar molecules



Prediction: ● Reliability: ★☆☆

Prediction is NON-Carcinogen, but the result may be not reliable. A check of the information given in the following section should be done, paying particular attention to the following issues:
- Accuracy of prediction for similar molecules found in the training set is not adequate
- similar molecules found in the training set have experimental values that disagree with the predicted value

✗ Global AD Index
AD index = 0
Explanation: The predicted compound is outside the Applicability Domain of the model.

✓ Similar molecules with known experimental value
Similarity index = 0.906
Explanation: Strongly similar compounds with known experimental value in the training set have been ..

✗ Accuracy of prediction for similar molecules
Accuracy index = 0
Explanation: Accuracy of prediction for similar molecules found in the training set is not adequate..

✗ Concordance for similar molecules
Concordance index = 0.487
Explanation: similar molecules found in the training set have experimental values that disagree with the predicted value..

Model's descriptors range check

Compound #1

CAS: 150-68-5
Dataset id:162 (Test Set)
SMILES: O=C(Nc1ccc(cc1)Cl)N(C)C
Similarity: 0.928
Experimental value : Carcinogen
Predicted value : NON-Carcinogen

Compound #2

CAS: 622-51-5
Dataset id:754 (Test Set)
SMILES: O=C(N)Nc1ccc(cc1)C
Similarity: 0.887
Experimental value : NON-Carcinogen
Predicted value : Carcinogen

Compound #3

CAS: 101-21-3
Dataset id:407 (Training Set)
SMILES: O=C(OC(C)C)Nc1cccc(c1)Cl
Similarity: 0.872
Experimental value : NON-Carcinogen
Predicted value : NON-Carcinogen

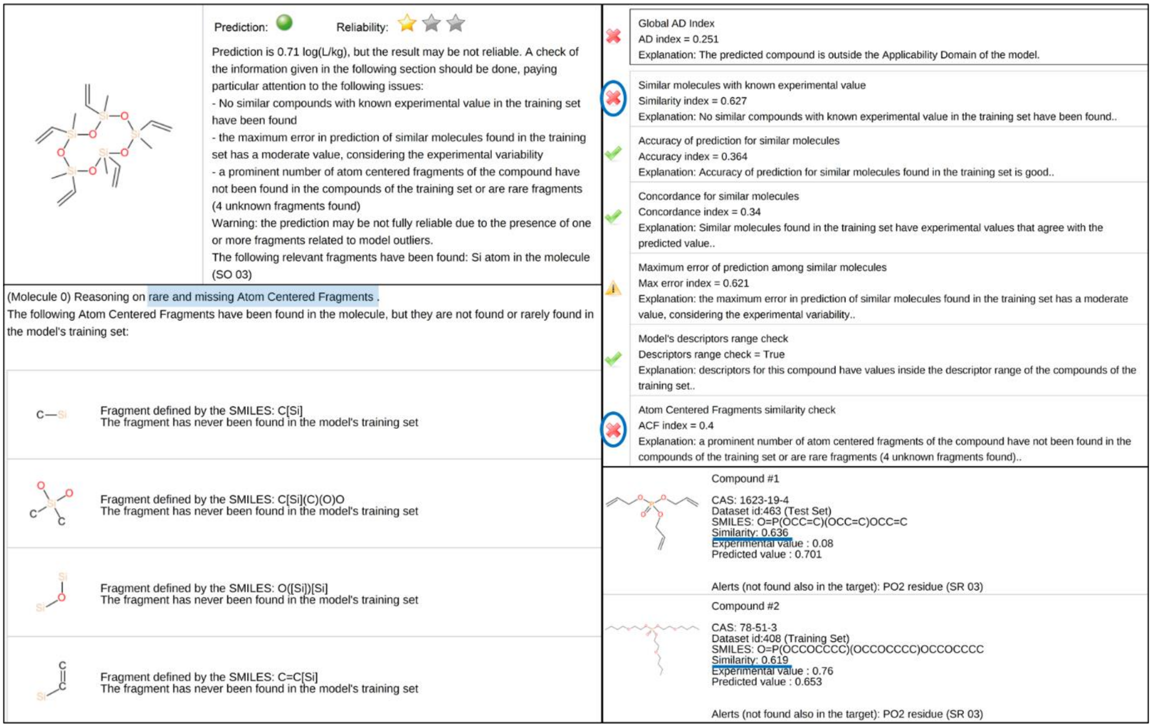
Compound #4

Figure B2. The accuracy of prediction for similar molecules

In this example (*Figure B2*), the target is predicted NON-Carcinogen with low reliability due to several reasons, included the low accuracy of prediction for similar molecules. Looking at the similar molecules, the first two are wrongly predicted. This leads to an accuracy index of 0. The third similar is correctly predicted but is less similar (it has a carbamate instead of the urea group). In this case, it is not possible to reach a reliable conclusion on this basis.

BOX 1

Atom Centered Fragments



Prediction: ● Reliability: ★☆☆

Prediction is 0.71 log(L/kg), but the result may be not reliable. A check of the information given in the following section should be done, paying particular attention to the following issues:
- No similar compounds with known experimental value in the training set have been found
- the maximum error in prediction of similar molecules found in the training set has a moderate value, considering the experimental variability
- a prominent number of atom centered fragments of the compound have not been found in the compounds of the training set or are rare fragments (4 unknown fragments found)

Warning: the prediction may be not fully reliable due to the presence of one or more fragments related to model outliers.
The following relevant fragments have been found: Si atom in the molecule (SO 03)

(Molecule 0) Reasoning on rare and missing Atom Centered Fragments .
The following Atom Centered Fragments have been found in the molecule, but they are not found or rarely found in the model's training set:

Fragment defined by the SMILES: C[Si]
The fragment has never been found in the model's training set

Fragment defined by the SMILES: C[Si](C)(O)O
The fragment has never been found in the model's training set

Fragment defined by the SMILES: O([Si])[Si]
The fragment has never been found in the model's training set

Fragment defined by the SMILES: C=C[Si]
The fragment has never been found in the model's training set

✗ Global AD Index
AD index = 0.251
Explanation: The predicted compound is outside the Applicability Domain of the model.

✗ Similar molecules with known experimental value
Similarity index = 0.627
Explanation: No similar compounds with known experimental value in the training set have been found..

✓ Accuracy of prediction for similar molecules
Accuracy index = 0.364
Explanation: Accuracy of prediction for similar molecules found in the training set is good..

✓ Concordance for similar molecules
Concordance index = 0.34
Explanation: Similar molecules found in the training set have experimental values that agree with the predicted value..

⚠ Maximum error of prediction among similar molecules
Max error index = 0.621
Explanation: the maximum error in prediction of similar molecules found in the training set has a moderate value, considering the experimental variability..

✓ Model's descriptors range check
Descriptors range check = True
Explanation: descriptors for this compound have values inside the descriptor range of the compounds of the training set..

✗ Atom Centered Fragments similarity check
ACF index = 0.4
Explanation: a prominent number of atom centered fragments of the compound have not been found in the compounds of the training set or are rare fragments (4 unknown fragments found)..

Compound #1

CAS: 1623-19-4
Dataset id:463 (Test Set)
SMILES: O=P(OCC=C)(OCC=C)OCC=C
Similarity: 0.636
Experimental value : 0.08
Predicted value : 0.701

Alerts (not found also in the target): PO2 residue (SR 03)

Compound #2

CAS: 78-51-3
Dataset id:408 (Training Set)
SMILES: O=P(OCCOCCCC)(OCCOCCCC)OCCOCCCC
Similarity: 0.619
Experimental value : 0.76
Predicted value : 0.653

Alerts (not found also in the target): PO2 residue (SR 03)

Figure B3. The Atom Centered Fragments

The ADI not only indicates if similar chemicals are present in the dataset of the model but identifies if rare or unknown fragments are present in the target. In the example above (BCF, *CAESAR* model), the target is a siloxane, a group of chemicals completely unknown to the model as indicated by the list of four unknown fragments. The presence of rare or unknown fragments represents a clear limit of the models, because it has few or no elements to evaluate the target chemical properly.

B2.2. The VEGA report (PDF version)

VEGA provides the user with all the information regarding the prediction and on how to interpret it in a single pdf report. This report provides the prediction, plus many other pieces of information, which have to be taken into account. On a conceptual point of view, there are three separate lines of evidence provided by VEGA:

- The prediction
- The similar substances – to be used for read-across
- The reasoning.

Each of these lines of evidence should be evaluated separately first, and then integrated looking at the complete set of lines of evidence. Please refer to the EFSA Guidance on Weight-of-evidence for a detailed description of the phases for this process (*https://www.efsa.europa.eu/en/efsajournal/pub/4971*) and to what we discussed in two sessions of Part A (The general theoretical context and Concluding remarks).

We suggest starting with the observations, the experimental data on similar substances. This element is very powerful: it refers to the real values, it is very convincing. If the model predicts not toxic, but there is similar substance which is toxic, this should be very carefully analysed, to explain while the similar substance is toxic: it may be that it contains a structural alert responsible for the effect, which is not present in the target substance, and in this case we may disregard the similar substance. See the example we discussed for the Concordance of the ADI).

It may be that the prediction is toxic, but the similar substance is not. In this case too we have to understand if there are reasons to explain the difference. If the similar substance is toxic, but the prediction for the target is not toxic, and we have not explanation to explain the different behaviour of the target, it is preferable to assume that the target is toxic too. Indeed, the observation prevails. At this point it becomes important to analyse the similarity. Similarity is a very local concept, it is powerful when the similarity is strong, but it declines rapidly. Indeed, on a mathematical point of view, similarity is not transitive: if A is similar to B, and B is similar to C, we cannot conclude that A is similar to C. In practical terms, the argumentation using VEGA is good if the similarity is 0.9 or higher. If it is 0.85 it may be questionable, and if it is lower than 0.85 it can be used only as support evidence. We have to underline that we are speaking about structural similarity only. In a complete reasoning, it is not sufficient, other aspects have to be considered (e.g. metabolism).

Let’s consider the reverse case, of a similar substance not toxic, and a prediction indicating toxicity. In this case, the point to be carefully verified is the basis of the *in silico* prediction, its reliability, if there is a theoretical explanation of the toxic effect. The choice between conflicting lines of evidence should be evaluated referring to the EFSA Guidance (*https://www.efsa.europa.eu/en/efsajournal/pub/4971*). Conflicting values increase uncertainty. The final decision may be adopted in a conservative way.

The reasoning is the second line of evidence that we should evaluate. If from a theoretical point of view there are reasons to assume that a substance is toxic, we should carefully evaluate this fact. This element may be represented by the presence of a structural alert (SA). We have to remember that there is a prevalence of toxic substances for each SA and that in some cases most of the substances with a SA are not toxic; thus, we have many false negatives. If we want to know the prevalence of toxic substances for the SA of interest, we can be seen it using ToxRead, if there is the model for that endpoint. The prevalence provides the value on the whole population of substances. What is more interesting is the local situation on similar substances. VEGA shows the similar substances and their experimental value for each SA in the target. This information should be used to evaluate the reliability of the SA for the target substance.

The third line of evidence provided by VEGA is the prediction. The ADI is an efficient way to evaluate the reliability of the prediction. This line of evidence is different from the two others, because it refers to a larger set of data (the training set is much larger than the few substances used for read-across); the knowledge extracted from the data quite probably contains elements not represented by the explicit rules or descriptors used for reasoning. Thus, the prediction has a broad basis.

If the three lines of evidence are in agreement, the uncertainty is very low. If some of the lines of evidence are missing, the uncertainty is higher. It may be that there are not similar substances. This is a weakness, because it is also likely that some particular features, related to the particular case represented by the target, may not be present in the *in silico* model. The information about the mechanism is not always present, because it depends on the endpoint and the algorithm. If there is disagreement between the different lines of evidence, the uncertainty is even higher.

We show some examples below.

B2.3. Examples on the general assessment of the results of the VEGA models

Example 1. The BCF easy example

Exercise 1

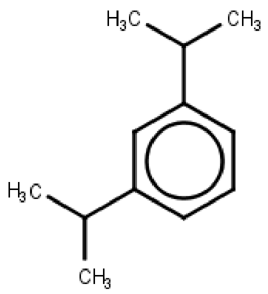
In the case of BCF the documentation in the **VEGA** report also involves the analysis of the logKow values. The description on the BCF and logKow values for the target compound and for known chemicals should be evaluated as well.

You can make as an exercise the prediction, before reading the part below. You have to copy and paste the SMILES and run the four BCF models in VEGA.

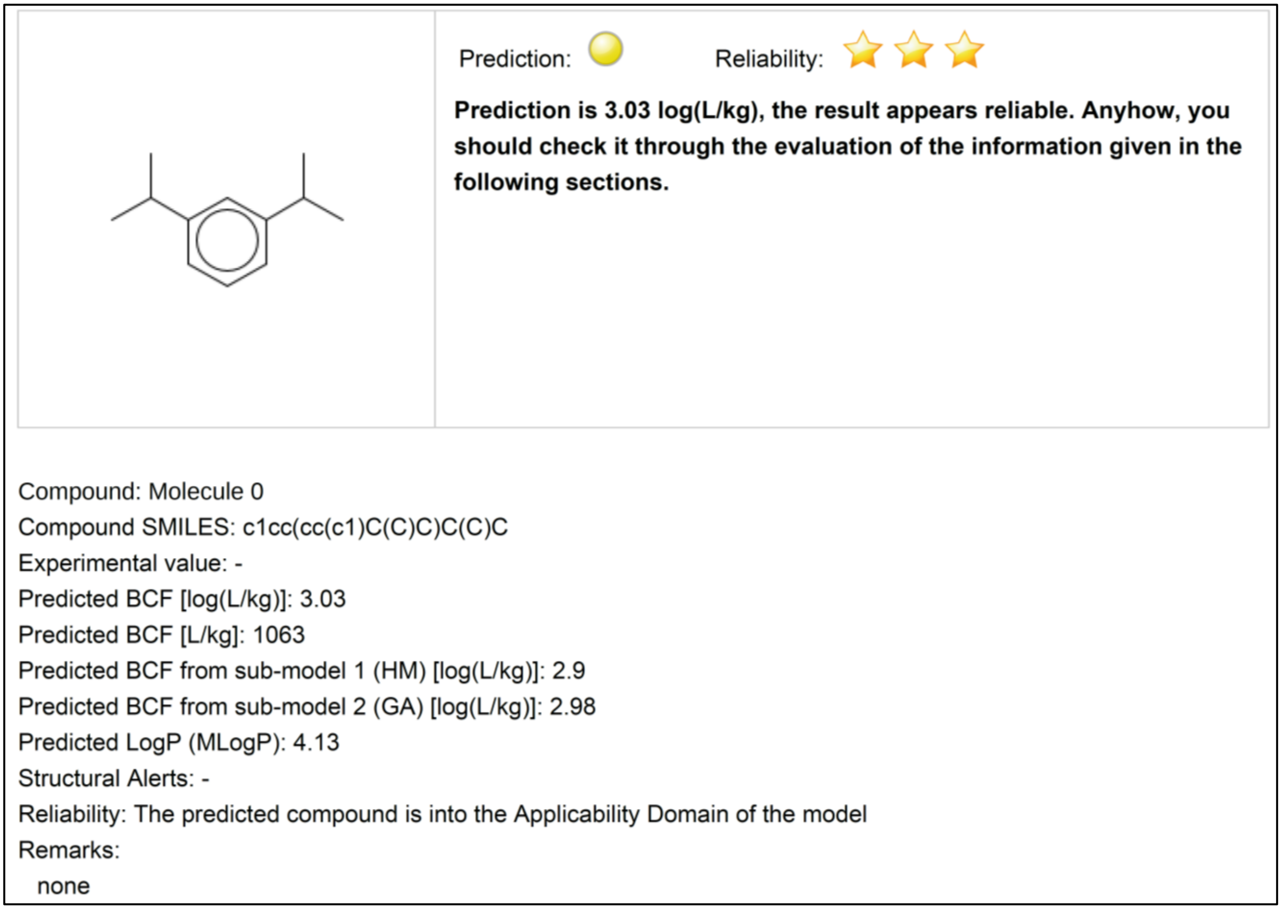
<u>The CAESAR model.</u>

Let' consider the results of the **CAESAR model** for the substance with the SMILES reported in *Table 2*.

Table 2. The input molecule for Exercise 1.

CAS No.	Name	SMILES	MOLECULE
99-62-7	1,3-Diisopropylbenzene	c1cc(cc(c1)C(C)C)C(C)C	

The cover of the report provides a summary of the results (*Figure 18*).



Prediction: ●

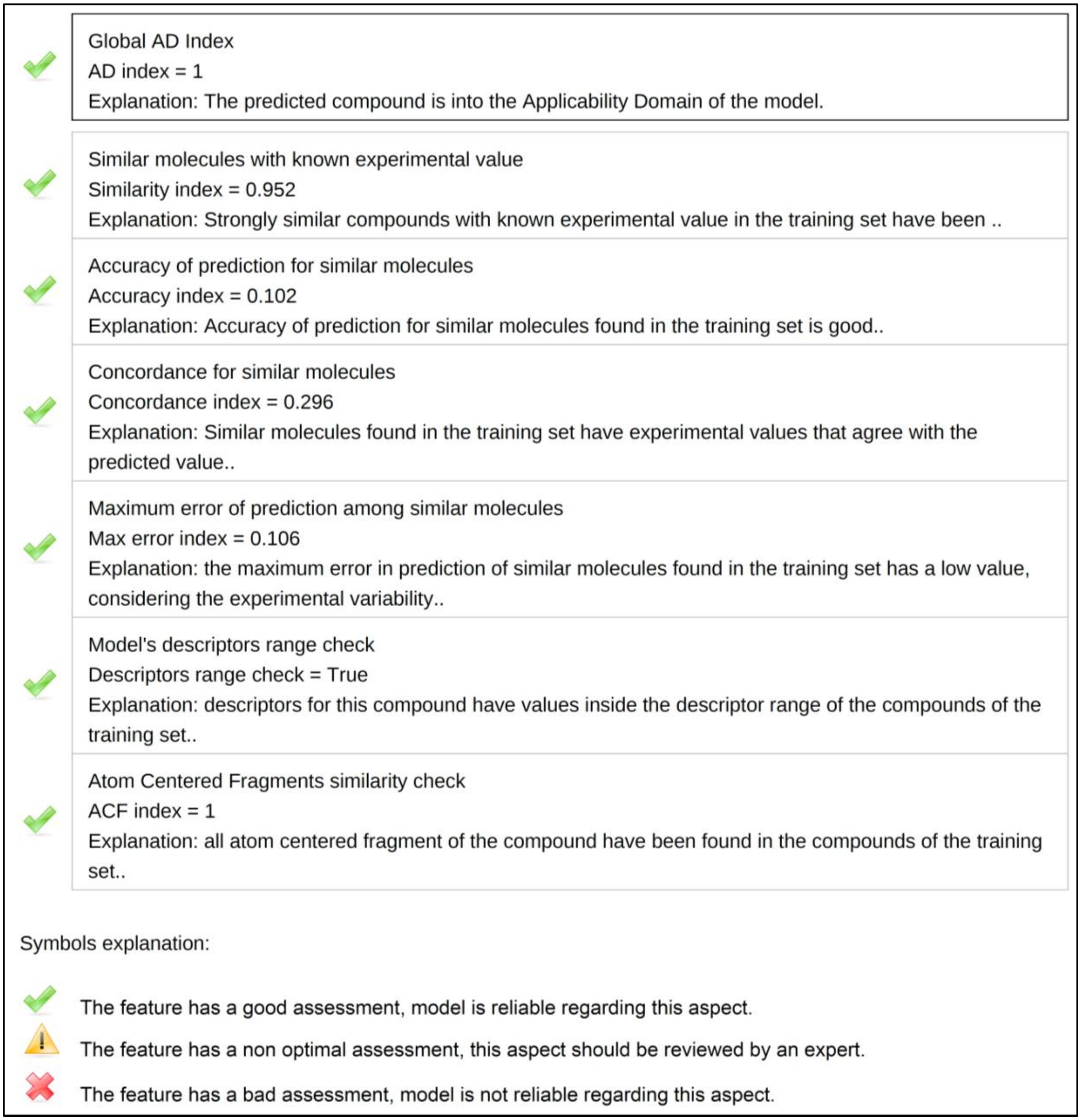
Reliability: ★★★

Prediction is 3.03 log(L/kg), the result appears reliable. Anyhow, you should check it through the evaluation of the information given in the following sections.

Compound: Molecule 0
Compound SMILES: c1cc(cc(c1)C(C)C)C(C)C
Experimental value: -
Predicted BCF [log(L/kg)]: 3.03
Predicted BCF [L/kg]: 1063
Predicted BCF from sub-model 1 (HM) [log(L/kg)]: 2.9
Predicted BCF from sub-model 2 (GA) [log(L/kg)]: 2.98
Predicted LogP (MLogP): 4.13
Structural Alerts: -
Reliability: The predicted compound is into the Applicability Domain of the model
Remarks:
none

Figure 18. The cover of the BCF CAESAR model.

The general **VEGA** evaluation of the applicability domain is good: three stars. The BCF value is close to the threshold value for labelling the substance as bioaccumulative (B): 3.3 in log unit. Indeed, the prediction is 3.03, and the colour is yellow. Thus, particular attention should be given to this evaluation. The CAESAR model is a hybrid model composed by two sub-models. The cover reports the results of both models: 2.9 and 2.98. These values are very close, thus we can expect that



✓ Global AD Index
AD index = 1
Explanation: The predicted compound is into the Applicability Domain of the model.

✓ Similar molecules with known experimental value
Similarity index = 0.952
Explanation: Strongly similar compounds with known experimental value in the training set have been ..

✓ Accuracy of prediction for similar molecules
Accuracy index = 0.102
Explanation: Accuracy of prediction for similar molecules found in the training set is good..

✓ Concordance for similar molecules
Concordance index = 0.296
Explanation: Similar molecules found in the training set have experimental values that agree with the predicted value..

✓ Maximum error of prediction among similar molecules
Max error index = 0.106
Explanation: the maximum error in prediction of similar molecules found in the training set has a low value, considering the experimental variability..

✓ Model's descriptors range check
Descriptors range check = True
Explanation: descriptors for this compound have values inside the descriptor range of the compounds of the training set..

✓ Atom Centered Fragments similarity check
ACF index = 1
Explanation: all atom centered fragment of the compound have been found in the compounds of the training set..

Symbols explanation:

✓ The feature has a good assessment, model is reliable regarding this aspect.

⚠ The feature has a non optimal assessment, this aspect should be reviewed by an expert.

✗ The feature has a bad assessment, model is not reliable regarding this aspect.

Figure 19. The ADI of the BCF CAESAR model.

the uncertainty of the CAESAR model is low. We notice that the prediction of CAESAR is 3.03, thus higher than the individual values of the two sub-models. This is a clear demonstration that the hybrid model is not simply the mean of the results of the two sub-models. For more details see:

- *http://www.caesar-project.eu/index.php?page=results&section=endpoint&ne=1*
- *https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2913327/pdf/1752-153X-4-S1-I1.pdf*

We evaluate now the details of the check done by the ADI tool for this prediction, as reported below (*Figure 19*).

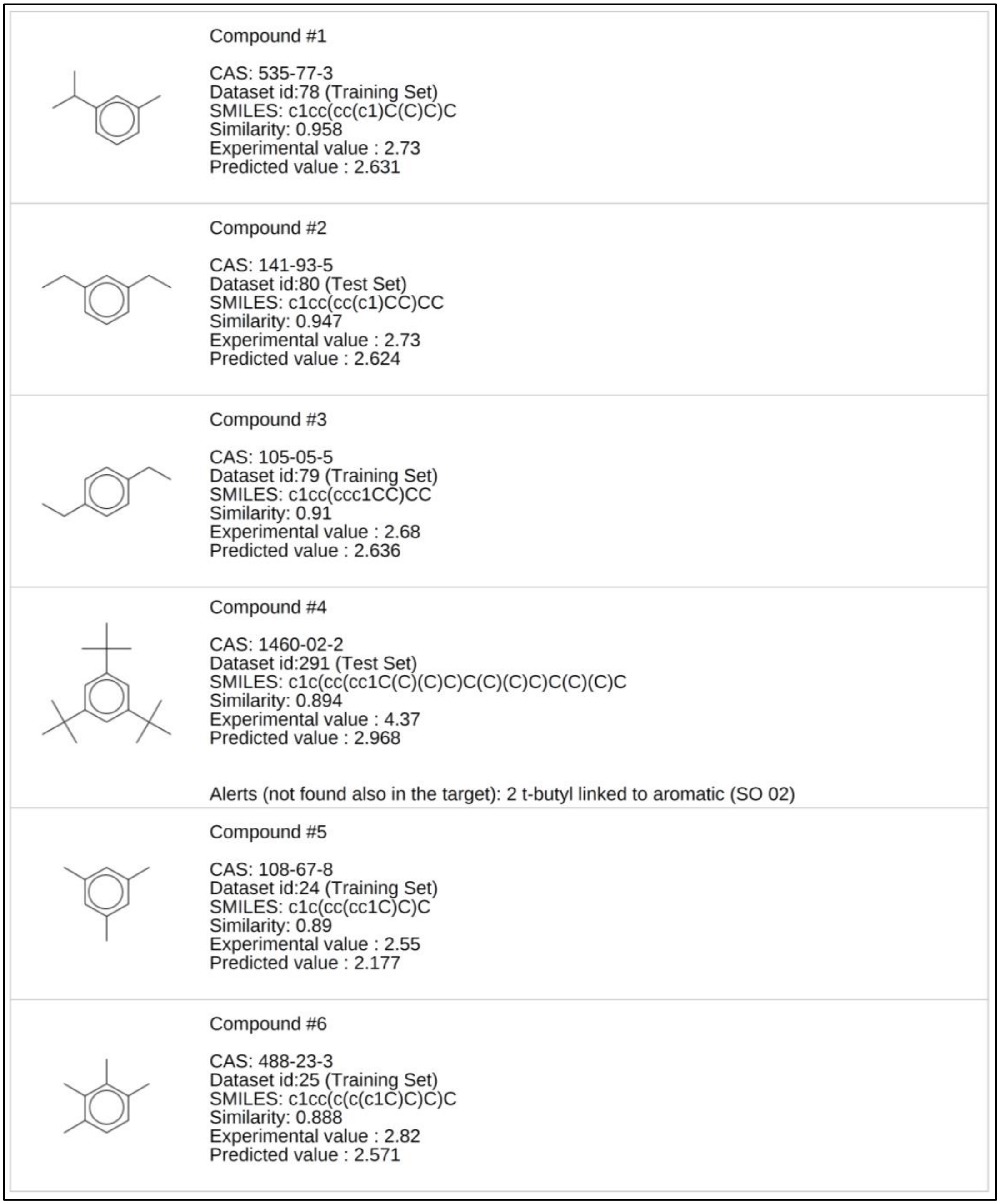
The ADI is 1 and all checks are good: there is no particular critical issue.

Now we look at the similar substances, as below (*Figure 20*).

All the six similar substances have a benzene ring, and substitutions with short aliphatic chains, with one or more carbon atoms, in some cases branched. Thus, they contain the same elements as the target substance. We expect that the differences will be related to the aliphatic chains, influencing the partitioning between water and the cell, simulated by the logKow value. The higher the number of carbon atoms, the higher the logKow and the BCF. In this phase of the evaluation, we start considering the experimental values of the similar substances, not the predicted value. Thus, we make a read-across study. The substance with three aliphatic carbons – similar number 5 - has the lowest BCF experimental value: 2.55. Then there are four substances with four aliphatic carbon atoms, with BCF values ranging from 2.68 to 2.82. There is a substance with 12 carbons, with the BCF value of 4.37. The target substance has six aliphatic carbons.

The three most interesting substances are the first three – this is often the case. These three similar substances have a similarity value higher than 0.9, while the other three similar substances have similarity lower than 0.9. All these three similar substances are smaller than the target. We can expect that the BCF value of the target substance will be between 2.82 (the highest value of the substances with four carbons) and 4.37, the value of similar number 4, but much closer to 2.82, since similar number 4 has the 12 aliphatic carbons instead of six, as in the target.

To complete the evaluation of the experimental values, VEGA provides two graphics. The first one, *Figure 21*, shows all the substances used in the CAESAR model, with their experimental BCF values, plotted versus the logKow value.



Compound #1

CAS: 535-77-3
Dataset id:78 (Training Set)
SMILES: c1cc(cc(c1)C(C)C)C
Similarity: 0.958
Experimental value : 2.73
Predicted value : 2.631

Compound #2

CAS: 141-93-5
Dataset id:80 (Test Set)
SMILES: c1cc(cc(c1)CC)CC
Similarity: 0.947
Experimental value : 2.73
Predicted value : 2.624

Compound #3

CAS: 105-05-5
Dataset id:79 (Training Set)
SMILES: c1cc(ccc1CC)CC
Similarity: 0.91
Experimental value : 2.68
Predicted value : 2.636

Compound #4

CAS: 1460-02-2
Dataset id:291 (Test Set)
SMILES: c1c(cc(cc1C(C)(C)C)C(C)(C)C)C(C)(C)C
Similarity: 0.894
Experimental value : 4.37
Predicted value : 2.968

Alerts (not found also in the target): 2 t-butyl linked to aromatic (SO 02)

Compound #5

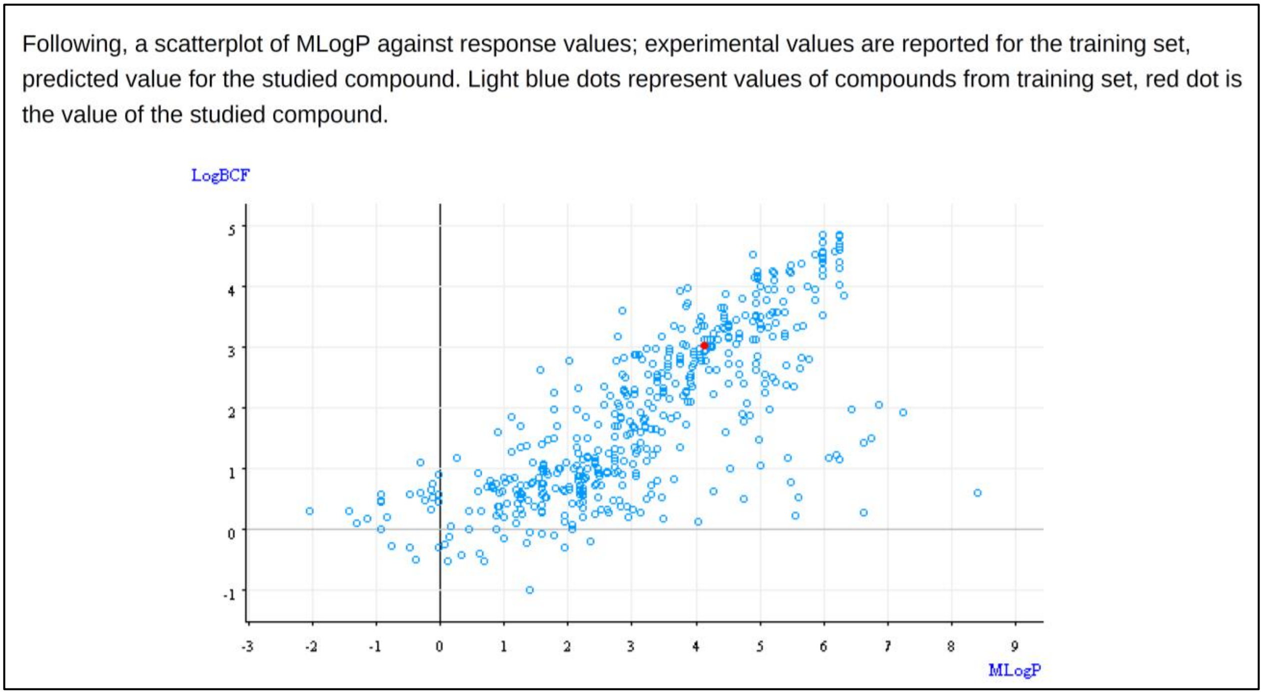
CAS: 108-67-8
Dataset id:24 (Training Set)
SMILES: c1c(cc(cc1C)C)C
Similarity: 0.89
Experimental value : 2.55
Predicted value : 2.177

Compound #6

CAS: 488-23-3
Dataset id:25 (Training Set)
SMILES: c1cc(c(c(c1C)C)C)C
Similarity: 0.888
Experimental value : 2.82
Predicted value : 2.571

Figure 20. The similar chemicals found by the BCF CAESAR model.

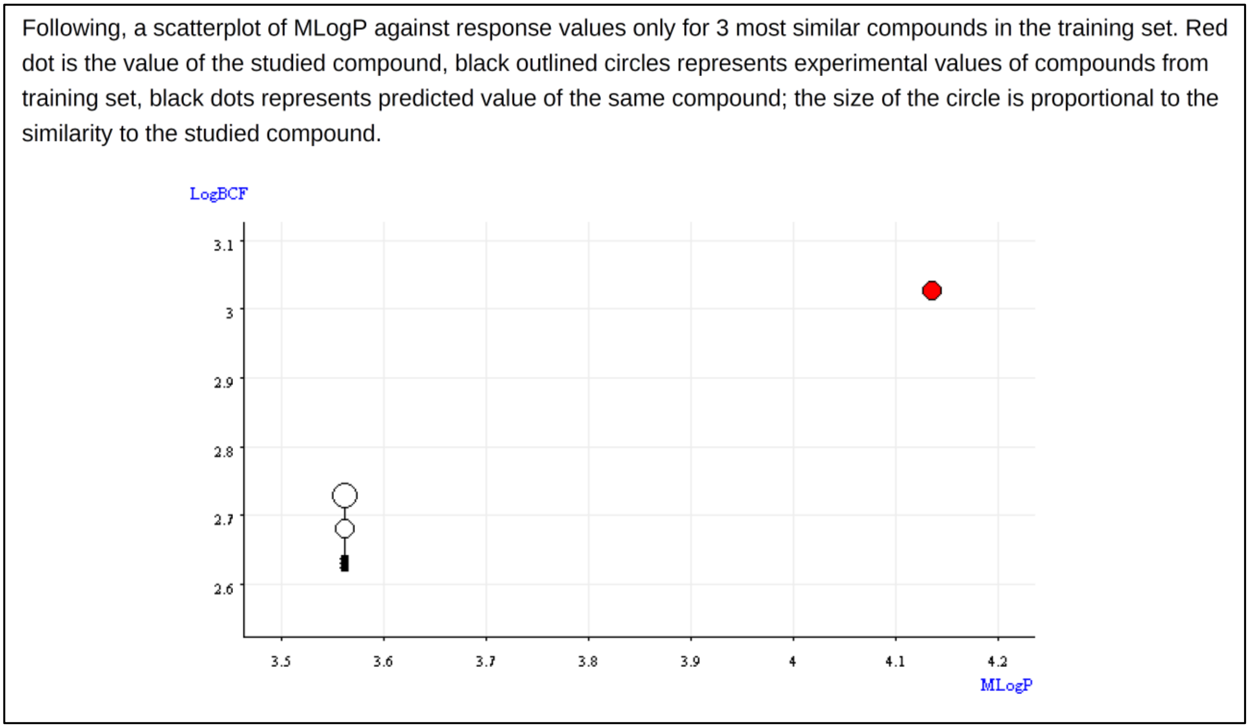
The red dot is the target substance. It is in the expected position, in agreement with the general behaviour of other substances, simply considering the descriptor MlogKow (which estimates the logKow).



Following, a scatterplot of MLogP against response values; experimental values are reported for the training set, predicted value for the studied compound. Light blue dots represent values of compounds from training set, red dot is the value of the studied compound.

Figure 21. The analysis of the molecular descriptors of the BCF CAESAR model (first part).

A closer look of the same picture is represented by the plot below (*Figure 22*), given by VEGA, which simply plots the three most similar substances.



Following, a scatterplot of MLogP against response values only for 3 most similar compounds in the training set. Red dot is the value of the studied compound, black outlined circles represents experimental values of compounds from training set, black dots represents predicted value of the same compound; the size of the circle is proportional to the similarity to the studied compound.

Figure 22. The analysis of the molecular descriptors of the BCF CAESAR model (second part).

The red circle is the target, and the white circles on the left indicate the experimental values of the three similar substances, as discussed above (similar number 1 and 2 overlap, because have the same experimental value).

Now we can evaluate the predicted values for the similar substances. In the last figure we discussed, the black dots indicate the predicted values. They are very close to the experimental one. When we evaluate the results of a predictive model, we should be aware of the uncertainty and variability of the experimental values, to evaluate if the difference between the predicted and experimental value is acceptable (see Part A of the eBook). For BCF the uncertainty of the experimental value is about 0.6 log unit (https://doi.org/10.1186/1752-153X-4-S1-S1). Thus, the predicted value in our case is quite accurate.

We may notice that the prediction of substance number 4 is much lower than the experimental value. However, VEGA already knows that there may be issues with substances with the terbutyl chain (and here we have three of these chains), as reported by the warning:

Alerts (not found also in the target): 2 t-butyl linked to aromatic (SO 02)

Thus, this similar substance should be disregarded.

Overall conclusion

We analyse the three lines of evidence provided by VEGA (see section The VEGA report).

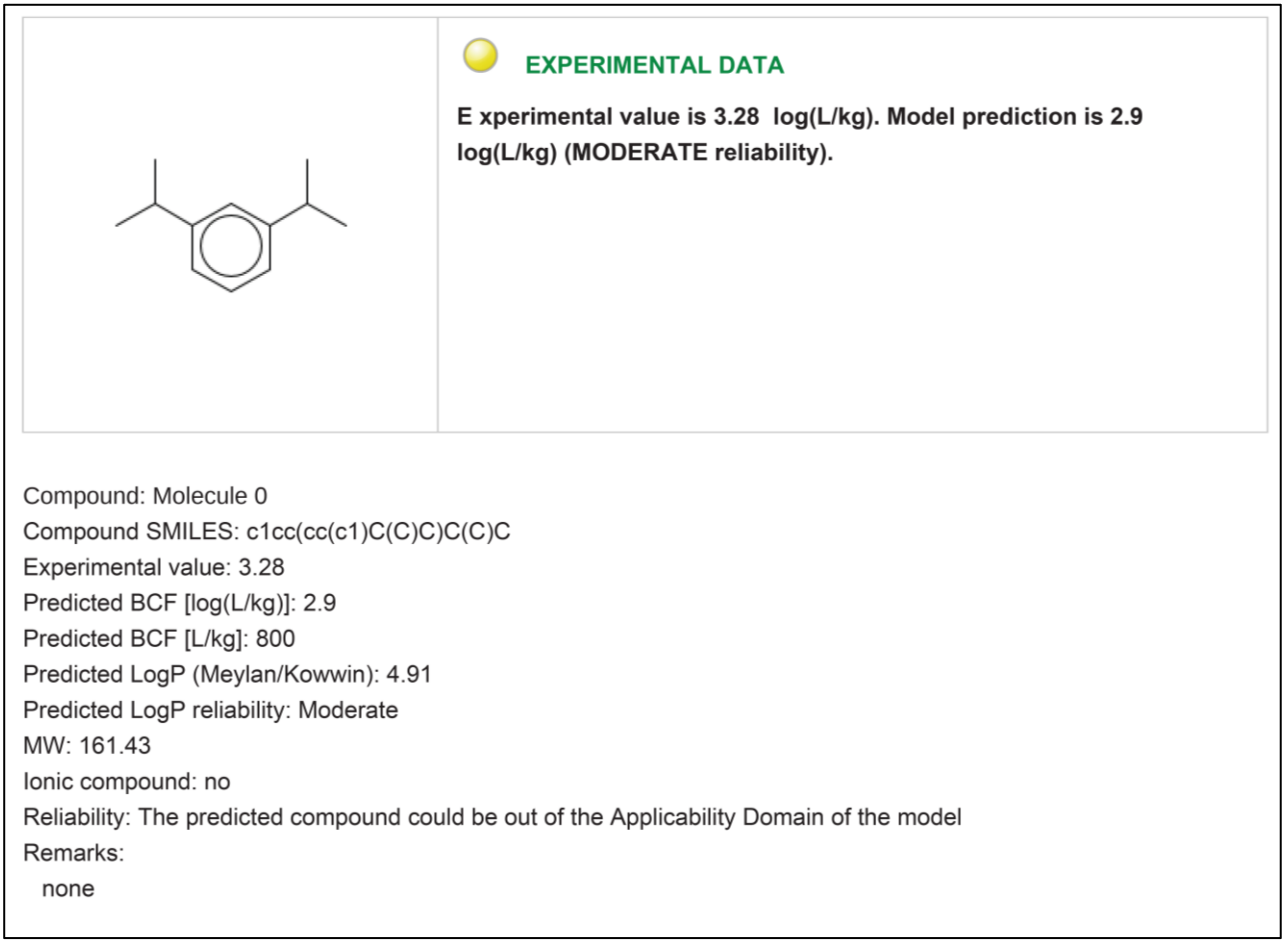
There are similar substances. The similar substances cover all the features present in the target substance. The parameter modulating the effect is associated with the number of carbons, which influences the logKow. The reliability of the prediction is high. The three lines of evidence are in agreement and the predicted value (about 3.0) is supported by the experimental values of similar substances, which are assumed to be lower.

We now analyse the same substance, with the SMILES c1cc(cc(c1)C(C)C)C(C)C, predicted with the other models in VEGA.

The Meylan model.

The results are summarized in *Figure 23*.

In this case the prediction has a moderate reliability, but VEGA shows the experimental value, present in the training set of the model, in this case the Meylan model from EPISuite. Thus, it is irrelevant to proceed with the prediction. Still, we evaluate the results to discuss the prediction.



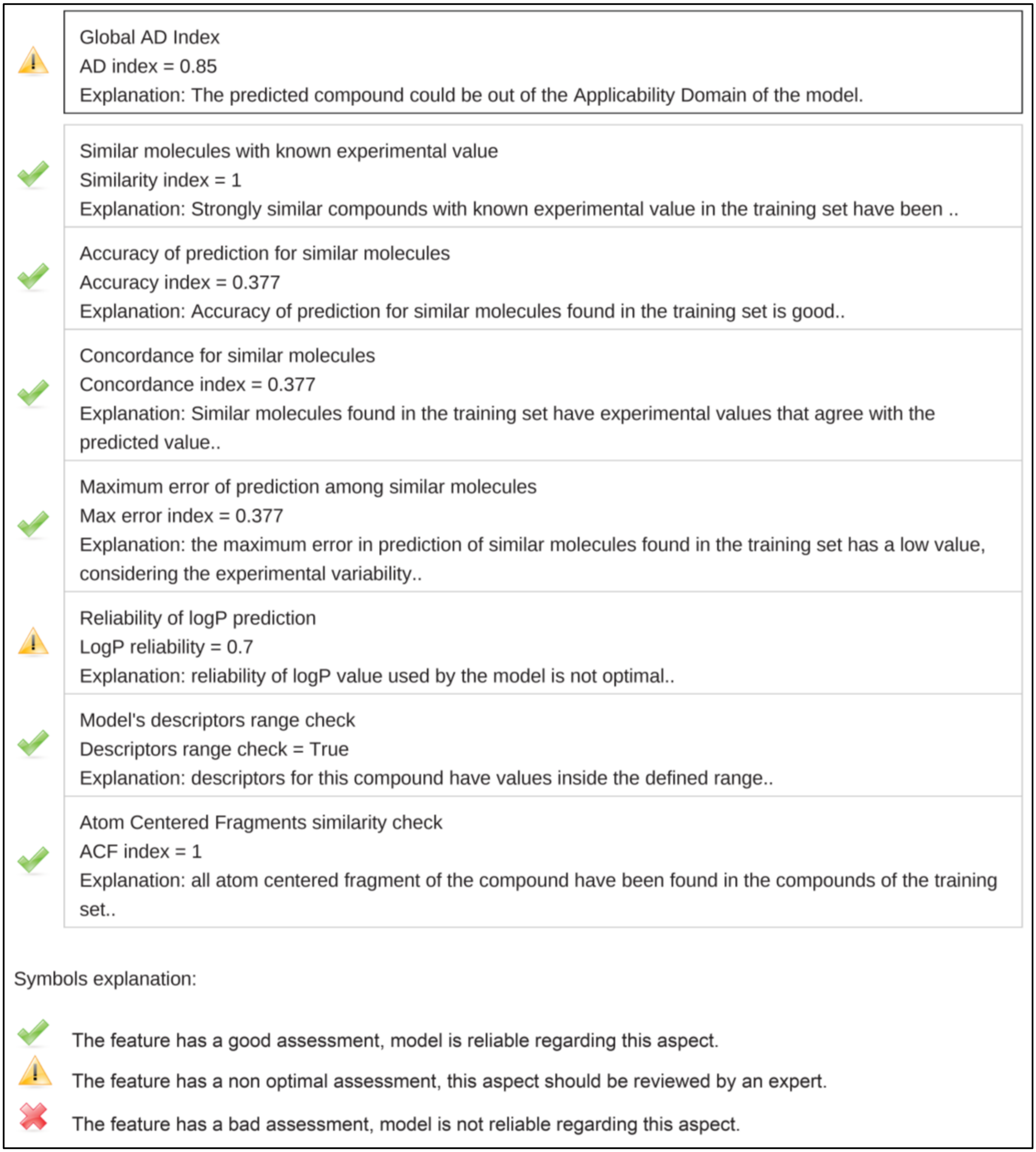
● **EXPERIMENTAL DATA**

E xperimental value is 3.28 log(L/kg). Model prediction is 2.9 log(L/kg) (MODERATE reliability).

Compound: Molecule 0
Compound SMILES: c1cc(cc(c1)C(C)C)C(C)C
Experimental value: 3.28
Predicted BCF [log(L/kg)]: 2.9
Predicted BCF [L/kg]: 800
Predicted LogP (Meylan/Kowwin): 4.91
Predicted LogP reliability: Moderate
MW: 161.43
Ionic compound: no
Reliability: The predicted compound could be out of the Applicability Domain of the model
Remarks:
none

Figure 23. The cover of the BCF Meylan model.

The results of the ADI are shown below (*Figure 24*).



⚠ Global AD Index
AD index = 0.85
Explanation: The predicted compound could be out of the Applicability Domain of the model.

✓ Similar molecules with known experimental value
Similarity index = 1
Explanation: Strongly similar compounds with known experimental value in the training set have been ..

✓ Accuracy of prediction for similar molecules
Accuracy index = 0.377
Explanation: Accuracy of prediction for similar molecules found in the training set is good..

✓ Concordance for similar molecules
Concordance index = 0.377
Explanation: Similar molecules found in the training set have experimental values that agree with the predicted value..

✓ Maximum error of prediction among similar molecules
Max error index = 0.377
Explanation: the maximum error in prediction of similar molecules found in the training set has a low value, considering the experimental variability..

⚠ Reliability of logP prediction
LogP reliability = 0.7
Explanation: reliability of logP value used by the model is not optimal..

✓ Model's descriptors range check
Descriptors range check = True
Explanation: descriptors for this compound have values inside the defined range..

✓ Atom Centered Fragments similarity check
ACF index = 1
Explanation: all atom centered fragment of the compound have been found in the compounds of the training set..

Symbols explanation:

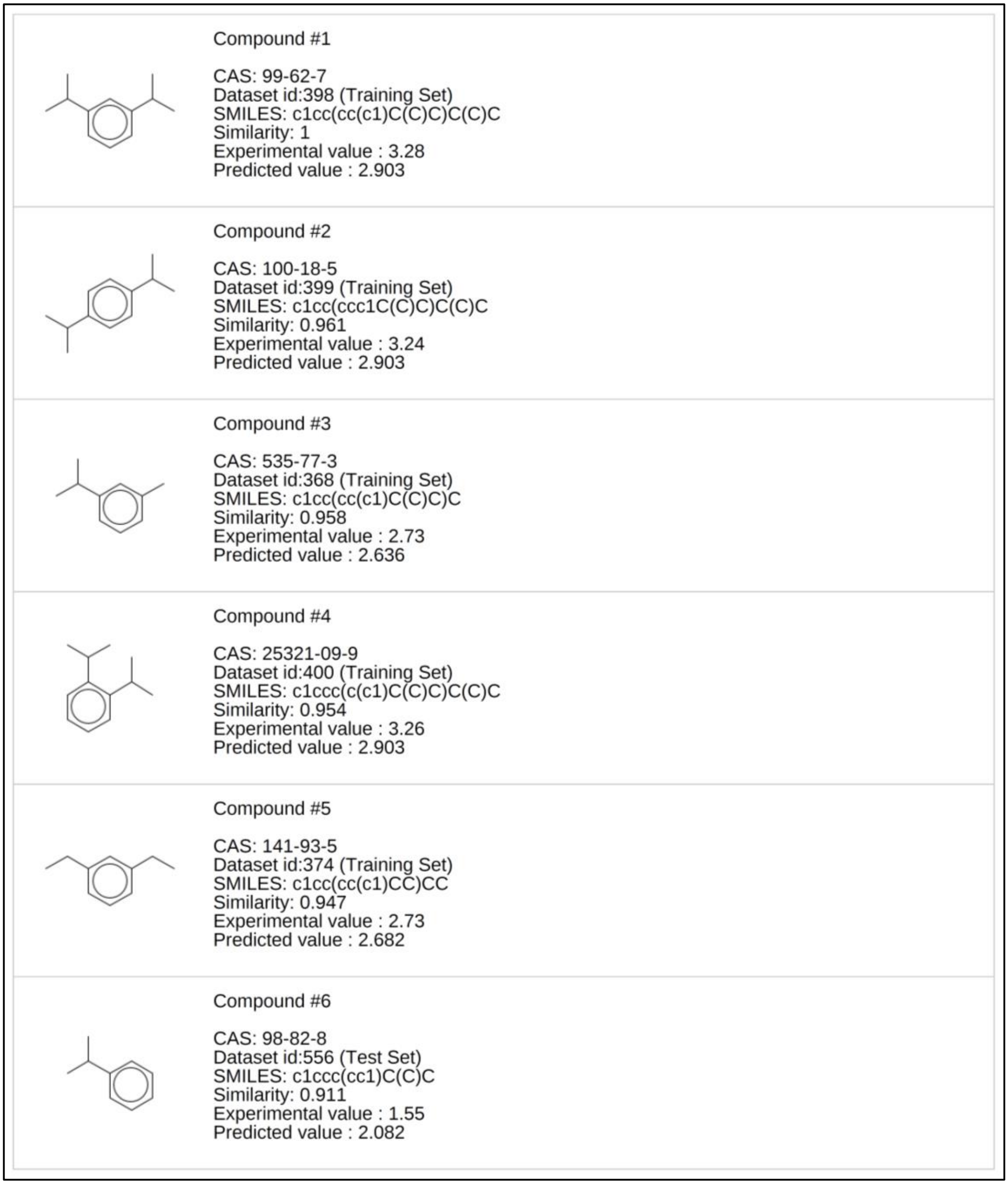
✓ The feature has a good assessment, model is reliable regarding this aspect.
⚠ The feature has a non optimal assessment, this aspect should be reviewed by an expert.
✗ The feature has a bad assessment, model is not reliable regarding this aspect.

Figure 24. The ADI of the BCF Meylan model.

The software identifies a potential weakness related to the logKow value.

Below, we show the most similar substances (*Figure 25*).

All the six most similar substances have a similarity value higher than 0.9. There is the target substance, as already mentioned. The second and fourth similar substances are isomers, with the two isopropyl residues in different positions. The experimental values are very similar, depending on the position. The values range from 3.28 to 3.24. Reducing the number of carbons, we observe lower BCF values: 2.73 with four aliphatic carbons, and 1.56, with three aliphatic carbons.



Compound #1

CAS: 99-62-7
Dataset id:398 (Training Set)
SMILES: c1cc(cc(c1)C(C)C)C(C)C
Similarity: 1
Experimental value : 3.28
Predicted value : 2.903

Compound #2

CAS: 100-18-5
Dataset id:399 (Training Set)
SMILES: c1cc(ccc1C(C)C)C(C)C
Similarity: 0.961
Experimental value : 3.24
Predicted value : 2.903

Compound #3

CAS: 535-77-3
Dataset id:368 (Training Set)
SMILES: c1cc(cc(c1)C(C)C)C
Similarity: 0.958
Experimental value : 2.73
Predicted value : 2.636

Compound #4

CAS: 25321-09-9
Dataset id:400 (Training Set)
SMILES: c1ccc(c(c1)C(C)C)C(C)C
Similarity: 0.954
Experimental value : 3.26
Predicted value : 2.903

Compound #5

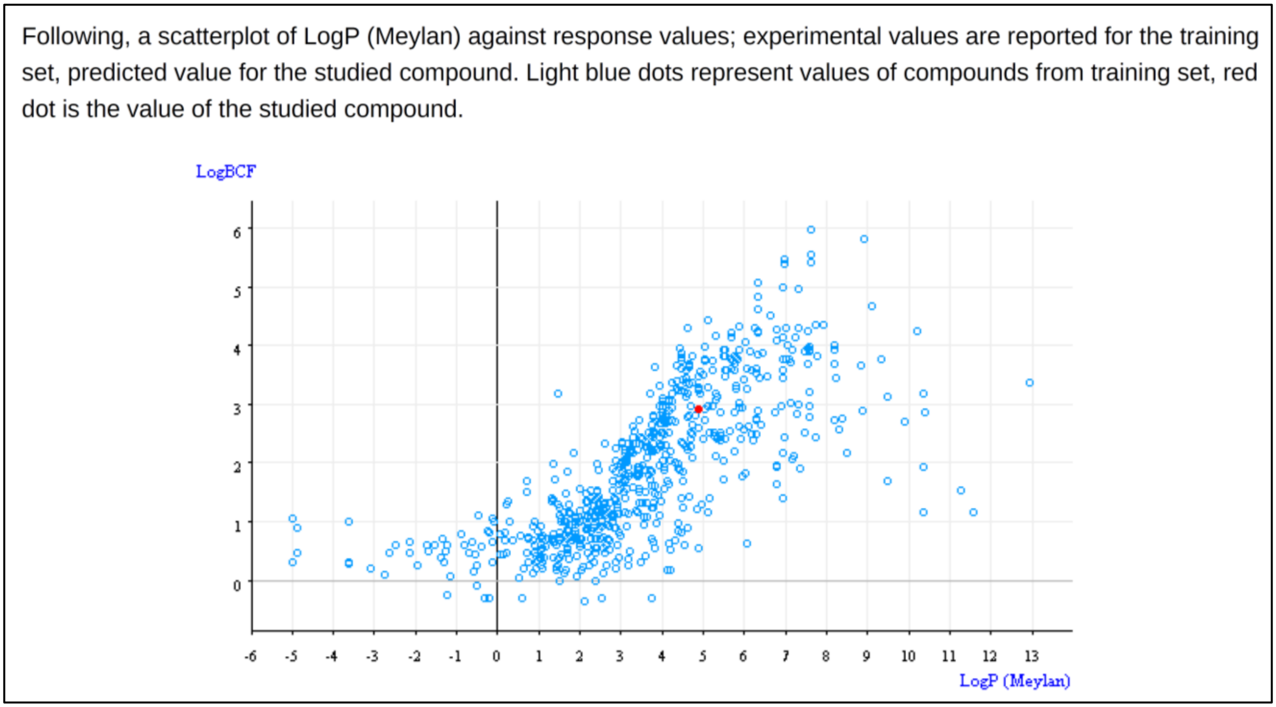
CAS: 141-93-5
Dataset id:374 (Training Set)
SMILES: c1cc(cc(c1)CC)CC
Similarity: 0.947
Experimental value : 2.73
Predicted value : 2.682

Compound #6

CAS: 98-82-8
Dataset id:556 (Test Set)
SMILES: c1ccc(cc1)C(C)C
Similarity: 0.911
Experimental value : 1.55
Predicted value : 2.082

Figure 25. The similar chemicals found by the BCF Meylan model.

The plot of the experimental BCF values versus logKow is presented below (*Figure 26*).



Following, a scatterplot of LogP (Meylan) against response values; experimental values are reported for the training set, predicted value for the studied compound. Light blue dots represent values of compounds from training set, red dot is the value of the studied compound.

Figure 26. The analysis of the molecular descriptors of the BCF Meylan model.

The red dot is in the cloud of the average values, a little bit lower compared to the same plot observed for the CAESAR model. This is somehow consistent with the potential issue indicated by the ADI algorithm for logKow. Still, the prediction is acceptable, and the reliability of the model can be considered high, regardless of the warning.

Note. Comparing the most similar substances of the CAESAR and Meylan model, it is clear that they are different. Indeed, each model contains its own collection of substances, with their experimental values.

Note. The Meylan model in VEGA is a reimplementation of the same model as in EPISuite. The user may appreciate the fact that in VEGA there are:

- An evaluation of the applicability domain expressed in a quantitative, automatic way
- The representation of the most similar substances

- Plots of the BCF value of the target compound compared with logKow, for all the substances and for the three most similar ones, facilitating the evaluation of the role of logKow.

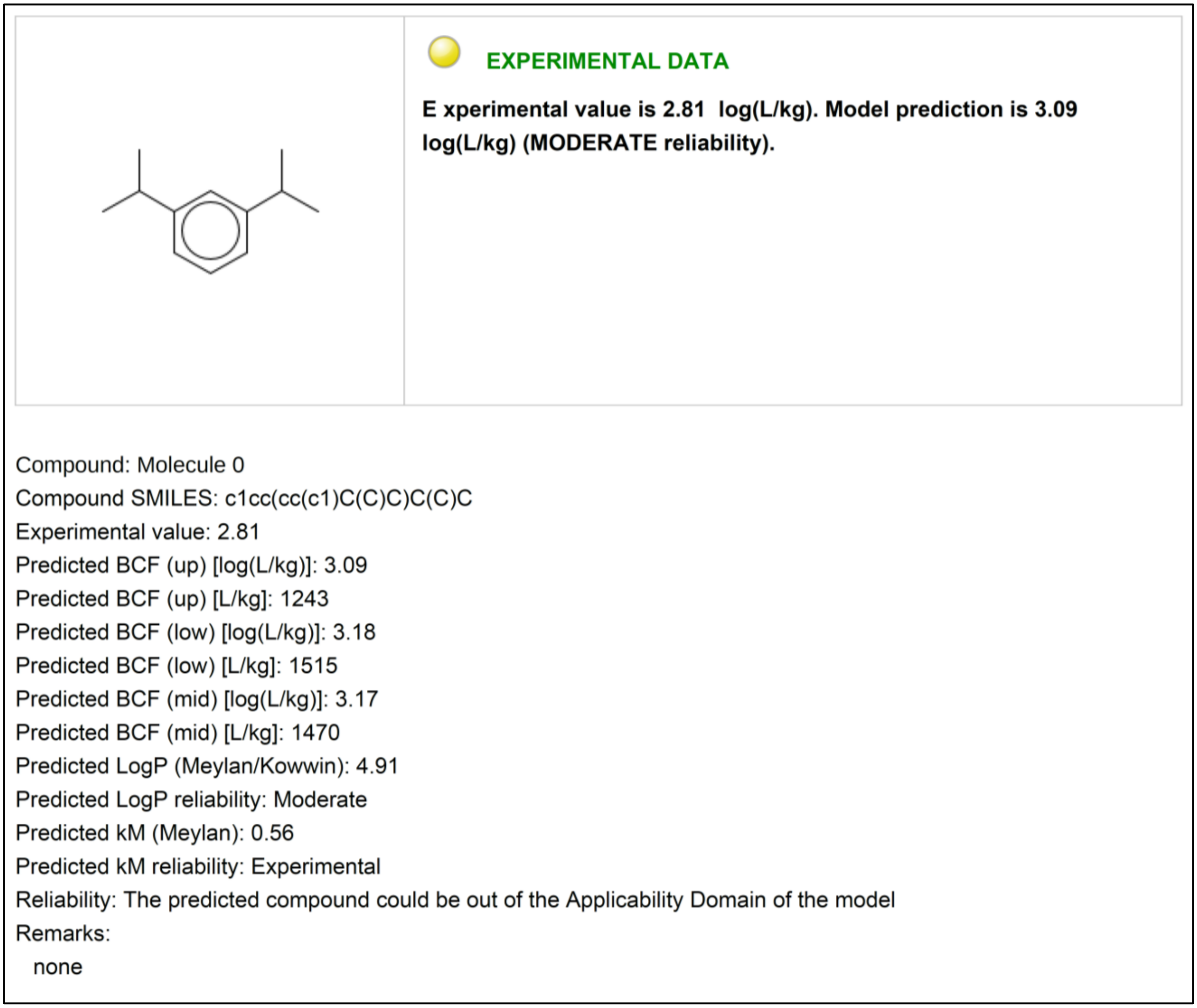
The Arnot-Gobas model.

The results are summarized in *Figure 27*.

In this case, the model provides different values for the different simulations done with three kinds of fish (upper, mid and lower trophic levels), and results vary, ranging from 3.09 to 3.18. This is a difference compared with the other models. Also in this case, we have the experimental value, as in the case of the Meylan model. Anyhow, we proceed with our evaluation.

Note. We observe that the experimental value from this model is different from the experimental value as in the Meylan model: 2.81 versus 3.28. We already commented that the experimental uncertainty and variability is about 0.6 log unit. Thus, these two values are within this range. As a consequence, two predictions within this range are equivalent.

The evaluation of the ADI indicates the same concern about logKow as for the Meylan model.



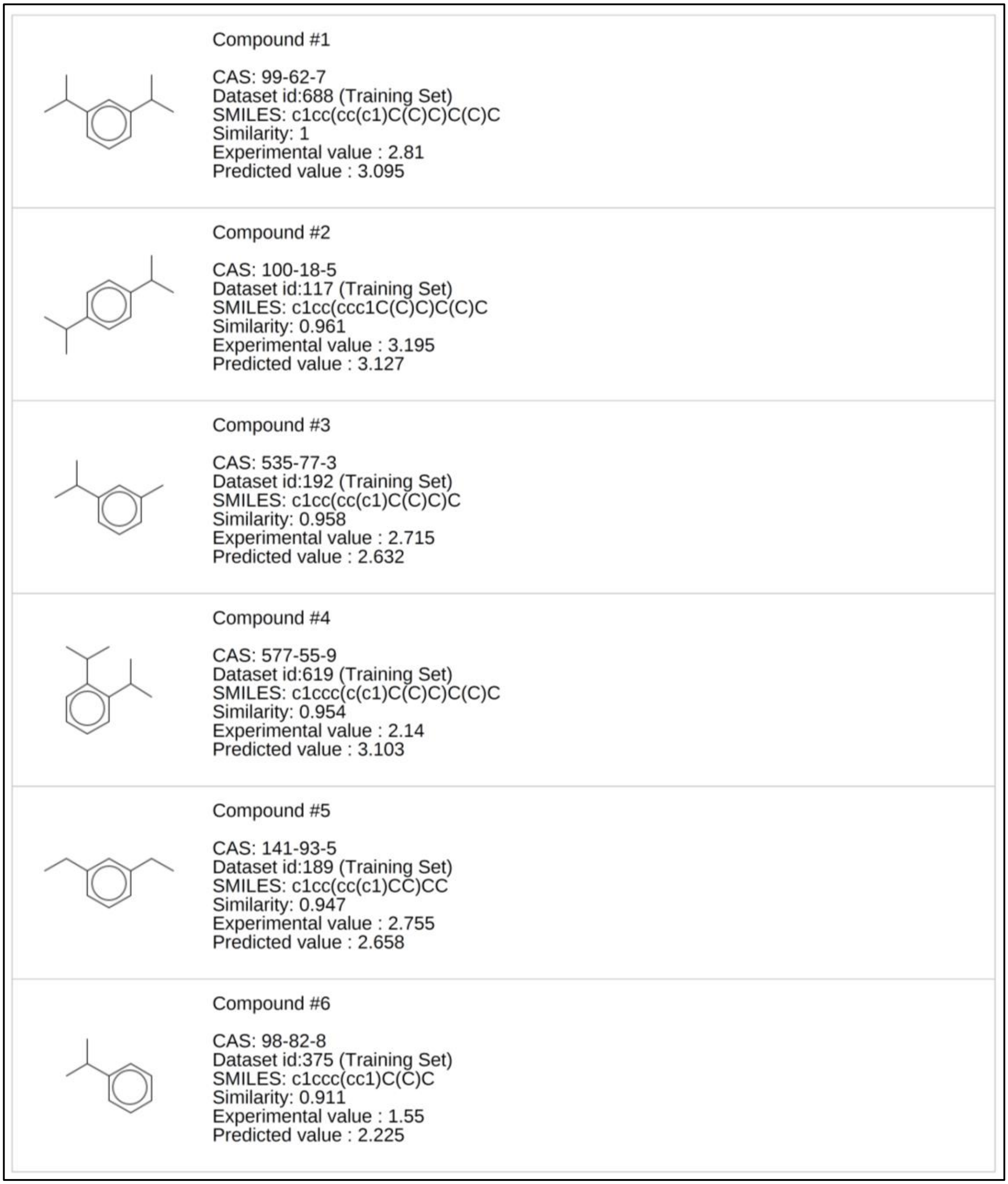
EXPERIMENTAL DATA

E xperimental value is 2.81 log(L/kg). Model prediction is 3.09 log(L/kg) (MODERATE reliability).

Compound: Molecule 0
Compound SMILES: c1cc(cc(c1)C(C)C)C(C)C
Experimental value: 2.81
Predicted BCF (up) [log(L/kg)]: 3.09
Predicted BCF (up) [L/kg]: 1243
Predicted BCF (low) [log(L/kg)]: 3.18
Predicted BCF (low) [L/kg]: 1515
Predicted BCF (mid) [log(L/kg)]: 3.17
Predicted BCF (mid) [L/kg]: 1470
Predicted LogP (Meylan/Kowwin): 4.91
Predicted LogP reliability: Moderate
Predicted kM (Meylan): 0.56
Predicted kM reliability: Experimental
Reliability: The predicted compound could be out of the Applicability Domain of the model
Remarks:
none

Figure 27. The cover of the BCF Arnot-Gobas model.

Below we provide the six most similar substances as in the Arnot-Gobas model (*Figure 28*).



Compound #1

CAS: 99-62-7
Dataset id:688 (Training Set)
SMILES: c1cc(cc(c1)C(C)C)C(C)C
Similarity: 1
Experimental value : 2.81
Predicted value : 3.095

Compound #2

CAS: 100-18-5
Dataset id:117 (Training Set)
SMILES: c1cc(ccc1C(C)C)C(C)C
Similarity: 0.961
Experimental value : 3.195
Predicted value : 3.127

Compound #3

CAS: 535-77-3
Dataset id:192 (Training Set)
SMILES: c1cc(cc(c1)C(C)C)C
Similarity: 0.958
Experimental value : 2.715
Predicted value : 2.632

Compound #4

CAS: 577-55-9
Dataset id:619 (Training Set)
SMILES: c1ccc(c(c1)C(C)C)C(C)C
Similarity: 0.954
Experimental value : 2.14
Predicted value : 3.103

Compound #5

CAS: 141-93-5
Dataset id:189 (Training Set)
SMILES: c1cc(cc(c1)CC)CC
Similarity: 0.947
Experimental value : 2.755
Predicted value : 2.658

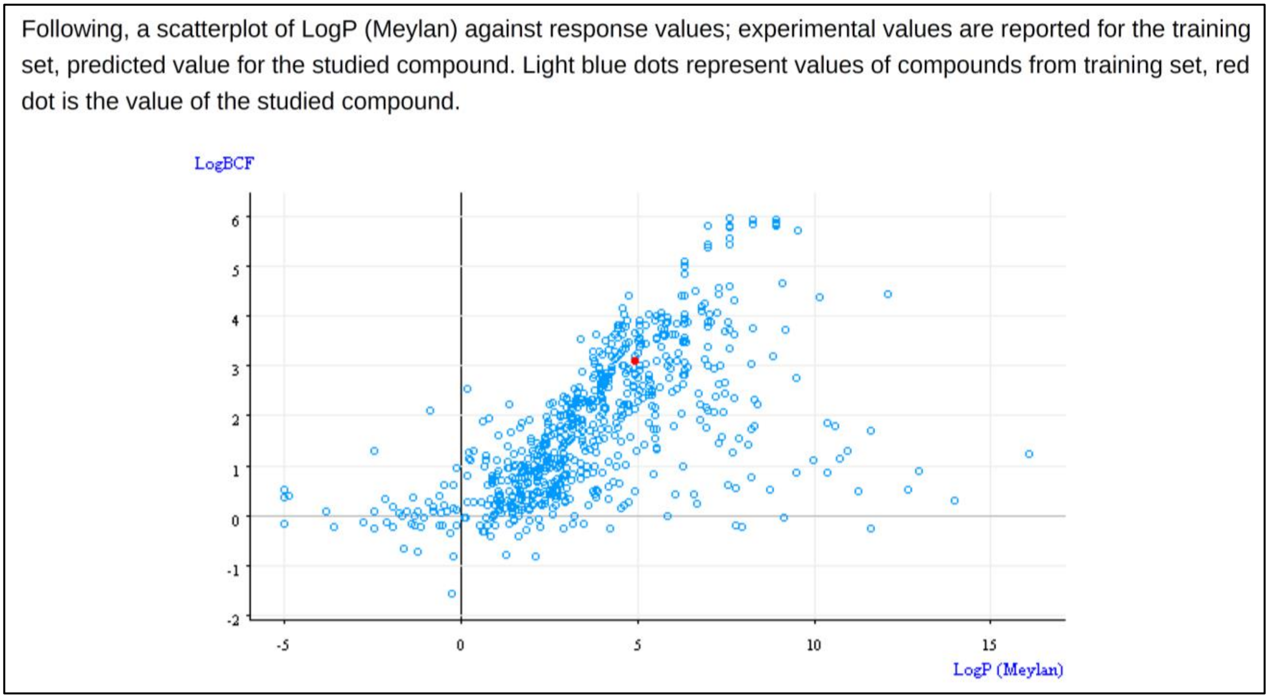
Compound #6

CAS: 98-82-8
Dataset id:375 (Training Set)
SMILES: c1ccc(cc1)C(C)C
Similarity: 0.911
Experimental value : 1.55
Predicted value : 2.225

Figure 28. The similar chemicals found by the BCF Arnot-Gobas model.

They are the same as those in the Meylan model, however the experimental values are in most of the cases different.

The plot of the experimental BCF values versus logKow is presented below (*Figure 29*).



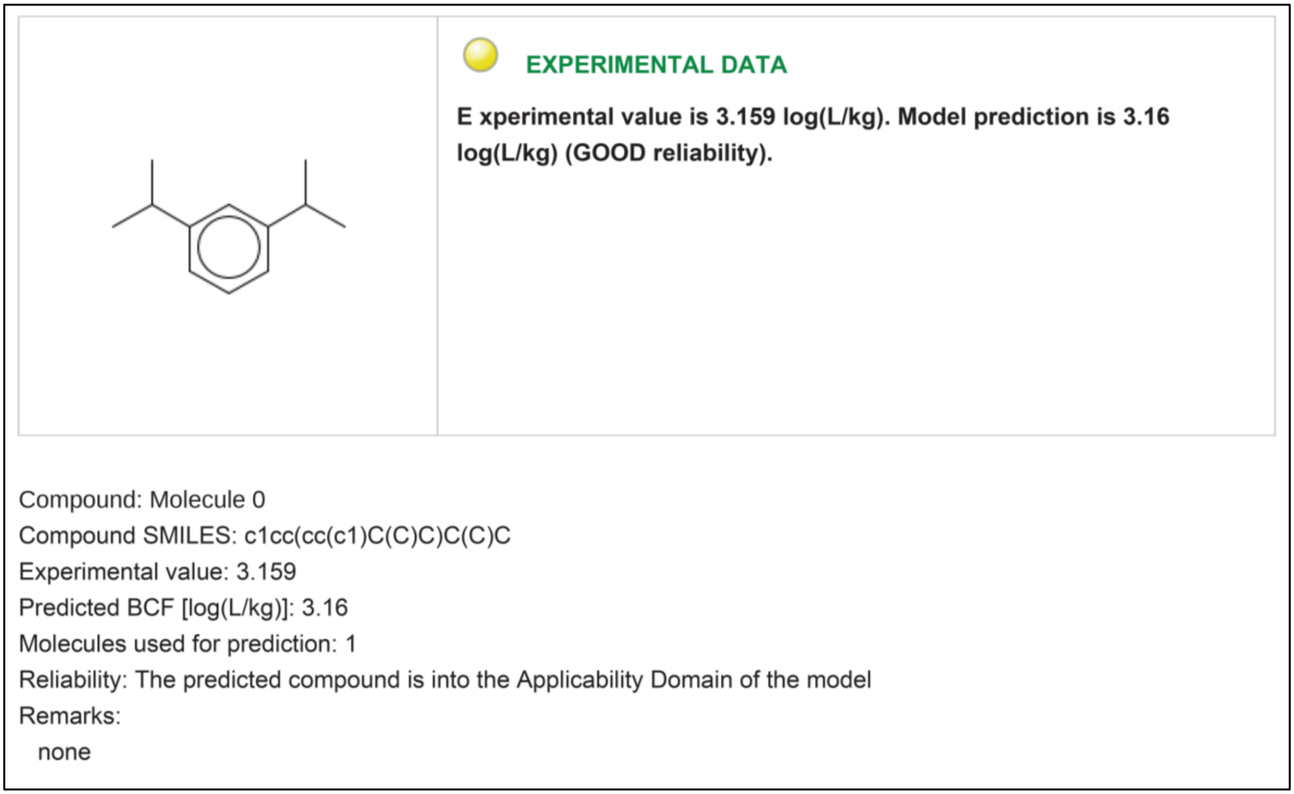
Following, a scatterplot of LogP (Meylan) against response values; experimental values are reported for the training set, predicted value for the studied compound. Light blue dots represent values of compounds from training set, red dot is the value of the studied compound.

Figure 29. The analysis of molecular descriptors of the BCF Arnot-Gobas model.

Thus, we can conclude that also for the Arnot-Gobas the prediction is with high reliability.

The KNN model.

The results are summarized in the figure below (*Figure 30*).



● EXPERIMENTAL DATA

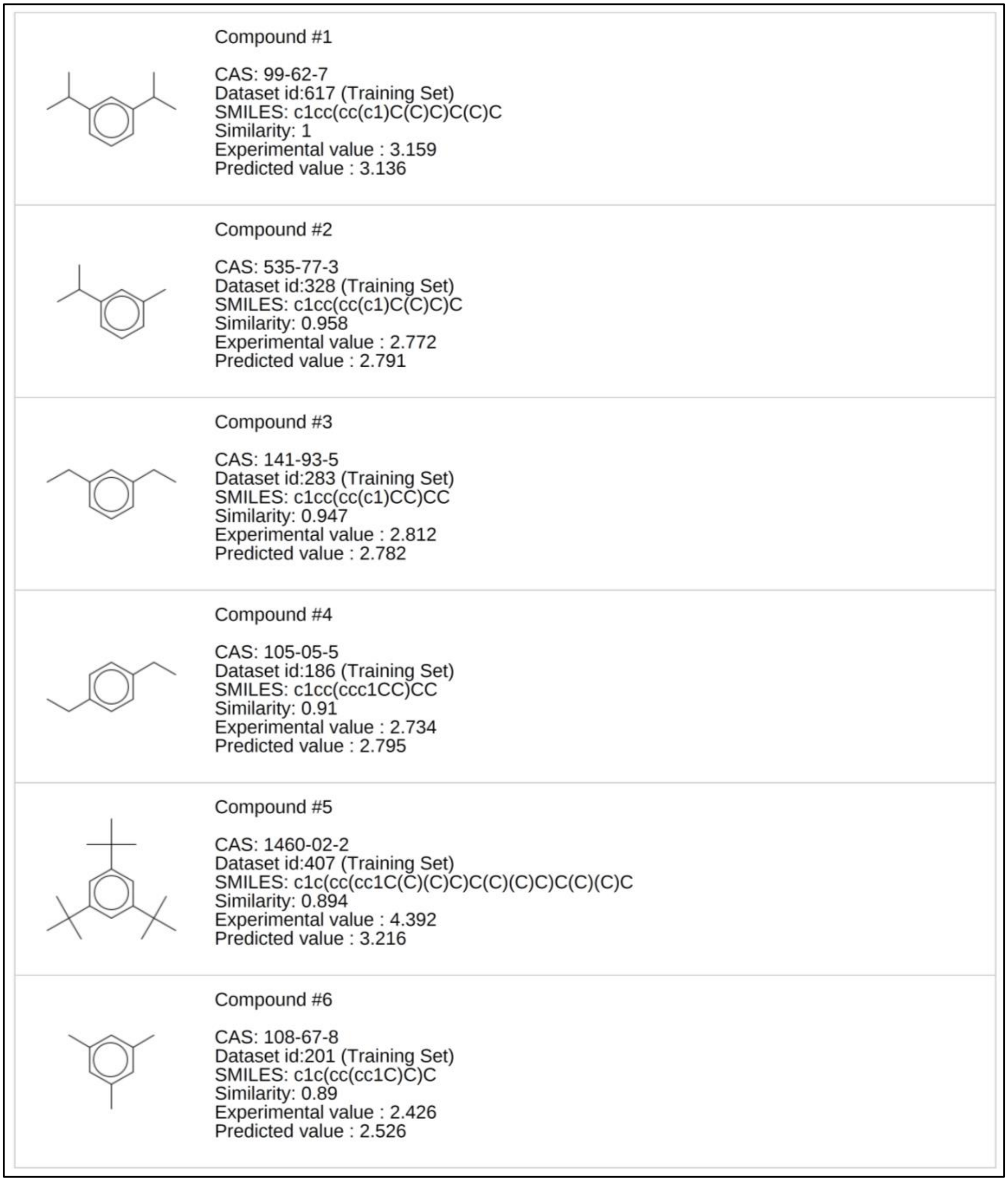
E xperimental value is 3.159 log(L/kg). Model prediction is 3.16 log(L/kg) (GOOD reliability).

Compound: Molecule 0
Compound SMILES: c1cc(cc(c1)C(C)C)C(C)C
Experimental value: 3.159
Predicted BCF [log(L/kg)]: 3.16
Molecules used for prediction: 1
Reliability: The predicted compound is into the Applicability Domain of the model
Remarks:
none

Figure 30. The cover of the BCF KNN-Read-Across model.

The ADI value is 1, and there is no issue indicated by the software. We have to highlight that the prediction is not a true prediction since it is based on the target itself.

Below we provide the six most similar substances (*Figure 31*).



Compound #1

CAS: 99-62-7
Dataset id:617 (Training Set)
SMILES: c1cc(cc(c1)C(C)C)C(C)C
Similarity: 1
Experimental value : 3.159
Predicted value : 3.136

Compound #2

CAS: 535-77-3
Dataset id:328 (Training Set)
SMILES: c1cc(cc(c1)C(C)C)C
Similarity: 0.958
Experimental value : 2.772
Predicted value : 2.791

Compound #3

CAS: 141-93-5
Dataset id:283 (Training Set)
SMILES: c1cc(cc(c1)CC)CC
Similarity: 0.947
Experimental value : 2.812
Predicted value : 2.782

Compound #4

CAS: 105-05-5
Dataset id:186 (Training Set)
SMILES: c1cc(ccc1CC)CC
Similarity: 0.91
Experimental value : 2.734
Predicted value : 2.795

Compound #5

CAS: 1460-02-2
Dataset id:407 (Training Set)
SMILES: c1c(cc(cc1C(C)(C)C)C(C)(C)C)C(C)(C)C
Similarity: 0.894
Experimental value : 4.392
Predicted value : 3.216

Compound #6

CAS: 108-67-8
Dataset id:201 (Training Set)
SMILES: c1c(cc(cc1C)C)C
Similarity: 0.89
Experimental value : 2.426
Predicted value : 2.526

Figure 31. The similar chemicals found by the BCF KNN-Read-Across model.

Overall evaluation of the four models for BCF for the target substance.

We found the experimental value of the target substance, actually more than one: 3.159, 2.81 and 3.28. Thus, in practice we can conclude that the substance is not BCF, but its value is very close to the threshold (3.3).

Let's imagine that we do not know these values and discuss only the other lines of evidence provided by the four models.

There are very similar substances, isomers of the target compound. Their experimental values are about 3.2 considering the values as in the Meylan and Arnot-Gobas similar substances, with the only exception of a value at 2.14 for the isomer similar number 4 of the Arnot-Gobas model. The other similar substances have values consistent with these values, thus we may expect that the target has the BCF value at about 3.2, disregarding the value at 2.14.

The role of the aliphatic carbons and logKow is consistent with the experimental values observed on the similar substances, as also represented by the plots.

The four predictions are 2.9, 3.03, 3.09, and 3.16, thus the four predictions are very consistent.

Considering the experimental values of the similar substances and the four predictions, we can conclude that the expected BCF value of the target is about 3-3.2.

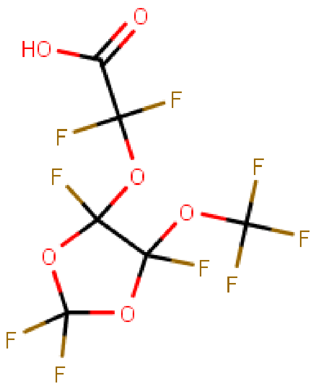
Example 2. The BCF difficult example

Exercise 2

Let's consider now a difficult substance, a perfluorinated compound. It is known that this category has a difficult behaviour. *Table 3* reports the substance to use.

You can make as an exercise the prediction, before reading the part below. You have to copy and paste the SMILES and run the four BCF models.

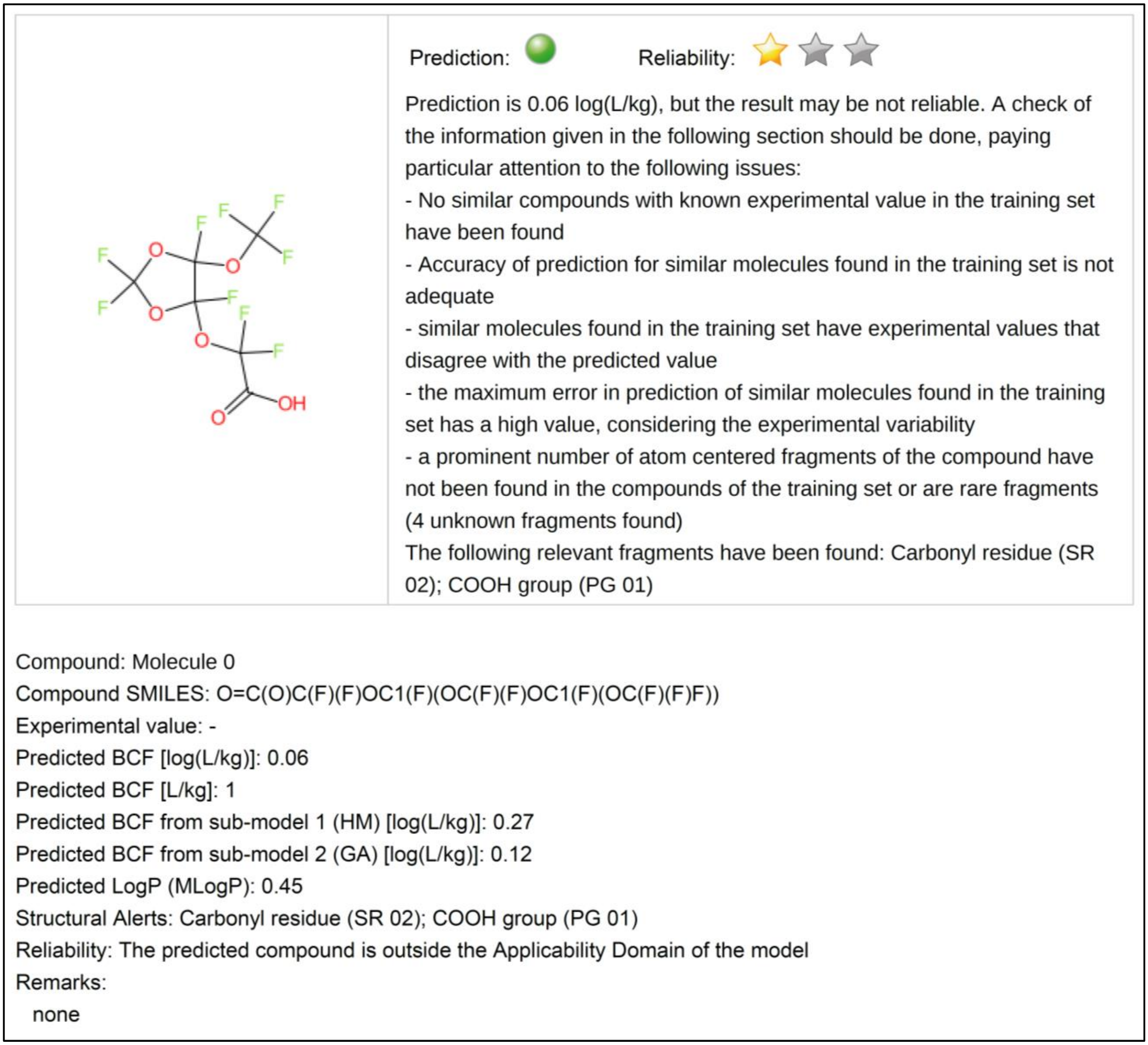
Table 3. The input molecule for Exercise 2.

CAS No.	Name	SMILES	MOLECULE
1190931-41-9	C604	C(=O)(C(OC1(C(OC(O1)(F)F)(OC(F)(F)F)F)F)(F)F)O	

The CAESAR model.

The cover of the report is shown below (*Figure 32*).

The prediction, 0.06, has a low reliability.



Prediction: ● Reliability: ★☆☆

Prediction is 0.06 log(L/kg), but the result may be not reliable. A check of the information given in the following section should be done, paying particular attention to the following issues:

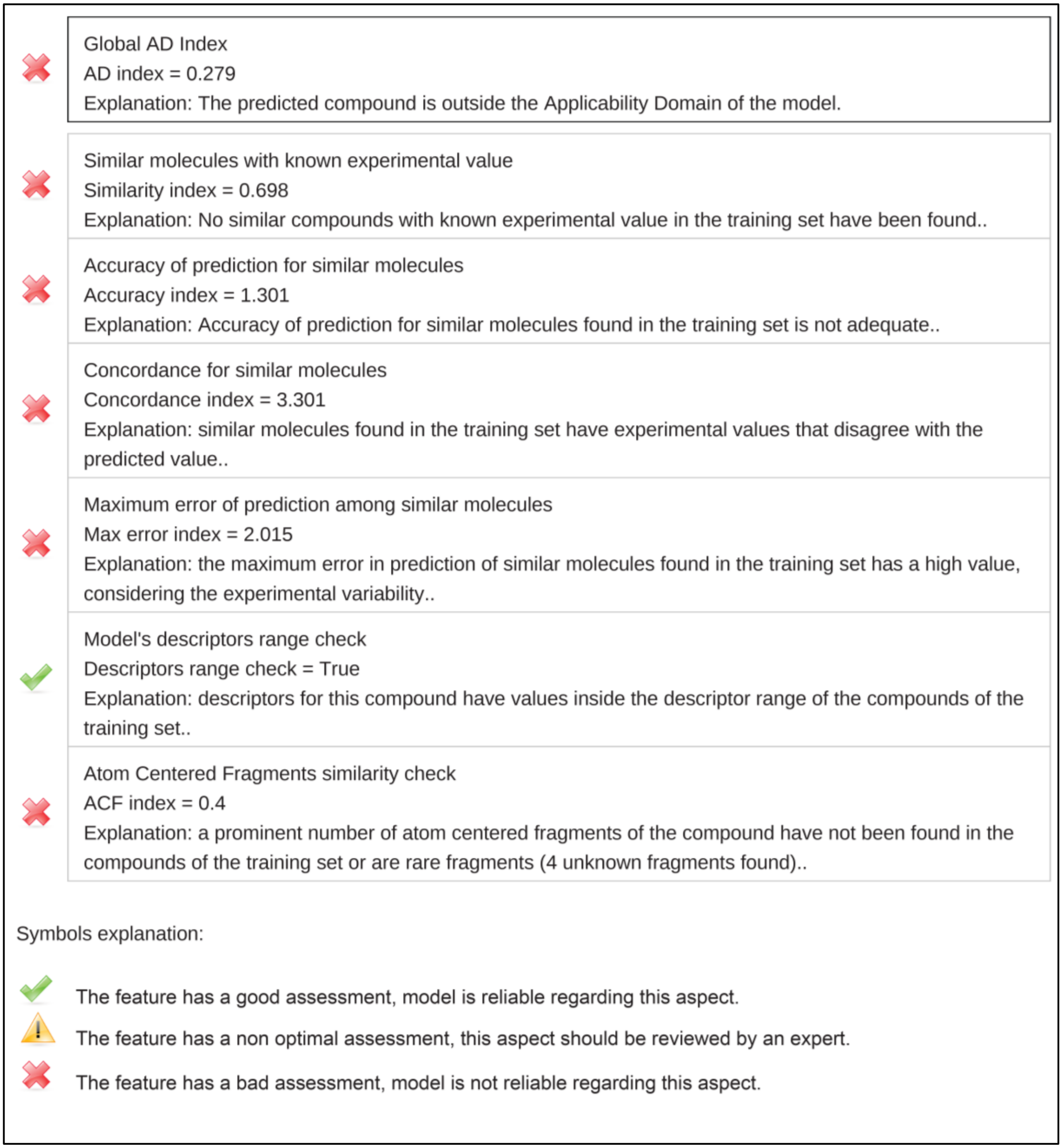
- No similar compounds with known experimental value in the training set have been found
- Accuracy of prediction for similar molecules found in the training set is not adequate
- similar molecules found in the training set have experimental values that disagree with the predicted value
- the maximum error in prediction of similar molecules found in the training set has a high value, considering the experimental variability
- a prominent number of atom centered fragments of the compound have not been found in the compounds of the training set or are rare fragments (4 unknown fragments found)

The following relevant fragments have been found: Carbonyl residue (SR 02); COOH group (PG 01)

Compound: Molecule 0
Compound SMILES: O=C(O)C(F)(F)OC1(F)(OC(F)(F)OC1(F)(OC(F)(F)F))
Experimental value: -
Predicted BCF [log(L/kg)]: 0.06
Predicted BCF [L/kg]: 1
Predicted BCF from sub-model 1 (HM) [log(L/kg)]: 0.27
Predicted BCF from sub-model 2 (GA) [log(L/kg)]: 0.12
Predicted LogP (MLogP): 0.45
Structural Alerts: Carbonyl residue (SR 02); COOH group (PG 01)
Reliability: The predicted compound is outside the Applicability Domain of the model
Remarks:
none

Figure 32. The cover of the BCF CAESAR model.

The report on the ADI is shown below (*Figure 33*).



✗	Global AD Index AD index = 0.279 Explanation: The predicted compound is outside the Applicability Domain of the model.
✗	Similar molecules with known experimental value Similarity index = 0.698 Explanation: No similar compounds with known experimental value in the training set have been found..
✗	Accuracy of prediction for similar molecules Accuracy index = 1.301 Explanation: Accuracy of prediction for similar molecules found in the training set is not adequate..
✗	Concordance for similar molecules Concordance index = 3.301 Explanation: similar molecules found in the training set have experimental values that disagree with the predicted value..
✗	Maximum error of prediction among similar molecules Max error index = 2.015 Explanation: the maximum error in prediction of similar molecules found in the training set has a high value, considering the experimental variability..
✓	Model's descriptors range check Descriptors range check = True Explanation: descriptors for this compound have values inside the descriptor range of the compounds of the training set..
✗	Atom Centered Fragments similarity check ACF index = 0.4 Explanation: a prominent number of atom centered fragments of the compound have not been found in the compounds of the training set or are rare fragments (4 unknown fragments found)..

Symbols explanation:

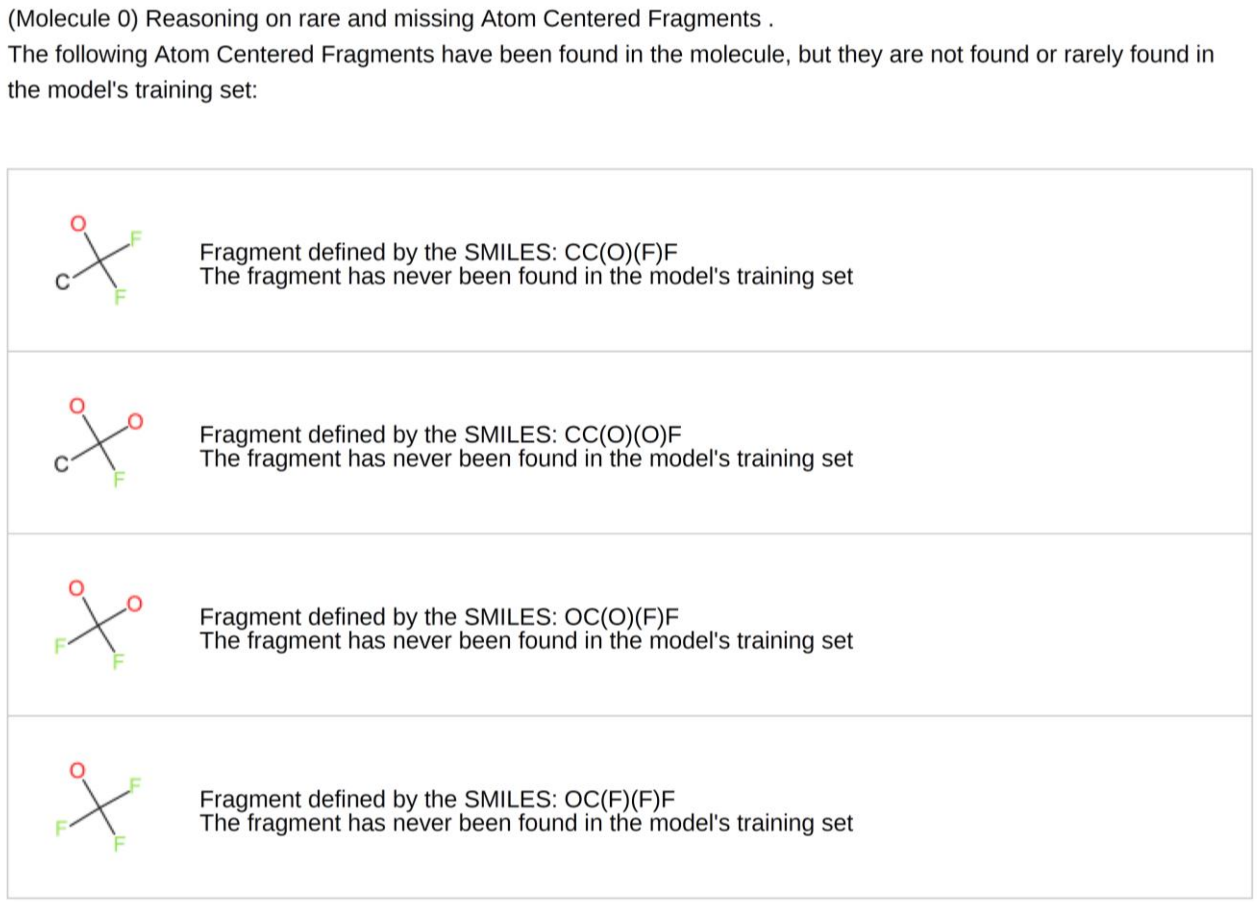
✓ The feature has a good assessment, model is reliable regarding this aspect.

⚠ The feature has a non optimal assessment, this aspect should be reviewed by an expert.

✗ The feature has a bad assessment, model is not reliable regarding this aspect.

Figure 33. The ADI of the BCF CAESAR model.

It contains many critical aspects for almost all the checks. There are not good similar substances. The predictions on the similar substances are not good. The experimental values of the similar substances disagree with the predicted value of the target compound, and in one or more cases the error is large. Furthermore, there are four unknown fragments in the target, which are shown below (*Figure 34*).



(Molecule 0) Reasoning on rare and missing Atom Centered Fragments .

The following Atom Centered Fragments have been found in the molecule, but they are not found or rarely found in the model's training set:

Fragment defined by the SMILES: CC(O)(F)F
The fragment has never been found in the model's training set

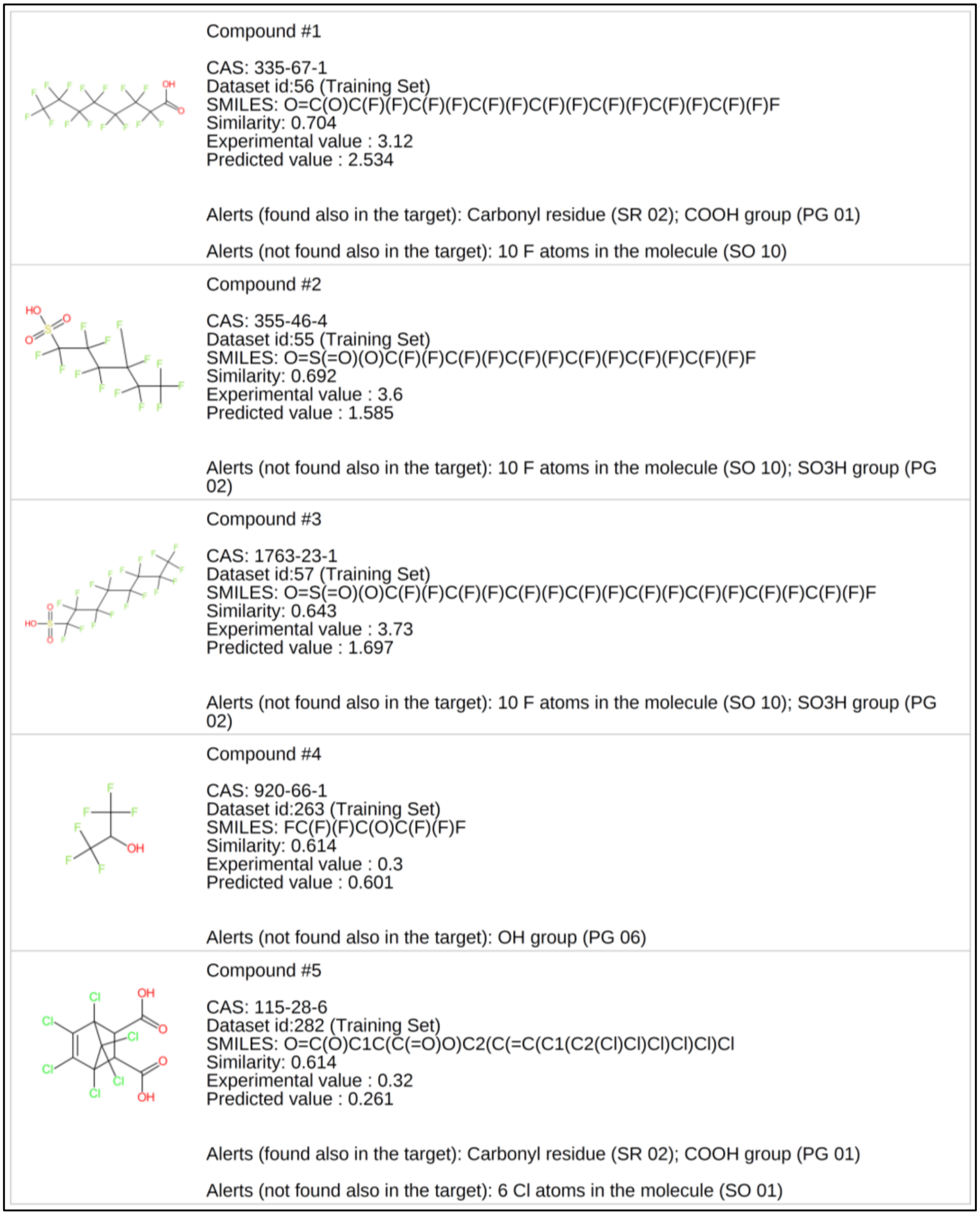
Fragment defined by the SMILES: CC(O)(O)F
The fragment has never been found in the model's training set

Fragment defined by the SMILES: OC(O)(F)F
The fragment has never been found in the model's training set

Fragment defined by the SMILES: OC(F)(F)F
The fragment has never been found in the model's training set

Figure 34. The rare and missing Atom Centered Fragments found by the BCF CAESAR model.

Below we show the most similar substances (*Figure 35*).



Compound #1

CAS: 335-67-1
Dataset id:56 (Training Set)
SMILES: O=C(O)C(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)(F)F
Similarity: 0.704
Experimental value : 3.12
Predicted value : 2.534

Alerts (found also in the target): Carbonyl residue (SR 02); COOH group (PG 01)

Alerts (not found also in the target): 10 F atoms in the molecule (SO 10)

Compound #2

CAS: 355-46-4
Dataset id:55 (Training Set)
SMILES: O=S(=O)(O)C(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)(F)F
Similarity: 0.692
Experimental value : 3.6
Predicted value : 1.585

Alerts (not found also in the target): 10 F atoms in the molecule (SO 10); SO3H group (PG 02)

Compound #3

CAS: 1763-23-1
Dataset id:57 (Training Set)
SMILES: O=S(=O)(O)C(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)(F)F
Similarity: 0.643
Experimental value : 3.73
Predicted value : 1.697

Alerts (not found also in the target): 10 F atoms in the molecule (SO 10); SO3H group (PG 02)

Compound #4

CAS: 920-66-1
Dataset id:263 (Training Set)
SMILES: FC(F)(F)C(O)C(F)(F)F
Similarity: 0.614
Experimental value : 0.3
Predicted value : 0.601

Alerts (not found also in the target): OH group (PG 06)

Compound #5

CAS: 115-28-6
Dataset id:282 (Training Set)
SMILES: O=C(O)C1C(C(=O)O)C2(C(=C(C1(C2(Cl)Cl)Cl)Cl)Cl)Cl
Similarity: 0.614
Experimental value : 0.32
Predicted value : 0.261

Alerts (found also in the target): Carbonyl residue (SR 02); COOH group (PG 01)

Alerts (not found also in the target): 6 Cl atoms in the molecule (SO 01)

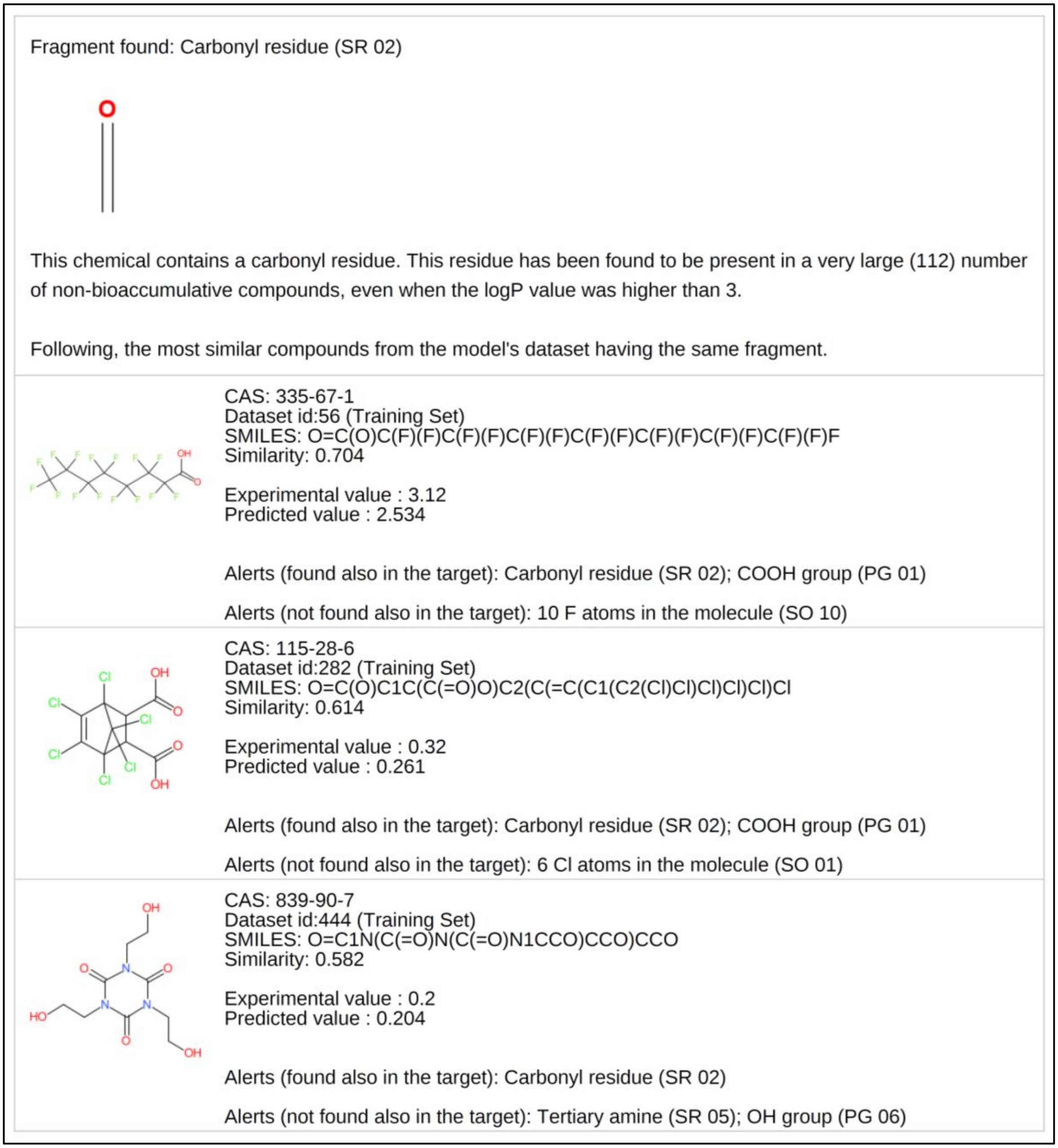
Figure 35. The five most similar chemicals found by the BCF CAESAR model.

The similarity values range from about 0.6 to about 0.7. These values are low. The first similar substance is a carboxylic acid with a perfluorinated carbon chain. It contains 10 carbon atoms and 15 fluorine atoms. Similar chemicals number 2 and 3 are less interesting, because they are perfluorinated sulphonic acids. The role of the fluorine atoms is already represented by similar 1. Thus, these two substances are not relevant. Similar number 4 is an alcohol, with 3 carbon and 6 fluorine atoms. The similar number 5 has a complex aliphatic cyclic structure, and two carboxylic groups, with chlorine atoms. Thus, it is not relevant. The same applies for similar number 6, not shown.

In the manual selection that we did, we considered structural elements, which are present in the target substance: the carboxylic acid with fluorine atoms, and the oxygen with fluorine atoms, in two substances with different number of carbons. The target compound has a carboxylic acid, 8 fluorine atoms, and four other groups. Unfortunately, the ether group linked to a carbon substituted with fluorine in a moiety which has been identified by VEGA as rare, thus we cannot expect similar substances with this group.

Furthermore, VEGA provides two elements for reasoning. The first one is the carbonyl moiety, as below (*Figure 36*).

VEGA indicates that this fragment is often associated to non bioaccumulative substances, and shows three similar substances (only the first one reported, here), both already shown among the most similar substances. The second moiety is the carboxylic acid, also associated with non bioaccumulative substances (not shown here).



Fragment found: Carbonyl residue (SR 02)

This chemical contains a carbonyl residue. This residue has been found to be present in a very large (112) number of non-bioaccumulative compounds, even when the logP value was higher than 3.

Following, the most similar compounds from the model's dataset having the same fragment.

CAS: 335-67-1
Dataset id:56 (Training Set)
SMILES: O=C(O)C(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)(F)F
Similarity: 0.704

Experimental value : 3.12
Predicted value : 2.534

Alerts (found also in the target): Carbonyl residue (SR 02); COOH group (PG 01)

Alerts (not found also in the target): 10 F atoms in the molecule (SO 10)

CAS: 115-28-6
Dataset id:282 (Training Set)
SMILES: O=C(O)C1C(C(=O)O)C2(C(=C(C1(C2(Cl)Cl)Cl)Cl)Cl)Cl
Similarity: 0.614

Experimental value : 0.32
Predicted value : 0.261

Alerts (found also in the target): Carbonyl residue (SR 02); COOH group (PG 01)

Alerts (not found also in the target): 6 Cl atoms in the molecule (SO 01)

CAS: 839-90-7
Dataset id:444 (Training Set)
SMILES: O=C1N(C(=O)N(C(=O)N1CCO)CCO)CCO
Similarity: 0.582

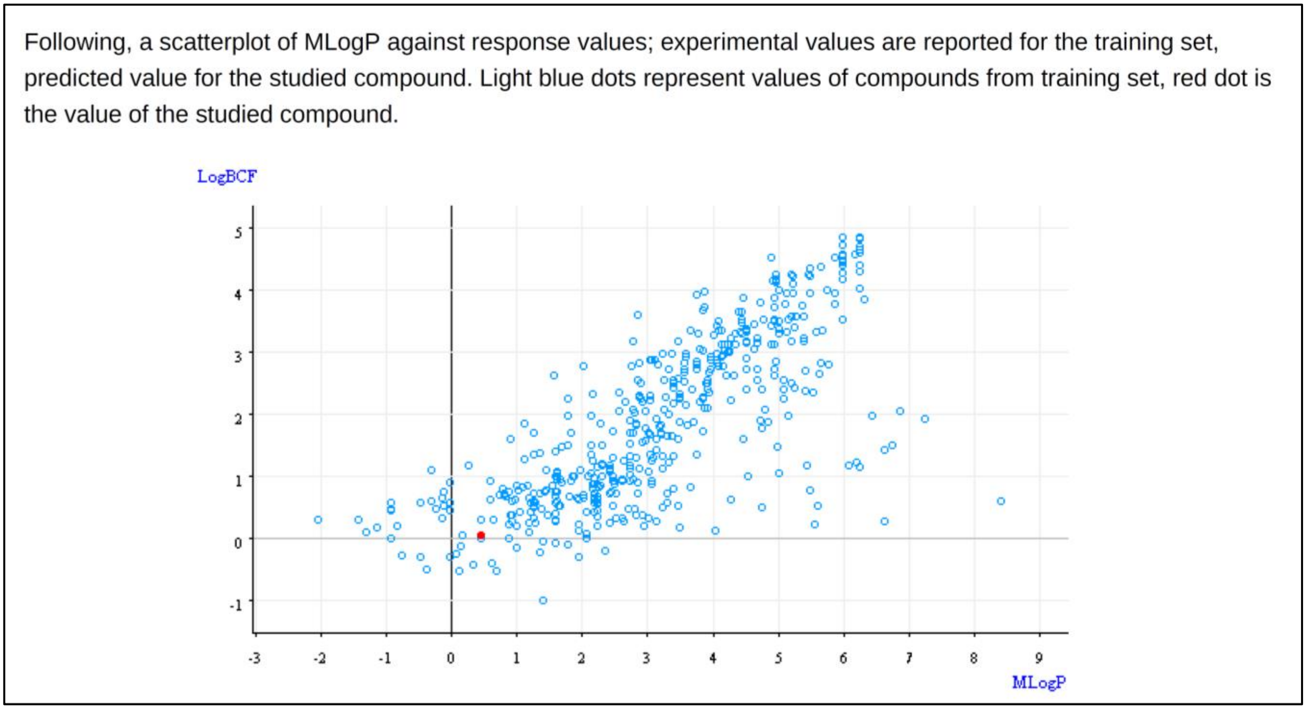
Experimental value : 0.2
Predicted value : 0.204

Alerts (found also in the target): Carbonyl residue (SR 02)

Alerts (not found also in the target): Tertiary amine (SR 05); OH group (PG 06)

Figure 36. The reasoning on the relevant fragments found by the BCF CAESAR model.

The plot of the experimental BCF values versus logKow is presented in *Figure 37*. It may be useful to relate the physico-chemical properties with the expected BCF value. From this plot we do not expect a high BCF value. However, the prediction of the logKow for the fluorinated substances, and for the target substance too, is expected to be uncertain. The reader may predict the logKow with VEGA for the target and verify that the predictions range from 0.45 to 6.2.

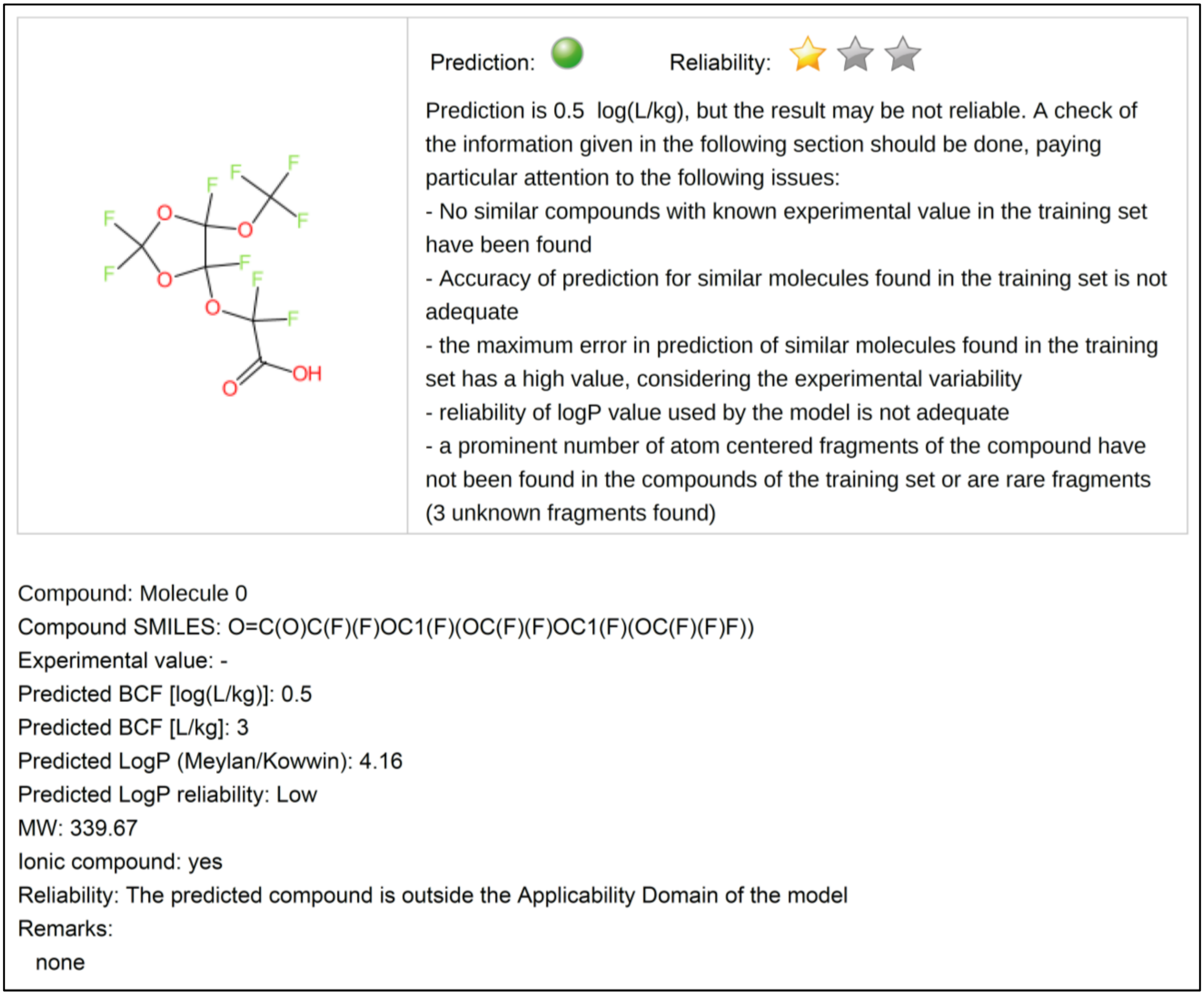


Following, a scatterplot of MLogP against response values; experimental values are reported for the training set, predicted value for the studied compound. Light blue dots represent values of compounds from training set, red dot is the value of the studied compound.

Figure 37. The analysis of molecular descriptors of the BCF CAESAR model.

<u>The Meylan model.</u>

The cover of the report is shown below (*Figure 38*).



Prediction: ● Reliability: ★☆☆

Prediction is 0.5 log(L/kg), but the result may be not reliable. A check of the information given in the following section should be done, paying particular attention to the following issues:
- No similar compounds with known experimental value in the training set have been found
- Accuracy of prediction for similar molecules found in the training set is not adequate
- the maximum error in prediction of similar molecules found in the training set has a high value, considering the experimental variability
- reliability of logP value used by the model is not adequate
- a prominent number of atom centered fragments of the compound have not been found in the compounds of the training set or are rare fragments (3 unknown fragments found)

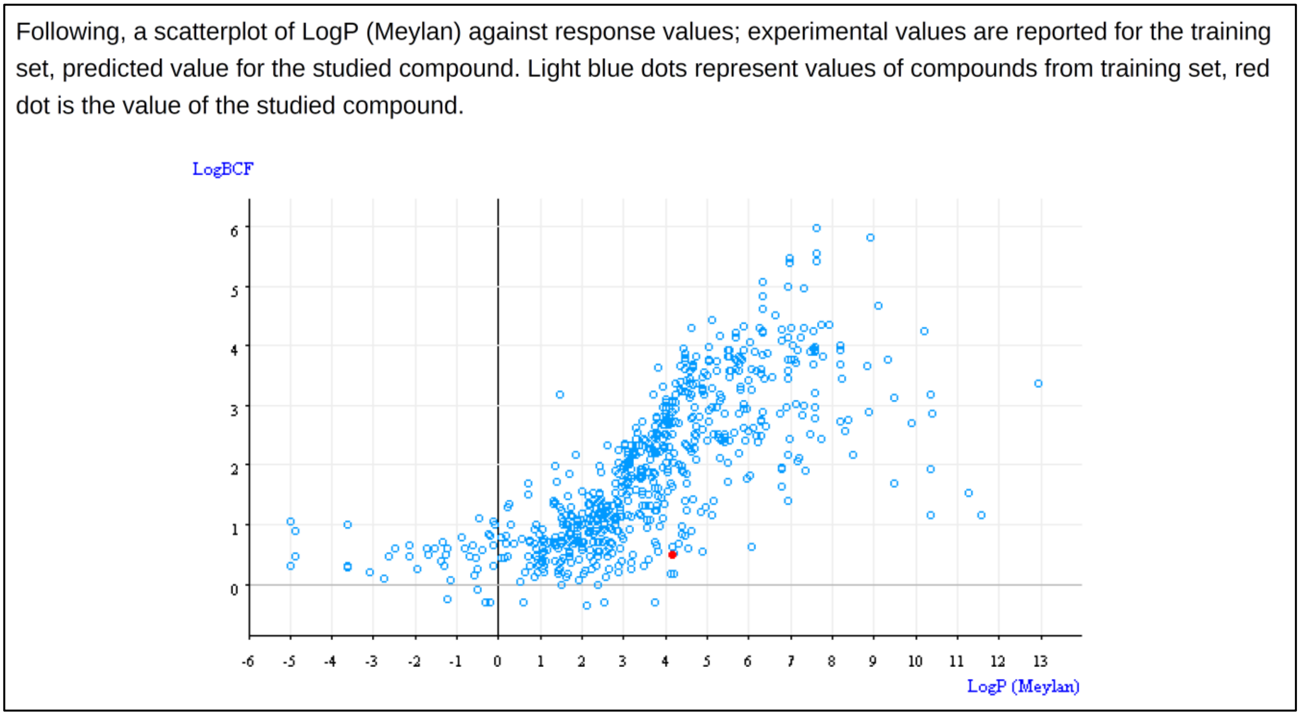
Compound: Molecule 0
Compound SMILES: O=C(O)C(F)(F)OC1(F)(OC(F)(F)OC1(F)(OC(F)(F)F))
Experimental value: -
Predicted BCF [log(L/kg)]: 0.5
Predicted BCF [L/kg]: 3
Predicted LogP (Meylan/Kowwin): 4.16
Predicted LogP reliability: Low
MW: 339.67
Ionic compound: yes
Reliability: The predicted compound is outside the Applicability Domain of the model
Remarks:
none

Figure 38. The cover of the BCF Meylan model.

The prediction is 0.5, and the value is uncertain. The ADI value is low, 0.247, indicating many critical aspects.

Similar substances have low similarity values, 0.623 or lower. The only similar substance of some interest is the same alcohol found in the CAESAR model. In this case, the reported experimental value is 0.4, instead of 0.3.

The plot of the experimental BCF values versus logKow is presented below (*Figure 39*).

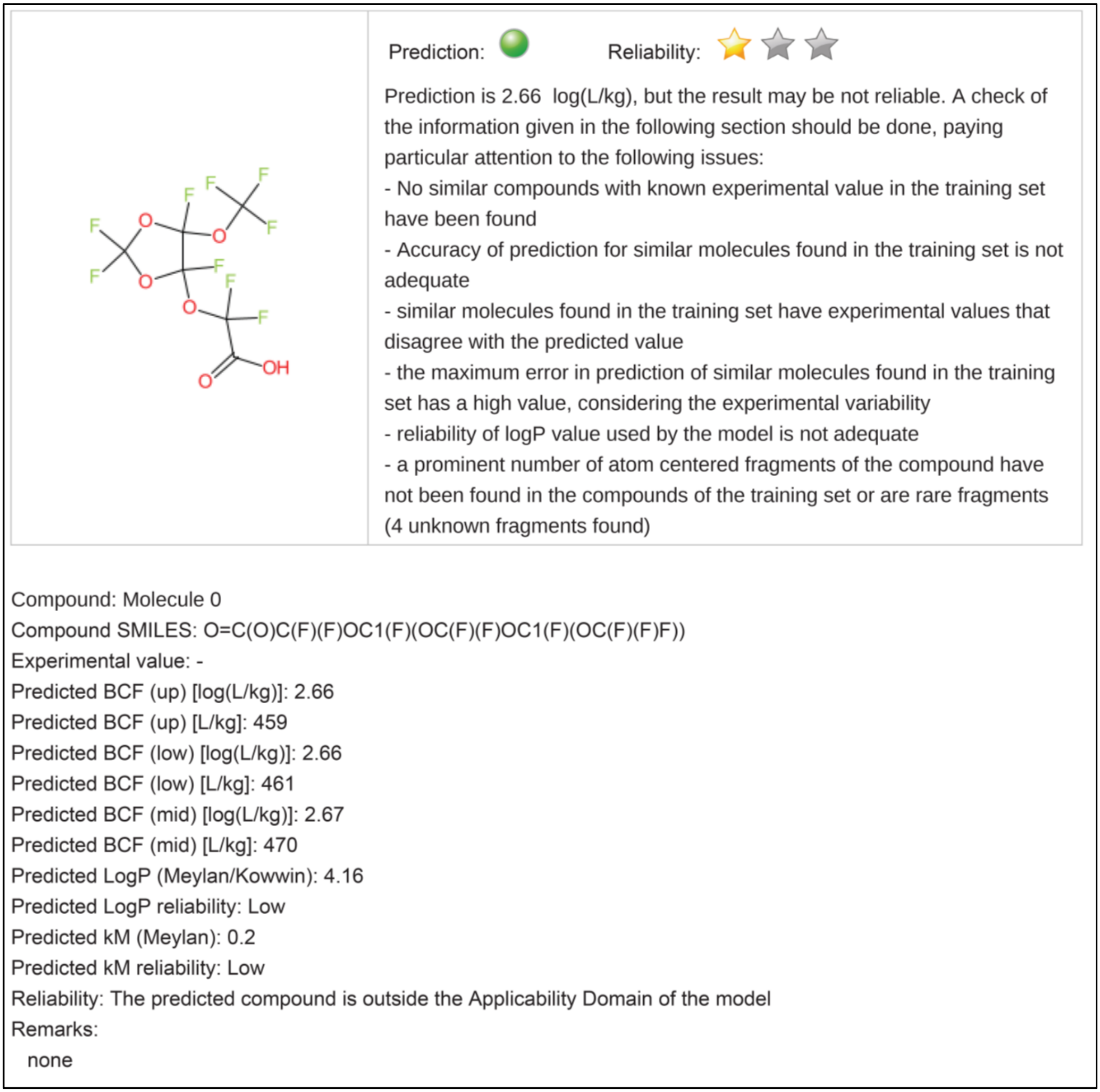


Following, a scatterplot of LogP (Meylan) against response values; experimental values are reported for the training set, predicted value for the studied compound. Light blue dots represent values of compounds from training set, red dot is the value of the studied compound.

Figure 39. The analysis of molecular descriptors of the BCF Meylan model.

<u>The Arnot-Gobas model.</u>

The cover of the report is shown below (*Figure 40*).

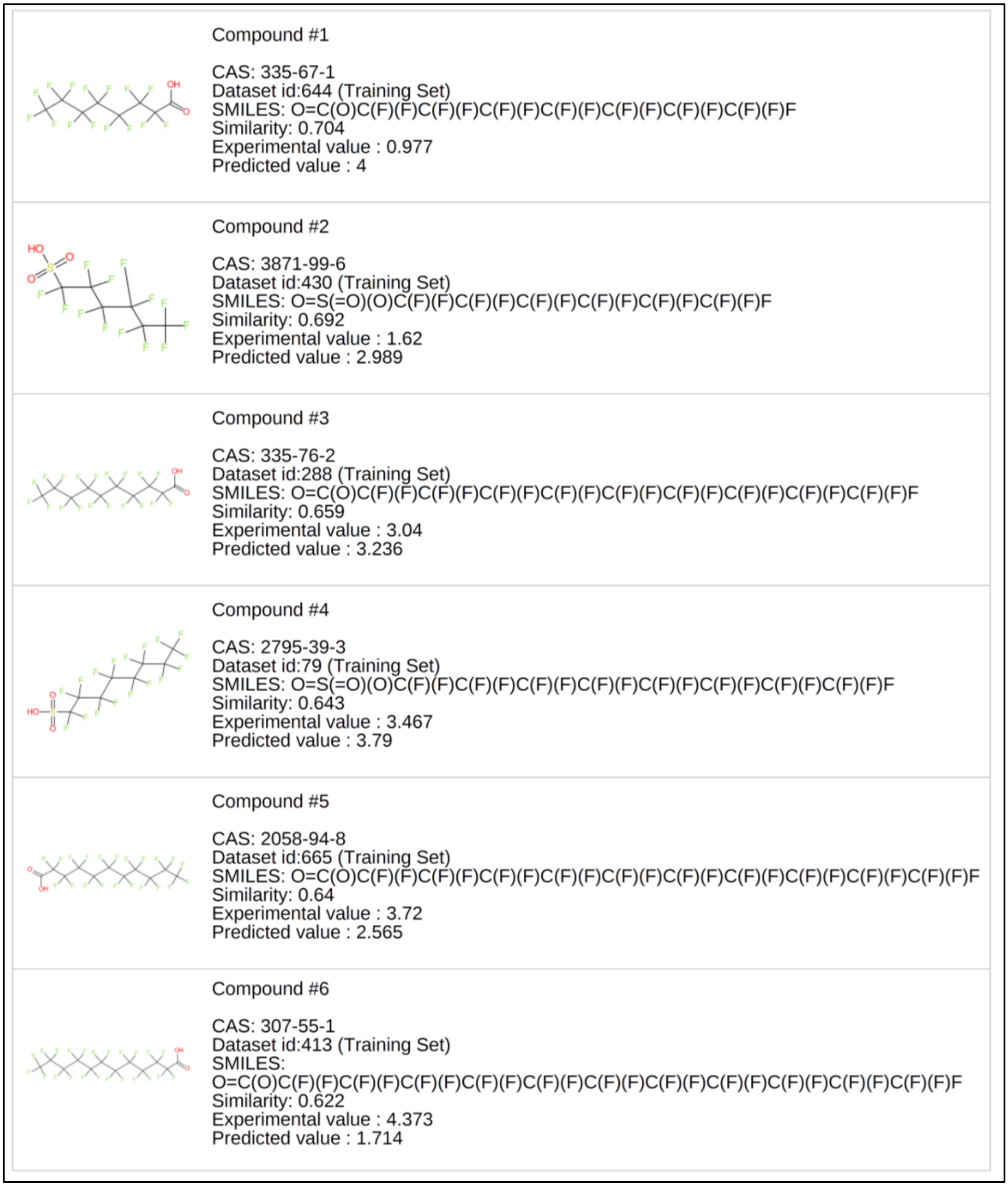


Prediction: ● Reliability: ★☆☆

Prediction is 2.66 log(L/kg), but the result may be not reliable. A check of the information given in the following section should be done, paying particular attention to the following issues:
- No similar compounds with known experimental value in the training set have been found
- Accuracy of prediction for similar molecules found in the training set is not adequate
- similar molecules found in the training set have experimental values that disagree with the predicted value
- the maximum error in prediction of similar molecules found in the training set has a high value, considering the experimental variability
- reliability of logP value used by the model is not adequate
- a prominent number of atom centered fragments of the compound have not been found in the compounds of the training set or are rare fragments (4 unknown fragments found)

Compound: Molecule 0
Compound SMILES: O=C(O)C(F)(F)OC1(F)(OC(F)(F)OC1(F)(OC(F)(F)F))
Experimental value: -
Predicted BCF (up) [log(L/kg)]: 2.66
Predicted BCF (up) [L/kg]: 459
Predicted BCF (low) [log(L/kg)]: 2.66
Predicted BCF (low) [L/kg]: 461
Predicted BCF (mid) [log(L/kg)]: 2.67
Predicted BCF (mid) [L/kg]: 470
Predicted LogP (Meylan/Kowwin): 4.16
Predicted LogP reliability: Low
Predicted kM (Meylan): 0.2
Predicted kM reliability: Low
Reliability: The predicted compound is outside the Applicability Domain of the model
Remarks:
none

Figure 40. The cover of the BCF Arnot-Gobas model.



Compound #1

CAS: 335-67-1
Dataset id:644 (Training Set)
SMILES: O=C(O)C(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)(F)F
Similarity: 0.704
Experimental value : 0.977
Predicted value : 4

Compound #2

CAS: 3871-99-6
Dataset id:430 (Training Set)
SMILES: O=S(=O)(O)C(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)(F)F
Similarity: 0.692
Experimental value : 1.62
Predicted value : 2.989

Compound #3

CAS: 335-76-2
Dataset id:288 (Training Set)
SMILES: O=C(O)C(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)(F)F
Similarity: 0.659
Experimental value : 3.04
Predicted value : 3.236

Compound #4

CAS: 2795-39-3
Dataset id:79 (Training Set)
SMILES: O=S(=O)(O)C(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)(F)F
Similarity: 0.643
Experimental value : 3.467
Predicted value : 3.79

Compound #5

CAS: 2058-94-8
Dataset id:665 (Training Set)
SMILES: O=C(O)C(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)(F)F
Similarity: 0.64
Experimental value : 3.72
Predicted value : 2.565

Compound #6

CAS: 307-55-1
Dataset id:413 (Training Set)
SMILES:
O=C(O)C(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)(F)F
Similarity: 0.622
Experimental value : 4.373
Predicted value : 1.714

Figure 41. The similar chemicals found by the BCF Arnot-Gobas model.

The prediction is 2.66, and the value is uncertain. The ADI value is low, 0.279, indicating critical aspects for all the components of the index.

Figure 41 shows the most similar substances.

As we discussed above, those of some interests are the perfluorinated carboxylic acids. There are four of them, with 8, 10, 11 and 12 carbons. The substance with 10 atoms is the same as in the CAESAR model, with a very similar BCF value as that in CAESAR: 3.04 versus 3.12. The analogue with 8 atoms has a BCF value of 0.977. This seems a quite low value compared to the analogue with 10 carbons and those with 11 and 12 carbons: adding two carbons, from 10 to 12, increases the value of about 1.3 log units.

The **KNN model** does not make any assessment, since the similar substances have a similarity value below the threshold required by the tool to make a reliable prediction. We notice that anyhow this model shows the most similar substances, which can be used for read-across. These substances are those which have already appeared in other models. The minor difference is that the experimental value of the perfluorinated isopropyl alcohol is 0.244.

Overall assessment.

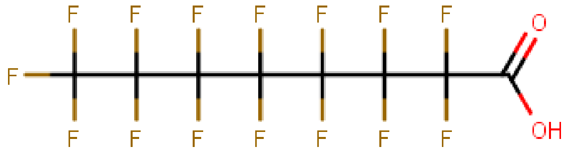
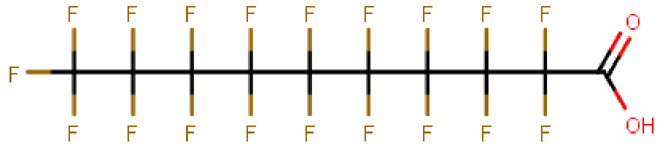
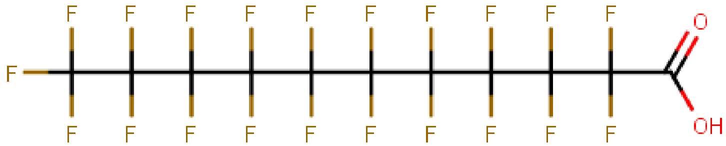
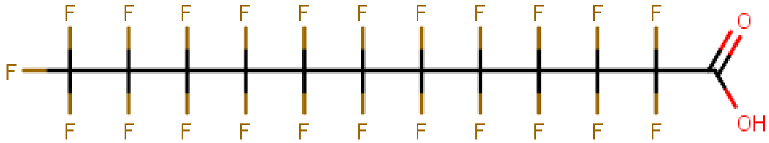
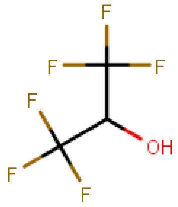
Let’s consider the three lines of evidence provided by VEGA.

- Similar substances. There are similar substances, reported in *Table 4*.

Considering the sequence of the BCF values for the analogue with 12, 11 and 10 carbons, the increase of the BCF is of about 0.65-0.68 log unit per added carbon. Proceeding with this sequence, the analogue with 8 carbons may have a BCF value at about 1.7. The experimental values for this substance are 0.977-3.12, with an average value of about 2. We can use the value of 1.85, which has a high level of uncertainty.

For the last analogue, there are multiple values, but we can use the value of 0.3. This value seems with low uncertainty.

Table 4. Similar substances found in four models.

	Perfluoroactanoic acid. Experimental values: 0.977-3.12
	Perfluorodecanoic acid. Experimental value: 3.04
	Perfluoroundecanoic acid. Experimental value: 3.72
	Perfluorododecanoic acid. Experimental value: 4.373
	Hexafluoro-2-propanol Experimental values: 0.4, 0.3, 0.244

To further analyse these similar substances, we should consider now the potential role of the different molecular components of these similar compounds, compared to the target, based on the mechanism. These useful components are listed in *Table 5*.

Table 5. Molecular component of the target and the two most similar substances.

Substance	N Carbon	N Fluorine	N Oxygen	N carboxylic group	BCF value
C604 (target)	6	9	6	1	
Perfluorooctanoic acid	8	15	2	1	1.85
Perfluoroisopropylalcohol	3	6	1	0	0.3

The number of carbons is increasing the BCF value, since it increases logKow. Thus, comparing the carbon atoms, we expect that the target will have the BCF between the two analogues, but closer to 1.85.

The influence of fluorine is expected to be lower. To explore it, we may look at the BCF value of the isopropyl alcohol, not fluorinated. Using VEGA, we do not have an experimental value, but the predicted values range from about 0 to 0.5. Thus, the value of the two alcohols, with and without fluorine, is expected to be similar.

The oxygen plays an opposite role, compared carbon, but this happens in particular for alcohols and carboxylic groups. Thus, we may expect that the BCF value will be closer to the value of the acid, 1.85, but lower, since the octanoic acid is larger.

As a conclusion, considering the analogues, we can assume that C604 has the BCF value between the values of the analogues, but closer to 1.85. The range of values of the predictive models is consistent with this, but with more uncertainty.

The VERMEER and CONCERT REACH LIFE projects collaborated with the LIFE PHOENIX project (*https://www.lifephoenix.eu/*). After we did this weight-of-evidence exercise on C604, within the PHOENIX LIFE project the BCF value of C604 was measured: 1.3. This shows that it is possible to use the results of VEGA, even if the uncertainty of the models is high. It is important to use all the three lines of evidence, as discussed above.

Based on what discussed with the last example, we can observe the following. In this particular case, the contribution of the *in silico* prediction is limited. We can imagine that each model provides a range of values, and we can see if there is an area where there is an overlap of the values predicted by the different models. This is what we did. Does it mean that the *in silico* models do not contribute to the overall assessment? No. Indeed, the results would be very different if, for instance, all the models provided predictions in the range from 3 to 6. In this case, the final uncertainty of the weight-of-evidence would be much higher, with the very different values from the read-across and the *in silico* models. Surely, the contribution of the *in silico* models is quite limited, as support information. From other sources (not given by VEGA) we know that the perfluorinated substances have a very peculiar behaviour. They do not like the water, but do not like the lipidic phase either. Thus, the common parameter to evaluate BCF, logKow, does not help very much. Furthermore, on a biochemical point of view the mechanisms are quite different from classical substances, too. These substances may prefer the proteins and the clearance processes vary. This makes very difficult this exercise. In cases like what we presented above, it may be preferable to rely on similar substances, even if these similar substances are not very similar. In our case we have two similar. One has BCF values quite accurate, based on the reproducibility of the results from different studies. The other substance has an uncertain value. Two observations, with one uncertain, is not the ideal situation. But we have seen how to introduce further elements, additional lines of evidence, based on the reasoning that we can make about the role of certain structural components. This provides further input, thus we have not only the observation, but also some theoretical basis. It is the combination of all the elements which can help in our assessment.

B.2.4. A short summary of the VEGA models

The description of the VEGA models is available from VEGAHUB in the download section, with QMRF of the individual models (when a new model is implemented, it may require some time to upload its QMRF): *https://www.vegahub.eu/portfolio-item/vega-qsar-models-qrmf/*.

Here we simplify provide some short comments.

If you want to have the integrated results of the different models for a specific endpoint, you can use the JANUS tool (see below).

Human Toxicity

Models for mutagenicity, bacterial reverse mutation test – Ames test.

VEGA provides 4 models, plus a consensus model, which integrates the overall results. There is another model specific for aromatic amines.

The predictions for mutagenicity generally are quite reliable. This is due to the large set of substances with experimental values, to the relatively simple endpoint, and to the availability of several good models. The experimental data are quire reproducible, 85-90%.

The **Mutagenicity (Ames test) CONSENSUS model** provides two scores, representing the reliability of the prediction as mutagenic and non-mutagenic. The label for the prediction is indicated based on the highest score. In the case that the two scores have the same value, the label is mutagenic since the conservative outcome is used. The score value ranges from 0 to 1. If the value is > 0.5, the overall prediction is reliable. If it is between 0.35 and 0.5, the prediction can be used, but the values should be checked carefully. If it is below 0.35, very careful check should be done, also using read-across. The four individual general models are quite different. The **Mutagenicity (Ames test) model (ISS)** is expert-based. The **Mutagenicity (Ames test) model (SarPy-IRFMN)** is a statistical one. However, from a practical point of view, most of the rules which are present in the ISS model are also present in the SARpy model. The **Mutagenicity (Ames test) model (CAESAR)** is hybrid integrating a statistical model with rules from the ISS model. The **Mutagenicity (Ames test) model (KNN-Read-Across)**, as all KNN models, is a non-parametric model, thus it is not a statistical one, but surely it is not expert-based.

- The **Mutagenicity (Ames test) model (ISS)** is the same as in Toxtree, and refers to the list of structural alerts known as Benigni-Bossa, after the names of the two researchers at the Istituto Superiore di Sanità (ISS), Rome, Italy. These rules are based on expert opinion. Thus, the model is expert-based. We took the database of mutagenic substance from the ISS, for the evaluation of the ADI. This model is expert-based, and the structural alerts (SAs) are explained with their mechanism by the authors (Benigni-Bossa). However, there are exception rules not explained. Some SAs are supported by very few substances to derive the mechanism, and some of them are present in substances which are mainly non-mutagenic, thus these SAs generate false positives. The ISS database is quite small, about 1000 substances, compared with the datasets used for other models for the same endpoint.
- The **Mutagenicity (Ames test) model (SarPy-IRFMN)** is based on more than 200 rules. These rules have been extracted using the SARpy software, thus this model is computer-based. The rules indicate both mutagenicity and lack of mutagenicity. Thus, SARpy is different from the ISS model which contains a few tens of SAs only for mutagenicity (however there are some exception rules, which indicate indeed lack of effect). Most of the SAs present in the ISS model are also present as SARpy rules. Actually, SARpy has more rules, because it specifies different conditions which are grouped within the same SA.

- The **Mutagenicity (Ames test) model (CAESAR)** is a hybrid model which combines a statistical model with a set of SAs of the ISS model.
- **Mutagenicity (Ames test) model (KNN-Read-Across)** is a KNN model, thus results are reliable when there are similar substances. The KNN models is not parametric, thus the "average" label is used, without any reasoning on other factors.

Mutagenicity (Ames test) model for aromatic amines (CONCERT/IRFMN) is specific for this class.

Note. Each model has its own dataset. Thus, for read-across, it is useful to look at all the models and the similar substances they provide. For this reason, it derives that the best assessment of the results is done, first considering the individual results, but then the overall picture, with feed-back from different models and datasets.

Note. It may be that the same substance has different call in two models, because the datasets refer to different collections of data.

The Developmental toxicity models

These endpoints are very complex. Only limited data are available. There are actually several separate endpoints generating this kind of toxicity, and the models can capture only some of them. Thus, it is very important to run both models, as below, and look if there is agreement and if the predictions have high reliability (thus if they are supported by similar substances – read-across perspective). If not, the results should be used very carefully.

- The **Developmental Toxicity model (CAESAR)**. It refers to developmental toxicity. It should be used only if the reliability is high, and preferably if there is at least one similar substance supporting the assessment. It is a conservative model, thus there may be false positives.
- The **Developmental/Reproductive Toxicity library (PG) model.** It addresses developmental and reproductive toxicity. It should be used only if the reliability is high, and preferably if there is at least one similar substance supporting the assessment. Compared to the CEASAR model, it indicates if the substance is reproductive toxicity. The model has been developed by P&G, and it is based on a long series of skeletons of molecules which are expected to by reproductive or developmental toxicity, and all the possible substituents in these skeletons related to the effects have been identified.

The carcinogenicity models

There are several carcinogenicity models in VEGA. Some are classifiers, others provide a continuous value for the carcinogenicity potency.

The classifiers include an expert-based model (ISS), and computer-based models.

Another important difference regards the definition of the endpoint. Indeed, some models consider as endpoint the results of in vivo studies, while other models define carcinogenicity based on the expert opinion, which considers not only the in vivo data, but also other elements.

The carcinogenicity endpoint is more complex than mutagenicity. The uncertainty and variability of this endpoint is quite high: the reproducibility is about 60%.

- The **Carcinogenicity model (ISS)** is an expert-based model, is a classification model and the definition of carcinogenicity is based not only on in vivo results. The ISS model includes some tens of SAs. It is the same model as implemented in Toxtree. All the SAs for mutagenicity as in the mutagenicity ISS model are SAs for carcinogenicity too. In addition to these SAs, there are a few others for non-genotoxic carcinogenicity.
- The **Carcinogenicity model (IRFMN-ISSCAN-CGX)** This model has been developed using SARpy. The list of substances comes from two sources, obtained by expert opinion: one is from ISS, and one from another expert, David Kirkland.
- The **Carcinogenicity model (IRFMN-Antares)** has been developed using SARpy. It refers to in vivo (rat, male and female) carcinogenicity.
- The **Carcinogenicity model (CAESAR)** is a computer-based model predicting in vivo carcinogenicity (in rats, male and female).
- The **Carcinogenicity inhalation classification model (IRFMN)** is a classifier. The route of exposure is inhalation. The model has been developed using data from the Risk Assessment Information System (RAIS) Toxicity values database (*https://rais.ornl.gov*).
- The **Carcinogenicity oral classification model (IRFMN)** is a classifier. The route of exposure is oral. The model has been developed using data from the Risk Assessment Information System (RAIS) Toxicity values database (*https://rais.ornl.gov*).
- The **Carcinogenicity inhalation Slope Factor model (IRFMN)** provides a continuous value based on the slope factor. This model should be used joined to the related classifier, if the prediction is positive.

- The **Carcinogenicity oral Slope Factor model (IRFMN)** provides a continuous value based on the slope factor. This model should be used joined to the related classifier, if the prediction is positive.
- The **Carcinogenicity in male rat (CORAL) model** is a classifier.
- The **Carcinogenicity in female Rat (CORAL)** model is a classifier.

The Acute Toxicity (LD50) model

Acute Toxicity (LD50) model (KNN) is a KNN model providing quantitative LD50 values. The KNN model is not parametric, thus the "average" label is used, without any reasoning on other factors.

The Skin Sensitization models

There are several models, which should be used jointly, to increase the overall reliability.

- **Skin Sensitization model (CAESAR)** is a classifier. The model provides a qualitative prediction of skin sensitization on mouse (Local Lymph Node Assay (LLNA)). The model consists of a logic procedure called Adaptive Fuzzy Partition (AFP) based on 8 descriptors. The model is over-conservative.
- **Skin Sensitization model (IRFMN-JRC)** is a classifier. It provides a qualitative prediction of skin sensitization on mouse (LLNA). The model consists in a decision tree based on 8 2D DRAGON descriptors. The model was built using the Recursive PARTitioning (rpart) module included in R software.
- **Skin Sensitization model (NCSTOX)** is a classifier. It was developed to evaluate 20.000 botanical compounds within the NCSTOX project. It is based on the combination of three different models: (a) CAESAR, (b) TOXTREE and (c) a new ad hoc fragment-based model.
- **Skin Sensitization model (TOXTREE)** identifies different mechanisms of action for chemicals defining the reactivity domains involved in the process. The model designs a series of SMARTS patterns capable of defining these reactivity domains. This was carried out using a large database of LLNA data.
- **Skin Sensitization (CONCERT/Kode)** is a classifier. It was developed using an Artificial Neural Network (ANN) approach. It is based on a LLNA dataset.
- **Skin Sensitization (CONCERT/SarPy)** is a classifier. It was developed using the SARpy software, extracting 89 active and inactive rules from a LLNA dataset.

The Skin Irritation models

Within the CONCERT REACH project three statistical models have been implemented, the first one using descriptors (**Skin Irritation (CONCERT/Kode)**), the second using CORAL (thus small molecular or atomic features derived from the SMILES string; **Skin Irritation (CONCERT/Coral)**), and the third using SARpy (thus using molecular fragments associated to the effect; **Skin Irritation model (CONCERT/SarPy)**.

The Eye Irritation models

Within the CONCERT REACH project three statistical models have been implemented, the first one using descriptors (**Eye Irritation (CONCERT/Kode)**), the second using a KNN approach (**Eye Irritation (CONCERT/KNN)**), and the third using SARpy (thus using molecular fragments associated to the effect; **Eye Irritation (CONCERT/SarPy)**).

The Chromosomal aberration model

Chromosomal aberration model (CORAL) has been done using CORAL (thus small molecular or atomic features derived from the SMILES string).

The micronucleus assays

- **In vitro Micronucleus activity** (IRFMN-VERMEER) has been developed with SARpy (thus using molecular fragments associated to the effect). It is a classifier.
- **In vivo Micronucleus activity (IRFMN)** is an integrated model: it combines the predictions from a model done with SARpy and another one using KNN**.**

The Estrogen Receptor effect models

- **Estrogen Receptor-mediated effect (IRFMN-CERAPP)** is based on the data provided by the CERAPP initiative which provided values on estrogen receptor effects. The call is based on a composite number of assays. It is done using SARpy**.**
- **Estrogen Receptor Relative Binding Affinity model (IRFMN)** is based on date related to the binding to the receptor. Thus, compared to the CERAPP model, is focused on the initial steps in the activity.

The Androgen Receptor effect model

Androgen Receptor-mediated effect (IRFMN-COMPARA) is based on the data provided by the COMPARA initiative which provided values on androgen receptor effects. The call is based on a composite number of assays. It is done using SARpy.

The Thyroid Receptor effect models

These models are two of the several models on nuclear receptors developed by the Nanjing University (Prof. Wei Shi). The models (**Thyroid Receptor Alpha effect (NRMEA)**, **Thyroid Receptor Beta effect (NRMEA)**) used the SARpy software to extract rules.

The Glucocorticoid Receptor effect models

Glucocorticoid Receptor (OBERON) has been developed by the Nanjing University (Prof. Wei Shi). The model used the SARpy software to extract rules.

The Thyroperoxidase Inhibitory Activity model

Thyroperoxidase Inhibitory Activity (OBERON) has been developed within the Oberon project. It is a KNN model based on DRAGON and other descriptors.

The Endocrine Disruptor activity model

Endocrine Disruptor activity screening (IRFMN) is a model based on the lists of substances suspected to be endocrine disruptors according to the EC or WHO. Based on these lists, chemical moieties have been extracted using SARpy and then manual processing. Compared to the other models towards specific endocrine disruptor endpoints, this model is a general one.

The NOAEL models

We recommend to run models for NOAEL and LOAEL jointly, to check the consistency of the results (LOAEL should occur at higher doses than NOAEL, and their difference should not be too large).

- **NOAEL (CONCERT/Coral)** is a quantitative model to predict NOAEL starting from data for oral subchronic toxicity in rats. It is a model based on optimal descriptors calculated by the Monte Carlo algorithm, using the CORAL software.
- **NOAEL (IRFMN-CORAL)** is a quantitative model to predict NOAEL for subchronic toxicity in rats. It is based on optimal descriptors calculated with simplified molecular input-line entry systems and the graph of atomic orbitals by the Monte Carlo method, using CORAL software.
- **Liver NOAEL (CORAL)** provides NOAEL predictions for liver toxicity. Also this model is based on rats subchronic toxicity data, and it was developed using CORAL software.

The LOAEL models

Similarly to what is described above, the LOAEL models were developed using the CORAL software and they are based on subchronic toxicity data on rats.

We recommend to run models for NOAEL and LOAEL jointly, to check the consistency of the results (LOAEL should occur at higher doses than NOAEL, and their difference should not be too large).

Two models are present, one for general toxicity and one for organ-specific (liver) toxicity:

- **LOAEL (CONCERT/Coral)**
- **Liver LOAEL (CORAL)**

The Cramer classification model

Cramer classification (TOXTREE) is a decision tree model. The model classifies chemicals, based on their level of oral toxicity, in one the three Cramer classes (Class I, Class II, Class III). The Cramer decision tree is the most commonly used approach is the application of the Threshold of Toxicological Concern (TTC) concept.

The Hepatotoxicity model

Hepatotoxicity model (IRFMN) is a classifier model individuating the presence in the molecule of fragments associated to the effects. Effects are derives from data on humans from the literature (DILI). It identifies structural alerts associated to the effect.

Ecotoxicity

The BCF models

There four BCF models in VEGA. The **BCF model (CAESAR)** model is a hybrid model based on two separate computer-based models. The **BCF model (Meylan)** and **BCF model (Arnot-Gobas)** models are reimplementation of those available within EPISuite. The **BCF model (KNN-Read-Across)** is a non-parametric model. This is the most different, compared to the previous models, because in the other models logKow has a main role.

Since logKow has a main role for three of the four models, it is important to understand the consequences of this. Let's consider the figure below (*Figure 42*), which is taken from a previous exercise. We notice three kinds of behaviour of the substances. Those on the left, with logKow below 0, have BCF value which is low. Actually, all of these substances are not bioaccumulative, and it is not so important to differentiate them. In the middle of the plot, where most of the substances are, there is a linear trend between logKow and BCF, until logKow 6-7. Then, the values with logKow values > 7 show an opposite trend, with a large spread of values. Here we have the largest uncertainty. We can imagine that the relationship between logKow and BCF is not linear.

There is a maximum, and then the curve may decrease. This may be due to several factors, but surely a substance with very high logKow value is poorly water-soluble, and thus, it does not reach the fish, in the experiment done in water (for bioconcentration only the uptakes form water through gills or skin are considered). We can also imagine that there are multiple Gaussian curves, for different families of chemical substances, and the maximum BCF value is reached at different logKow values. Thus, what we observe in the figure below, it the overlap of many curves, for different clusters of substances, and indeed in some cases the maximum can be reached at quite low logKow value, so that we can observe substances with BCF values lower than expected even in the central part of the plot.

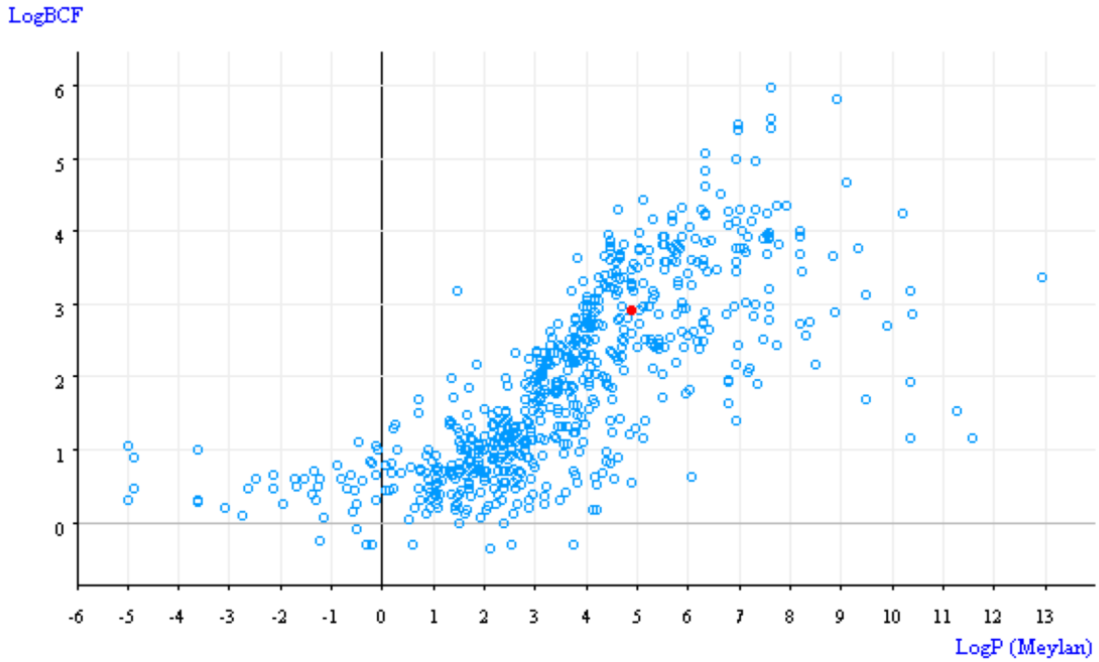


Figure 42. The analysis of molecular descriptors.

http://www.caesar-project.eu/index.php?page=results&section=endpoint&ne=1

Full details of the paper:

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2913327/pdf/1752-153X-4-S1-I1.pdf.

The Aquatic Acute Toxicity models

There are several models for fish acute toxicity, some are for a specific fish species and others refer to fish in general (i.e. they were obtained using a dataset with data for more than one species). In particular, two models refer to *Pimephales promelas* (Fathead minnow): **Fathead Minnow LC50 96h (EPA), Fathead Minnow LC50 model (KNN-IRFMN)**; the first is the reimplementation of the single model present in the US EPA T.E.S.T tool, the second a KNN. The specie-specific model that refers to *Poecilia reticulata* (Guppy, **Guppy LC50 model (KNN-IRFMN)**) is another KNN. The two specie-specific KNN models are based on small datasets, therefore they often give no prediction. Another KNN model is the **Fish Acute (LC50) Toxicity model (KNN-Read-Across)** (https://doi.org/10.1007/978-3-030-16443-0_18) The **Fish Acute (LC50) Toxicity model (NIC)** (https://doi.org/10.1186/s13321-017-0218-y)is a Counter Propagation Artificial Neural Network (CP ANN). The **Fish Acute (LC50) Toxicity model (IRFMN)** is a Tree Ensemble Random Forest (https://doi.org/10.3390/molecules26226983). Unlike the previous models, based on general industrial chemicals, the **Fish Acute (LC50) Toxicity model (IRFMN-Combase)** is based on a set of biocides and developed using CORAL. The last model available for fish acute toxicity, **Fish Acute (LC50) Toxicity classification (SarPy-IRFMN)**, is a classification model built using the SARpy tool (doi.org/10.23937/2572- 4061.1510016). It classifies the chemicals in four classes: < 1 mg/L, 1 - 10 mg/L, 10 - 100 mg/L, and > 100 mg/L.

For Daphnia acute toxicity there are four models. One, the **Daphnia Magna Acute (EC50) Toxicity model (IRFMN-Combase)** (doi: 10.1007/978-1- 4939-7899-1_27), is based on a set of biocides and developed using CORAL, the others on general industrial chemicals. The **Daphnia Magna Acute (EC50) Toxicity model (IRFMN)** is a statistical model based on chemical descriptors; the **Daphnia Magna LC50 48h (EPA)** is the reimplementation of the single model present in the US EPA T.E.S.T. tool; the **Daphnia Magna LC50 48h (DEMETRA)** (doi: 10.1021/es071430t. PMID: 18284152) is the reimplementation of the model (a hybrid model based on multiple linear regressions) present in the DEMETRA tool.

For algae acute toxicity three models are available, tow quantitative and one qualitative. The quantitative models are the **Algae Acute (EC50) Toxicity model (IRFMN)** (https://doi.org/10.3390/molecules26226983), which is a Tree Ensemble Random Forest model specific for *Raphidocelis subcapitata*, and the **Algae Acute (EC50) Toxicity model (ProtoQSAR-Combase)** (https://doi.org/10.3390/molecules26226983), specific for biocides and based on chemical descriptors and support vector machine. It can be used in combination with the qualitative model, the **Algae Classification Toxicity model (ProtoQSAR-Combase)** (https://doi.org/10.1002/9781119681397.ch27) that is a Support Vector Machine (SVM) specific for biocides.

Table 6. Measure units for the acute aquatic toxicity models.

Fish	Fish Acute (LC50) Toxicity model (NIC)	log neg (mmol/L)
	Guppy LC50 model (KNN-IRFMN)	log neg (mmol/L)
	Fathead Minnow LC50 96h (EPA)	log neg (mol/L)
	Fathead Minnow LC50 model (KNN-IRFMN)	log neg (mmol/L)
	Fish Acute (LC50) Toxicity model (KNN-Read-Across)	log neg (mg/L)
	Fish Acute (LC50) Toxicity model (IRFMN)	adim[*]
	Fish Acute (LC50) Toxicity model (IRFMN-Combase)	log neg (mmol/L)
Daphnia	Daphnia Magna Acute (EC50) Toxicity model (IRFMN-Combase)	log (mmol/L)
	Daphnia Magna LC50 48h (EPA)	log neg (mol/L)
	Daphnia Magna LC50 48h (DEMETRA)	log neg (mol/L)
	Daphnia Magna Acute (EC50) Toxicity model (IRFMN)	log (mol/L)
Algae	Algae Acute (EC50) Toxicity model (IRFMN)	adim[§]
	Algae Acute (EC50) Toxicity model (ProtoQSAR-Combase)	log (mg/L)

[*] *The conversion in mg/L can be done using the formula (MW is the molecular weight)*
((Prediction * 0.11 + 1) ^ (1.0 / 0.11)) * MW

[§] *The conversion in mg/L can be done using the formula (MW is the molecular weight)*
((Prediction * 0.07 + 1) ^ (1.0 / 0.07)) * MW

Each quantitative model uses different measure units (see the *Table* 6). This is due the fact that, even if original data are expressed in mg/L, to obtain better models the values are usually converted into mol/L (or mmol/L) and in many cases also transformed in log units. This could represent a difficulty if the user wants to aggregate the results of all the models for the same endpoint. For this reason, we suggest using JANUS.

The Aquatic Chronic Toxicity models

There is only one model for each aquatic species to evaluate the chronic toxicity. The models **Fish Chronic (NOEC) Toxicity model (IRFMN)** and **Algae Chronic (NOEC) Toxicity model (IRFMN)**) are Tree Ensemble Random Forest (https://doi.org/10.3390/molecules26226983) and their output is adimensional (see the *Table 7* for the conversion). **Daphnia Magna Chronic (NOEC) Toxicity model (IRFMN)** is a statistical model based on chemical descriptors.

Table 7. Measure units for the chronic aquatic toxicity models.

Specie	Model	Measure unit	Conversion to mg/L
Fish	Fish Chronic (NOEC) Toxicity model (IRFMN)	adim	((Value * 0.01 + 1) ^ (1.0 / 0.01)) * MW
Daphnia	Daphnia Magna Chronic (NOEC) Toxicity model (IRFMN)	log (mmol/L)	10^(value*MW)
Algae	Algae Chronic (NOEC) Toxicity model (IRFMN)	adim	((Value * 0.03 + 1) ^ (1.0 / 0.03)) * MW

The Mode of Action models

- **Verhaar classification (TOXTREE)** is a decision tree constructed using a series of structural alerts designed to enable organic chemicals to be assigned to one of the four Verhaar categories. The Verhaar classification is a consolidated approach used worldwide to classify substances according to their ecotoxicological mode of action.
- **MOA fish toxicity classification (EPA T.E.S.T.)** implements the model available in T.E.S.T. indicating the mode of action (MOA).
- **MOA pesticide classification (IRFMN)** is a statistical model based on descriptors.

The Terrestrial Acute Toxicity models

- **Bee acute toxicity model (KNN-IRFMN)** is a classification indicating oral toxicity towards bees.
- **Earthworm Toxicity (CONCERT)** is a partial least squares regression (PLS) model for the prediction of reproductive toxicity induced by organic compounds in *Folsomia candida* using 28 days NOEC data (https://doi.org/10.1016/j.jhazmat.2021.127236)

The Sludge Toxicity models

There are two models (**Sludge Classification Toxicity model (ProtoQSAR-Combase), Sludge (EC50) Toxicity model (ProtoQSAR-Combase)**), both statistical ones using molecular descriptors. They have been built up considering biocide activity within the COMBASE project on biocides. One model provides a categorical output, the second one a quantitative value.

The Zebrafish Embryo Activity model

Zebrafish embryo AC50 (IRFMN-CORAL) is a model based on CORAL. It relates to the fish, thus, it is interesting for aquatic toxicity. In addition, zebrafish embryo is also used as a model providing data useful for developmental toxicity in human too.

Fate and Distribution

The Ready Biodegradability model

Ready Biodegradability model (IRFMN) is a SARpy-based model, with a collections of rules associated to ready biodegradability.

The Persistence (sediment) models

For persistence VEGA has models for sediment, soil and water, as requested by the European legislation. For each of these compartments, VEGA has a model providing a classification, and a quantitative model, providing a continuous value. The two models should be used jointly, to verify that there is not conflict.

- **Persistence (sediment) model (IRFMN)** is a model based on KNN and a set of rules extracted with SARpy.
- **Persistence (sediment) quantitative model (IRFMN)** is a counter-propagation neural network using molecular descriptors.

The Persistence (soil) models

- **Persistence (soil) model (IRFMN)** is a model based on KNN and a set of rules extracted with SARpy.
- **Persistence (soil) quantitative model (IRFMN)** is a counter-propagation neural network using molecular descriptors.

The Persistence (water) models

- **Persistence (water) model (IRFMN)** is a model based on KNN and a set of rules extracted with SARpy.
- **Persistence (water) quantitative model (IRFMN)** is a counter-propagation neural network using molecular descriptors.

The Persistence (air) model

Air Half-Life (IRFMN-CORAL) is a CORAL model, based on SMILES attributes obtained from SMILES.

Physical-Chemical properties

The Octanol/Water partition coefficient (logKow) models

- **LogKow model (Meylan-Kowwin)** implements the EPISuite model. Compared to the original model, using VEGA you can visualize the similar substances and measure the reliability of the prediction; VEGA provides an automatic evaluation of the applicability domain.
- **LogKow model (MLogKow)** is the implementation of the MlogKow model, using regression equation with structural parameters.
- **LogKow model (ALogKow)** is the implementation of the AlogKow model, using regression equation based on 120 atom types.

The Water solubility model

Water solubility model (IRFMN) is a neural network model based on 15 Dragon descriptors.

The Vapour pressure model

Vapour Pressure (CONCERT/Kode) is a neural network model based on molecular descriptors. It has been developed with the CONCERT REACH project.

The Melting point models

- **Melting Point (CONCERT/Kode)** is a neural network model based on molecular descriptors. It has been developed with the CONCERT REACH project.
- **Melting Point (CONCERT/KNN)** is a KNN model. It has been developed with the CONCERT REACH project.

The Hydrolysis model

Hydrolysis (IRFMN-CORAL) is a CORAL model, based on SMILES attributes obtained from SMILES.

The Henry’s law constant model

Henry's Law model (OPERA) is the implementation in VEGA of the OPERA model. We thank Kamel Mansouri for his assistance.

The Octanol/air partition coefficient (KOA) model

KOA model (OPERA) is the implementation in VEGA of the OPERA model. We thank Kamel Mansouri for his assistance.

The Soil adsorption coefficient of organic compounds (KOC) model

KOC model (OPERA) is the implementation in VEGA of the OPERA model. We thank Kamel Mansouri for his assistance.

Human PBPK

The Plasma Protein Binding models

- **Plasma Protein Binding - LogK (IRFMN)** is a random forest model; it addresses the fraction unbound.
- **Plasma Protein Binding - sqFU (CORAL)** is a CORAL model, based on SMILES attributes obtained from SMILES. It addresses the fraction unbound**.**

The Aromatase activity models

- **Aromatase activity model (IRFMN)** is a random forest model, using data from Tox21 on aromatase activity. It uses a number of molecular descriptors.
- **Aromatase activity model (TOX21)** is a SARpy model with a number of rules, mainly related to triazoles, and related substances with rings with heteroatoms including a different number of nitrogen, sulfur and oxygen atoms.

The p-Glycoprotein activity model

P-Glycoprotein activity model (NIC) is a models developed by the National Institute of Chemistry (Slovenia); it is a counter-propagation neural network using molecular descriptors. It has been developed within the IN3 project.

The Hepatic Steatosis MIE models

There are four models for Hepatic Steatosis MIE (**Hepatic Steatosis MIE assay for PXR up (TOXCAST), Hepatic Steatosis MIE assay for PPARg up (TOXCAST), Hepatic Steatosis MIE assay for PPARa up (TOXCAST),** and **Hepatic Steatosis MIE assay for NRF2 (TOXCAST)**). Data at the basis of the models derive from ToxCast, specific assays. These are statistical models using molecular descriptors. They are models integrating, for each endpoint, results from individual models developed with random forest and balanced random forest.

The Skin permeation (logKp) models

- **Skin Permeation (LogKp) model (Potts and Guy)** is a model to predict skin permeability coefficient (Kp), which defines the rate of chemical penetrating across the stratum corneum. It is based on the existing equation of Potts and Guy. It is a simple linear equation based on two molecular descriptors: Molecular Weight (MW) and $\log K_{ow}$.
- **Skin Permeation (LogKp) model (Ten Berge),** similarly to the Potts and Guy model, it predicts the skin permeability coefficient. This model is based on the already existing equation of Ten Berge, based on the same descriptors used within the Potts and Guy equation.

The Adipose tissue-blood partition model

Adipose tissue - blood model (INERIS) is a statistical model using molecular descriptors. The algorithm is random forest. The model has been built up by INERIS (France).

The Body elimination half-life model

Total body elimination half-life (QSARINS) is a multiple linear regression model (OLS) using molecular descriptors. It has been developed by QSARINS.

Ecological PBPK

The kM/Half-Life model

kM/Half-Life model (Arnot-EpiSuite) is the implementation of the EPISuite Biotransformation (kN) BCFBAF model. It is a linear regression model based on logKow, molecular weight and contribution from specific fragments**.**

B.3. The read-across tools in VEGAHUB

There are different tools for read-across in VEGAHUB:

- VEGA
- KNN
- ToxRead
- RAXPY
- ToxDelta
- aiQSAR
- toDIVINE
- VERA

VEGA should be used as an integrated tool, for the *in silico* models and for read-across. In the Section about VEGA we discussed this with examples.

Note. Not all the *in silico* tools within VEGA have the associated set of compounds used to build up the model. Indeed, in the case of expert-systems, there is not a training set, thus, some models cannot be used for read-across. This is the case of the Cramer model, for instance.

We discuss below the main tools: KNN, ToxRead and VERA. toDIVINE is not publicly available, it is for regulatory bodies, since it contains data on REACH registered substances.

B.3.1. KNN models

Several models apply the **KNN** algorithm. This program finds the nearest neighbors (the k substances closer to the target one). This is the list of these models:

- Mutagenicity (Ames test) model (KNN-Read-Across)
- Acute Toxicity (LD50) model (KNN)
- Eye Irritation (CONCERT/KNN)
- BCF model (KNN-Read-Across)
- Fish Acute (LC50) Toxicity model (KNN-Read-Across)
- Fathead Minnow LC50 model (KNN-IRFMN)

- Guppy LC50 model (KNN-IRFMN)
- Bee acute toxicity model (KNN-IRFMN)
- Persistence (sediment) model (IRFMN)
- Persistence (soil) model (IRFMN)
- Persistence (water) model (IRFMN)
- Melting Point (CONCERT/KNN)

The algorithm is described in *https://doi.org/10.1016/j.chemosphere.2015.10.054*.

The KNN algorithm is not-parametric. It does not use molecular descriptors to develop the predictive model. It searches for the closest similar substances, and then the property value of the target is obtained considering the experimental values of the analogues. Thus, it is not possible to use mechanistic concepts in this case. This is a disadvantage. The concept is very close to the read-across strategy, and indeed it is an automatic read-across. Thus, this tool has the advantages and disadvantages of read-across. In particular, if there are similar substances (for instance similarity higher than 0.85), the prediction may be quite good. If there are not similar substances, it does not work. In the worst case, VEGA does not make the prediction, but anyhow it shows the similar.

Another important aspect to be considered, is that, since the tool has no mechanistic knowledge, and no link to descriptors (as for the other *in silico* models), it cannot recognize if the similar substances belong to a category different from the target one; for instance, in the case of BCF prediction, if all the similar substances have logKow higher than the target, this will shift the prediction towards a higher value than it should be associated to the target substance. This is a negative behaviour in particular with continuous values, while this issue is not so critical for classifier models.

B3.2. ToxRead

ToxRead is a tool specific for read-across, to analyse similar substances. It is a tool to investigate different scenarios, based on the number of similar substances used. Furthermore, it does not work in batch mode. Thus, these characteristics make it valid to proceed manually in the evaluation, while for automatic assessment done on many substances, other tools, as VERA, should be used.

ToxRead integrates the structural similarity with other metrics which refer to features relevant for the specific property of interest. Thus, there are tens of different ToxRead tools, for the different properties.

The structural similarity is evaluated using the same algorithm used by VEGA. Similarity depends on many factors, and there is no absolute measurement for it. We optimized the algorithm of structural similarity used in VEGA on the basis of a check with 4 million compounds, and this is an advantage of VEGA compared with other programs. The similarity is calculated as described (*http://jcheminf.springeropen.com/articles/10.1186/s13321-014-0039-1*). The software calculates how similar the similar compound is providing a score between 1 (in case of identity) and 0. Values of 0.9 for similarity indicate a good similarity. Usually values lower than 0.75 indicate that the similar compound has important differences compared to the target.

While VEGA shows the 6 most similar substances, with ToxRead the user can choose this number. Six in our experience is the good default value. If the training set is small, there may be less similar substances. If the structure of the target substance is particular, there will be less similar substances. If the substance has a common structure and for the property of interest there are many compounds, it may be good to look at more substances, but we do not expect that the result will vary using more than 6 substances. To search for many similar substances may be interesting if there is the need to explore an unusual behaviour of the target.

In many cases, ToxRead uses structural alerts (SAs) to identify a possible mechanism or effect of lack of effect. These SAs are particularly numerous in the case of mutagenicity (Ames test): more than 800.

The SAs can be of different nature; they can:

- be associated with a positive effect, such as genotoxicity.
- be exception rules of the previous SA.
- reduce the positive effect, in a generic way. While the previous rules block the effect provoked by the SAs, these SAs modulate the effect, and may also cancel the effect, because a parallel process may prevail.
- be neutral (these may be called rules, instead of SAs, due to the lack of effect). The interest in the information about these SAs is that the user may want to analyse the possible reason of effect in a certain molecule; if a certain fragment is known to be non-toxic, it may be skipped from the evaluation.

The SAs can be associated to known plausible effect. In several cases there are papers explaining the possible mechanism, which may also apply to the chemical under investigation. There are levels of uncertainty associated with these SAs as well. The user should not consider the presence of an alert of this kind as a demonstration of the effect. Indeed, most of these SAs do not have an accuracy of 100%, and those with this high level of accuracy have a limited number of chemicals.

There are different levels of accuracy depending on the "sub-families" of SAs. Thus, if the percentage of positive chemicals associated with aromatic amines for genotoxicity, for instance, is X %, it may be higher for a subset of aromatic amines. Conversely, it may be reduced for other subsets. It is convenient to look for the most specific alert, which better fits the structure of the substance of interest. ToxRead contains the largest list of these SA for mutagenicity.

There are different methods to derive SAs. In **VEGA** and ToxRead there are collections of SAs derived by studies done by human experts, like those present in Toxtree, the so-called Benigni-Bossa rules for genotoxicity. These SAs are also present in **VEGA** and ToxRead. There are SAs derived by human experts of Mario Negri (*doi: 10.1080/10590501.2015.1096881*), present in ToxRead. There are SAs derived from computer programs, like SARpy (*http://www.vega-qsar.eu/research.html*), developed by Politecnico di Milano, or by another program developed by CRS4. All these SAs contain both active and inactive fragments, with the exception of those from Toxtree (these are only to identify toxic effects).

The occurrence of concordant structural alerts in the similar compounds is considered too. This relates to the information on the toxicity/property value. This information is associated with the information on the concordance above discussed. This information is important to decide if the similar compound is relevant or not. If the similar compound has the same SAs of the target compound, the similar compound is relevant. Of course, this applies only in the case that the effect can be explained by the SAs, thus it does not apply to all cases. Even in cases in which there is the same SAs in both the similar and the target compound, the user should evaluate if there are reasons to modulate or cancel the positive effect. See above, the SAs which are exception rules, or modulate the effect. For more details see below.

The ToxRead model for mutagenicity has a large database, larger than any individual model, since it integrates the values of all models, plus adds new substances. Also for this reason, it is convenient to use ToxRead for read-across.

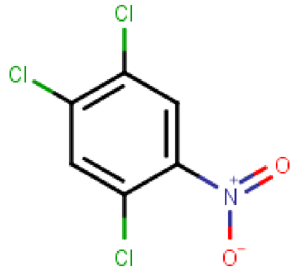
Note. Since it merges different databases, it is possible that for the same substance there are two, opposite experimental values.

Example 1 of an evaluation done with ToxRead

Exercise 3

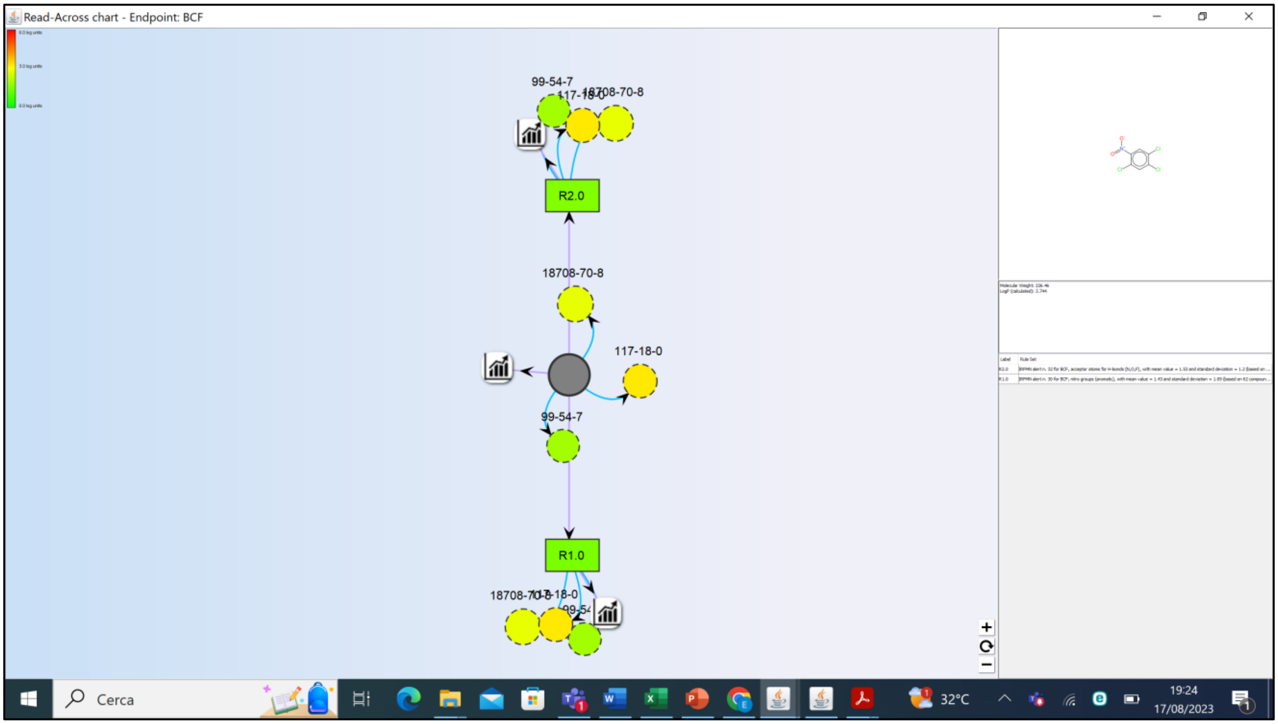
You should predict the BCF value of the substance reported in the *Table 8* below. Use different number of similar substances.

Table 8. The input molecule for the exercise 3.

CAS No.	Name	SMILES	MOLECULE
89-69-0	2,4,5-trichloro-1-nitrobenzene	C1=C(C(=CC(=C1Cl)Cl)Cl)[N+](=O)[O-]	

With three number of similar substances, you obtain this result (*Figure 43*).

The software shows in the middle the target substance (grey circle), linked to three similar substances, represented by three circles. Please refer to the Guide for the meaning of all the graphical elements. (*https://www.vegahub.eu/wp/wp-content/uploads/2019/06/ToxRead-0.17-beta-Guide.pdf*) Briefly the colour of the circle indicates the BCF value of each of the substances, and the size of the circle refers to the similarity value. Clicking on each similar substance, you can see its structure. These three substances are shown below.



Molecular Weight: 226.46
LogP (calculated): 3.744

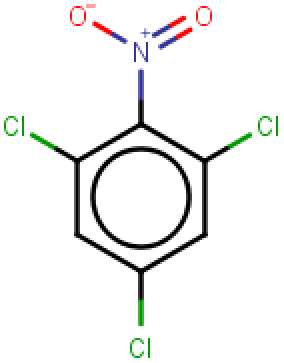
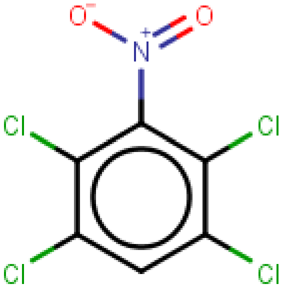
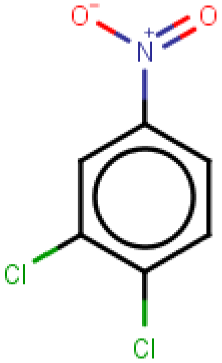
Label	Rule Set
R2.0	IRFMN alert n. 32 for BCF, acceptor atoms for H-bonds (N,O,F), with mean value = 1.53 and standard deviation = 1.2 (based on ...
R1.0	IRFMN alert n. 30 for BCF, nitro groups (aromatic), with mean value = 1.43 and standard deviation = 1.09 (based on 62 compoun...

Figure 43. The results of ToxRead for BCF with three similar chemicals.

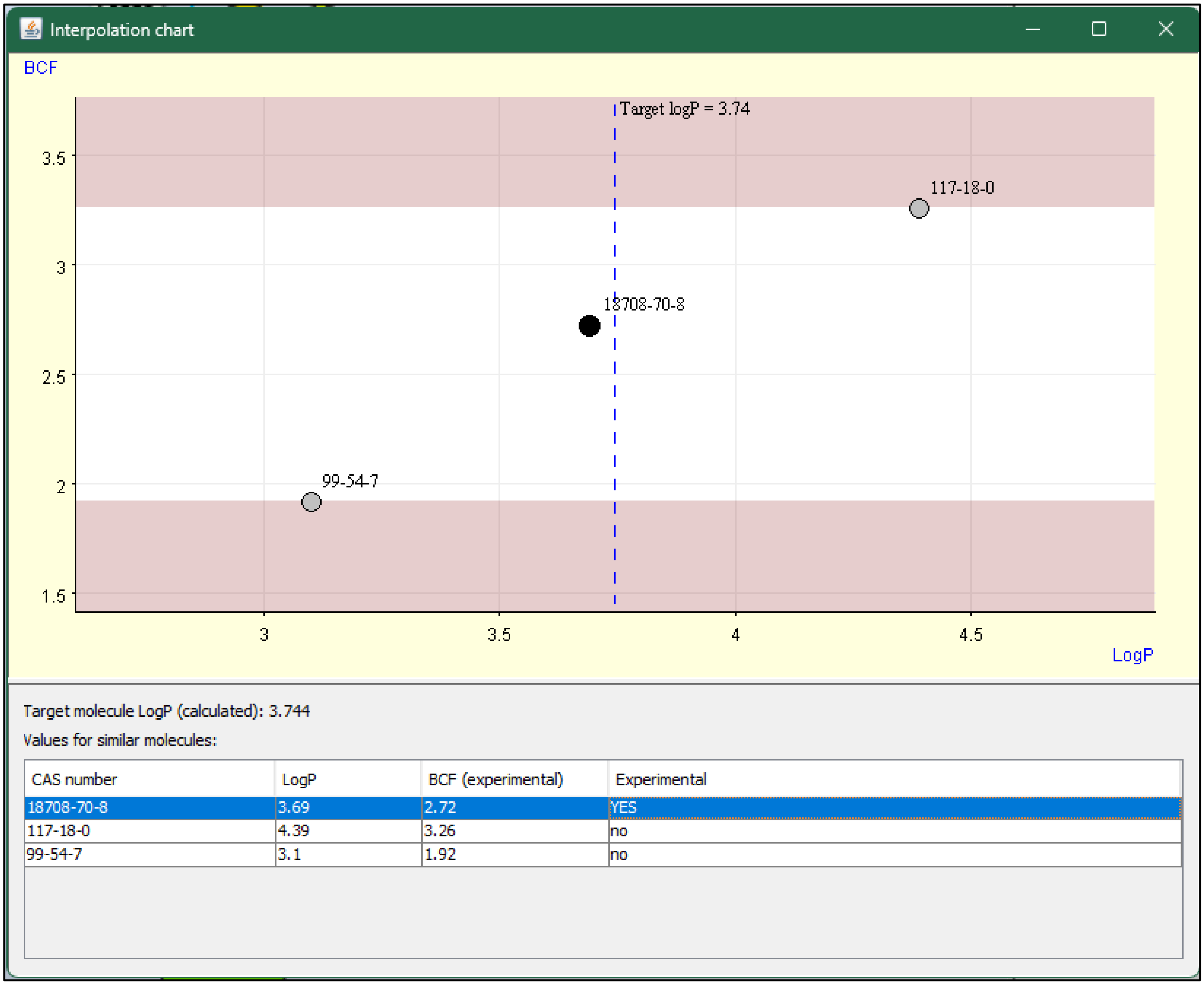
The first one is an isomer of the target substance, with 3 chlorine atoms in different positions. The second similar has 4 chlorine atoms, and the third one has two chlorine atoms (see *Table 9*).

If we click on the icon showing a plot at the centre of the screen, you can visualize the plot as below (*Figure 44*).

Table 9. The three similar substances.

CAS No. 18708-70-8	CAS No. 117-18-0	CAS No. 99-54-7
		
Similarity to target: 0.981 Experimental activity: 2.72 Molecular Weight: 226.46 LogKow (experimental): 3.69	Similarity to target: 0.955 Experimental activity: 3.26 Molecular Weight: 260.9 LogKow (calculated): 4.389	Similarity to target: 0.94 Experimental activity: 1.92 Molecular Weight: 192.02 LogKow (calculated): 3.1

The BCF values of the three similar substances (y axis) are plotted versus their logKow. For one of them, the logKow is experimental (black dot), for the other two the values are predicted. The three dots follow an ideal line, with BCF values increasing progressively with the logKow, and with the number of chlorine atoms. *Table 10* the plot shows the exact values. The dashed vertical line indicates the predicted logKow value of the target substance. From this graph, you can assume that the BCF value of the target compound is very similar to the value of its isomers. The same conclusion could be obtained simply using the isomer. But this graph further supports this conclusion. Indeed, to make read-across using one single substance is always risky – the value of the similar may be wrong.



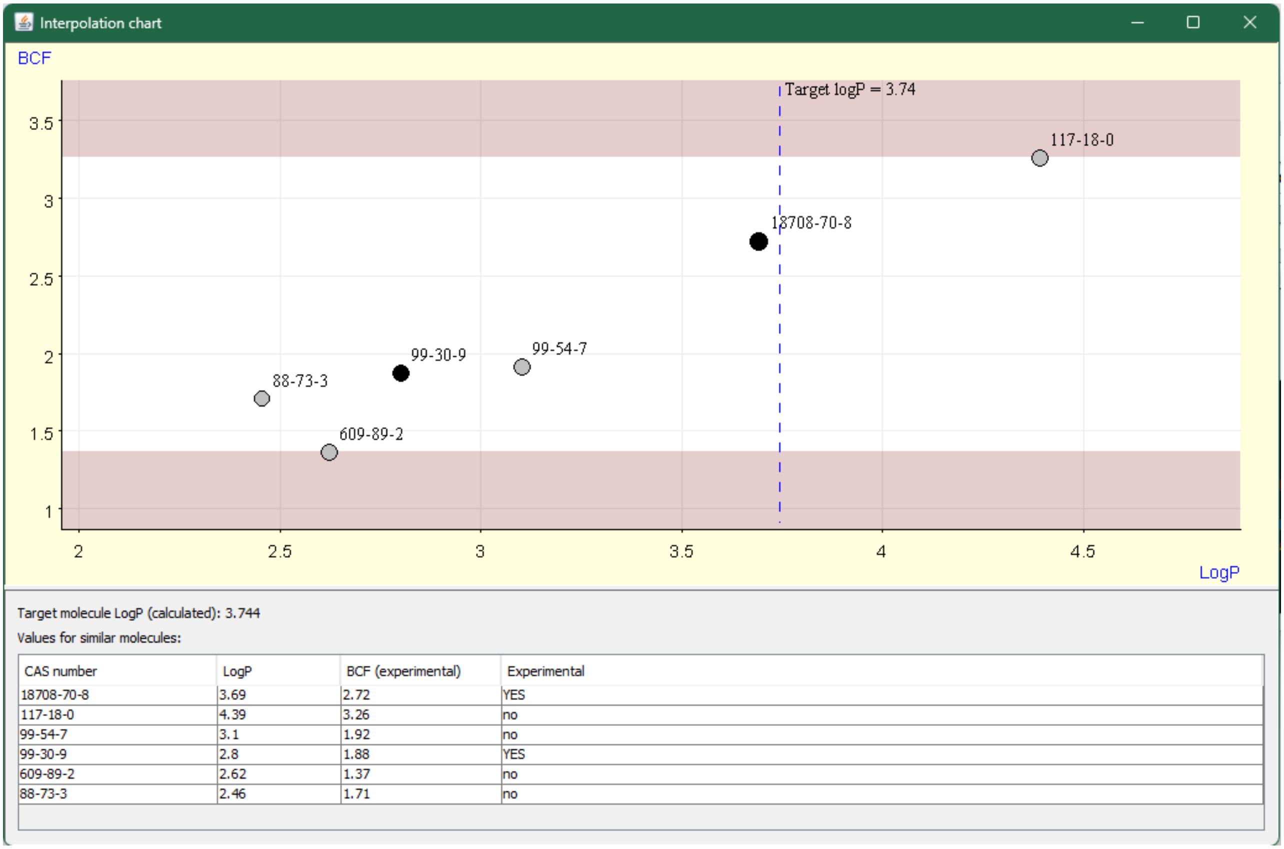
Target molecule LogP (calculated): 3.744

Values for similar molecules:

CAS number	LogP	BCF (experimental)	Experimental
18708-70-8	3.69	2.72	YES
117-18-0	4.39	3.26	no
99-54-7	3.1	1.92	no

Figure 44. The plot of the three most similar chemicals in ToxRead.

Let’s use now ToxRead asking for the six most similar substances. We obtain this plot (*Figure 45*):



Target molecule LogP (calculated): 3.744

Values for similar molecules:

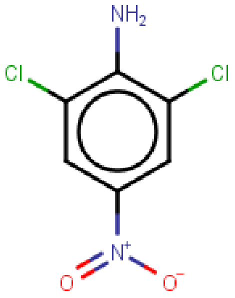
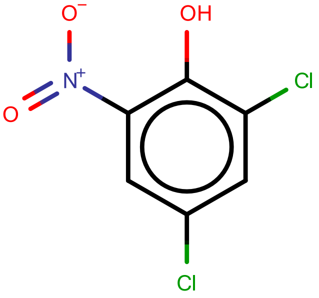
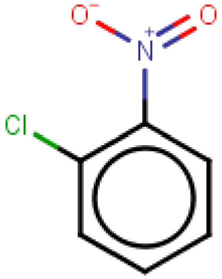
CAS number	LogP	BCF (experimental)	Experimental
18708-70-8	3.69	2.72	YES
117-18-0	4.39	3.26	no
99-54-7	3.1	1.92	no
99-30-9	2.8	1.88	YES
609-89-2	2.62	1.37	no
88-73-3	2.46	1.71	no

Figure 45. The plot of the six most similar chemicals in ToxRead.

Now we have three new similar substances, which are shown below, in the order of similarity.

Now the plot with the 6 similar substances is noisier, but still we can reach a similar conclusion. If we increase the number of substances, we introduce more substances introducing “noise”, i.e. information which is affected by other components, and are progressively with a behaviour deviating from the behaviour of polychlorinated nitrobenzenes.

Table 10. The three similar substances.

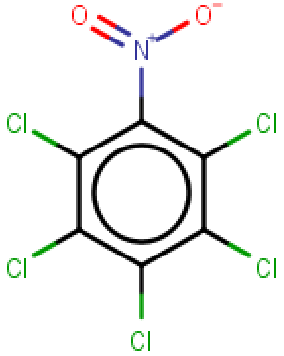
CAS No. 99-30-9	CAS No. 609-89-2	CAS No. 88-73-3
		
Similarity to target: 0.916 Experimental activity: 1.88 Molecular Weight: 207.04 LogKow (experimental): 2.8	Similarity to target: 0.904 Experimental activity: 1.37 Molecular Weight: 208.02 LogKow (calculated): 2.619	Similarity to target: 0.881 Experimental activity: 1.71 Molecular Weight: 157.57 LogKow (calculated): 2.455

Example 2 of an evaluation done with ToxRead and VEGA

Exercise 4

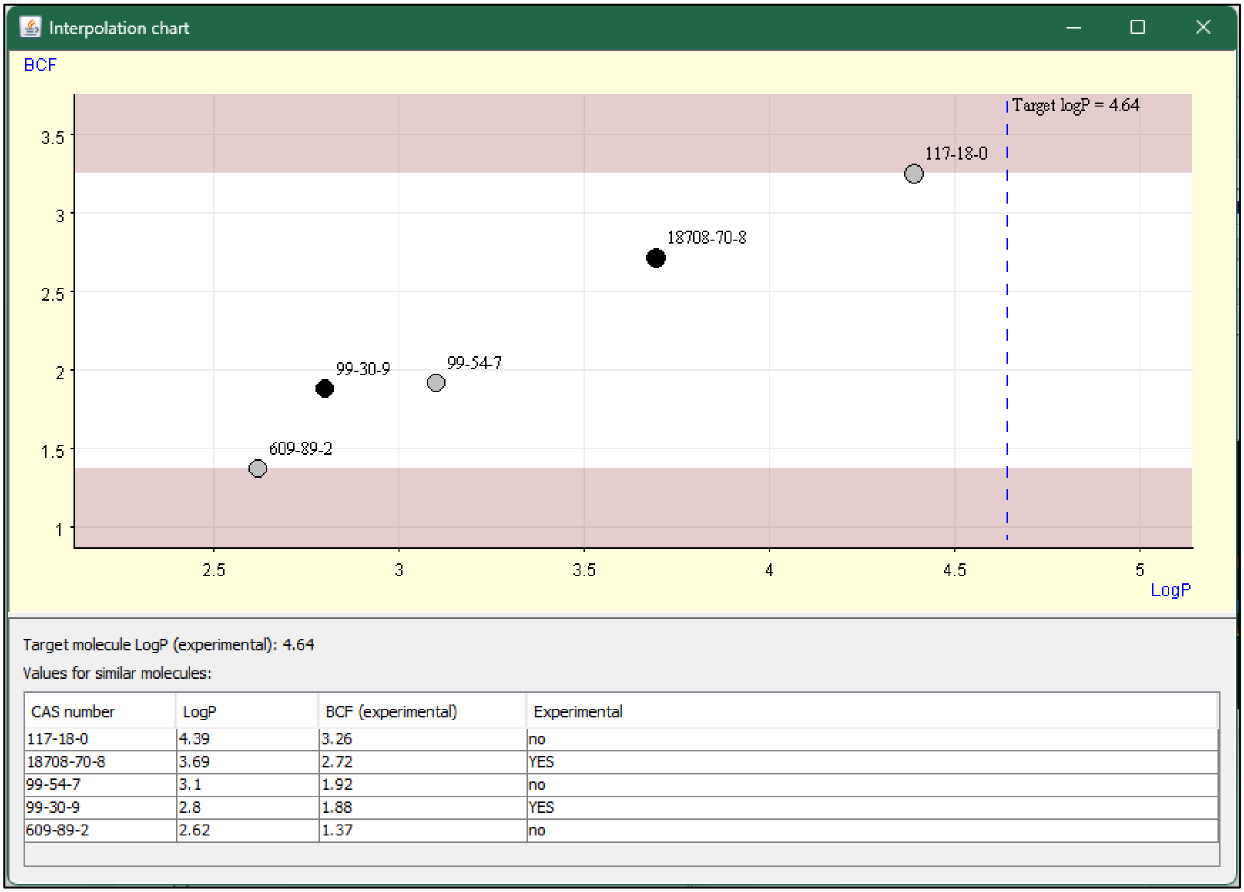
You should predict the BCF value of the substance in the *Table 11*, choosing 5 similar with ToxRead, and also using VEGA.

Table 11. The input molecule for the exercise 4.

CAS No.	Name	SMILES	MOLECULE
82-68-8	pentachloronitrobenzene	O=[N+]([O-])c1c(c(c(c(c1Cl)Cl)Cl)Cl)Cl	

With ToxRead we obtain this plot (*Figure 46*).

From this plot, we can assume that the BCF value is probably 3.5, anyhow higher than 3.2, the value of the analogue with only four chlorine atoms. The other substances are less and less chlorinated, and this figure is very consistent with the figures that we obtained with exercise 3.



Target molecule LogP (experimental): 4.64

Values for similar molecules:

CAS number	LogP	BCF (experimental)	Experimental
117-18-0	4.39	3.26	no
18708-70-8	3.69	2.72	YES
99-54-7	3.1	1.92	no
99-30-9	2.8	1.88	YES
609-89-2	2.62	1.37	no

Figure 46. The plot of the five most similar chemicals in ToxRead.

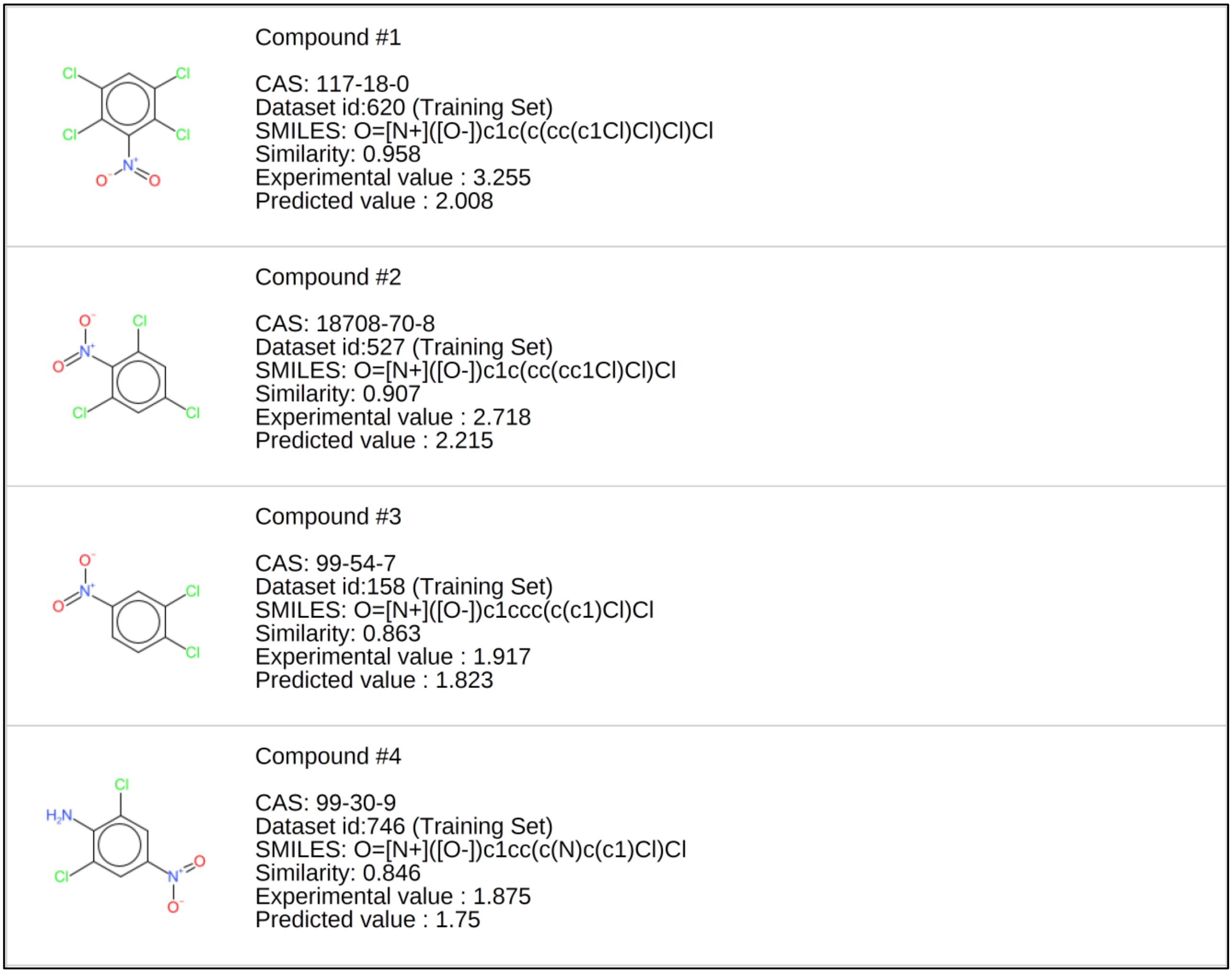
Now we consider the results with VEGA.

<u>The KNN model.</u>

It provides this prediction:

- Predicted BCF [log(L/kg)]: 2.53
- Molecules used for prediction: 4

Figure 47 shows the table with these 4 molecules.



Compound #1

CAS: 117-18-0
Dataset id:620 (Training Set)
SMILES: O=[N+]([O-])c1c(c(cc(c1Cl)Cl)Cl)Cl
Similarity: 0.958
Experimental value : 3.255
Predicted value : 2.008

Compound #2

CAS: 18708-70-8
Dataset id:527 (Training Set)
SMILES: O=[N+]([O-])c1c(cc(cc1Cl)Cl)Cl
Similarity: 0.907
Experimental value : 2.718
Predicted value : 2.215

Compound #3

CAS: 99-54-7
Dataset id:158 (Training Set)
SMILES: O=[N+]([O-])c1ccc(c(c1)Cl)Cl
Similarity: 0.863
Experimental value : 1.917
Predicted value : 1.823

Compound #4

CAS: 99-30-9
Dataset id:746 (Training Set)
SMILES: O=[N+]([O-])c1cc(c(N)c(c1)Cl)Cl
Similarity: 0.846
Experimental value : 1.875
Predicted value : 1.75

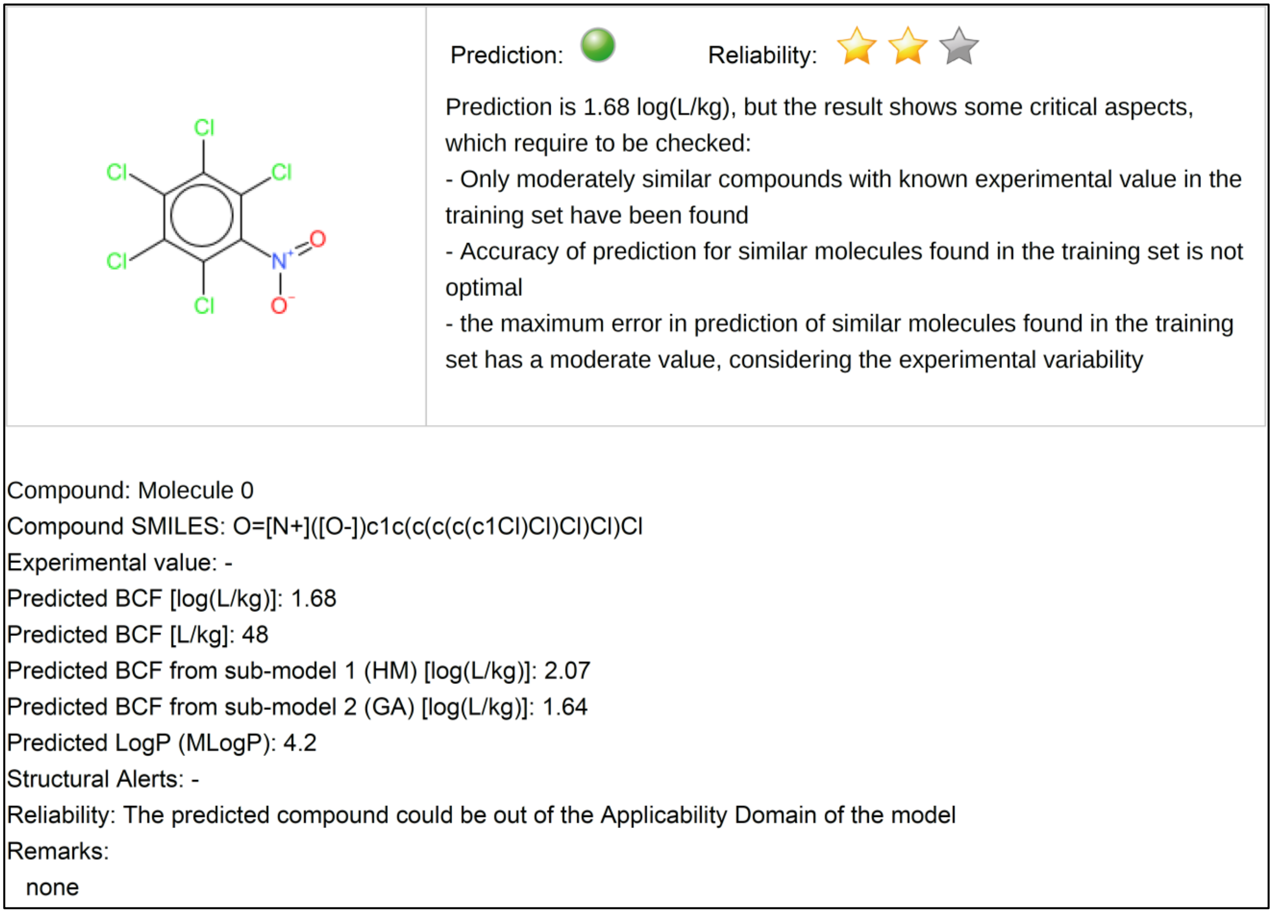
Figure 47. The similar chemical found by the BCF KNN/Read-Across model.

We can observe that the most similar substance has logBCF value at 3.255, and it has four chlorine atoms. The similar number 2 has three chlorine atoms, and its logKow is 2.18, lower than that of the analogue with four chlorines, and even lower is the logBCF value of similar number 3, with two chlorine atoms. Based on these facts, we can conclude that the prediction at 2.53 is not correct. To get this conclusion we can use the reasoning about the expected mechanism. Indeed, we said that the KNN is a non-parametric model, which does not use any descriptor (thus no info on logKow).

The ADI tool correctly indicates that the reliability of the KNN model is low. Actually, what KNN does is to take somehow the average of the logBCF values of the most similar substances

The CAESAR model.

Below we show the summary of the CAESAR model (*Figure 48*).



Prediction: ● Reliability: ★★☆

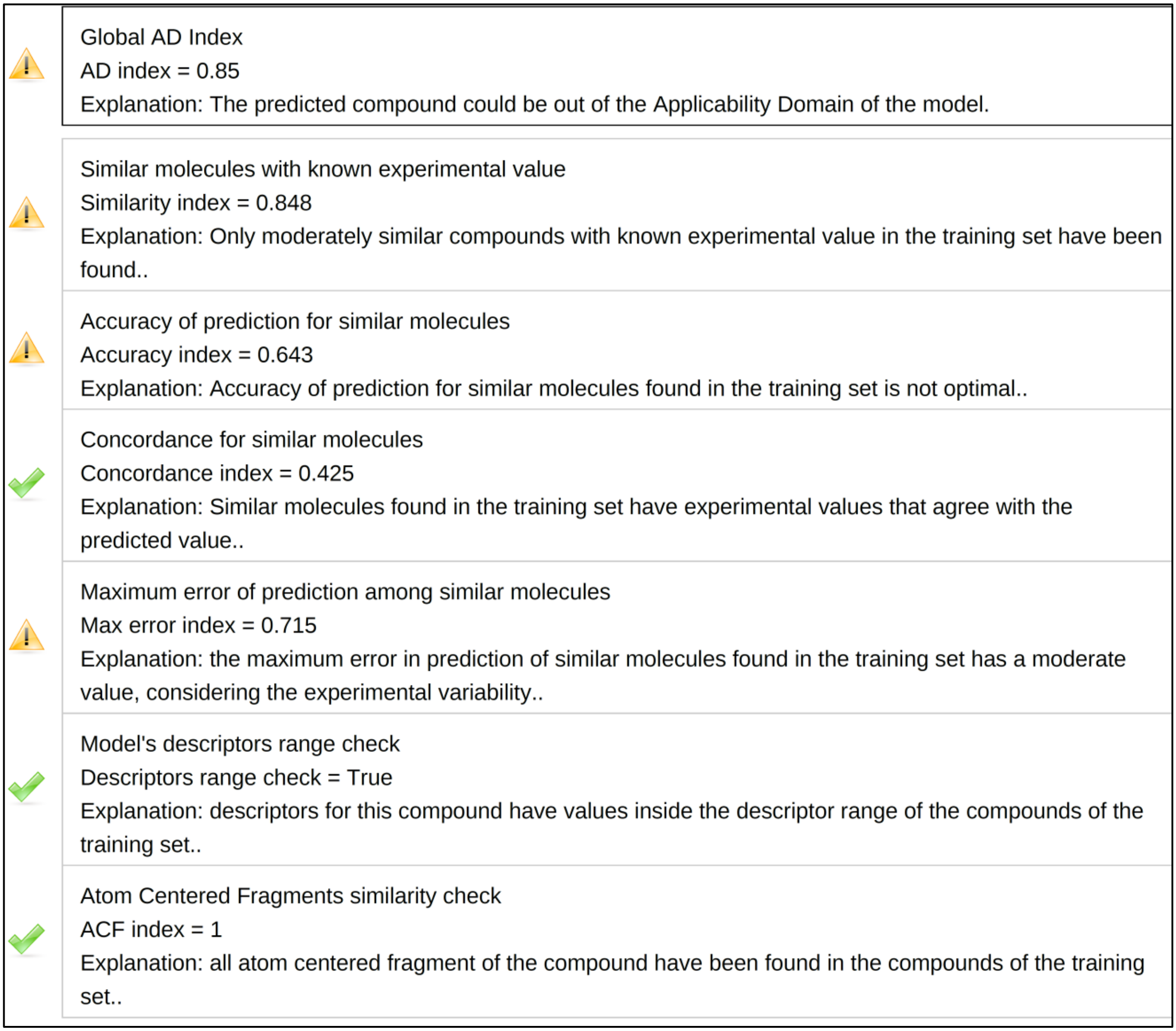
Prediction is 1.68 log(L/kg), but the result shows some critical aspects, which require to be checked:
- Only moderately similar compounds with known experimental value in the training set have been found
- Accuracy of prediction for similar molecules found in the training set is not optimal
- the maximum error in prediction of similar molecules found in the training set has a moderate value, considering the experimental variability

Compound: Molecule 0
Compound SMILES: O=[N+]([O-])c1c(c(c(c(c1Cl)Cl)Cl)Cl)Cl
Experimental value: -
Predicted BCF [log(L/kg)]: 1.68
Predicted BCF [L/kg]: 48
Predicted BCF from sub-model 1 (HM) [log(L/kg)]: 2.07
Predicted BCF from sub-model 2 (GA) [log(L/kg)]: 1.64
Predicted LogP (MLogP): 4.2
Structural Alerts: -
Reliability: The predicted compound could be out of the Applicability Domain of the model
Remarks:
none

Figure 48. The cover of the BCF CAESAR model.

The predicted value is quite low, lower than that of the KNN model, and this may appear unexpected. However, the reliability is higher than that of the KNN model.

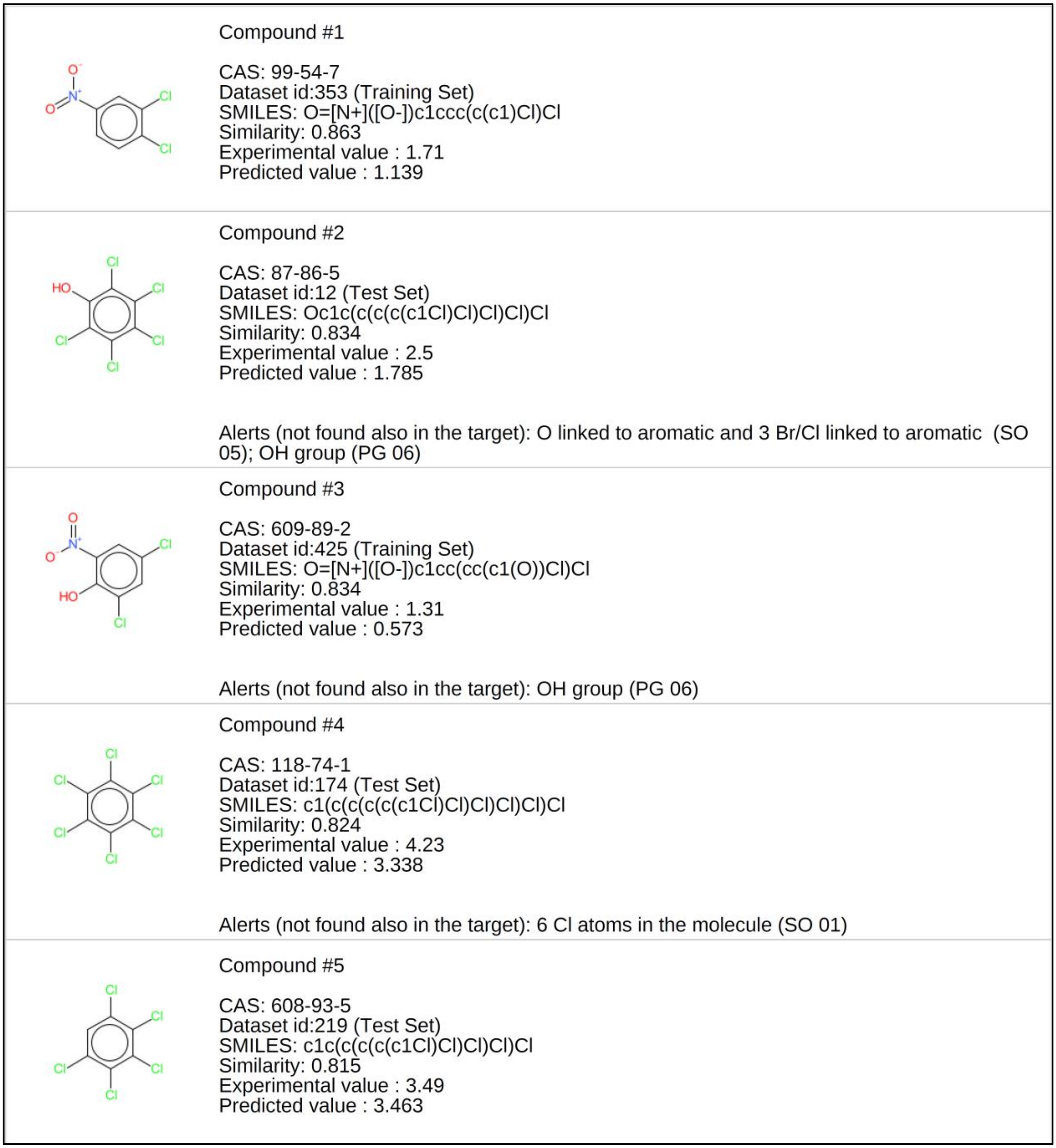
The description of the ADI evaluation is reported below (*Figure 49*).



⚠	Global AD Index AD index = 0.85 Explanation: The predicted compound could be out of the Applicability Domain of the model.
⚠	Similar molecules with known experimental value Similarity index = 0.848 Explanation: Only moderately similar compounds with known experimental value in the training set have been found..
⚠	Accuracy of prediction for similar molecules Accuracy index = 0.643 Explanation: Accuracy of prediction for similar molecules found in the training set is not optimal..
✓	Concordance for similar molecules Concordance index = 0.425 Explanation: Similar molecules found in the training set have experimental values that agree with the predicted value..
⚠	Maximum error of prediction among similar molecules Max error index = 0.715 Explanation: the maximum error in prediction of similar molecules found in the training set has a moderate value, considering the experimental variability..
✓	Model's descriptors range check Descriptors range check = True Explanation: descriptors for this compound have values inside the descriptor range of the compounds of the training set..
✓	Atom Centered Fragments similarity check ACF index = 1 Explanation: all atom centered fragment of the compound have been found in the compounds of the training set..

Figure 49. The ADI of the BCF Caesar model.

The most similar substances are similar but not very similar: the most similar has similarity at about 0.86, and the other two at about 0.83 (see *Figure 50).*



Compound #1

CAS: 99-54-7
Dataset id:353 (Training Set)
SMILES: O=[N+]([O-])c1ccc(c(c1)Cl)Cl
Similarity: 0.863
Experimental value : 1.71
Predicted value : 1.139

Compound #2

CAS: 87-86-5
Dataset id:12 (Test Set)
SMILES: Oc1c(c(c(c(c1Cl)Cl)Cl)Cl)Cl
Similarity: 0.834
Experimental value : 2.5
Predicted value : 1.785

Alerts (not found also in the target): O linked to aromatic and 3 Br/Cl linked to aromatic (SO 05); OH group (PG 06)

Compound #3

CAS: 609-89-2
Dataset id:425 (Training Set)
SMILES: O=[N+]([O-])c1cc(cc(c1(O))Cl)Cl
Similarity: 0.834
Experimental value : 1.31
Predicted value : 0.573

Alerts (not found also in the target): OH group (PG 06)

Compound #4

CAS: 118-74-1
Dataset id:174 (Test Set)
SMILES: c1(c(c(c(c(c1Cl)Cl)Cl)Cl)Cl)Cl
Similarity: 0.824
Experimental value : 4.23
Predicted value : 3.338

Alerts (not found also in the target): 6 Cl atoms in the molecule (SO 01)

Compound #5

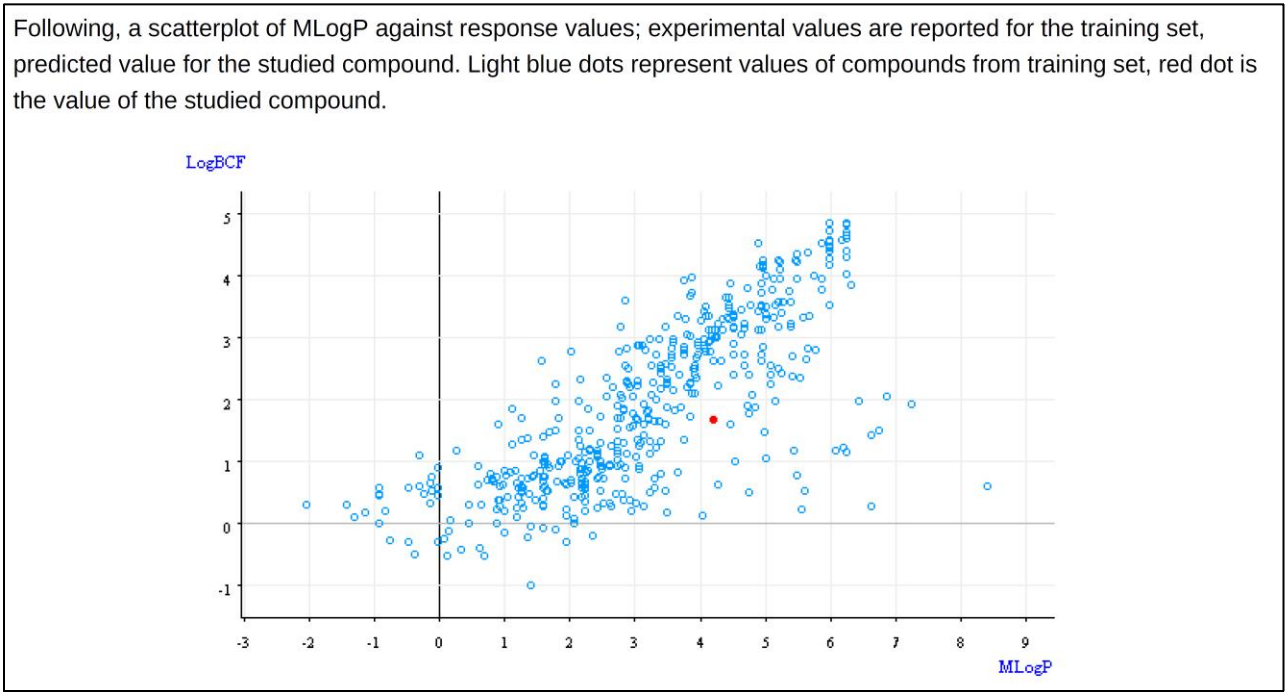
CAS: 608-93-5
Dataset id:219 (Test Set)
SMILES: c1c(c(c(c(c1Cl)Cl)Cl)Cl)Cl
Similarity: 0.815
Experimental value : 3.49
Predicted value : 3.463

Figure 50. The similar chemicals found by the BCF Caesar model.

The accuracy of the prediction is not very good too. The error may be of 0.7 unit or higher.

Considering the presence of the nitro group and chlorine atoms, there are two analogues, one of them has an additional phenolic group. (Similar substance number 6 has the nitro group and one single chlorine.)

The plot of BCF versus logKow values is reported below (*Figure 51*).

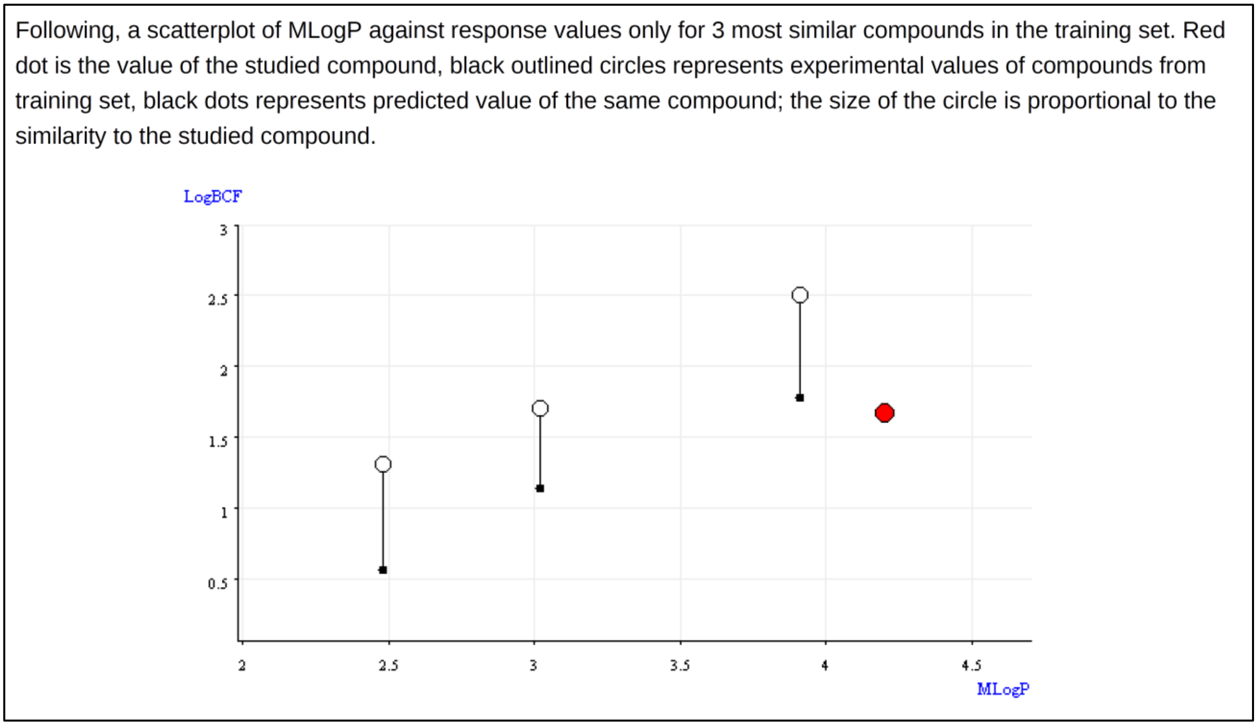


Following, a scatterplot of MLogP against response values; experimental values are reported for the training set, predicted value for the studied compound. Light blue dots represent values of compounds from training set, red dot is the value of the studied compound.

Figure 51. The analysis of molecular descriptors of the BCF CAESAR model (first part).

It is possible that CAESAR underestimates the BCF value, which may be higher.

Let’s consider the same plot with the three most similar substances (*Figure 52*).



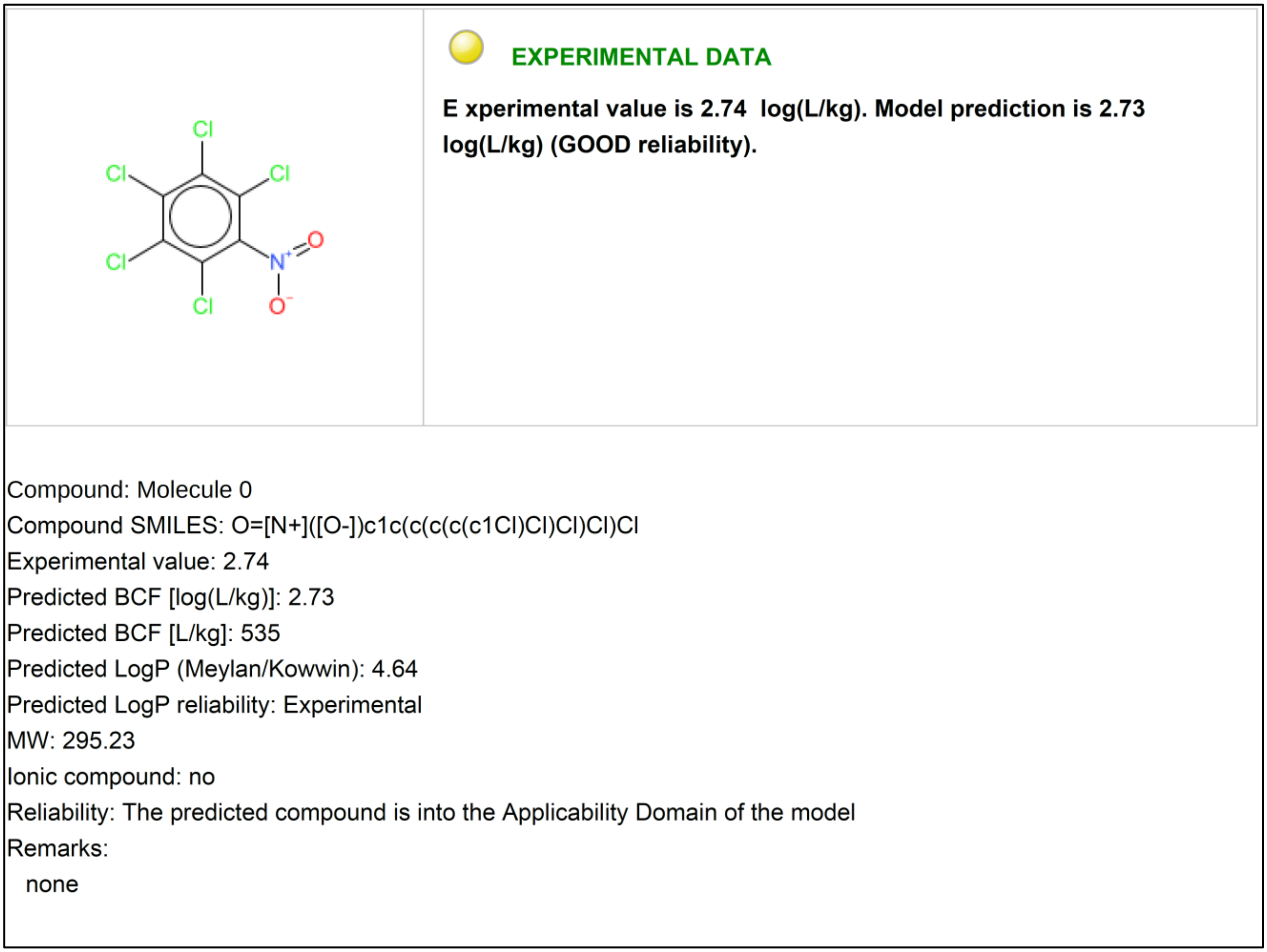
Following, a scatterplot of MLogP against response values only for 3 most similar compounds in the training set. Red dot is the value of the studied compound, black outlined circles represents experimental values of compounds from training set, black dots represents predicted value of the same compound; the size of the circle is proportional to the similarity to the studied compound.

Figure 52. The analysis of molecular descriptors of the BCF CAESAR model (second part).

Now we clearly see that for the similar substances CAESAR underestimates the BCF value. From this plot we can imagine that the correct value for the target substance is between 2.5 and 3.

The Meylan model.

Figure 53 shows the summary of the prediction.



EXPERIMENTAL DATA

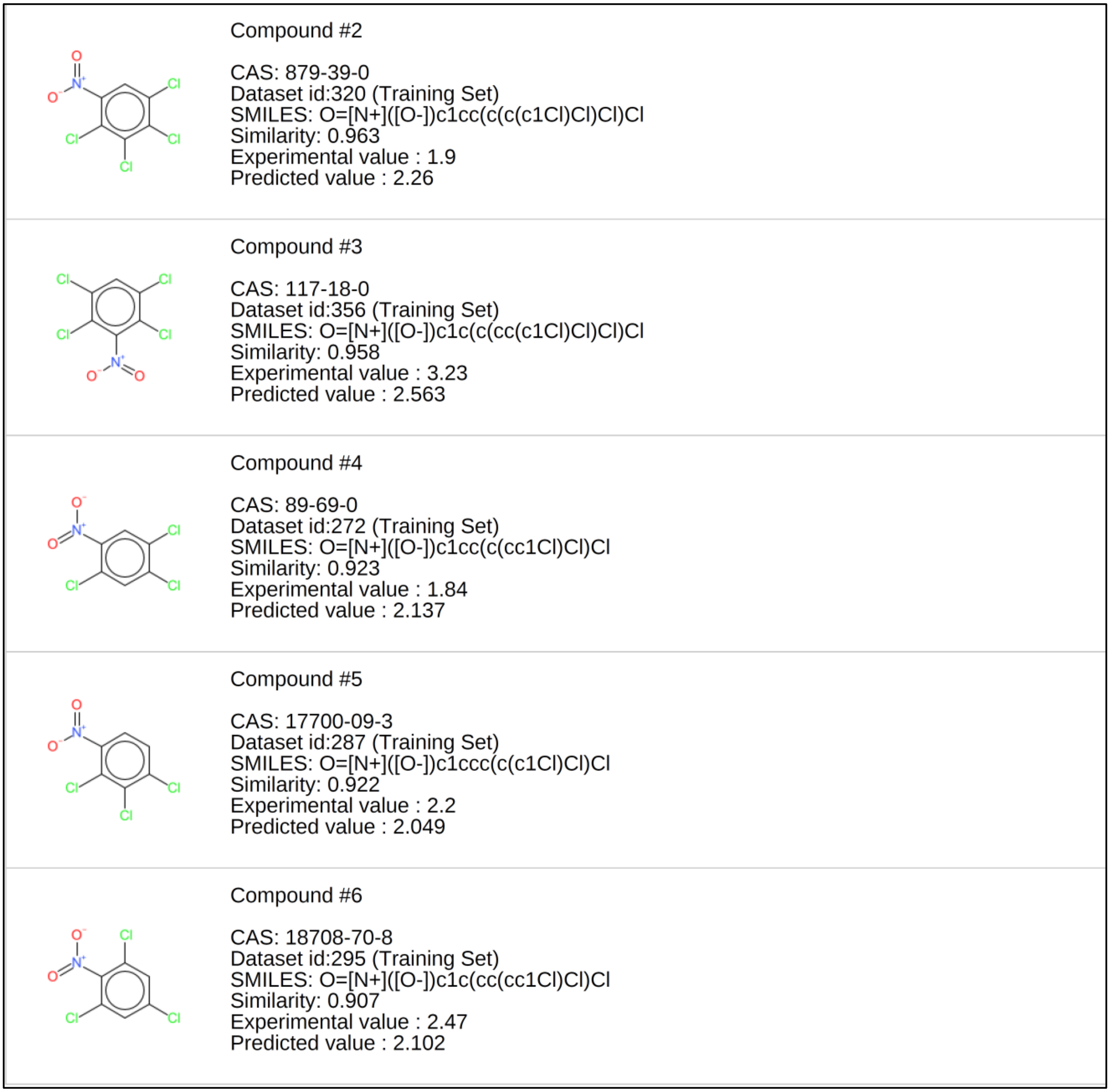
E xperimental value is 2.74 log(L/kg). Model prediction is 2.73 log(L/kg) (GOOD reliability).

Compound: Molecule 0
Compound SMILES: O=[N+]([O-])c1c(c(c(c(c1Cl)Cl)Cl)Cl)Cl
Experimental value: 2.74
Predicted BCF [log(L/kg)]: 2.73
Predicted BCF [L/kg]: 535
Predicted LogP (Meylan/Kowwin): 4.64
Predicted LogP reliability: Experimental
MW: 295.23
Ionic compound: no
Reliability: The predicted compound is into the Applicability Domain of the model
Remarks:
none

Figure 53. The cover of the BCF Meylan model.

Actually, we also have the experimental value, in this case. However, disregard now the experimental value, and continue in our evaluation, as if we do not know it. The prediction is at 2.73.

So, we disregard the most similar substance provided by the model, because it is the target substance, and we consider the following ones, as in *Figure 54*.



Compound #2

CAS: 879-39-0
Dataset id:320 (Training Set)
SMILES: O=[N+]([O-])c1cc(c(c(c1Cl)Cl)Cl)Cl
Similarity: 0.963
Experimental value : 1.9
Predicted value : 2.26

Compound #3

CAS: 117-18-0
Dataset id:356 (Training Set)
SMILES: O=[N+]([O-])c1c(c(cc(c1Cl)Cl)Cl)Cl
Similarity: 0.958
Experimental value : 3.23
Predicted value : 2.563

Compound #4

CAS: 89-69-0
Dataset id:272 (Training Set)
SMILES: O=[N+]([O-])c1cc(c(cc1Cl)Cl)Cl
Similarity: 0.923
Experimental value : 1.84
Predicted value : 2.137

Compound #5

CAS: 17700-09-3
Dataset id:287 (Training Set)
SMILES: O=[N+]([O-])c1ccc(c(c1Cl)Cl)Cl
Similarity: 0.922
Experimental value : 2.2
Predicted value : 2.049

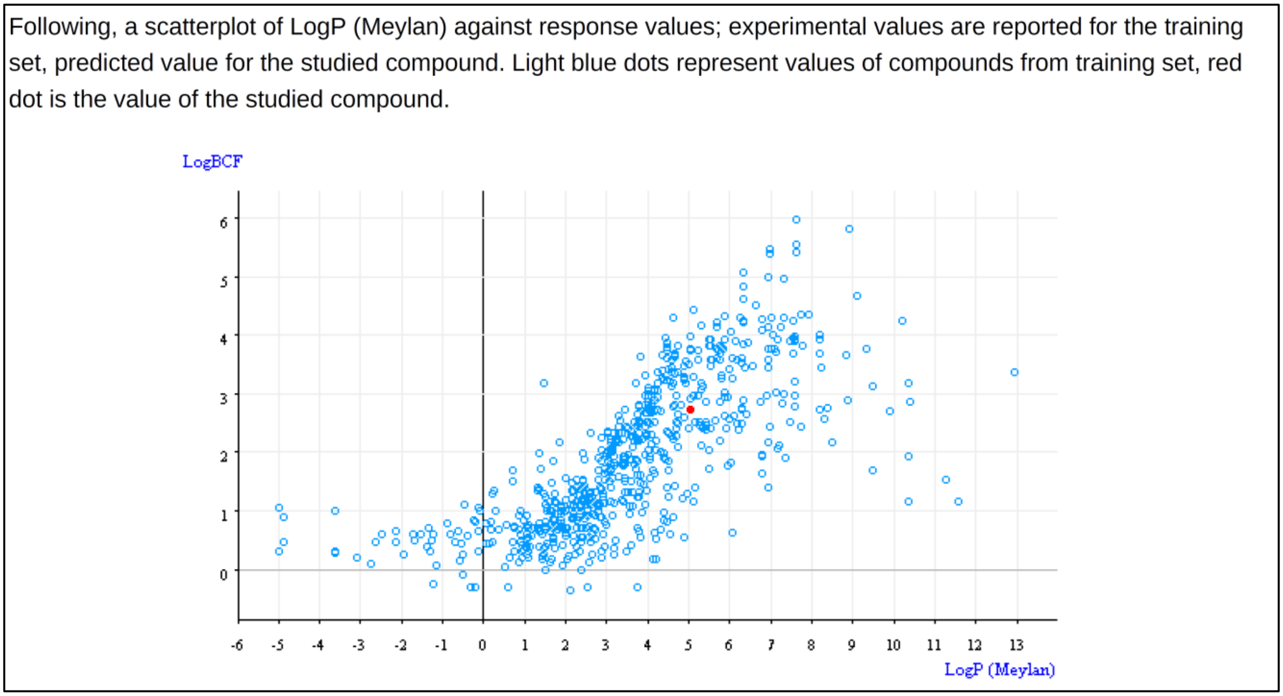
Compound #6

CAS: 18708-70-8
Dataset id:295 (Training Set)
SMILES: O=[N+]([O-])c1c(cc(cc1Cl)Cl)Cl
Similarity: 0.907
Experimental value : 2.47
Predicted value : 2.102

Figure 54. The similar chemicals found by the BCF Meylan model.

The training set of this model is quite rich in substances similar to the target. There are two of the three analogues containing four chlorine and the nitro group. There are three analogues with three chlorine atoms.

The plot of the BCF versus logKow is reported in *Figure 55*.

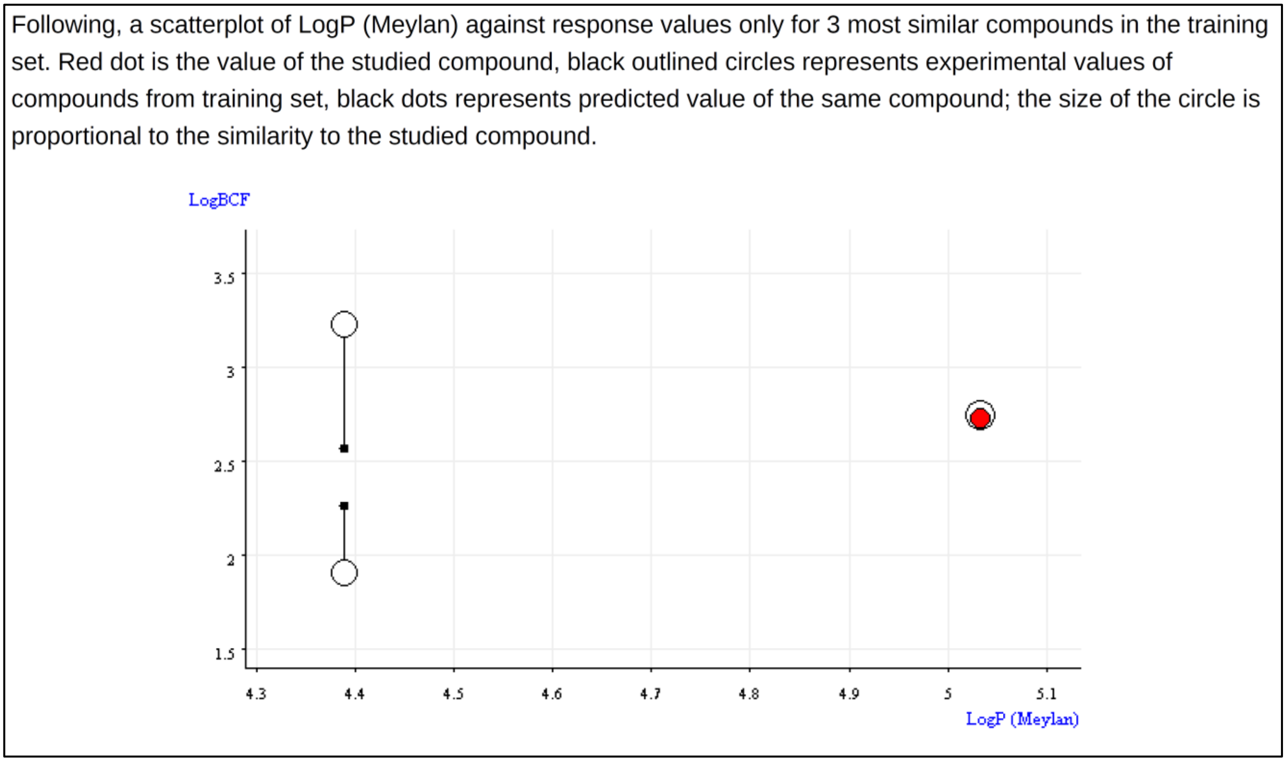


Following, a scatterplot of LogP (Meylan) against response values; experimental values are reported for the training set, predicted value for the studied compound. Light blue dots represent values of compounds from training set, red dot is the value of the studied compound.

Figure 55. The analysis of molecular descriptors of the BCF Meylan model (first part).

In this case, the target is somehow lower than expected, but anyhow in the “common” position.

We now consider the same plot with the most similar substances (*Figure 56*).



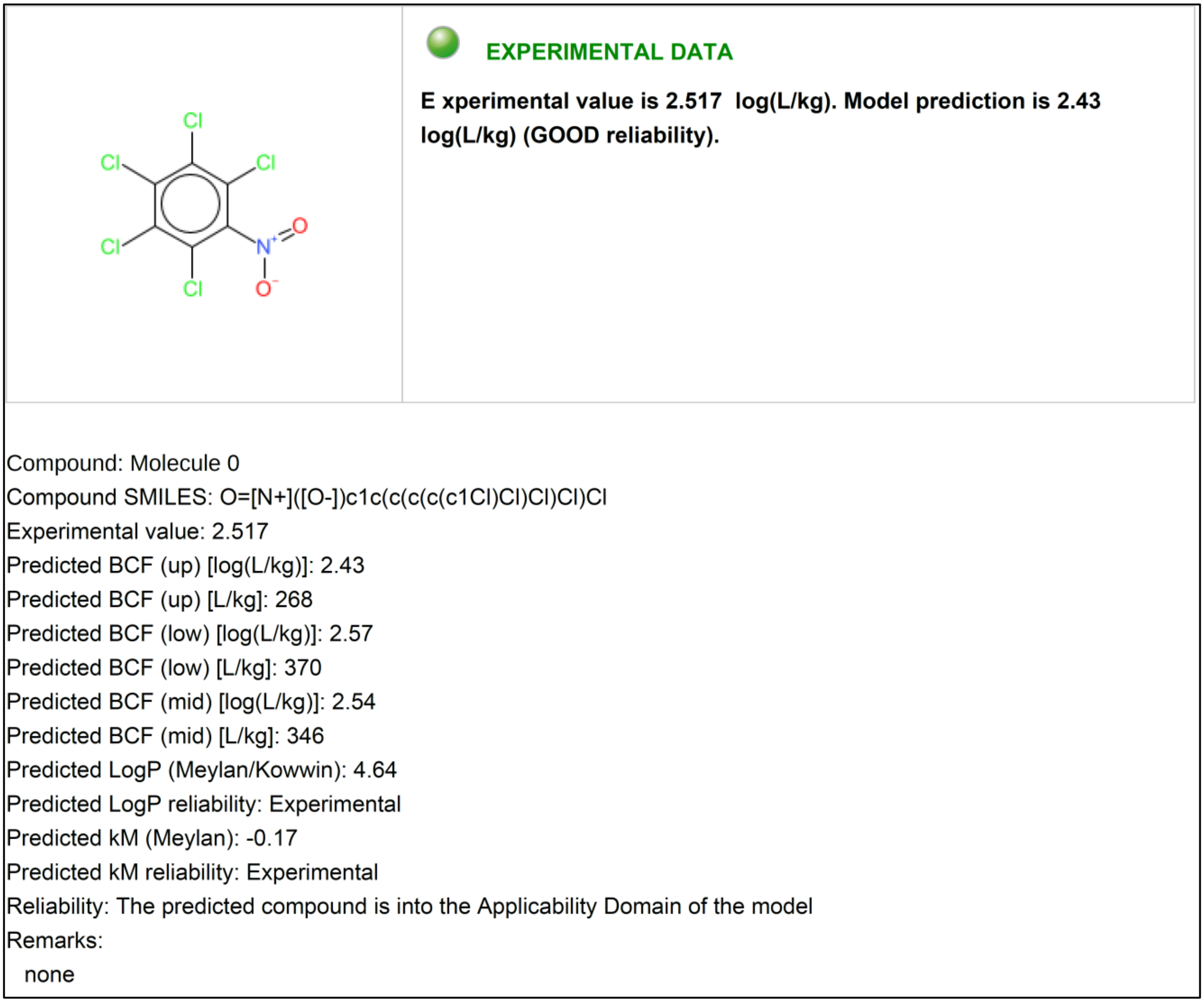
Following, a scatterplot of LogP (Meylan) against response values only for 3 most similar compounds in the training set. Red dot is the value of the studied compound, black outlined circles represents experimental values of compounds from training set, black dots represents predicted value of the same compound; the size of the circle is proportional to the similarity to the studied compound.

Figure 56. The analysis of molecular descriptors of the BCF Meylan model (second part).

Here we have to disregard the “similar” substance overlapping the target, because it is the target itself. Thus, we have the two analogues with four chlorine atoms. The predicted logKow is the same, but their experimental value is different (the open circles), actually very different, 1.4 log unit! We may suspect that the experimental values are not so accurate.

The Arnot-Gobas model.

Figure 57 shows the summary of the Arnot-Gobas prediction.



● EXPERIMENTAL DATA

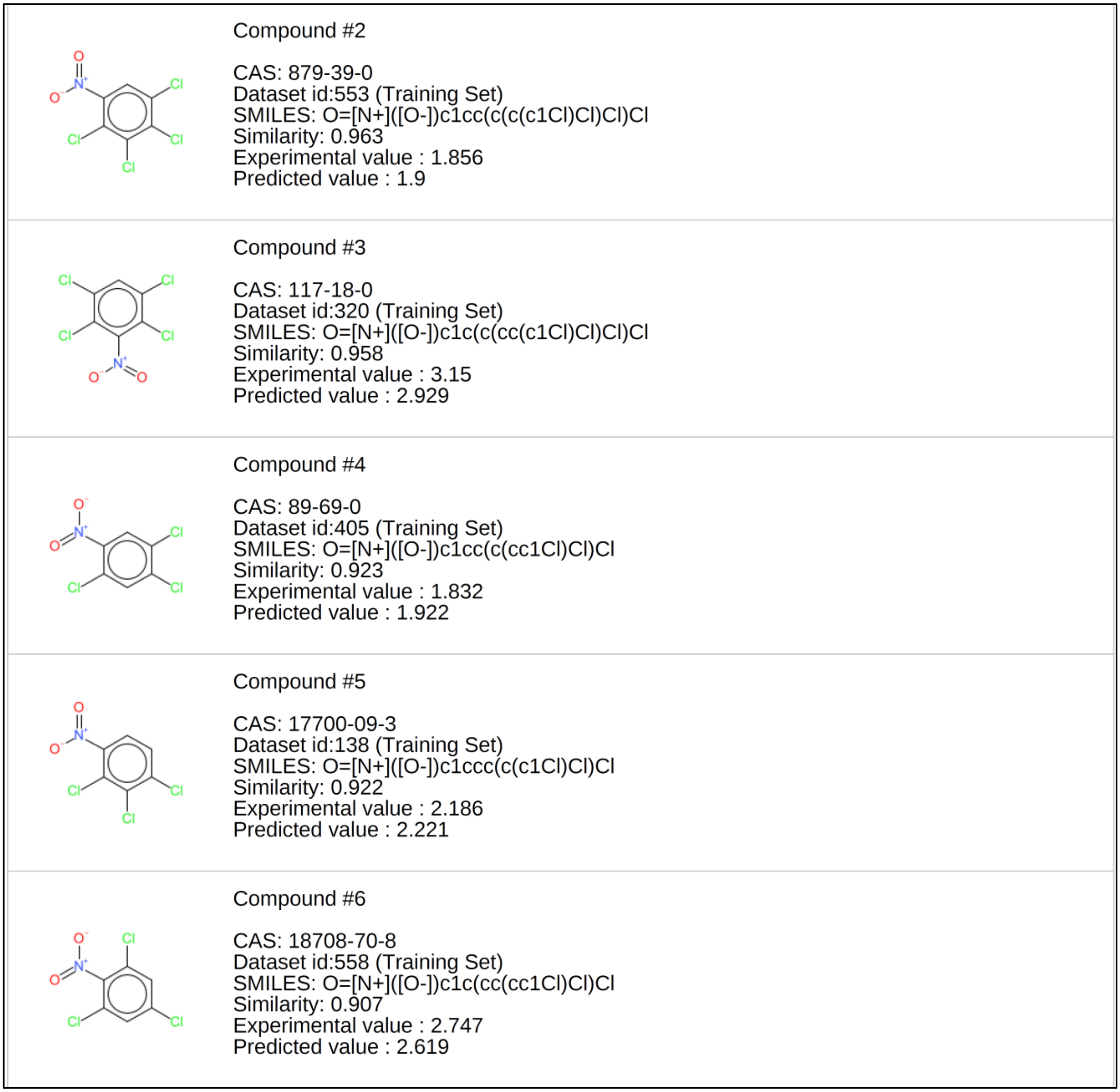
E xperimental value is 2.517 log(L/kg). Model prediction is 2.43 log(L/kg) (GOOD reliability).

Compound: Molecule 0
Compound SMILES: O=[N+]([O-])c1c(c(c(c(c1Cl)Cl)Cl)Cl)Cl
Experimental value: 2.517
Predicted BCF (up) [log(L/kg)]: 2.43
Predicted BCF (up) [L/kg]: 268
Predicted BCF (low) [log(L/kg)]: 2.57
Predicted BCF (low) [L/kg]: 370
Predicted BCF (mid) [log(L/kg)]: 2.54
Predicted BCF (mid) [L/kg]: 346
Predicted LogP (Meylan/Kowwin): 4.64
Predicted LogP reliability: Experimental
Predicted kM (Meylan): -0.17
Predicted kM reliability: Experimental
Reliability: The predicted compound is into the Applicability Domain of the model
Remarks:
none

Figure 57. The cover of the BCF Arnot-Gobas model.

Also in this case we have the experimental value, as 2.5. It is a different experimental value, compared with that provided by the Meylan model, but it is quite close to the value provided by the other model: 2.74. Also in this case, for the exercise' sake, we disregard the experimental value. The prediction is 2.43. The ADI is good.

In *Figure 58* the most similar substances (disregarding the identity: the target substance).



Compound #2

CAS: 879-39-0
Dataset id:553 (Training Set)
SMILES: O=[N+]([O-])c1cc(c(c(c1Cl)Cl)Cl)Cl
Similarity: 0.963
Experimental value : 1.856
Predicted value : 1.9

Compound #3

CAS: 117-18-0
Dataset id:320 (Training Set)
SMILES: O=[N+]([O-])c1c(c(cc(c1Cl)Cl)Cl)Cl
Similarity: 0.958
Experimental value : 3.15
Predicted value : 2.929

Compound #4

CAS: 89-69-0
Dataset id:405 (Training Set)
SMILES: O=[N+]([O-])c1cc(c(cc1Cl)Cl)Cl
Similarity: 0.923
Experimental value : 1.832
Predicted value : 1.922

Compound #5

CAS: 17700-09-3
Dataset id:138 (Training Set)
SMILES: O=[N+]([O-])c1ccc(c(c1Cl)Cl)Cl
Similarity: 0.922
Experimental value : 2.186
Predicted value : 2.221

Compound #6

CAS: 18708-70-8
Dataset id:558 (Training Set)
SMILES: O=[N+]([O-])c1c(cc(cc1Cl)Cl)Cl
Similarity: 0.907
Experimental value : 2.747
Predicted value : 2.619

Figure 58. The similar chemicals found by the BCF Arnot-Gobas model.

These are the same as those from the Meylan model, with some modifications in the predictions and in the experimental values too.

Overall conclusion

The most reliable predictions come from the Meylan and Arnot-Gobas models, according to the ADI. The predictions (we do not consider the experimental values) are at about 2.4 and 2.7. We discussed the prediction from CAESAR and concluded that a reasonable value is between 2.5 and 3. The KNN model provides the prediction at about 2.5. The last two predictions are less reliable. Anyhow, the four *in silico* model agree with values in a close range.

The ToxRead model does not provide a prediction, but apparently it may be possible to conclude that the value is higher than 3.3, based on the trend of values of the similar values, thus, referring to a plausible relationship between logKow and BCF (mechanistic reasoning). We have to better investigate this apparently conflictual value, compared with the *in silico* models. We have already commented that the relationship between logKow and BCF is a complex one Thus, the apparent mechanism, supported by the trend observed from the ToxRead plot, does not take into account that the BCF trend is not a linear one, as we discussed. In this case, the *in silico* models are more informative and reliable.

Further discussion

- We have seen that it is useful, for regression models, which provide continuous values, to start from the line of evidence which gives higher reliability.
- We have seen that when the *in silico* models are good, they provide reliable predictions. The advantage also come from the fact that there are multiple predictions, two of them with ADI values, but overall they get to the same conclusion, thus two predictions can be considered as key studies, and thus as supporting evidence.
- By the way, multiple values in agreement are always good, also in the case of experimental values, and this was an example. Within the JANUS software, the system considers agreement between experimental values as the first step, for instance.
- If the user thinks that it is convenient to use only the *in silico* models, this can be done, without using further lines of evidence. To provide data for REACH, it is easier if the evidence comes from a single kind of tools, such as *in silico* models, while it is more complex to provide data from heterogeneous values, such as read-across and *in silico*

models. Note. Multiple *in silico* models can be grouped within the same description and are not considered weight-of-evidence.

- The use of read-across is useful, as a way to have a general view of the values for similar substances. We have seen that values may vary changing the number of chlorine atoms, and that it is not always clear the relationship between the number of chlorine atoms and BCF.
- It is useful to look at the similar substances within different data sets, since each model has its own collection of data.
- It is always preferable to use interpolation and not extrapolation, which is more risky, as shown in the present case. We arrived at a wrong conclusion with extrapolation from ToxRead.
- The use of mechanism, of reasoning, may be biased when the reasoning is partial. We have seen that the assumption that the relationship between number of chlorine atoms and BCF was wrong. We have seen that looking at more data on other analogues, this relationship was not clearer. Thus, it is always recommended to double-check the correctness of our assumption using experimental data. If the data confirm the assumption, we can use our mechanistic hypothesis. If the data does not confirm our theory, a more complex theory is necessary. This is the great Galileo’s lesson. In the present case, as we discussed introducing the BCF models in VEGA, a more complex relationship, not linear, applies.
- It is useful to apply a feedback process where the individual lines of evidence are evaluated again based on the other elements arising from a new line of evidence.

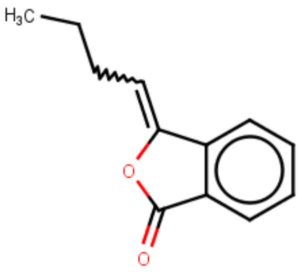
Example of an evaluation done with ToxRead and VEGA

Exercise 5

You should predict mutagenicity of the substance reported in the *Table 12*, using both VEGA and ToxRead. Below its structure.

VEGA provides 4 models, plus a consensus model, which integrates the overall results. This is what should be used, the whole set of models (there is another model for aromatic amines, not relevant now).

Table 1. Input molecule for the exercise 5.

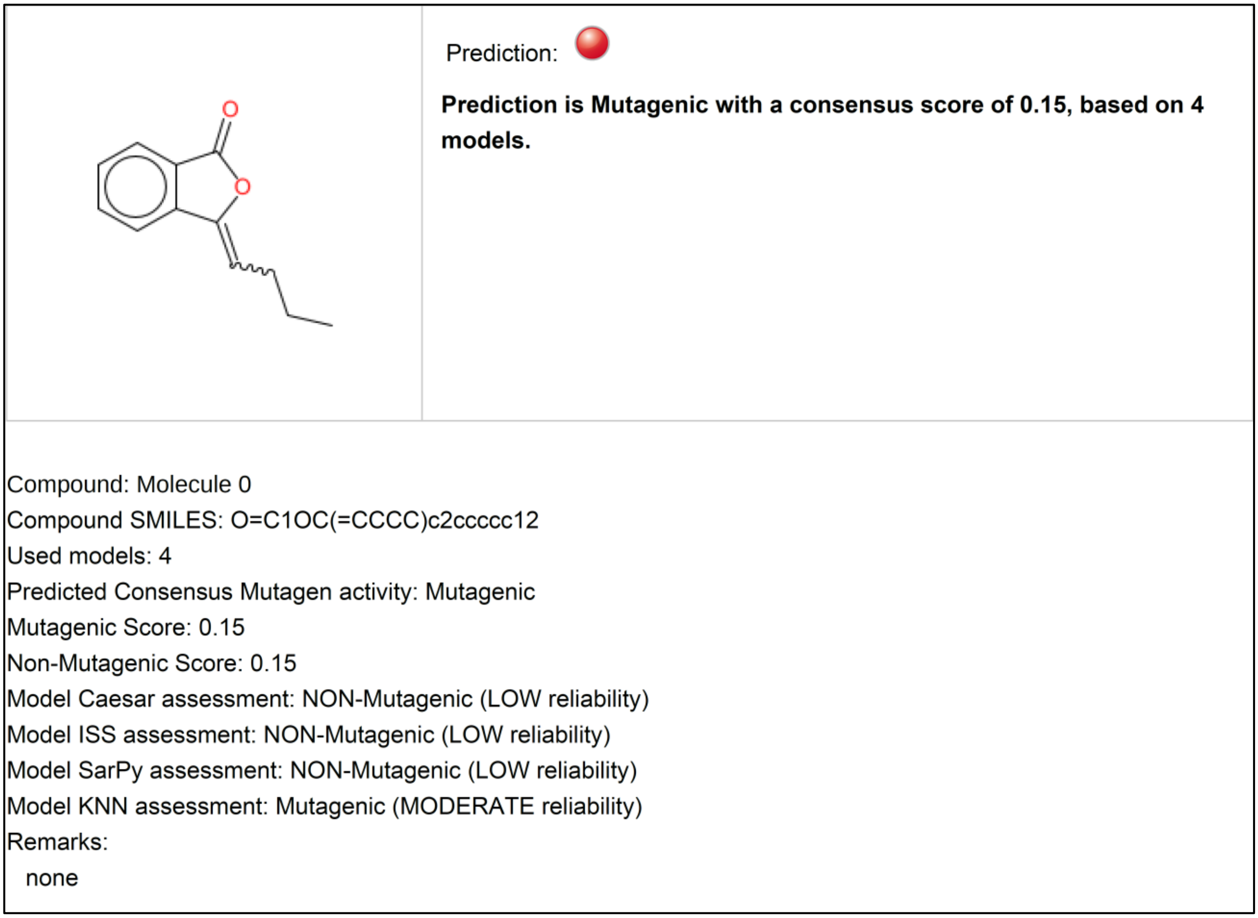
CAS No.	Name	SMILES	MOLECULE
551-08-6	3-Butylidenephthalide	O=C1OC(=CCCC)c2ccccc12	

The consensus model.

The **consensus model** provides a first overview (*Figure 59*).

The assessment indicates mutagenic, but the reliability is very low (0.15) and the prediction not reliable. The consensus model provides the same score value for non-mutagenicity too.

The results of the four models integrated within the Consensus model are reported too. All of them have low reliability, but one, the KNN model, with moderate reliability. Three models indicate non-mutagenicity with low reliability, while KNN indicates the opposite.



Prediction: ●

Prediction is Mutagenic with a consensus score of 0.15, based on 4 models.

Compound: Molecule 0
Compound SMILES: O=C1OC(=CCCC)c2ccccc12
Used models: 4
Predicted Consensus Mutagen activity: Mutagenic
Mutagenic Score: 0.15
Non-Mutagenic Score: 0.15
Model Caesar assessment: NON-Mutagenic (LOW reliability)
Model ISS assessment: NON-Mutagenic (LOW reliability)
Model SarPy assessment: NON-Mutagenic (LOW reliability)
Model KNN assessment: Mutagenic (MODERATE reliability)
Remarks:
none

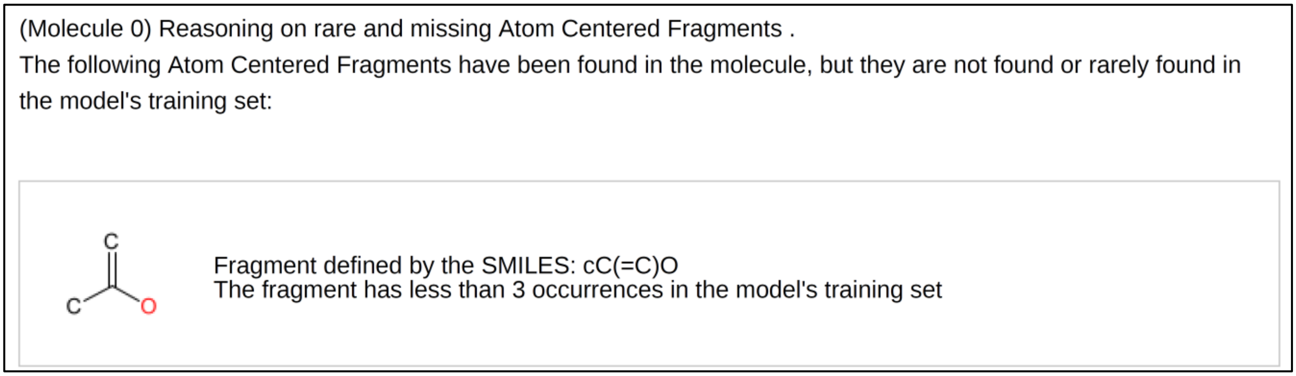
Figure 59. The cover of the Mutagenicity Consensus model.

The CAESAR model

The low ADI value of the CAESAR model is mainly due to the low concordance and accuracy values.

The ISS model

The low ADI value of the ISS model is mainly due to the low concordance value and for the presence of a rare fragment, as indicated in *Figure 60*.



(Molecule 0) Reasoning on rare and missing Atom Centered Fragments .

The following Atom Centered Fragments have been found in the molecule, but they are not found or rarely found in the model's training set:

Fragment defined by the SMILES: cC(=C)O
The fragment has less than 3 occurrences in the model's training set

Figure 60. The reasoning on rare and missing fragments of the Mutagenicity ISS model.

Thus, both models indicate as non-mutagenic the target, but the critical aspects are due to different reasons. As we said, describing the Ames models, the ISS has a dataset quite small and different from that of CAESAR. The CAESAR model contains a substance very similar to the target, reported in *Figure 61*.

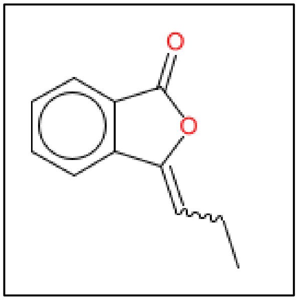


Figure 61. Similar substance found by the Mutagenicity CAESAR model.

It is exactly as the target, with one carbon less in the aliphatic chain. This, and another similar mutagenic substance, generate the low concordance value: there are two substances experimentally mutagenic, while the prediction is for non-mutagenicity.

Conversely, the lack of this similar substance in the ISS model generates the warning related to the rare fragment.

The SARpy model

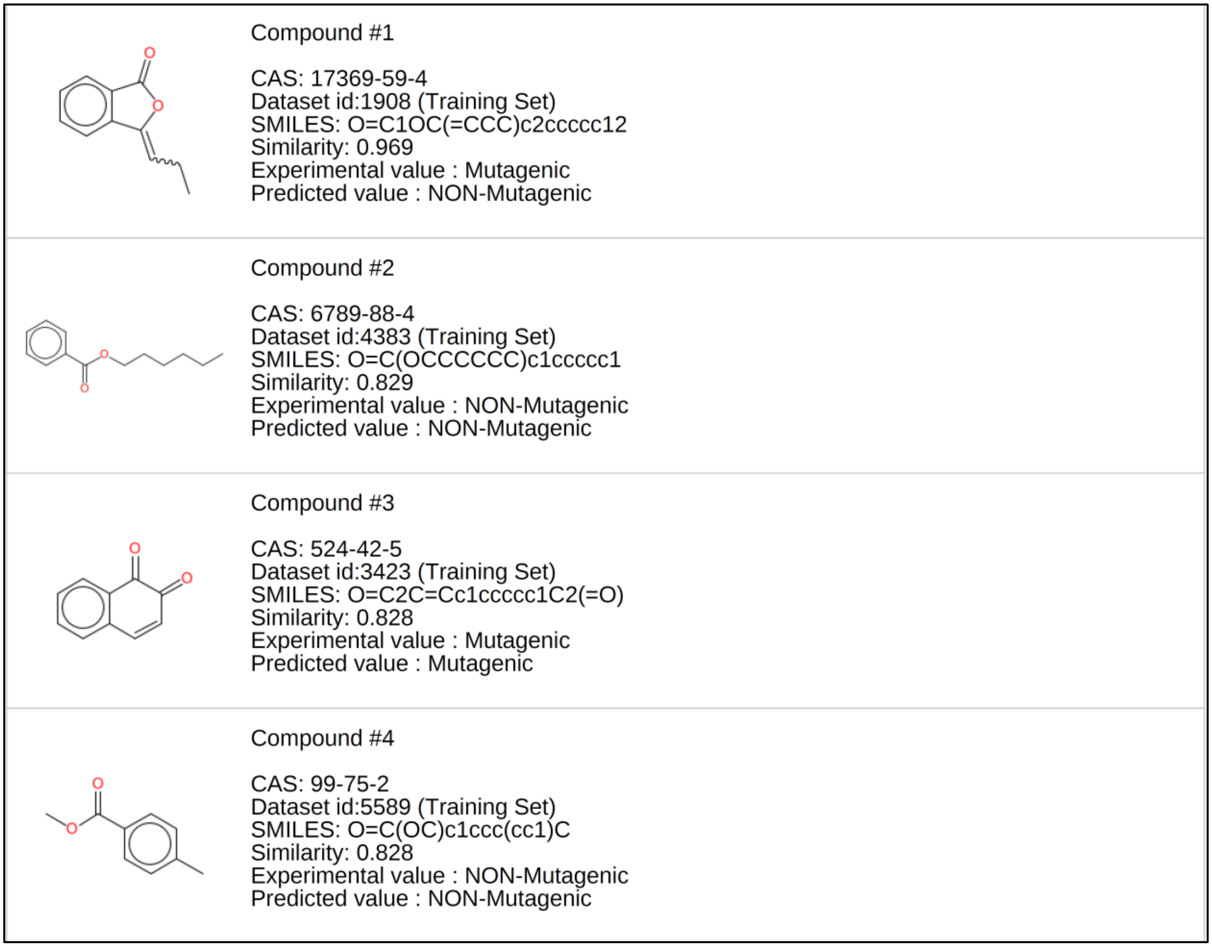
The low ADI value of the SARpy model is mainly due to the low concordance and accuracy values.

Regarding the mechanism, the ISS model, of course, does not find any structural alert (in the presence of a SA the prediction is mutagenic). SARpy, similarly, does not find any toxic SA, but identified a SA, the benzoic acid ester moiety, associated to lack of mutagenicity. Thus, we cannot use the mechanistic information.

The role of the SARpy *in silico* model is equivocal, and very uncertain.

The KNN model

The KNN model indicates mutagenicity, with moderate reliability. This model is a read-across model.



Compound #1

CAS: 17369-59-4
Dataset id:1908 (Training Set)
SMILES: O=C1OC(=CCC)c2ccccc12
Similarity: 0.969
Experimental value : Mutagenic
Predicted value : NON-Mutagenic

Compound #2

CAS: 6789-88-4
Dataset id:4383 (Training Set)
SMILES: O=C(OCCCCCC)c1ccccc1
Similarity: 0.829
Experimental value : NON-Mutagenic
Predicted value : NON-Mutagenic

Compound #3

CAS: 524-42-5
Dataset id:3423 (Training Set)
SMILES: O=C2C=Cc1ccccc1C2(=O)
Similarity: 0.828
Experimental value : Mutagenic
Predicted value : Mutagenic

Compound #4

CAS: 99-75-2
Dataset id:5589 (Training Set)
SMILES: O=C(OC)c1ccc(cc1)C
Similarity: 0.828
Experimental value : NON-Mutagenic
Predicted value : NON-Mutagenic

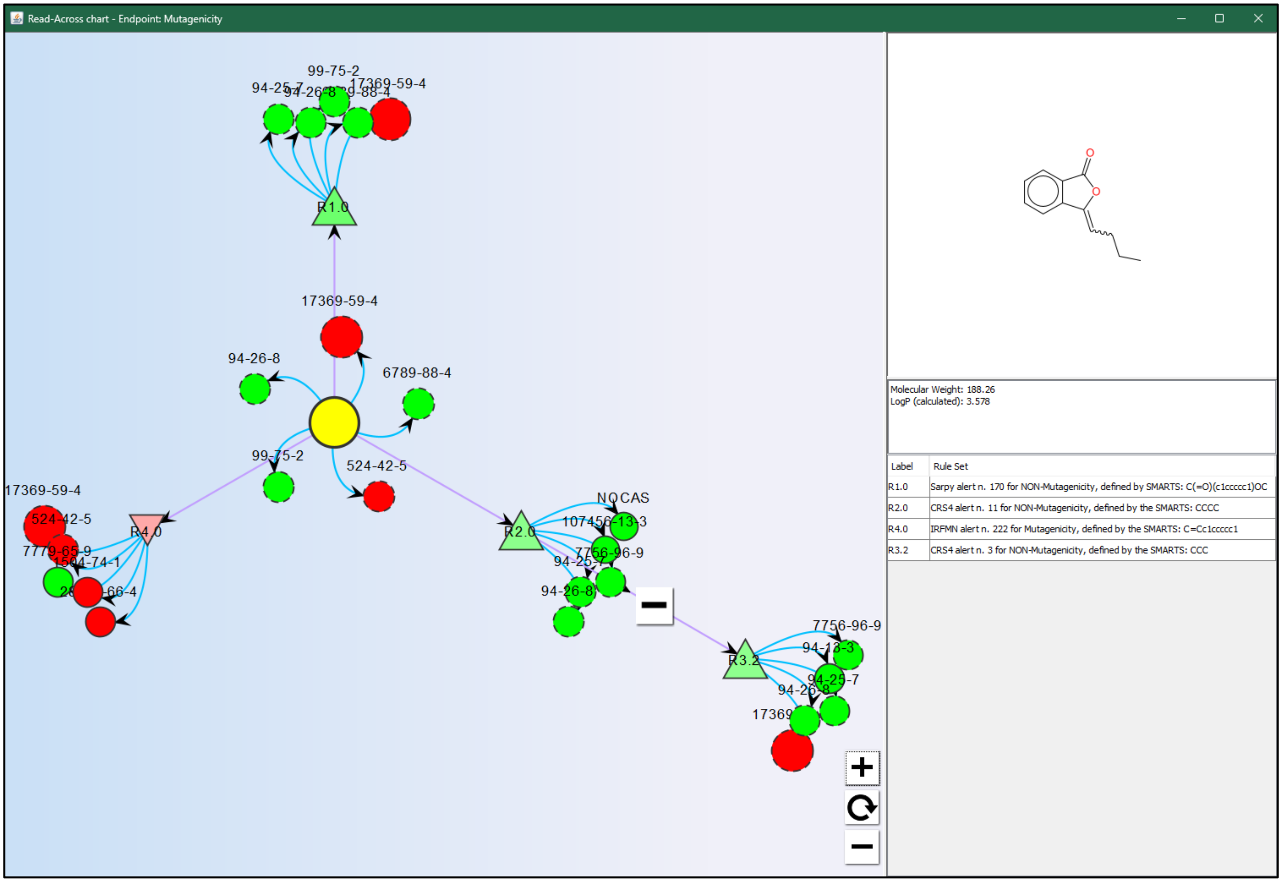
Figure 62. Similar substance found by the Mutagenicity KNN model.

For the assessment, it uses the first similar substances, four in this case (*Figure 62*).

Since KNN is an automatic tool, it may be convenient to look at the ToxRead results, where we have a better graphical representation, and more elements for reasoning.

ToxRead

Figure 63 shows the results using five similar substances.



Read-Across chart - Endpoint: Mutagenicity

Molecular Weight: 188.26
LogP (calculated): 3.578

Label	Rule Set
R1.0	Sarpy alert n. 170 for NON-Mutagenicity, defined by SMARTS: C(=O)(c1ccccc1)OC
R2.0	CRS4 alert n. 11 for NON-Mutagenicity, defined by the SMARTS: CCCC
R4.0	IRFMN alert n. 222 for Mutagenicity, defined by the SMARTS: C=Cc1ccccc1
R3.2	CRS4 alert n. 3 for NON-Mutagenicity, defined by the SMARTS: CCC

Figure 63. The output of Tox-Read for mutagenicity with six similar chemicals.

There are three rules, one indicating mutagenicity, and two lack of mutagenicity. These rules are represented by triangles, of different colour and orientation.

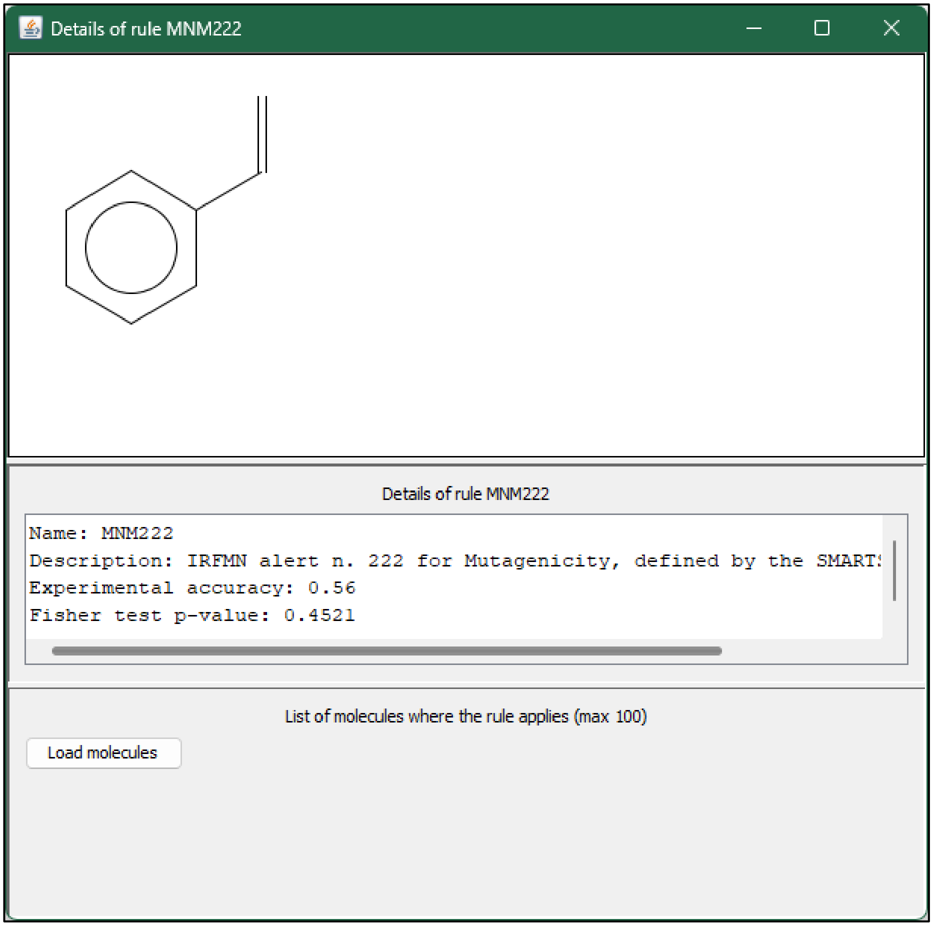
The first triangle, the green one in upper part of the figure, is the same as the SA identified by SARpy: the benzoic ester moiety. 82% of the substances with this fragment are not mutagenic. Still, we see that there is one mutagenic compound, which is the same identified by the VEGA models CEASAR; SARpy and KNN, represented in *Figure 61*.

Only one carbon is the difference between this mutagenic substance and the target. If we cannot explain why the presence of an additional carbon may cancel mutagenicity, we have to conclude that the target substance is mutagenic.

It is known that adding the chain may modify mutagenicity. The short aliphatic aldehydes are mutagenic, but with a longer chain that are no more mutagenic. But in this case, we do not know if this is the same, and when the mutagenicity disappears.

The second rule is simply a chain of four carbons. All the substances containing this fragment are not toxic, but they are not so similar to the target. If we further explore the graph, we have a rule associated to this one, in the lower part of the figure at the right. This fragment is a chain of three carbons. In this case we identify again the mutagenic similar substance shown above.

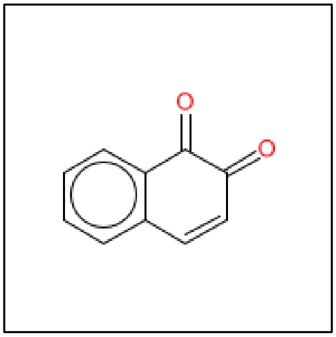
The third rule is for mutagenicity (*Figure 64*).



```
Name: MNM222
Description: IRFMN alert n. 222 for Mutagenicity, defined by the SMART
Experimental accuracy: 0.56
Fisher test p-value: 0.4521
```

Figure 64. The third rule found by Tox-Read for mutagenicity for the target substance.

Linked to this, there are four mutagenic substances.



The first one, is the same we presented repeatedly, very similar to the target. All the others are not so similar, and they may be mutagenic for the presence of other SAs. The second one is the substance reported by KNN (*Figure 65*).

Figure 65. The second similar found by Tox-Read for mutagenicity for the target substance.

The third one has an unsaturated aldehyde, and the fourth one has a polycyclic aromatic structure.

Overall conclusion

We have a very similar mutagenic substance. This is the key line of evidence. Since we cannot find any reason why the presence of an additional methyl in the chain cancels the mutagenicity, we conclude that the substance is mutagenic. Unfortunately, we have only one similar useful substance. The other substances are not so useful, because they miss the rare fragment, as already indicated as an issue by the ISS model in VEGA. This case is as the example of germanium that we discussed in Part A.

The mechanism provides some support. One possible explanation is the presence of the SA reported above.

The *in silico* models are not useful, because equivocal and of low reliability.

Discussion on this exercise.

- This exercise indicates that in some cases the read-across evaluation may be the best solution.
- If we have a classification model and a very similar substance (similarity > 0.9), this is quite probably the key element. Here, we have a very peculiar situation with a substance which is almost identical to the target. When this happens, it is recommended to start from this line of evidence.

B.4. VERA

We have seen with exercise 4 that read-across can be very useful. However, the evaluation has been done manually. The KNN model, among the different VEGA models, can provide an automatic way to identify the good similar substances. However, as we commented, KNN has not toxicological knowledge. It simply uses structural similarity. Thus, similar substances may be not relevant, and often they are merged without any chemical or toxicological consideration.

ToxRead helps to navigate clustering substances according to different features, not only according to structural similarity. However, it does not work in batch mode, it is for manual exploration.

VERA offers advantages over the tools so far discussed.

At the basis of VERA there is the concept of multiple metrics, and comparison between different clusters of similar substances. We have seen above, for instance, in Exercise 4, that there were different clusters of toxic or non-toxic substances depending on the SAs. ToxRead, in case of ambiguity (the *in silico* models were equivocal too), was not able to take a decision, and thus the circle at the centre, representing the target substance, was yellow: not red, not green.

VERA looks if there are similar substances present in both conflicting clusters and tries to decide based on the read-across data on the substances in the clusters.

Initially, the substances are gathered according to the VEGA similarity (i.e. structural similarity), but using a low value: 0.65. As we have seen, this value is not what is used within VEGA to identify similar substances. This is to allow for further pruning and cluster substances, according to different metrics for similarity. SAs or other factors representing the adverse effect are used to get substances useful to verify if the presence in the molecule of the SAs is always associated with adverse effect. Then, other metrics are used to cluster substances, such as a collection of molecular groups, to explore the role of the different substituents in the molecule to modulate the effect.

To learn more on VERA, you can see this video (*https://www.life-concertreach.eu/resources-item/web-seminars-17-31-may-2023/*).

B.5. SWAN

SWAN represents the integration of the information related to the read-across assessment, provided by VERA, and the assessment based on the *in silico* models, as provided by VEGA. To get the value from the *in silico* models already integrated for the same endpoint, the outcome of JANUS is used.

To learn more on SWAN, you can see this video (*https://www.life-concertreach.eu/resources-item/life-concert-reach-project-final-workshop-training-session-20-06-2023/*).

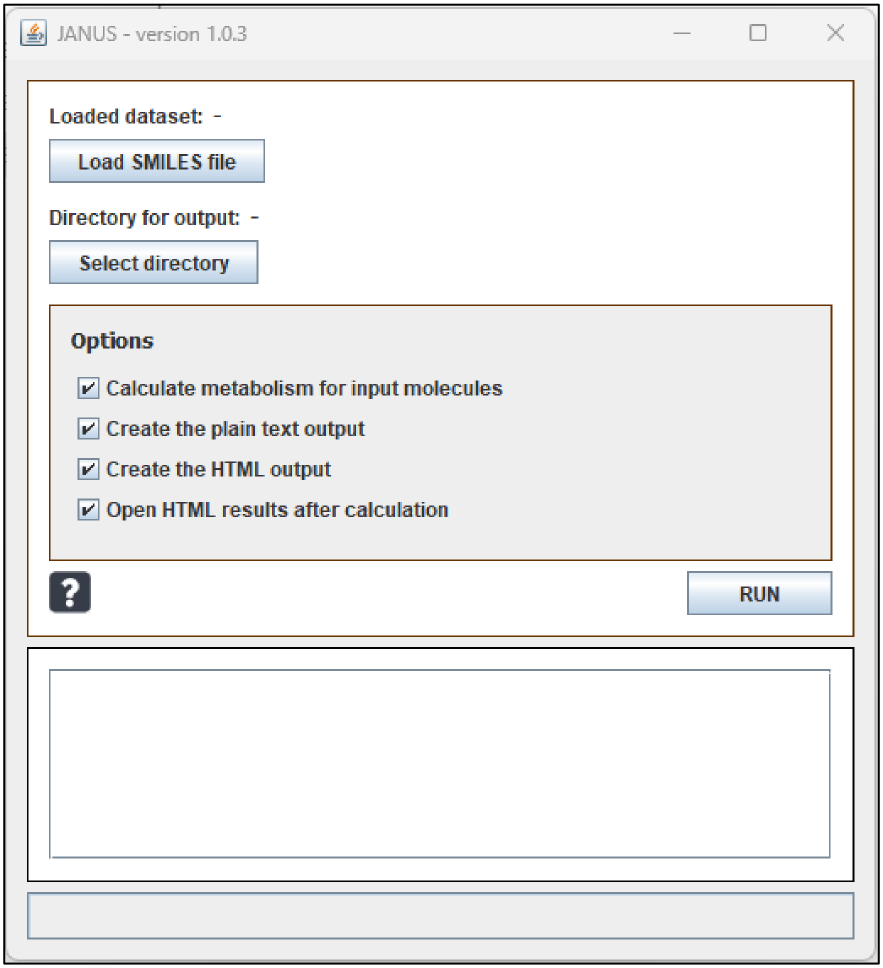
B.6. JANUS

As mentioned above VEGA offers more than one hundred of models to estimate several toxicological, ecotoxicological, environmental, and physico-chemical properties. The user can run them one-by-one or in-batch, for a single or for many chemicals. In some cases more models are available in VEGA for the same property (e.g. the four models for BCF described before). In these cases, the user has to combine the results manually, which can be complicated if many substances have to be evaluated. JANUS offers the opportunity to run 48 models of VEGA and combine the results automatically. The models refer to REACH requirements and thresholds: persistence (P; it evaluates P in three compartments: water, sediment and soil), bioaccumulation (B; evaluated through the BCF in fish), aquatic toxicity (T; in details, it evaluates acute and chronic toxicity towards algae, *Daphnia*, and fish), carcinogenicity (C), mutagenicity (M), reproductive/development toxicity (R), and endocrine disruption (ED; in details, it focus on androgen and estrogen binding activities). In addition, the user can run a model to predict the potential microbial metabolites of each substance in input and process the resulting metabolites with the 48 models.

The purpose of JANUS is the prioritization according to the REACH PBT, CMR, and ED properties. It gives three scores, one bases on P and B (named vPvB), one on the human related properties CMR and ED (named SVHC) and the last on all the properties (named PBT). For each property, it also gives a property score and the details of the prediction of each model run. These property scores combine the predictions available and their reliability (starting from the VEGA ADI). All the scores range from 0 (no concern) to 1 (maximum concern). Around 0.5 there is a "grey zone" in which converge the chemicals with moderate concern and the chemicals with low reliability (regardless of the predicted concern).

Example 1 – prioritization

To see how JANUS works we need a list of chemicals. In this example we used some chemicals classified as substances of very high concern (SVHC) by ECHA (*https://echa.europa.eu/candidate-list-table*) and other that are considered 'safe' by the US EPA (*https://www.epa.gov/saferchoice/safer-ingredients*) and flaged as 'GreenCircle'. You can run in JANUS this list flagging the 'Calculate metabolism for input molecules' (*Figure 66*).



JANUS - version 1.0.3

Loaded dataset: -

Directory for output: -

Options

☑ Calculate metabolism for input molecules

☑ Create the plain text output

☑ Create the HTML output

☑ Open HTML results after calculation

Figure 66. The input window of JANUS.

Looking at the final scores, one can see that the score SVHC has a higher number of chemicals with a high score, whereas the other two has more chemicals with a low score. This is due to the different approaches adopted. The human related properties are evaluated with a conservative approach. Indeed, for the human-related properties most models are in classification but, for the prioritization purpose, we need a quantitative value. To assign a quantitative value we have to use surrogate models; therefore, we adopted a conservative approach. The environmental and ecotoxicological endpoints, which are quantitative, adopt a more balanced approach, to better discriminate the different concerns and not to consider of concern all the chemicals. In addition, the human-related properties and the aquatic toxicity count for one parameter (the T of the PBT). For this reason, the vPvB and the PBT scores are quite similar. The *Table 13* shows the results sorted by the score PBT. The original class represents the reason for the inclusion in the SVHC or in safer chemical ingredient lists. The two PBT/vPvB chemicals are predicted with a high vPvB and PBT scores (above 0.6). They also have high SVHC score due to its conservative approach. The three CMR/ED chemicals have low vPvB and PBT, but high SVHC scores. The green circles have all the three scores below 0.5. Anyway, the summarized and detailed output for each property evaluated are available to allow the user a deeper analysis.

The three metabolites generated can add useful information. Chemical 25973-55-1 has one metabolite that results of lower concern for the environmental/ecotoxicological properties, comparable toxicity for the human-related properties. Chemicals 108-46-3 has two metabolites, both with comparable environmental/ecotoxicological concern. One has lower SVHC score, but the other has higher SVHC score. This information can help the user in the assessment, identifying chemicals that can be generated during degradation. Even if usually metabolism reduces the toxicity of a chemical, in some cases this is not. We have also to consider that these are microbial metabolites, which may be less toxic for bacteria but more or equally toxic for the other living organisms.

Table 13. The three final scores for the input chemicals and the generated metabolites. The original classification is also reported.

CAS No.	Original Class	SCORE(vPvB)	SCORE(SVHC)	SCORE(PBT)
25973-55-1	PBT;vPvB	0.628	0.723	0.66
36437-37-3	vPvB	0.613	0.727	0.651
25973-55-1 [M-01]*	-	0.318	0.727	0.419
106-94-5	R	0.283	0.867	0.412
2601-33-4	GreenCircle	0.248	0.449	0.305
108-46-3 [M-01]	-	0.155	0.927	0.282
64-67-5	CM	0.153	0.93	0.279
110-40-7	GreenCircle	0.167	0.183	0.226
108-46-3 [M-02]	-	0.13	0.585	0.216
108-46-3	ED	0.09	0.865	0.191
127-19-5	R	0.09	0.769	0.184
1117-86-8	GreenCircle	0.125	0.142	0.176
669-90-9	GreenCircle	0.125	0.142	0.167

**First metabolite of the chemical 25973-55-1.*

If the user is interested in one property, he/she can sort the chemicals using the property score of the property of interest, which combines the value assigned to the property and its reliability. For instance, look at the chemicals in the *Table 14*. The first three chemicals have the same reliability, but the second and third chemicals are predicted more bioaccumulative than the first, therefore they have higher scores. The third and fourth chemicals are predicted with similar bioaccumulation potential, but the third, has a lower score due to the higher reliability of the prediction.

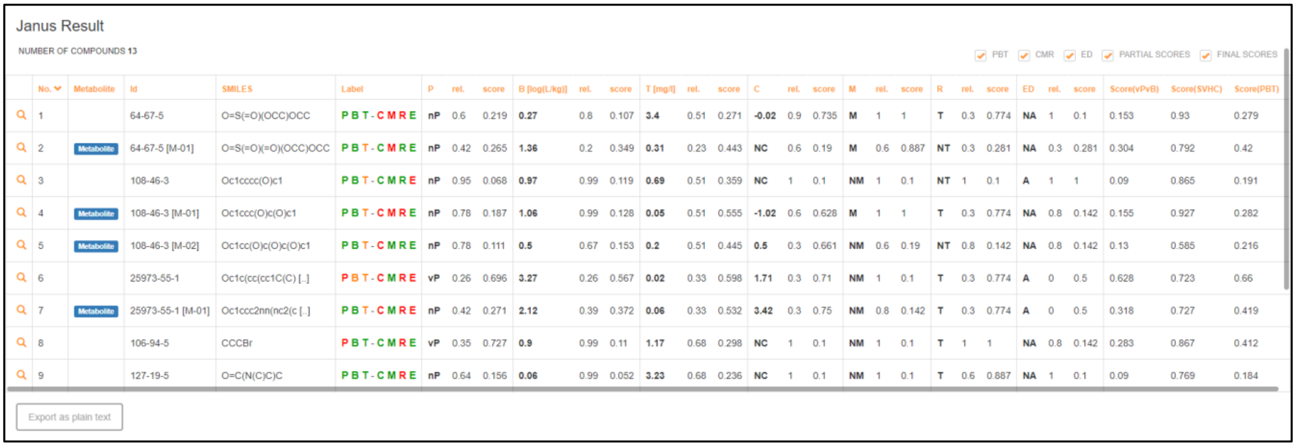
Table 14. Prediction details for bioaccumulation for four chemicals (assessment, reliability, and property score).

CAS	SMILES	B assessment [log units]	B reliability	B score
127-19-5	N,N-Dimethylacetamide (DMAC)	0.064	0.99	0.052
106-94-5	1-bromopropane (n-propyl bromide)	0.899	0.99	0.11
108-46-3	Resorcinol	0.97	0.99	0.119
2601-33-4	Myristyl betaine	0.9	0.29	0.288

Example 2 – single property 1

The user may use the scores to prioritize a list of chemicals, as in the example above, or can use JANUS to automatically combine the models for the same property available in VEGA (for the PBT, CMR and ED properties). For this example we focus on the chemical with the CAS No. 25973-55-1.

Figure 67 shows the main page of the results in html format (flag the option ‘Create the HTML output’). The user can decide what to visualize in addition to the identification of the substance and the label: the PBT, CMR, and/or ED properties, the partial scores (the scores of each property) and/or the final scores. Some chemicals are the metabolites generated by JANUS (e.g. the molecule No. 2 is a metabolite of the No. 1).



Janus Result

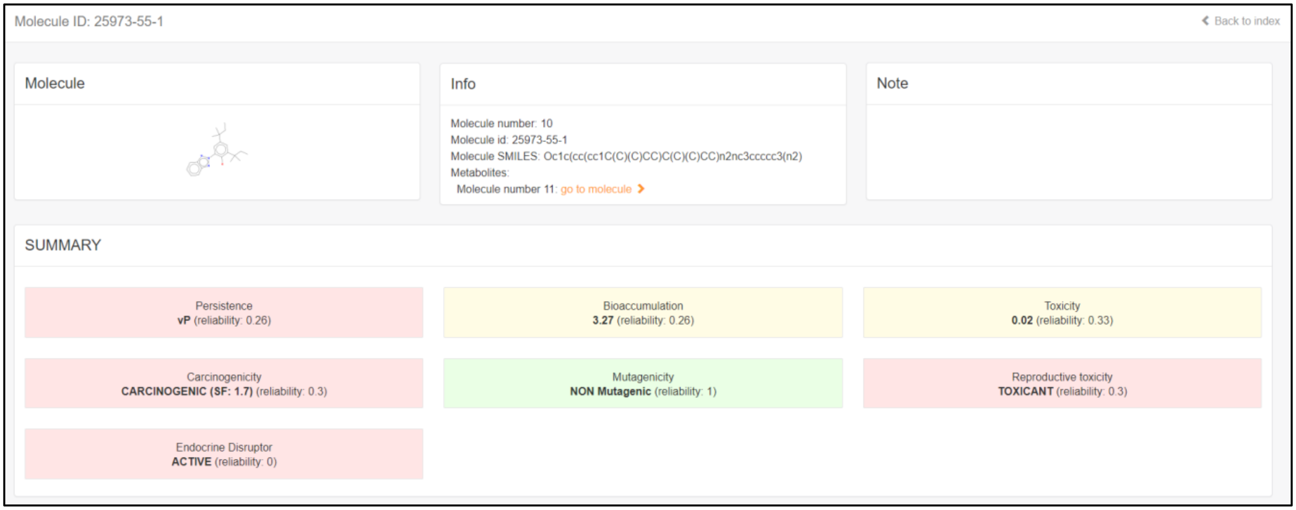
NUMBER OF COMPOUNDS 13

☑ PBT ☑ CMR ☑ ED ☑ PARTIAL SCORES ☑ FINAL SCORES

	No.	Metabolite	Id	SMILES	Label	P	rel.	score	B [log(L/kg)]	rel.	score	T [mg/l]	rel.	score	C	rel.	score	M	rel.	score	R	rel.	score	ED	rel.	score	Score(vPvB)	Score(SVHC)	Score(PBT)
🔍	1		64-67-5	O=S(=O)(OCC)OCC	PBT-CMRE	nP	0.6	0.219	**0.27**	0.8	0.107	**3.4**	0.51	0.271	**-0.02**	0.9	0.735	**M**	1	1	**T**	0.3	0.774	**NA**	1	0.1	0.153	0.93	0.279
🔍	2	Metabolite	64-67-5 [M-01]	O=S(=O)(=O)(OCC)OCC	PBT-CMRE	nP	0.42	0.265	**1.36**	0.2	0.349	**0.31**	0.23	0.443	**NC**	0.6	0.19	**M**	0.6	0.887	**NT**	0.3	0.281	**NA**	0.3	0.281	0.304	0.792	0.42
🔍	3		108-46-3	Oc1cccc(O)c1	PBT-CMRE	nP	0.95	0.068	**0.97**	0.99	0.119	**0.69**	0.51	0.359	**NC**	1	0.1	**NM**	1	0.1	**NT**	1	0.1	**A**	1	1	0.09	0.865	0.191
🔍	4	Metabolite	108-46-3 [M-01]	Oc1ccc(O)c(O)c1	PBT-CMRE	nP	0.78	0.187	**1.06**	0.99	0.128	**0.05**	0.51	0.555	**-1.02**	0.6	0.628	**M**	1	1	**T**	0.3	0.774	**NA**	0.8	0.142	0.155	0.927	0.282
🔍	5	Metabolite	108-46-3 [M-02]	Oc1cc(O)c(O)c(O)c1	PBT-CMRE	nP	0.78	0.111	**0.5**	0.67	0.153	**0.2**	0.51	0.445	**0.5**	0.3	0.661	**NM**	0.6	0.19	**NT**	0.8	0.142	**NA**	0.8	0.142	0.13	0.585	0.216
🔍	6		25973-55-1	Oc1c(cc(cc1C(C) [..]	PBT-CMRE	vP	0.26	0.696	**3.27**	0.26	0.567	**0.02**	0.33	0.598	**1.71**	0.3	0.71	**NM**	1	0.1	**T**	0.3	0.774	**A**	0	0.5	0.628	0.723	0.66
🔍	7	Metabolite	25973-55-1 [M-01]	Oc1ccc2nn(nc2(c [..]	PBT-CMRE	nP	0.42	0.271	**2.12**	0.39	0.372	**0.06**	0.33	0.532	**3.42**	0.3	0.75	**NM**	0.8	0.142	**T**	0.3	0.774	**A**	0	0.5	0.318	0.727	0.419
🔍	8		106-94-5	CCCBr	PBT-CMRE	vP	0.35	0.727	**0.9**	0.99	0.11	**1.17**	0.68	0.298	**NC**	1	0.1	**NM**	1	0.1	**T**	1	1	**NA**	0.8	0.142	0.283	0.867	0.412
🔍	9		127-19-5	O=C(N(C)C)C	PBT-CMRE	nP	0.64	0.156	**0.06**	0.99	0.052	**3.23**	0.68	0.236	**NC**	1	0.1	**NM**	1	0.1	**T**	0.6	0.887	**NA**	1	0.1	0.09	0.769	0.184

Figure 67. The main page of the html output of JANUS.

The target chemical (not shown in the figure above) has a PBT score of 0.66, a vPvB score of 0.628 and a SVHC score of 0.723. This means that it is a substance of concern, even if scores are not particularly high. Let's examine the details of the predictions. Clicking on the magnifying glass in the html output or looking at the txt files saved in the output folder, the user can analyse the results of each molecule. *Figure 68* shows the summary of the predictions.



Molecule ID: 25973-55-1

Molecule

Info

Molecule number: 10
Molecule id: 25973-55-1
Molecule SMILES: Oc1c(cc(cc1C(C)(C)CC)C(C)(C)CC)n2nc3ccccc3(n2)
Metabolites:
Molecule number 11: go to molecule ›

Note

SUMMARY

Persistence
vP (reliability: 0.26)

Bioaccumulation
3.27 (reliability: 0.26)

Toxicity
0.02 (reliability: 0.33)

Carcinogenicity
CARCINOGENIC (SF: 1.7) (reliability: 0.3)

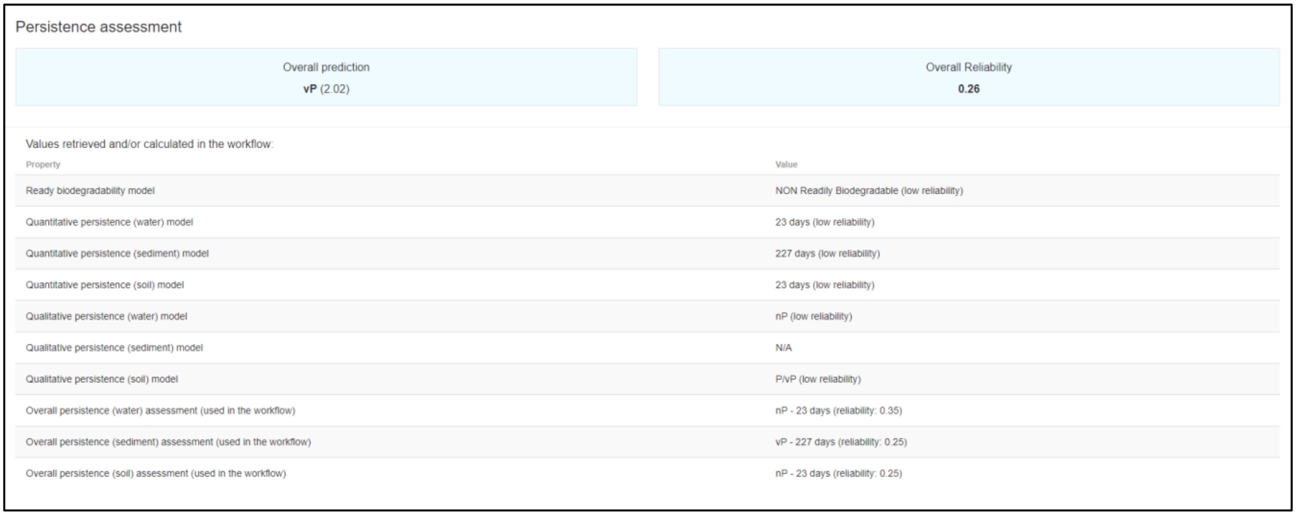
Mutagenicity
NON Mutagenic (reliability: 1)

Reproductive toxicity
TOXICANT (reliability: 0.3)

Endocrine Disruptor
ACTIVE (reliability: 0)

Figure 68. The first part of the detail page of the html output of JANUS for the target chemical.

The target is predicted as vP but with a low reliability. For persistence, according to REACH tree compartments have to be analysed: water, soil and sediment. For each there are one qualitative and one quantitative model. Since the threshold to classify the chemicals are different depending on the compartment, the output of each evaluation is converted into an index (the 2.02 of the final prediction shown in *Figure 69*). It does not represent a half-life value. All the predictions have low reliability. Only for sediment the half-life (in days) exceeds the threshold for vP chemicals, but to be P or vP in one compartment is sufficient to be classified accordingly. An additional model predicts the target non readily biodegradable (even if with low reliability), supporting the prediction of persistence.



Persistence assessment

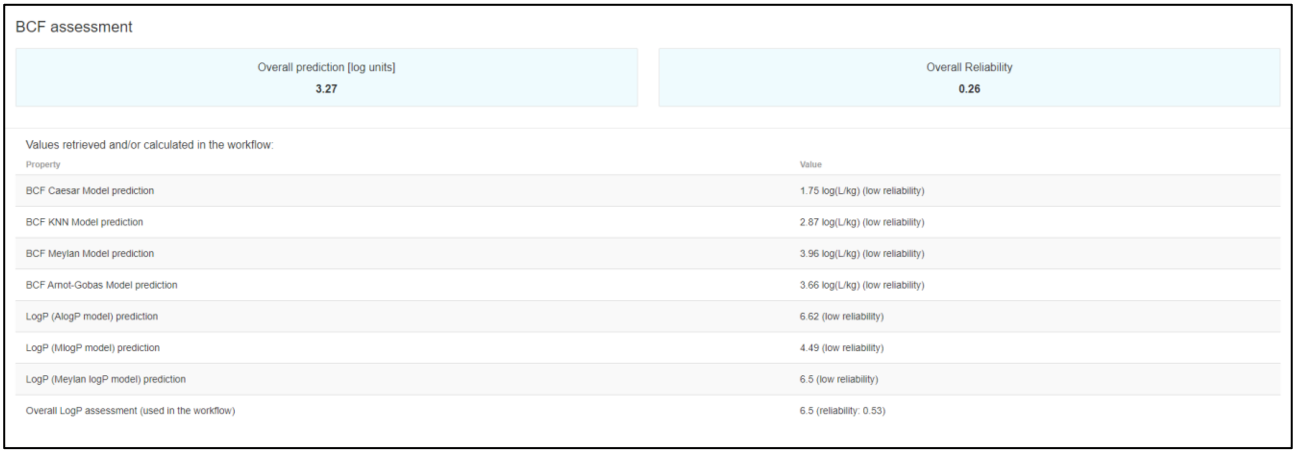
Overall prediction	Overall Reliability
vP (2.02)	**0.26**

Values retrieved and/or calculated in the workflow:

Property	Value
Ready biodegradability model	NON Readily Biodegradable (low reliability)
Quantitative persistence (water) model	23 days (low reliability)
Quantitative persistence (sediment) model	227 days (low reliability)
Quantitative persistence (soil) model	23 days (low reliability)
Qualitative persistence (water) model	nP (low reliability)
Qualitative persistence (sediment) model	N/A
Qualitative persistence (soil) model	P/vP (low reliability)
Overall persistence (water) assessment (used in the workflow)	nP - 23 days (reliability: 0.35)
Overall persistence (sediment) assessment (used in the workflow)	vP - 227 days (reliability: 0.25)
Overall persistence (soil) assessment (used in the workflow)	nP - 23 days (reliability: 0.25)

Figure 69. The details for persistence in the html output page of JANUS for the target chemical.

The target is predicted nB with low reliability (see *Figure 70*). The predicted logBCF (3.27 l.u.), that is the combination of the four models available in VEGA, is very close to the B threshold (of 3.3 l.u.). All the predictions have low reliability and spread from 3.96 (vB) to 1.75 (nB). To support the decision, the logKow is also analysed. Again, the tree models give quite diverse predictions, all with low reliability. If the user is interested in a deeper analysis, he/she can run the models in VEGA. The low reliability and the value sufficiently high generate a partial score of 0.567, slightly above the 'grey zone'.



BCF assessment

Overall prediction [log units]	Overall Reliability
3.27	**0.26**

Values retrieved and/or calculated in the workflow:

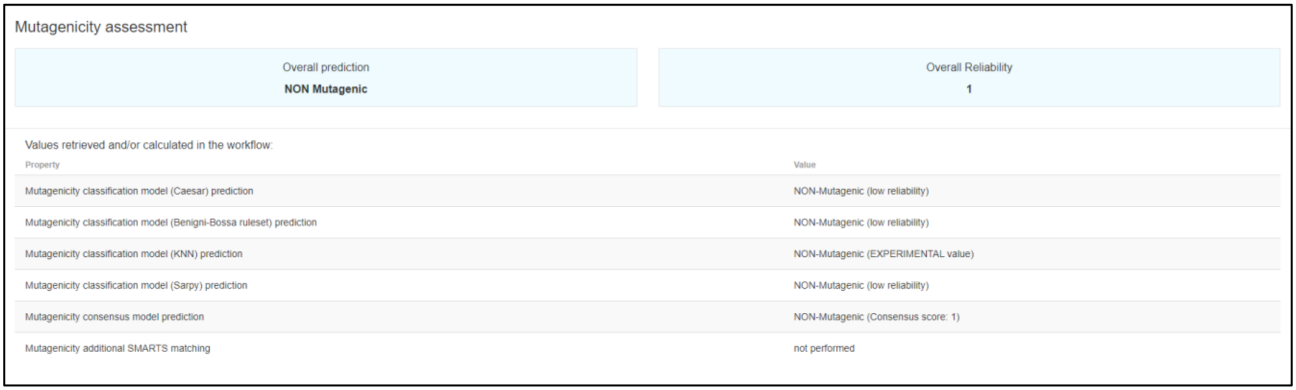
Property	Value
BCF Caesar Model prediction	1.75 log(L/kg) (low reliability)
BCF KNN Model prediction	2.87 log(L/kg) (low reliability)
BCF Meylan Model prediction	3.96 log(L/kg) (low reliability)
BCF Arnot-Gobas Model prediction	3.66 log(L/kg) (low reliability)
LogP (AlogP model) prediction	6.62 (low reliability)
LogP (MlogP model) prediction	4.49 (low reliability)
LogP (Meylan logP model) prediction	6.5 (low reliability)
Overall LogP assessment (used in the workflow)	6.5 (reliability: 0.53)

Figure 70. The details for bioaccumulation in the html output page of JANUS for the target chemical.

For aquatic toxicity the situation is similar, with a variability in the predictions, all with low reliability. Aquatic toxicity combines three representative organisms (algae, *Daphnia*, and fish), using acute and chronic toxicity (even if the prediction is based on the chronic values). The prediction (0.02 mg/L) is very close to the threshold of 0.01 mg/L. Again, a prediction close to the threshold but with low reliability generates a partial score (0.598) slightly above the ‘grey zone’ (even if closer to 1 than to 0).

Except mutagenicity, the other properties are predicted as C, R and ED with low reliability. The user can do the same reasoning as above.

For mutagenicity one model has an experimental value (M), therefore the evaluation is based on it, even if the other models are concordant (*Figure 71*).



Mutagenicity assessment

Overall prediction	Overall Reliability
NON Mutagenic	1

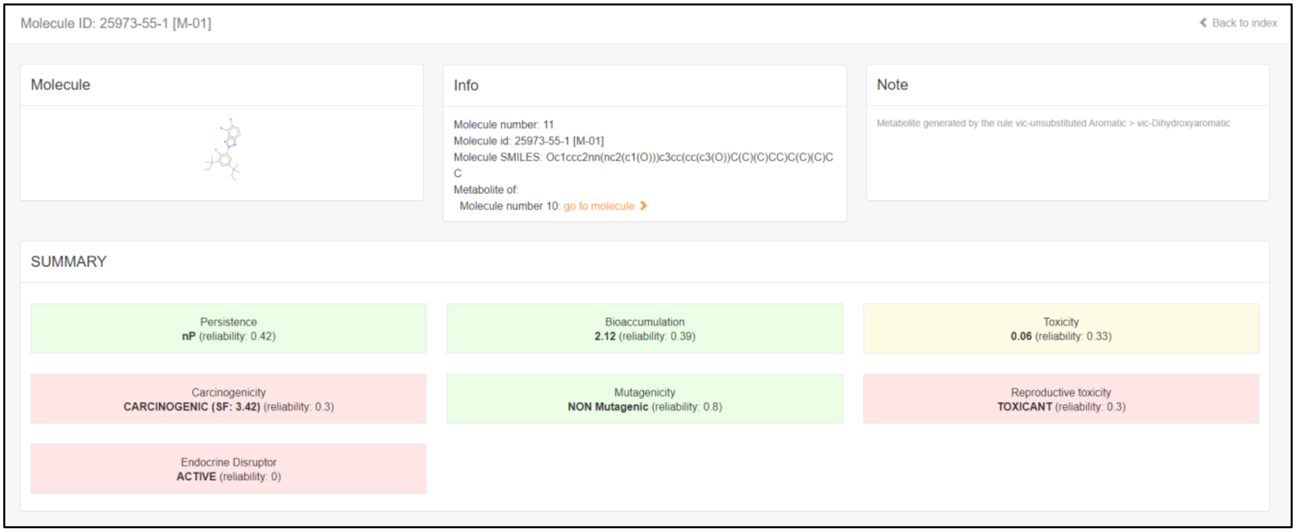
Values retrieved and/or calculated in the workflow:

Property	Value
Mutagenicity classification model (Caesar) prediction	NON-Mutagenic (low reliability)
Mutagenicity classification model (Benigni-Bossa ruleset) prediction	NON-Mutagenic (low reliability)
Mutagenicity classification model (KNN) prediction	NON-Mutagenic (EXPERIMENTAL value)
Mutagenicity classification model (Sarpy) prediction	NON-Mutagenic (low reliability)
Mutagenicity consensus model prediction	NON-Mutagenic (Consensus score: 1)
Mutagenicity additional SMARTS matching	not performed

Figure 71. The details for mutagenicity in the html output page of JANUS for the target chemical.

The target is listed in the SVHC candidate list because PBT and vPvB. It is correctly identified as vP (even if with low reliability), and nM. The other properties are underestimated (B and T) or overestimated (C, R, and ED); all have low reliability. The user can consider this chemical of concern but the prediction have to be verified.

The target has a metabolite (*Figure 72* shows the summary of the predictions) that can be analysed as the parental.



Molecule ID: 25973-55-1 [M-01]

Molecule

Info

Molecule number: 11
Molecule id: 25973-55-1 [M-01]
Molecule SMILES: Oc1ccc2nn(nc2(c1(O)))c3cc(cc(c3(O))C(C)(C)CC)C(C)(C)C
C
Metabolite of:
Molecule number 10: go to molecule ›

Note

Metabolite generated by the rule vic-unsubstituted Aromatic > vic-Dihydroxyaromatic

SUMMARY

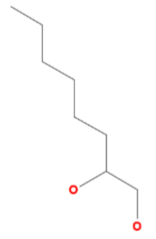
Persistence **nP** (reliability: 0.42)	Bioaccumulation **2.12** (reliability: 0.39)	Toxicity **0.06** (reliability: 0.33)
Carcinogenicity **CARCINOGENIC (SF: 3.42)** (reliability: 0.3)	Mutagenicity **NON Mutagenic** (reliability: 0.8)	Reproductive toxicity **TOXICANT** (reliability: 0.3)
Endocrine Disruptor **ACTIVE** (reliability: 0)		

Figure 72. The first part of the output details in the html output page of JANUS for the metabolite of the target chemical.

Example 3 – single property 2

Another example is the chemical with the CAS No. 1117-86-8. The *Table 15* shows the summary of the results as automatically saved in the txt files (if the option ‘Create the plain text output’ is flagged). It is nP, nB (logBCF < 3.3 l.u.), nT (toxicity > 0.01 mg/L), nC, nM, nR, and nED.

Table 15. Summary results for the target chemical.

Molecule		CAS No.	1117-86-8
		M assessment	NON Mutagenic
P assessment [class]	nP	M reliability	0.8
P assessment [index]	0.15	M score	0.142
P reliability	0.82	R assessment	NON Toxicant
P score	0.097	R reliability	0.8
B assessment [log units]	0.678	R score	0.142
B reliability	0.67	ED assessment [class]	Inactive
B score	0.163	ED assessment [index]	0.1
T assessment [mg/l]	0.7711	ED reliability	0.8
T reliability	0.33	ED score	0.142

T score	0.381	SCORE(vPvB)	0.125
C assessment	NON Carcinogenic	SCORE(SVHC)	0.142
C reliability	0.8	SCORE(PBT)	0.176
C score	0.142	Remarks	-

With the only exceptions of B and T, the other properties are based on reliable values and the partial scores are close to 0. B prediction has a moderate reliability; therefore, the partial score is slightly higher than for the other properties. Anyway, it is quite low, due to the very low potential of bioaccumulation predicted. The highest partial score is for T that has a prediction with low reliability. More in detail (see *Table 16*), the evaluation is based on the lowest chronic value. In this case it is the prediction for fish (i.e. the fish is predicted to be the most sensitive specie). Chronic values are based on one model (one for each specie), that is compared with the acute values and the water solubility. If the acute and/or the water solubility are lower than the chronic value, its reliability is reduced. This is not the case, but the chronic toxicity has low reliability. As already mentioned, the user can run the chemical in VEGA to examine the details of the prediction.

Table 16. Details on the aquatic toxicity predictions.

Fish Acute (LC50) Toxicity model (KNN)	11.95 mg/L (good reliability)	Overall Fish Acute (LC50) Toxicity assessment (used in the workflow)	11.9 mg/L (reliability: 0.79)	Overall Fish assessment (used in the workflow)	NON Toxic, 0.7711 mg/L (reliability: 0.25)
Fish Acute (LC50) Toxicity model (NIC)	1.88 mg/L (moderate reliability)				
Fish Acute (LC50) Toxicity	27.56 mg/L (moderate reliability)				

model (IRFMN)					
Fathead Minnow Acute (LC50) Toxicity model (EPA)	93.57 mg/L (moderate reliability)				
Fathead Minnow Acute (LC50) Toxicity model (KNN)	11.78 mg/L (moderate reliability)				
Guppy Acute (LC50) Toxicity model (KNN)	9.44 mg/L (moderate reliability)				
Fish Chronic (NOEC) Toxicity model (IRFMN)	0.7711 mg/L (low reliability)	Overall Fish Chronic (NOEC) Toxicity assessment (used in the workflow)	0.7711 mg/L (reliability: 0.25)		
Daphnia Magna Acute (LC50) Toxicity model (DEMETRA)	10.3 mg/L (moderate reliability)	Overall Daphnia Magna Acute (LC50) Toxicity assessment	32 mg/L (reliability: 0.63)	Overall Daphnia Magna assessment (used in the workflow)	NON Toxic, 4.54 mg/L (reliability: 0.45)

Daphnia Magna Acute (LC50) Toxicity model (EPA)	1057.88 mg/L (moderate reliability)	(used in the workflow)			
Daphnia Magna Acute (LC50) Toxicity model (IRFMN)	32.05 mg/L (good reliability)				
Daphnia Magna Chronic (NOEC) Toxicity model (IRFMN)	4.54 mg/L (moderate reliability)	Overall Daphnia Magna Chronic (NOEC) Toxicity assessment (used in the workflow)	4.54 mg/L (reliability: 0.45)		
Algae Acute (EC50) Toxicity model (IRFMN)	35.49 mg/L (moderate reliability)	Overall Algae Acute (EC50) Toxicity assessment (used in the workflow)	35.5 mg/L (reliability: 0.45)	Overall Algae assessment (used in the workflow)	NON Toxic, 26 mg/L (reliability: 0.25)
Algae Chronic (NOEC) Toxicity	25.97 mg/L (low reliability)	Overall Algae Chronic (NOEC) Toxicity	26 mg/L (reliability: 0.25)		

model (IRFMN)		assessment (used in the workflow)			
Water Solubility model (IRFMN)	5666.4 mg/L (moderate reliability)				

This chemical is listed in the Safer chemicalingredient list as green circle (i.e. safe chemical).

B.7. The CONCERT GATEWAY

There are several other platforms that you may want to use. The LIFE project CONCERT REACH promoted a single platform with the main European tools freely available for *in silico* models: VEGA, Danish QSAR Database, OCHEM and AMBIT. There is a collection of hundreds of models, organized for the different endpoints suitable for REACH mainly, but most of them are also useful for other regulations. This collection of models is available at the link https://www.life-concertreach.eu/results/.

In the Gateway, at the link above, there are both predictive models and tools for read-across.