Umweltforschungsplan des
Bundesministeriums für Umwelt,
Naturschutz, Bau und Reaktorsicherheit

Forschungskennzahl 3713 63 414

PROMETHEUS – PRioritization Of chemicals: a METHodology Embracing PBT parameters into a Unified Strategy

von

Emilio Benfenati, Claudia Ileana Cappelli, Maria Ifigenia Petoumenou, Fabiola Pizzo, Anna Lombardo, Federica Albanese, Alessandra Roncaglioni
Istituto di Ricerche Farmacologiche Mario Negri, Milan

Alberto Manganaro
Kode, Lodi

Frank Lemke
KnowledgeMiner Software, Berlin

Istituto di Ricerche Farmacologiche "Mario Negri"
Via la Masa 19
20156 Mailand
Italien

Im Auftrag des Umweltbundesamtes

Mai 2015

Kurzbeschreibung

Ziel des PROMETHEUS-Projektes ist es, ein Konzept für ein Programm zur Priorisierung von Substanzen innerhalb des PBT-Assessments zu entwickeln. Grundlage hierfür sind die Nutzung und die Implementierung einer Vielzahl von in-silico Modellen für P, B und T (in Zusammenarbeit mit dem CALEIDOS LIFE Projekt), um in naher Zukunft ein Pilotprogramm für das PBT-Assessment lauffähig zur Verfügung stellen zu können.

Das entwickelte Softwarekonzept beruht auf der Zusammenfassung vielfältiger Modelle, die aus entsprechenden experimentellen Daten ausgewählter Endpunkte algorithmisch erhalten wurden. Das so integrierte System wurde schrittweise einer Validierung unterzogen, indem seine Vorhersagefähigkeit für eine Reihe von Chemikalien überprüft wurde, die entweder von Behörden als PBT eingestuft wurden, oder solche die bezüglich PBT in der Literatur als unbedenklich gelten oder zu denen nur zu wenigen Eigenschaften Informationen vorliegen. Die Ergebnisse der Validierung des integrierten Systems zeigen eine erfolgreiche Identifizierung und Priorisierung von PBT- und vPvB-Verbindungen gegenüber entsprechend unbedenklich geltenden Verbindungen.

Abstract

The aim of the PROMETHEUS project is to create a conceptual scheme suitable for a program for prioritization of substances to be evaluated for PBT-assessment. To do this, it is possible to use and integrate a battery of in silico models for P, B and T (in collaboration with the CALEIDOS-life project), in order to implement a pilot program for PBT-assessment in the near future.

The final software was obtained by the aggregation of models, built on suitable experimental data related to the endpoints chosen. Subsequently the system was subjected to a validation test to verify its performance on a set of chemicals containing molecules labelled as PBT by Authorities and others (non PBT or with information on only a few properties) obtained from the literature. The results of validation demonstrated that the integrated model for PBT (vPvB) prioritization separates successfully PBT from non-PBT compounds.

Table of Contents

Table of Figures

List of Tables

List of abbreviations

ACR	Acute chronic ratio
AD	Applicability domain
ADI	Applicability domain index
B	Bioaccumulation
BCF	Bioconcentration factor
CAS	Chemical abstracts service
CBA	Cost-benefit analysis
ChV	Chronic Value
Des	Desirability index
DT50	Degradation half-life
EC	European Community
EGS	Ecosystem Goods and Services
FDA	Food and drugs administration
FELS	Fish, early-life stage
FHM	Feathead-Minnow
HELCOM	Kommission zum Schutz der Meeresumwelt im Ostseeraum (Convention on the Protection of the Marine Environment of the Baltic Sea Area).
HL	Half life
LC50	Lethal dose 50
Log KOW	Logarithm of octanol-water partition coefficient (Kow)
NOEC	No observed effect concentration
OECD	Organisation for Economic Co-operation and Development
P	Persistence
P	Normalized property score
PBT	Persistence, bioaccumulation, toxicity
QSAR	Quantitative structure-activity relationship
R	Reliability
RB	Ready biodegradability
REACH	Registration, Evaluation, Authorisation and Restriction of Chemicals
RIVM	National Institute for Public Health and the Environment, The Netherlands
SA	Structural alert
SDF	Standard data file
SMILES	Simplified molecular-input line-entry system

T	Toxicity
US-EPA	United States-Environmental Protection Agency
USGS	United States Geological Survey
vPvB	Very persistent, very bioaccumulative

Zusammenfassung

Das Projekt PROMETHEUS hat das Ziel, eine Strategie zur Identifizierung und Priorisierung von für Umwelt und menschliche Gesundheit besonders besorgniserregenden Chemikalien zu erarbeiten. Der Fokus liegt hierbei auf allen Substanzen, die als PBT (persistent, bioakkumulativ oder toxisch) zu identifizieren sind. Dies ist ein komplexes Problem und die Strategie, die PROMETHEUS verfolgt, beginnt mit der Recherche und sorgfältigen Sammlung von Daten. In einer Reihe weiterer, meist zusammenhängender Aktivitäten, wurde nach Wegen gesucht, die Ursachen für die schädlichen Wirkungen zu verstehen und vorherzusagen. Ziel war es das Wissen über die Gründe, die zu schädlichen Effekten führen, zu erweitern und das PBT-Verhalten neuer Substanzen vorherzusagen.

Das Projekt wurde aus technologischer Sicht unterstützt durch die Neuentwicklung von konzeptionellen oder implementierten Modellen, die idealerweise zukünftig in einer Plattform integriert werden können.

Ziel von PROMETHEUS aber war es ein konzeptionelles Schema zur Zusammenfassung verschiedener Evaluierungswerkzeuge innerhalb des PBT-Assessments zu entwerfen, auch im Hinblick auf eine später zu erfolgende Softwareimplementierung.

Die Schwerpunkte des PBT-Assessments liegen auf den drei Eigenschaften Persistenz, Bioakkumulation und Toxizität, die alle auch in Kombination mit anderen "Eigenschaften" oder in Form von spezifischen zugehörigen Endpunkten wie DT50 oder Biokonzentrationsfaktor (BCF) auftreten können. So kann die Eigenschaft der leichten biologischen Abbaubarkeit zur ersten Prüfung auf Persistenz verwendet werden: eine Verbindung, die als leicht biologisch abbaubar gilt, ist mit Sicherheit nicht persistent. Darüber hinaus wird logKow (der Logarithmus des Octanol-Wasser-Verteilungskoeffizienten) in vielen Komponenten der Gesamtstrategie als grundlegende Eigenschaft zur Beurteilung der Bioakkumulation verwendet.

Es liegen keine ausreichenden experimentellen Daten vor, sowohl für bekannte PBT-Chemikalien als auch für PBT-unbedenkliche Chemikalien, um basierend auf diesen Daten ein neues Modell für die "PBT"-Eigenschaft direkt zu entwickeln. Wir waren daher gezwungen separate Herangehensweisen für jede der drei Eigenschaften zu verfolgen und diese Module dann mit Hilfe geeigneter Gewichte in ein gemeinsames System zu integrieren. Die begrenzte Anzahl tatsächlich bekannter PBT- und nicht-PBT-Chemikalien konnte dann dazu verwendet werden, die Korrektheit der gewählten, auf Expertenschätzungen beruhenden Gewichtung zu überprüfen.

Es ist anzumerken, dass aus den Daten zu allen Endpunkten (P, logP, BCF und akute Toxizität bei Fisch) anorganische Verbindungen und Gemische eliminiert und Salze neutralisiert wurden. Zusammenfassend liegen folgende Ergebnisse vor.

Persistenz

Für die Überprüfung der Persistenz wird konzeptionell mit der Prüfung auf leichte biologische Abbaubarkeit begonnen. Dazu werden ein Klassifizierungsmodell und sein erzeugter kontinuierlicher (reeller) Wert verwendet. Ist die Chemikalie biologisch abbaubar, dann ist sie nicht persistent. Im anderen Fall, wenn die Chemikalie nicht biologisch abbaubar ist, werden Modelle für die Persistenz in Wasser, Boden und Sediment genutzt. Ist die Chemikalie persistent in einem der drei Lebensräume, können wir sie als persistent klassifizieren.

Im Einzelnen wurden in einem ersten Schritt Daten zur biologischen Abbaubarkeit für jeden Lebensraum und aus unterschiedlichen Quellen zusammen getragen. Cheng et al. (2012) stellt kontinuierliche Daten hierzu bereit, die wir für unsere Zwecke genutzt haben. Weitere Daten wurden aus dem Datensatz des EU-Projektes ANTARES erhalten (Lombardo et al., 2014). In diesem Projekt wurden die mögliche Nutzung und Leistungsfähigkeit alternativer Testverfahren für REACH verifiziert, sowie geprüft, welche in silico Modelle bessere Ergebnisse liefern, wenn sie für eine große Anzahl von Substanzen getestet werden. Letztlich diente die QSAR Toolbox (Version 3.1) als Quelle, aus der relevante Daten extrahiert wurden. Alle diese Daten

wurden zu einem einzigen Datensatz mit kontinuierlichen experimentellen Werten zur biologischen Abbaubarkeit von 1207 organischen Verbindungen zusammengefasst. Zudem wurden Daten mit Halbwertszeit aus verschiedenen Quellen (Gouin et al., 2004; Gramatica und Papa, 2007) zusammengetragen und so ein Datensatz mit 297 Verbindungen für Persistenz in Sedimenten sowie 298 Verbindungen für Persistenz in Boden und Wasser erhalten. Diese Daten waren in Kategorien angegeben (in semi-dekadischer logarithmischer Skalierung, ausgedrückt in "Stunden"). Für die Persistenz im Boden war eine weitere Datenquelle das *United States Geological Survey* (USGS). Die Verbindungen aus dieser Quelle wurden überprüft und dann zum bestehenden Datensatz hinzugefügt, wodurch ein Datensatz von 537 Verbindungen für die Prüfung der Persistenz im Boden entstand. Für die Persistenz in Wasser und Boden war eine weitere Quelle der Report des *National Institute for Public Health and the Environment* (Linders et al., 1994) mit DT50 Daten zu Pestiziden. Diese Daten wurden ebenfalls geprüft, z.B. bezüglich der Übereinstimmung von Strukturen und CAS-Nummern, kategorisiert und zu den bestehenden Datensätzen hinzugefügt, so dass am Ende Datensätze mit 351 Verbindungen zur Prüfung der Persistenz in Wasser und 568 Verbindungen für Persistenz im Boden zur Verfügung standen. Um diese Daten nutzen zu können, mussten zunächst die Originalwerte an die unter REACH vorgegebenen Kriterien angepasst werden. Diese Anpassung bezog sich nicht auf die Einheiten der Werte, sondern war der Tatsache geschuldet, dass unter REACH andere Grenzwerte gelten als die, die von den kanadischen Autoren in ihren Originaldaten verwendet wurden. Die unterschiedlichen Kategorien und relativen Grenzwerte erschwerten diese Arbeit insofern, dass es hier keine gute Übereinstimmung zwischen der kanadischen und europäischen Kategorisierung gibt.

Basierend auf den beschriebenen Datensätzen entwickelten wir verschiedene Modelle mit Hilfe unterschiedlicher Modellbildungsverfahren. K-NN Modelle (k-Nearest Neighbor) wurden für Persistenz in Wasser, Boden und Sediment entwickelt. Der k-NN Algorithmus ist ein ähnlichkeitsbasiertes Verfahren, der das Ergebnis einer zu untersuchenden Verbindung auf der Basis bekannter experimenteller Werte einer Anzahl ihr am meisten ähnlichen Verbindungen eines gegebenen Trainingsdatensatzes abschätzt. Hierzu wird der Mittelwert der experimentellen Werte der gesuchten Eigenschaft über die erhaltenen k ähnlichen Substanzen ermittelt. Die Anzahl k ähnlicher Substanzen wird durch die Software selbst festgelegt, wobei der Nutzer verschiedene Optionen wählen kann. Bei dem von uns eingesetzten k-NN Verfahren wird im Gegensatz zu anderen Implementierungen ein den jeweiligen Ähnlichkeiten entsprechender gewichteter Mittelwert gebildet.

Die Ergebnisse dieser Modelle werden dann durch andere Modelle, die mit SARpy und istChemFeat erhalten wurden, überprüft. SARpy ist ein Programm, das die chemische Struktur in Fragmente aufteilt und dann nach den Fragmenten sucht, die für die gesuchte Aktivität oder Eigenschaft von Relevanz sind. SARpy wird in Abschnitt 2.2.4 näher beschrieben und wurde bereits für die leichte biologische Abbaubarkeit genutzt. Das Programm istChemFeat wird in Abschnitt 2.2.5 ausführlicher beschrieben. Die Ergebnisse dieser zusätzlichen Tools werden dann zu einem Konsensergebnis kombiniert, um eine erhöhte Aussagefähigkeit über die Zuverlässigkeit der Vorhersage zu bekommen.

Bioakkumulation

Die Prüfung der Aktivität auf Bioakkumulation kann mit den vorhandenen Modellen für den BCF und dem Octanol-Wasser-Verteilungsquotienten ausgedrückt als $\log K_{ow}$ (logP) vorgenommen werden. $\text{Log}K_{ow}$ ist in vielen Modellen enthalten und ist konzeptionell ein wichtiger Faktor für den BCF. Experimentelle Daten zum $\log K_{ow}$ von 2482 Chemikalien, die bereits für REACH registriert wurden, konnten aus der ECHA CHEM Datenbank im Projekt CALEIDOS abgerufen werden. CALEIDOS ist das Nachfolgeprojekt von ANTARES. Während ANTARES Datensammlungen aus der Literatur verwendet hat, um die Resultate von in silico Modelle zu überprüfen, nutzt CALEIDOS Daten über Chemikalien, die im Rahmen von REACH bis 2013 offiziell registriert worden sind. Diese Daten wurden dafür genutzt um zu überprüfen, ob die in Frage stehenden in silico Modelle in der Lage sind, die von den Registranten eingereichten Werte vorherzusagen. Ein weiterer Unterschied zwischen diesen beiden Projekten ist z.B., dass sie nicht genau die gleichen Endpunkte untersuchten. PROMETHEUS selbst hat von der Arbeit, die in den ANTARES und PROMETHEUS

Projekten geleistet wurden insofern profitiert, dass es möglich war, die jeweils bessere Datensammlung und das Wissen über die besten Modelle nutzen zu können.

Nach Vorverarbeitung und begrenzender Auswahl enthält der finale Datensatz Daten zu 729 Molekülen. Weitere experimentelle Daten von 10.005 Verbindungen sind in der VEGA Datenbank verfügbar. Nach entsprechender Prüfung und Konsolidierung dieser Daten, stand ein finaler Datensatz von 9.961 Chemikalien für das PROMETHEUS Projekt zur Anwendung bereit. Eine Vielzahl von Modellen wurde hinsichtlich ihrer logP-Vorhersagekraft geprüft, sowohl individuell als auch in kombinierter Form. Die besten von ihnen wurden dann für das PROMETHEUS Schema verwendet.

851 chemische Verbindungen aus der ANTARES Datenbank bildeten die Grundlage für den Bioakkumulations-Endpunkt, wobei auf die implementierten Modelle der VEGA Plattform zurückgegriffen wurde.

Toxizität

Die Toxizitätskomponente ist komplexer als Persistenz und Bioakkumulation. Sie kann aus mehreren Faktoren bestehen, einschließlich dieser fünf: akute und chronische Ökotoxizität, akute und chronische Toxizität für den Menschen sowie endokrine Störungen. Prinzipiell existieren Modelle für jeden dieser Faktoren, allerdings nicht in gleicher und gleichbleibender Qualität. Wir haben uns daher auf die akute und chronische Ökotoxizität bei Fischen konzentriert. Diese Beschränkung wurde zu Beginn des PROMETHEUS Projektes mit dem UBA besprochen und basiert im Wesentlichen auf dem Umstand, dass die Qualität existierender Modelle für die Toxizität bei wirbellosen Wassertieren und Algen als eher schlecht zu bezeichnen ist sowie deren Verfügbarkeit zudem begrenzt ist. Sobald bessere Modelle zur Verfügung stehen, können in der Zukunft Modelle für Daphnia und Algen hinzugefügt werden.

Experimentelle Daten über die akute Toxizität bei Fischen wurden aus einer Reihe von Quellen abgerufen. Auf einen ersten Satz von LC50 Daten bei 96 Stunden bei Fathead Minnow für 567 Verbindungen konnte aus dem ANTARES Projekt zurückgegriffen werden. Diese Daten stammen aus der frei zugänglichen Datenbank der U.S. EPA. Innerhalb des Projektes CALEIDOS sind experimentelle Daten ausgewählt worden, die aus Registrierungen für REACH stammen, um die Adäquatheit verschiedener QSAR Modelle zu evaluieren. Dieser Datensatz umfasst nach entsprechender Konsolidierung 718 chemische Verbindungen mit insgesamt 1081 Messwerten. Da die Daten teilweise nicht reellwertig sind und daher nicht in nachfolgenden Schritten verwendet werden konnten, ergab sich ein effektiver Datensatz von nur 455 Verbindungen. In der neueren Literatur finden sich größere Datensätze, wie z.B. von Su et al., erschienen im April 2014. Dieser Datensatz enthält Daten, die aus verschieden Studien zu unterschiedlichen Fischarten aus Publikationen, die peer-reviewed wurden, und Onlinedatenbanken zusammengetragen wurden. Nach einer ersten Konsolidierung der Daten enthält dieser Datensatz 953 Verbindungen zu einer oder mehreren Fischarten und ein Mittelwert wurde daraus zu jeder Verbindung ermittelt. Dieser Datensatz bildete die Grundlage für die Entwicklung neuer Modelle mit der k-NN Methode.

Innerhalb von PROMETHEUS nutzten wir folgende (Q)SAR-Modelle zur Überprüfung der akuten Toxizität für Fische: AlogP, MlogP, VEGA KOWWIN, SARpy, istCHEMfeat, ECOSAR, T.E.S.T., VEGA und k-NN. Zusätzlich zur akuten Toxizität wurden mögliche chronische Effekte durch eine Reihe von Regeln evaluiert um zu prüfen, inwieweit chronische Toxizität auftreten kann.

PBT Assessment

Für das PBT assessment werden unter Anwendung mehrkriterieller Entscheidungsmethoden (MCDM) alle Informationen und Modelle, die für jeden der betrachteten Endpunkte erhalten wurden, in ein einziges System integriert. In einem ersten Schritt werden Prüf- und Zuverlässigkeitswerte zu einem einzigen Wert zusammengefasst. Zuvor werden diese jedoch, wie von der MCDM-Methode gefordert, in das Intervall [0, 1] transformiert, wobei der Wert 1 die maximale Ausprägung einer Eigenschaft bedeutet. Als optimalen

Schwellwert zur Separierung von besorgniserregenden Stoffen (PBT-Verbindungen) von weniger bedenklichen Stoffen (keine PBT-Verbindungen) wird der Wert 0,5 zu Grunde gelegt.

Für die Persistenz erfolgt diese Transformation (Normierung) wie folgt. Das Ergebnis der Überprüfung der Persistenz einer chemischen Substanz ist eine von vier Kategorien: nP, nP/P, P/vP oder vP. Jeder dieser Kategorien wird ein definierter numerischer Wert zugeordnet, wobei für die binäre Zuordnung PBT vs. nicht PBT der Wert 0,5 gewählt wurde, der somit auch unbekannten Ergebnissen zugewiesen wird (Tabelle Z1). Unbekannte Ergebnisse können entstehen, wenn kein hinreichend guter Vorhersagewert berechnet werden kann.

Tabelle Z1: Zuordnung von numerischen Werten zu Kategorien der Persistenz.

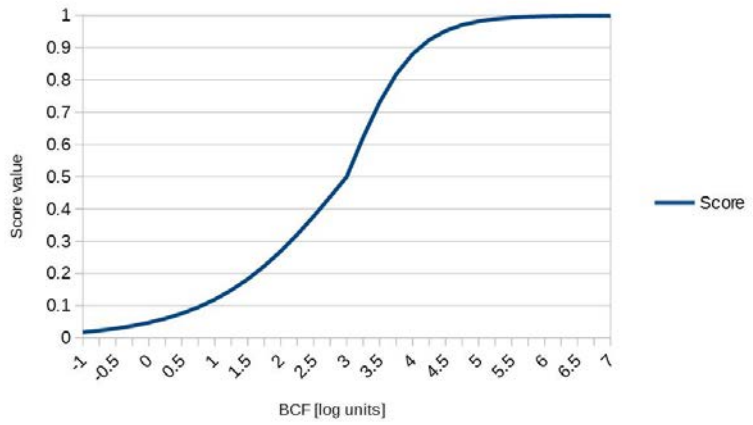
Kategorie	Wert
nP	0,3
nP/P	0,6
P/vP	0,8
vP	1,0
unbekannt	0,5

Die Normierung des Bioakkumulationswertes basiert auf dem logarithmierten Vorhersagewert für den BCF der untersuchten chemischen Substanz. Hier sind zwei Werte von speziellem Interesse: 3,3 als Grenzwert für B und 3,7 als Grenzwert für vB. Die vorhergesagten Bioakkumulationswerte werden mit Hilfe folgender logistischen Funktion in einen Score transformiert: $Normalized\ BCF\ value = \left\{ \begin{matrix} \frac{1}{1+e^{-(BCF-3)}}\ for\ BCF < 3.0\ log\ units \\ \frac{1}{1+e^{-2(BCF-3)}}\ for\ BCF \geq 3.0\ log\ units \end{matrix} \right\}$

$$Normierter\_BCF\_Wert = \left\{ \begin{matrix} \frac{1}{1+e^{-(BCF-3)}} BCF < 3.0 \log Einheiten \\ \frac{1}{1+e^{-2(BCF-3)}} BCF >= 3.0 \log Einheiten \end{matrix} \right\}$$

Für einen BCF-Wert von 3,0 ergibt sich somit ein Score von 0,5 als Schwellwert zwischen B und nB. Ebenso wird deutlich, dass für Werte > 3,0 die Transformationskurve einen steileren Anstieg besitzt mit dem Ziel, eine klarere Differenzierung zwischen B (3,3) und vB (3,7) Verbindungen zu erhalten. Abbildung Z1 zeigt die Transformationskurve für logarithmierte BCF-Werte im Bereich -1 bis 7.

Abbildung Z1: Transformationskurve für vorhergesagte BCF-Werte in einen normierten Score.



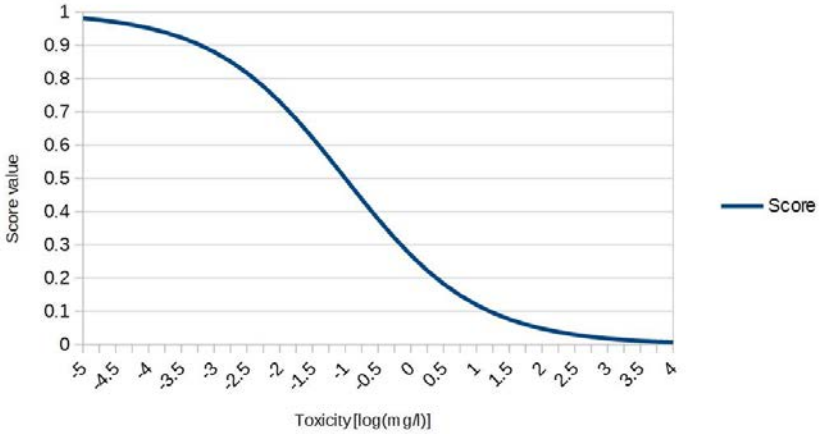
Der Ausgabewert des Moduls für Toxizität besteht aus der berechneten Toxizität für Fische in mg/l. Hierbei handelt es sich im Allgemeinen um akute Toxizität. Zur Berücksichtigung auch chronischer Toxizität wird eine Option im entsprechenden Modul zur Verfügung gestellt.

Für die Toxizität beträgt der Grenzwert von Interesse 0,01 mg/l, unterhalb dessen chemische Verbindungen als toxisch im Rahmen der PBT-Bewertung betrachtet wird. Die berechneten Toxizitätswerte werden mit Hilfe folgender logistischen Funktion in einen normierten Score transformiert: $Normalized\ T\ value = 1 - \frac{1}{1+e^{-(log(TOX)+1)}}$

$$Normierter\_T\_Wert = 1 - \frac{1}{1+e^{-(\log(TOX)+1)}}$$

Hier ergibt sich für einen logarithmierten Wert von -1 (0,1 mg/l) der neutrale Score von 0,5. Dieser eher konservativ gewählte Wert im Vergleich zum tatsächlichen Grenzwert von -2 (0,01 mg/l) soll die vorhandene Unsicherheit von experimentellen, berechneten und Grenzwerten abbilden helfen. Es ist bekannt, dass chemische Substanzen bereits mit Toxizitäten um oder unter 0,1 mg/l toxische Effekte zeigen, obwohl sie über dem regulatorischen Grenzwert von 0,01 mg/l liegen.

Abbildung Z2: Transformationskurve für vorhergesagte Toxizitätswerte in einen normierten Score.



Ähnlich wie bei der Normierung der Bioakkumulation zeigt die logistische Funktion eine gute Differenzierung um den Grenzwert herum, während deutlich höhere (um 0,001 mg/l) bzw. niedrigere (um 10 mg/l) Toxizitätswerte weniger stark differenziert werden.

Abbildung Z2 zeigt die Transformationskurve für Toxizität im Bereich -5 (0,00001 mg/l) bis 4 (10.000 mg/l).

Nach diesem ersten Normierungsschritt liegen drei einzelne Scores für P, B und T einschließlich ihrer Zuverlässigkeitswerte ebenfalls im Intervall [0, 1] vor, die wiederum mittels MCDM kombiniert werden. Der sich ergebende Index (desirability index DES) stellt dann den endgültigen PBT-Wert der bewerteten chemischen Verbindung dar.$DES_{PBT} = P^{0.4} \cdot B^{0.4} \cdot T^{0.2}$

$$DES_{PBT} = P^{0.4} B^{0.4} T^{0.2}$$

Hohe Werte (nahe 1) repräsentieren mit hoher Wahrscheinlichkeit (Zuverlässigkeit) besonders besorgniserregende PBT-Verbindungen, während niedrige Werte (nahe 0) mit hoher Wahrscheinlichkeit weniger bedenkliche PBT-Verbindungen indizieren. Insbesondere für das vPvB Screening wird folgender reduzierte Index zur Verfügung gestellt:$DES_{vPvB} = P^{0.5} \cdot B^{0.5}$

$$DES_{vPvB} = P^{0.5} B^{0.5}$$

Diese Art der Zuordnung stellt sicher, dass hohe PBT bzw. vPvB Scores die am meisten besorgniserregenden Verbindungen anzeigen, sowohl bezüglich des Grades ihrer schädigenden Wirkungen als auch bezüglich der Zuverlässigkeit ihrer Zuordnung. Das entwickelte Konzept ist hinsichtlich dieser beiden Kriterien optimiert worden.

Nach der Softwareimplementierung des Konzeptes wurden Validierungstests durchgeführt um die Vorhersagekraft des Ansatzes bekannte PBT-Verbindungen von Nicht-PBT-Verbindungen zu trennen, zu überprüfen. Zu diesem Zweck wurde ein Datensatz von in der Literatur oder von Behörden klassifizierten PBT-Verbindungen zusammengestellt (Tabelle Z2).

Tabelle Z2: Datenquellen zur Akkumulierung des Validierungsdatensatzes.

#	Endpunkt	Quelle
27	PBT	Candidate List; UBA list
869	logBCF	Read-across dataset (VEGA)
351 297 568	P_{water} $P_{sediment}$ P_{soil}	Gouin et al., 2004 and Gramatica and Papa, 2007; RIVM (Linders et al., 1994); and USGS (Prioritizing Pesticide Compounds for Analytical Methods Development, 2012)
91	Fish Chronic Tox	OECD QSAR toolbox (ECOTOX database)

Unsere Prüfung des entwickelten Konzeptes hat demonstriert, dass das implementierte integrierte Modell zur PBT-Priorisierung erfolgreich PBT- von Nicht-PBT-Verbindungen unterscheiden und klassifizieren kann. Die Möglichkeit zur Prüfung auf vPvB-Eigenschaften besteht.

Summary

The project PROMETHEUS is aimed to elaborate a strategy to identify and prioritize chemicals, which may be of concern for the environment and the human health using a screening approach. The output of the project is a program, which lists chemicals, and in top of the list there are those assessed as most critical. The focus is on all chemicals that can be identified as PBT (persistent, bioaccumulative and toxic). The influence on the specific reason for concern, as P, B, or T, is reported by the program. The matter is complex and the strategy planned within PROMETHEUS started from an effort to gather suitable data, and then continued into a series of activities, parallel in some cases, or closely related most commonly, addressing ways to understand and predict the reasons for hazardous effects.

This project was supported from a technological point of view by the development of models (both conceptual and implemented ones), which are integrated into a single innovative program.

The aim of PROMETHEUS was to obtain a conceptual scheme for integrating different evaluation tools within a PBT framework, with a reference to software tools, which may be implemented in an integrated platform within a future initiative.

The key targets of the PBT assessment are the three properties: persistence (P), bioaccumulation (B), and toxicity (T). Each of them may be composed of other "properties" or endpoints, such as DT50 and bioconcentration factor (BCF). For instance, the ready biodegradability property may be used as first assessment for persistence: a compound that is considered ready biodegradable is for sure non-persistent. Furthermore the $\log K_{ow}$ (the logarithm of the partition coefficient between octanol and water) is a property, which is used in many components of the overall strategy, to drive the assessment on bioaccumulation.

We do not have sufficient experimental values on a set of known PBT-chemicals and a similar set of other chemicals, which are non-PBT, to be used to assess "PBT" as a property itself using these experimental data to build up a new model for the "PBT"-feature. Moreover, there are no individual in silico models, which are error-free in the predictions of the properties of interest. (Here and below we refer to model as an in silico model, i.e. a computer program which predicts the property based on the chemical structure of the substance). Several studies showed that depending on the chemical of interest one predictive model may be appropriate, but changing the target chemical a different model may be more appropriate. This is well-known issue, related to the so-called applicability domain of the model. Based on the fact that multiple models may exist for the same endpoint, there have been many studies showing that combining them improves the results. Also regulators in several cases require application of more than one model, when available. For this reason we tested our specific case the results obtained combining the results from multiple models addressing a single endpoint, such as P, B and T. Thus, our philosophy has been to rely on separate approaches, which address the three properties separately, and then integrate the results of the models using proper weights.

This overall philosophy, producing the program, which combines the different models, should be checked based on chemicals with experimental values. There is no "experimental value" for PBT, but there are individual values for the individual endpoints, such as P, B, and T, and actually the reality is even more complicated, because the evaluation of the persistence, for instance, is based on a series of individual assessment for the behaviour in water, soil and sediment. Thus, we have gathered a number of experimental values of chemicals which have been labelled as PBT, or which cannot be PBT. We clarify that if a substance is not at the same time P, and B, and T, it is not defined PBT. The limited number of officially identified PBT-chemicals has been used to check the correctness of the results of our program, and a similar number of chemicals non-PBT has been used to check for negative results.

Below we summarize the results. Please note that for data about all the endpoints (P, logP, BCF and fish acute toxicity) inorganic compounds and mixtures were eliminated and salts were neutralised.

Persistence (P) assessment

For the P-assessment the conceptual scheme starts with the ready biodegradability (RB) assessment. The scheme uses the classifier model, and the continuous value, for further evaluation. If the chemical is ready biodegradable, the chemical is not B. In the other case, when the chemical is not ready biodegradable, the scheme uses the models for the persistence in the different compartments, water, soil and sediment. If the chemical is persistent in one of the compartments, we classify the chemical as persistent.

In a first step for each of these compartments and for RB, the first information was collected in the following way: Data on RB were collected considering several sources. Cheng et al. (2012) provides continuous data on RB that we used for our purposes. More data were obtained from the dataset used in the ANTARES EC Project and already described in Lombardo et al. (2014). This project aimed to verify the possible use and performance of the non-testing methods for REACH legislation and to check which in silico models provide better results, when tested on large collections of substances. Finally the QSAR Toolbox (version 3.1) was used to extract further data. All these datasets were merged and we finally obtained a dataset containing 1207 organic compounds with continuous experimental data on RB.

For persistence data with half-life (HL) were collected from different sources (Gouin et al., 2004; Gramatica and Papa, 2007) to obtain a dataset with 297 compounds for sediment, and 298 for soil and water. These data were only available in categories (in a semi-decade log scale basis, expressed as "hours"). For soil only, another source was available from the United States Geological Survey (USGS). These compounds were checked and then added to the soil dataset, obtaining a dataset of 537 compounds. For water and soil another source of data for pesticides was the National Institute for Public Health and the Environment (RIVM) Report (Linders et al., 1994). Also these compounds were checked (e.g. correspondence between structures and CAS numbers), categorized as explained before, and added to water and soil datasets obtaining a final datasets of 351 data for water and 568 for soil. In order to use these data we had to adapt the original values to those used under REACH. This does not refer to the units, but to the fact that under REACH the threshold criteria are different from those used by the Canadian authors in their original data set. The different categories and relative thresholds complicated the work, because there is no good correspondence between the sets of categories used by the Canadian authors and the REACH criteria.

Using these data we developed several models. Different methods have been combined. A k-nearest neighbour (K-NN) model has been developed for persistence, for the three compartments (in total three models). The kNN algorithm is a similarity-based approach and can estimate the outcome of a target compound on the basis of the known experimental values of its most similar chemicals present within the training set of the model. In practice, this models searches for the most similar compounds and assigns the property value of the target compounds based on the average of the property values of the k most similar chemicals. The number k is found by the software and the developer checks different possibilities. In the particular kNN model we applied, the software is more sophisticated than other kNN models. Indeed, the property values keeps into account how similar are the most similar compounds, and provide weights based on this.

Further check of the results is done using models based on SARpy and istChemFeat. SARpy is a program which cuts the chemical structure into fragments, and then searches for fragments which are related to the activity or property of interest. SARpy is described in more details in Section 2.2.4. SARpy has been already used for predicting RB. istChemFeat is described in more details in Section 2.2.5. These additional tools are combined within a consensus approach, to increase or decrease the certainty of the result.

Bioaccumulation (B) assessment

The B-assessment is done with BCF- and $\log K_{ow}$ models. $\text{Log}K_{ow}$ (the logarithm of the partition coefficient between octanol and water) is present in many models and also conceptually is an important factor for BCF. Experimental data on $\log K_{ow}$ of 2482 chemicals already registered for REACH have been retrieved from ECHA CHEM database (http://www.echemportal.org/echemportal/participant/participantinfo.action?participantID=140) within the CALEIDOS project. CALEIDOS is the project, which continued ANTARES. While ANTARES used

collections of data from the literature to check the results of the in silico models, CALEIDOS started when the REACH regulation was in force, and thus could use the data on chemicals registered for REACH until 2013. These data were used to check if the in silico models were able to predict the values submitted by registrants. There are other differences between these two projects, and for instance they did not addressed exactly the same endpoints. PROMETHEUS took advantage of the work done by ANTARES and CALEIDOS, using the collection of data from one or the other project, and the knowledge on the best models to use.

Experimental data (10,005 compounds) on $\log K_{ow}$ are available within the VEGA database and they have been used also for PROMETHEUS project. VEGA is platform (explained in Section 2.2) with in silico tools, and it also includes, as part of the platform, a database with data on many thousands of substances, for many endpoints. The set has been processed and cleared from compounds that were replicated or that had problems with the provided molecule structure. This final dataset has 9,961 chemicals. Many models have been checked for their performance for $\log K_{ow}$ prediction, used alone or combined, also using KnowledgeMiner Insights. The best ones have been used within the PROMETHEUS scheme.

851 compounds from ANTARES database were used for the bioconcentration factor (BCF). In this case the chosen models are those based on VEGA. The final dataset with BCF values from CALEIDOS, after processing and pruning, contains data on 729 molecules.

Toxicity (T) assessment

The T-assessment is more complex than those of the other two properties as it toxicity includes different endpoints, including these five: acute and chronic ecotoxicity, acute and chronic human toxicity, and endocrine disruption, which does not fall under the classical chronic endpoints, and is addressed both for human and environmental endpoints. In principle, for each of them there exist models, but these are not of homogeneous quality. In this project we limited the evaluation on acute and chronic ecotoxicity focused on fish. The focus on fish has been discussed at the beginning of the PROMETHEUS project with UBA, and it is mainly based on the fact that the quality of the existing models on aquatic invertebrates and algae is quite poor, and the availability of the models also limited. In the future, when better models will be developed, models for daphnia and algae can be added.

Experimental data on fish acute toxicity were retrieved from different sources. A first set of 567 compounds with LC50 data on Fathead Minnow at 96 hours was retrieved from the work done within the project ANTARES. Data were obtained from the database freely available on U.S. EPA website. Within another project *(*CALEIDOS*)* we selected experimental data coming from registrants to evaluate the adequacy of different QSAR models. The dataset, after a pruning activity, was composed by 718 compounds with a total of 1081 data points. Not all the values were continuous ones; in some cases the value was referred as greater as or lower than a value and could not be used for the next steps. In view of this, the dataset used in this work is composed of only 455 compounds. In the literature there is also a new and larger set of compounds available published by Su et al. (2014). It contains data coming from different studies on multiple fish species, taken from peer-reviewed publications and online databases. After pruning, the dataset consists of 953 compounds with data on one or more species and the average value for each compound was calculated. This set was used to develop new models.

Within PROMETHEUS we used the following models to assess fish acute toxicity: VEGA KOWWIN, SARpy, istChemFeat, ECOSAR, T.E.S.T., VEGA and k-NN. In addition to acute toxicity, the possible chronic effects have been evaluated, through a series of rules, to assess if excess chronic toxicity may appear.

PBT assessment

In this work we decided to use multiple criteria decision making (MCDM) methods to integrate in a single system all information obtained from the different models used for each endpoint. The MCDM approach provides methods where a set of input variables are transformed and aggregated. It will be more thoroughly

be discussed in Section 3.1. We added a first step, in which assessment and reliability values for each endpoint are aggregated into a unique value. Before combining the assessment and reliability values, they have to be transformed - as required by MCDM - to obtain a score ranging between 0 and 1, where 1 represents the optimum. For each endpoint we decided to keep the transformed value of 0.5 as an ideal threshold to separate assessment values going in the direction of major concern (PBT compounds) and of less concern (non PBT).

After these steps, the result of the overall evaluation consists in three unique scores for P, B and T, which have to be combined using a MCDM technique. This index represents the final score used to rank compounds, so that high (towards 1) values are related to compounds with a PBT prediction, performed with good reliability, and they should be the compounds of major concern. On the other side, low (towards 0) values are related to compounds with a non-PBT prediction, performed with good reliability. The ranking ensures that the first compounds will be those with more concern PBT prediction based on the high “PBT” score joined with the high reliability of the results. The scheme was optimized with this goal.

After the development of the software, we carried out a validation test to check its ability to differentiate compounds labelled as PBT compounds from those are non-PBT.

For this purpose we built up a set of chemicals with molecules labelled as PBT and non-PBT taken from the literature and assessed by regulatory authorities. The results of validation demonstrated that the integrated model for PBT prioritization separates successfully PBT vs non-PBT compounds. A similar model can evaluate vPvB-substances.

1 INTRODUCTION

The project aimed at developing a new strategy to prioritize those molecules that can be a threat to the environment. The purpose is to identify chemicals of higher concern using a fast program which can screen a large number of substances; successively, the substances of concern may be manually assessed to verify their adverse effects. A related objective is to increase knowledge about the reasons that lead to damaging effects and, thanks to these, make predictions about new substances. The reasons for the adverse effects are provided in the form of fragments, associated to the effect. These fragments have been obtained through software like SARpy, and can be used to extend the knowledge about the causes of the adverse effect, since these fragments can be read and understood quite easily by human experts, while this is not always the case when we evaluate the molecular descriptors used in the in silico models.

Indeed, SARpy works using the experimental data, and not prior information on the mechanism. Additional information may be acquired by both study of phenomena and mechanisms of action, using also computational methods, to be incorporated in a more general scheme of modelling. The aim is to use them to develop a general program for prioritization composed by multiple modules.

Particular attention is given to those substances that can be identified as PBT (persistent, bioaccumulative and toxic). Each of the three properties persistence, bioaccumulation and toxicity, can be modelled using a series of other parameters.

To cope with this issue, an integrated strategy is needed, addressing the correct endpoint in a suitable way, and combining different perspectives.

To reach the goal of a developing a robust program integrating many models and the associated uncertainty, the overall objective of this project is to use a series of existing and new tools for the evaluation of P, B, and T in a combined assessment framework.

The framework will provide an innovative weight of evidence architecture for each parameter: P, B, T, vP, and vB. Finally, all these tools, generating data and support information, should be integrated into a unified framework, which merges the values associated to different properties into a single index, suitable to prioritize the chemicals of concern for regulatory activities. This framework includes a series of workflows for each P, B and T property. These workflows have been implemented into programs, which use as inputs the experimental values, or the values predicted by a series of models. Finally, the results of the programs implementing each workflow are used as inputs by another program, which integrated the results into the final score. We notice that in this framework we always want to have a value for the PBT assignment, even when the experimental value of the P, B or T properties are missing, which may be quite frequent. For this purpose, in many cases we have to use models, existing or developed within this project. However, these models should have an acceptable reliability. A critical issue for this is the low availability of reliable models for terrestrial toxicity (the one that exist are quite poor, and thus not used in the present project) and for aquatic toxicity endpoints. The availability of models of good quality for algae and daphnia is also very limited. This reduces the overall assessment of the aquatic toxicity to that on fish, with strong limitations also in this case in particular for chronic toxicity. On the basis of these facts, the overall reliability of the T-component in the framework is lower than the reliability of the other two components, P and B, and thus, as we will see, the relevance assigned to T into the combined assessment is lower.

The complexity of the problem requires a combination of the results of different workflows, one for each of the three PBT properties. In addition, we keep into account the uncertainty associated to each value, such as the value for P, B, or T. The overall picture surely is complex, but we underline that the solution we adopted makes it transparent the identification of the information related to each property. In this way we can also imagine an evolution of the program we developed that allows the user to visualize results according to parameters defined by the user, for instance only with an uncertainty that is lower than a certain threshold. In any case, even now, the program allows already the visualization of the results for each individual PBT property, and the uncertainty associated to each value. Indeed, this overall evaluation should be transparent

enough to allow a clear understanding of the cause of concern. We will show in the results how this can be used for substances with different property values and different levels of uncertainty.

The PROMETHEUS project introduces improvements compared to previous approaches to integrate values of heterogenous nature, and is based on the original experience of the partner of the consortium. The key computational strategy to integrate data of heterogeneous nature at the basis of the present software refers to studies of the coordinating partner published in the first version in 2010 (Boriani et al., 2010), and from even earlier studies for the algorithm which takes into account the uncertainty of the results (Porcelli et al., 2008). The other partners also provided their original contribution, referring to tools to integrated results from multiple models within a hybrid strategy (Amoury et al., 2007), for instance relative to the collaboration between KnowledgeMiner and the coordinating partner. Similarly, Kode exploited its experience on multiple criteria decision making (MCDM) approach, mainly related to the development of the DART software (Manganaro et al., 2008) which has been commissioned by the European Joint Research Centre (JRC) for the specific goal of applying such techinques for environmental safety assessment. A few approaches for PBT (or only PB) have been presented in the past, for instance from US EPA (http://www.pbtprofiler.net/), from Canada with the RAIDAR model (CEMC Report No. 200703, 2007), from the JRC (based on the usage of the above mentioned DART software) (Pavan and Worth, 2008) and later from the Dutch authorities (RIVM, 2011). The main differences with past approaches refer to the introduction of a robust procedure to deal with uncertainty of the predicted values, the use of experimental values for PBT as starting point, the use of a very large battery of multiple models for the same property, and the more advanced evaluation of chronic fish toxicity.

This project is supported from a technological point of view by the development of models (both conceptual and implemented ones), which may be ideally integrated into a single platform. This single platform is not part of PROMETHEUS, but a perspective for further development of PROMETHEUS in a later project. In other words, the general framework developed within PROMETHEUS is a feasibility study, producing already a program for internal use, and not as a user-friendly program. Its implementation for the general use would need further activities. Thus, PROMETHEUS codified scientific concepts, associated to environmental and toxicological properties, into a set of software models, which may be implemented in a single platform in a possible future project.

The overall task of the project was divided into a sequence of subtasks where different models were used or new models developed. This was in particular necessary when the existing models proved to be not robust enough in order to get a reliable assessment of the chemicals.

In the following we describe in details the work done.

Section 2 reports the sources of the data, but also how the data have been processed. In the same section we also describe the software used. Some models existed already, specific for a certain endpoint. We also employed programs to develop new models or define categories to be used as intermediate steps in the modelling process.

In Section 3 we present the workflows that have been developed for each property we considered. Besides P, B and T we also addressed $\log K_{ow}$ because this is a useful parameter for successive models. Within each workflow several models are used, of different nature, to be applied for all chemicals, or for specific cases.

Finally, in Section 4 we describe how the results of the three workflows have been integrated into a single score value, used to list chemicals, and thus to screen substances. This section also shows the results obtained when we applied the overall program to the substances with known PBT label, both positive and negative, obtained from the literature and Authorities.

2 SOURCE OF THE DATA AND USED MODELS

The following sections describe the sources of data of all endpoints chosen and a general description of each model used.

2.1 Data

2.1.1 The data on persistence

For persistence the suitable and available data refer to ready biodegradability (RB), which can be used only as an initial step in the PBT assessment, and data on persistence for water, soil and sediment compartments. Indeed, the information about RB is immediately applicable if the substance is readily biodegradable, and thus not persistent (nP).

Data on RB were collected considering several sources. Cheng et al. (2012) provides continuous data on ready biodegradability that we used for our purposes. More data were obtained from the dataset used in ANTARES EC Project. Finally the QSAR Toolbox (version 3.1) was used to extract further data. Many tests are described in the OECD guideline 301 for the assessment of RB of substances. In order to have more homogeneous data and obtain more reliable model, we considered only studies that followed OECD Guideline 301 C (MITI test) with a duration test of 28 days, because data from obtained with this protocol were more numerous. If a study reported a duration test shorter than 28 days, the value was retained only if BOD was greater than 60%, otherwise it was deleted from the dataset.

Inorganic compounds and mixtures were eliminated and salts were neutralised. Moreover, compounds with non-concordant experimental data between different sources were deleted. For each compound the correspondence between CAS numbers and chemical structures was double-checked using ChemID plus (http://chem.sis.nlm.nih.gov/chemidplus/) and Pubchem compound (https://pubchem.ncbi.nlm.nih.gov/). All the datasets considered were merged and we finally obtained a dataset containing 1.207 organic compounds with experimental data on ready biodegradability given as numerical, continuous values, thus not as a category (such as positive or negative).

After an initial screening using ready biodegradability (Annex XIII, REACH Regulation), for persistence we used data (as half-life (HL)) from these sources. We started from Gouin et al., 2004, which contains information on HL, in hours, for 233 organic compounds categorized in 9 classes (in a semi-decade log scale basis). It covers four environmental compartments: water (not specified if marine or freshwater), sediment (not specified if marine or freshwater), soil and air. To each class a mean value of half-life has been assigned by authors, that is the only value available for each chemical.

Another source was the paper by Gramatica and Papa 2007 that contains 250 organic compounds with data for the same environments and categorized in the same classes as Gouin et al., 2004. Since no thresholds for air are set for PBT assessment, we considered only water, soil and sediment. In particular, for water and sediment we considered all data as for freshwater environment. All the compounds, in both sources, were double-checked with ChemID plus (http://chem.sis.nlm.nih.gov/chemidplus/) and Pubchem (https://pubchem.ncbi.nlm.nih.gov/). Salts, mixtures, doubtful compounds and duplicates were eliminated. Also compounds present in both datasets but with different values were eliminated. In this way, a dataset with 297 compounds for sediment and 298 substances for soil and water was obtained. For soil only, another source was available from USGS (Prioritizing Pesticide Compounds for Analytical Methods Development, 2012). It contains 318 pesticide compounds with HL for soil. These compounds were checked as explained above: the continuous values were categorized following the same criteria reported in Gouin et al., 2004 and then added to the soil dataset, obtaining a dataset of 537 compounds. For water and soil another source of DT50 data for pesticides was a RIVM Report (Linders et al., 1994). Also these compounds were checked, categorized and added to water and soil datasets, as explained before, obtaining a final datasets of 351 data for water and 568 for soil.

Each source of data used has problems: RIVM and USGS contain only pesticides, RIVM data are quite old, Gouin et al., 2004 and Gramatica and Papa, 2007, contains only categorized data. With categories only models for classification can be obtained, even though the number of categories is quite numerous. Moreover with these categories it is impossible to discriminate P compounds (see table 1 and 2).

Thus, there is no univocal, ideal source of data filling the conditions to have continuous values (with substances different from pesticides), or values, which are split according to the categories adopted under REACH.

For this reason we adapted the data from Gouin et al., 2004 and Gramatica and Papa, 2007 into the classes suitable for the EU regulation: vP, P, nP. Table 1 shows how the original classes have been transformed into the REACH vP, P, nP classes. We can see that some classes can be translated into the REACH classes, but other classes in the Table 1 appear mixed, because they include substances, which may refer to more than one class. The classes according to the REACH regulation are presented in Table 2.

Table 1: Half-life classes on the basis of the available data

Class	Class range (h)	Value assigned (h)	Persistence classes for soil	Persistence classes for freshwater	Persistence classes for sediment (fresh water)
1	0 - 10	5	nP	nP	nP
2	10 - 30	17			
3	30 - 100	55			
4	100 - 300	170			
5	300 - 1000	550		nP/P	
6	1000 - 3000	1700	nP/P	P/vP	nP/P
7	3000 - 10000	5500	P/vP	vP	P/vP
8	10000 - 30000	17000	vP		vP
9	30000 - 100000	55000			
10	100000 - 300000	170000			
11	300000 - 1000000	550000			

As a result, some of the so obtained classes contain substances “pure”, relative to a single REACH class, but unfortunately, there are some of the classes in Table 1 which cannot be univocally assigned to a single class, because we do not have continuous value, but ranges as defined by the authors of the studies. The number of original classes, as split by the authors of the original studies, was in some cases even too detailed for the purpose of the REACH classification and thus were merged. Table 1 presents the way used to “translate” the original classes into the REACH classes.

Table 2: P and vP thresholds considered, both in days and in hours

Environment	P (DT50, d)	vP (DT50, d)	P (DT50, h)	vP (DT50, h)
Marine water	60	60	1440	1440
Estuarine or fresh water	40	60	960	1440
Marine sediment	180	180	4320	4320
Estuarine or fresh water sediment	120	180	2880	4320
Soil	120	180	2880	4320

We added two categories (10 and 11, see Table 1) since we had a number of data points from the pesticides results (data from RIVM and USGS) with values in a range longer than the values from those in Gouin et al., 2004 and Gramatica and Papa, 2007. In this ways, we classify the values in four classes: nP compounds (i.e. compounds with values below the P threshold), nP/P compounds (i.e. compounds which cannot be sorted into a single REACH class on the basis of the original range used by the Canadian authors, range which includes the P threshold of REACH), P/vP compounds (i.e. compounds in the class range in which are included both P and vP compounds because includes the vP threshold) and vP compounds (i.e. compounds above the vP threshold). We recognize that this situation is not the ideal one, but the splitting of the original data has not been done by us, and there is no information about the continuous values.

Another critical point for model building, for soil and water datasets, is that they are unbalanced, with a prevalence of nP compounds. The percentage of compounds included in the four classes (nP, nP/P, P/vP, vP) for each compartment is shown in table 3.

Table 3: Data distribution in the four classes for the three compartments

	sediment	water	soil
nP	25.2%	50.1%	54.4%
nP/P	23.2%	22.2%	22.7%
P/vP	20.8%	13.9%	11.6%
vP	30.6%	13.6%	11.2%

2.1.2. The data on logK_{ow}

Experimental data on logK_{ow} of 2482 chemicals already registered for REACH have been retrieved from ECHA CHEM database within the CALEIDOS project (thus, formally this is not part of PROMETHEUS).

The dataset has been pruned, considering organic monoconstituents (excluding inorganic chemicals, mixtures and UVCBs), studies with reliability 1 and 2 according to the Klimisch score (Klimisch et al., 1997), the test guidelines recommended by the endpoint specific guidance for the implementation of REACH legislation (EC, 2012), the temperature of the experimental test, pH (to be sure that the experimental values of the substances are referred to the non-ionized form) and purity (high purity: > 80%). Salts have been excluded from the analysis.

The correspondence among SMILES, structures and CAS numbers has been checked and the chemicals without CAS number nor/or SMILES have been identified.

The final dataset of CALEIDOS contains data on 729 chemicals. Few of them have more than one experimental value. In case of multiple experimental data for the same substance, we have calculated the arithmetic mean.

Further experimental data (10,005 compounds) are available within VEGA databases and they have been used also for PROMETHEUS project. This dataset has been built merging the data available in the OECD QSAR Toolbox and the ones used as training and test set of KOWWIN v1.68 (in EPIsuite™). The set has been processed and cleared from compounds that were replicated or that had problems with the provided molecule structure. The correspondence between CAS number, name and structure were automatically checked using several softwares and websites: KOWWIN v1.68, OECD QSAR Toolbox, ChemSpider and GChem. In case of multiple values, we have calculated their arithmetic mean. Starting from this dataset of 10005 compounds, a new dataset were obtained excluding also salts and mixtures. This final dataset has 9,961 chemicals.

2.1.3. The data on bioconcentration factor (BCF)

851 compounds with experimental values from ANTARES database were used for this endpoint. These experimental data were collected from 3 different sources: Arnot and Gobas, 2006, Dimitrov et al., 2005 and CEFIC.

2.1.4. The data on fish acute toxicity

Experimental data on fish acute toxicity were retrieved from different sources.

A first set of 567 compounds with LC50 data on Fathead Minnow at 96 hours was retrieved from the work done within the project ANTARES.

Data were obtained from the database compiled by the MED-Duluth group and freely available in the United States Environmental Protection Agency (U.S. EPA) website. MED-Duluth tested a series of industrial organic compounds using Fathead minnow (FHM) for the purpose of developing an expert system to predict the acute mode of toxic action from chemical structure. The results were also used to develop QSAR models.

We selected compounds that had experimental toxicity values related to FHM and filtered the database as reported in Maran et al. in 2007, and then we eliminated one additional compound because it was an isomeric mixture. Thus the final set, derived from the work done within ANTARES project, is composed of 567 compounds (Cappelli et al., 2015).

Within the project CALEIDOS we selected experimental data coming from registrants to evaluate the adequacy of different QSAR models. The original dataset contained all data about aquatic toxicity extracted from the ECHA CHEM database in the OECD Toolbox version 3.1. A total of 87572 data points on aquatic toxicity (only Klimisch score 1 and 2) on about 4000 compounds were retrieved. Starting from this huge database, a strict selection of the specific endpoint to address was performed (see Table 4). We found a large number of data coming from limit test, whose results are not exact values, but only indications that the substance is considered non-toxic because the LC50 is higher than a certain concentration (usually 100 mg/L). These values are not continuous and could not be used for the next steps. For compounds with more than one data, the geometric mean was calculated. The dataset after this pruning activity was composed by 491 compounds.

These two datasets were used to perform an evaluation of a first approach based on the integration of different models according to chemical classes. In particular, for the CALEIDOS dataset only the 455 compounds were used, with a predicted value for all the three models (ECOSAR v4.1, TerraQSAR™ v. 1.1, T.E.S.T. v4.0.1 and VEGA v1.0.8). For details on this work see CALEIDOS Deliverable 07 (Report on the performance of 16 models), 2014.

Then, we found in the literature a new and larger set of compounds published by Su et al. in April 2014. It contains data coming from different studies on multiple fish species, taken from peer-reviewed publications and online databases. After a brief pruning (see Table 4), the dataset consists of 953 compounds with data on one or more species and the average value for each compound was calculated. This set was used to develop new models with the k-NN approach.

Table 4: Pruning criteria for the three datasets.

	Dataset 1*	Dataset 2	Dataset 3
Original source	Maran U et al., 2007	ECHA CHEM database in the OECD Toolbox version 3.1	Su et al., 2014
Original No. of compounds	568	Ca. 4000	965
Final No. of data	567	455	948

Elimination rules			
Mixtures	x	X	-
Inorganic compounds	-	X	x
Non exact values (e.g. values with < or >)	-	X	-
Disconnected compounds (included salts)	-	X	x
Selection rules/datum details			
Chemical	-	Organic and organometallic mono-constituent	-
Endpoint	LC50	LC50	LC50
Effect	Mortality	Mortality	Mortality
Fresh/salt water	Freshwater	Freshwater	Freshwater
Duration time	96 hr	96 hr	96 hr
Species selected	Pimephales promelas	Species accepted in REACH guidance for aquatic toxicity	Poecilia reticulata, Oncorhynchus mykiss, Pimephales promelas, Oryzias latipes
Guideline followed	-	Only those accepted in the REACH guidance for aquatic toxicity (OECD 203, US-EPA, EU method, ASTM, DIN, etc.)	-
Measure unit	mmol/l (log neg)	mg/l	mmol/l (log neg)

* This dataset was already pruned. Only one compounds was eliminated because it is an isomeric mixture.

2.1.5. The data on fish chronic toxicity for ACR

A dataset of 91 compounds with both acute and chronic experimental toxicity data for fish was created to extract rule for the acute-to-chronic ratio (ACR) calculation. These data were extracted from two sources:

1. OECD QSAR Toolbox v3.2 Aquatic OASIS, Aquatic ECETOC, Aquatic Japan MoE and ECOTOX databases were checked but only data from ECOTOX database were available.
2. ECHEM portal. Only experimental values from ECHA registrations were extracted (February 2013).

From both the sources only organic compounds were considered. Compounds with dissociated structures (including salts) were eliminated.

Firstly chronic data were analysed. To build a homogeneous dataset we decided to consider only data obtained according to one test, the OECD 210 (FELS). Detailed criteria of selection are reported in Tab. 5. All the data were combined into a unique dataset in which the minimum datum was selected in case of multiple data. The toxicity data above water solubility were eliminated. Water solubility was extracted from WSKOWWIN v1.42. The experimental values were selected and only if they were not available, the predicted value was used. Calculation was performed using CAS number., SMILES and SDF file (generated by SMILES). The applicability domain (AD) was manually checked and the values considered outside it were

eliminated. A dataset with 112 compounds with chronic data were obtained. Acute toxicity data for these compounds were searched following the same criteria. A final dataset of 91 compounds with both acute and chronic toxicity data for fish were obtained.

Table 5: Specific criteria selection for both acute and chronic toxicity endpoint.

Chronic toxicity data selection	Acute toxicity data selection
NOEC 28 -90 days (depending on fish)	LC50 96h (4 days)
Fish species: OECD 210	Fish species: OECD 203
Purity ≥ 80% or not available	Purity ≥ 80% or not available
Initial age/life stage < 48h (eggs, embryos, blastula, gastrula, morula, eyed eggs, eyed embryo, larvae)	Initial age/life stage: juvenile (checked length where possible)
Exposure type: flow-throw, semi-static (renewal)	Exposure type: flow-throw, semi-static (renewal)
No. of doses ≥ 5 or not available	No. of doses ≥ 5 or not available
T (depending on fish)	T (depending on fish)
	Eliminated salt water and field test

A second dataset was used to confirm the ACR. The source is a list of 222 compounds with both acute and chronic toxicity data for daphnia and fish. These data were not used to identify new rules, but only to check the existing ones because these dataset contains data for several kind of fish (but not defined) and obtained with different test guidelines, (e.g. OECD 210, 215, 212, life cycle). Disconnected structures (included salt) and inorganic compounds were eliminated as well as compounds in common with the dataset used to generate the structural alerts. A final dataset of 108 compounds were obtained.

2.2 Software

Within PROMETHEUS we used, optimized and developed some tools, which can be applied to many datasets. In general, when models were available and of good quality we used them. We verified the quality of the RB model with a new set of compounds, to increase our reliability on this model. In other cases we developed new models, in order to get better predictions. The general strategy has been to have available a series of models to be used in a battery. We tried to apply methodologies which are based on different strategies, and in particular statistical methods and methods based on fragments. This should increase the robustness of the combined approach, because it covers different perspectives. For this reason we developed, for instance, a number of models using SARpy, which produce fragments related to the property.

Below we describe the models we tested, both the existing ones and those new, developed within PROMETHEUS.

Within the successive phase of the development of the workflow, we used the models, which provided better results when integrated. Thus, not all the models described below passed to the successive phase of the inclusion into the workflow, because before we did a preliminary phase to check and select those preferable.

Descriptions of the models used for each software is discussed below.

2.2.1 AlogP v.1.0.0.

The model provides a quantitative prediction of water/octanol partition coefficient and it is implemented inside the VEGA online platform. It is based on the Ghose-Crippen-Viswanadhan LogP and consists of a regression equation based on the hydrophobicity contribution of 120 atom types (A.K. Ghose et al. 1986; V.N. Viswanadhan et al., 1993; A.K. Ghose et al. 1998).

The Applicability Domain (AD) specifies the scope of the QSAR models and defines the model limitations with respect to its structural domain and response space. If an external compound is beyond the defined scope of a model, it is considered outside that model's AD and cannot be associated with a reliable prediction.

Within the VEGA platform, the applicability domain of predictions is assessed using the Applicability Domain Index (ADI) that has values from 0 (worst case) to 1 (best case). The ADI is calculated by grouping several other indices, each one taking into account a particular issue of the applicability domain. Most of the indices are based on the calculation of the most similar compounds found in the training and test set of the model, calculated by a similarity index that consider molecule's fingerprint and structural aspects (count of atoms, rings and relevant fragments).

For each index, including the final ADI, three intervals for its values are defined, such that the first interval corresponds to a positive evaluation, the second one corresponds to a suspicious evaluation and the last one corresponds to a negative evaluation.

Below, all applicability domain components are reported along with their explanation and the intervals used.

- Similar molecules with known experimental value. This index takes into account how similar are the first two most similar compounds found. Values near 1 mean that the predicted compound is well represented in the dataset used to build the model, otherwise the prediction could be an extrapolation. Defined intervals are:
  - $1 \geq$ index > 0.9 strongly similar compounds with known experimental value in the training set have been found
  - $0.9 \geq$ index > 0.75 only moderately similar compounds with known experimental value in the training set have been found
  - index ≤ 0.75 no similar compounds with known experimental value in the training set have been found
- Accuracy (average error) of prediction for similar molecules. This index takes into account the error in prediction for the two most similar compounds found. Values near 0 mean that the predicted compounds falls in an area of the model's space where the model gives reliable predictions, otherwise the greater is the value, the worse the model behaves. Defined intervals are:
  - index < 0.5 accuracy of prediction for similar molecules found in the training set is good
  - $0.5 \leq$ index < 1.0 accuracy of prediction for similar molecules found in the training set is not optimal
  - index > 1.0 accuracy of prediction for similar molecules found in the training set is not adequate
- Concordance with similar molecules (average difference between target compound prediction and experimental values of similar molecules). This index takes into account the difference between the predicted value and the experimental values of the two most similar compounds. Values near 0 mean that the prediction made disagrees with the values found in the model's space, thus the prediction could be unreliable. Defined intervals are:
  - index < 0.5 similar molecules found in the training set have experimental values that agree with the target compound predicted value
  - $0.5 \leq$ index < 1.0 similar molecules found in the training set have experimental values that slightly disagree with the target compound predicted value

- index > 1.0 similar molecules found in the training set have experimental values that completely disagree with the target compound predicted value

- Maximum error of prediction among similar molecules. This index takes into account the maximum error in prediction between the two most similar compounds. Values near 0 means that the predicted compounds falls in an area of the model's space where the model gives reliable predictions without any outlier value. Defined intervals are:

  - index < 0.5 the maximum error in prediction of similar molecules found in the training set has a low value, considering the experimental variability
  - 0.5 ≤ index < 1.0 the maximum error in prediction of similar molecules found in the training set has a moderate value, considering the experimental variability
  - index ≥ 1.0 the maximum error in prediction of similar molecules found in the training set has a high value, considering the experimental variability

- Global AD Index. The final global index takes into account all the previous indices, in order to give a general global assessment on the applicability domain for the predicted compound. Defined intervals are:

  - 1 ≥ index > 0.85 predicted substance is into the Applicability Domain of the model
  - 0.85 ≥ index > 0.75 predicted substance could be out of the Applicability Domain of the model
  - index ≤ 0.75 predicted substance is out of the the Applicability Domain of the model

This scheme is adopted by all the models implemented in VEGA, and thus we illustrate in more detail here, but this can be used also for the other models. Please notice, however, that the thresholds to define the three intervals above mentioned may vary depending on the model, because they are based on the number of compounds.

Within the output of VEGA for each molecule, results are organized in sections with the following order:

- *Prediction summary*: reported a depiction of the compound and the final assessment of the prediction. A graphical representation of the evaluation of the prediction and of its reliability is also provided, using green for low $\log K_{ow}$ value (less than 3.0); yellow for high $\log K_{ow}$ value (more than 3.0 and less than 8.0) and red for very high $\log K_{ow}$ value (more than 8.0). The reliability is expressed with a maximum of three stars for the best case (chemical into the AD) or with a minimum of one star for the worse case (chemical is out of the AD).
- *Applicability Domain (Similar compound):* report of the list of the six most similar compounds found in the training and test set of the model, along with their depiction and relevant information (mainly experimental value and predicted value).
- *Applicability Domain scores*: list of all Applicability Domain scores, starting with the global Applicability Domain Index (ADI). Note that the final assessment on prediction reliability is given on the basis of the value of the ADI. For each index, it is reported its value and a brief explanation of the meaning of that value.
- *Reasoning (fragments and moieties)*: If some rare and/or missing Atom Centered Fragments are found, they are reported here with a depiction of each fragment.

The model accepts as input two molecule formats: SDF (multiple MOL file) and SMILES. All molecules found as input are pre-processed before the calculation of molecular descriptors, in order to obtain a standardized representation of compound. For this reason, some cautions should be taken.

Results given as PDF file consists of a document containing all the information about the prediction.

On the pruned training set from EPIsuite™ KowWin module (9,961 compounds), the logP model has the following statistics: test set: n = 9961; R^2 = 0.84; RMSE = 0.72.

In this study, the value of ADI is used as a measure of the reliability of the predictions, such that each compound can be ordered in the prioritization process also on the basis of its uncertainty and not only on the basis of the (experimental or predicted) value of the property.

2.2.2 MlogP 1.0.0.

The model provides a quantitative prediction of water/octanol partition coefficient ($logK_{ow}$). It is implemented inside the VEGA online platform. The model is based on the Moriguchi LogP (MLogP) and consists of a regression equation based on 13 structural parameters (I. Moriguchi et al. 1992; I. Moriguchi et al. 1994). For the purpose of applicability domain assessment, the training set of the Meylan LogP model (9,961 compounds) has been considered, setting all molecules as belonging to the test set. The applicability domain of predictions is assessed using an ADI as described above. As for the previous model, for each index three intervals for its values are defined, such that the first interval corresponds to a positive evaluation, the second one corresponds to a suspicious evaluation and the last one corresponds to a poor evaluation.

The model accepts as input two molecule formats: SDF (multiple MOL file) and SMILES.

Results given as PDF file consists of a document containing all the information about the prediction. For each chemical results are organized in sections in the same way described for the model above (AlogP).

2.2.3 VEGA KOWWIN (Meylan) v. 1.1.3.

The model provides a quantitative prediction of water/octanol partition coefficient ($logK_{ow}$). It is implemented inside the VEGA online platform. It is based on the Atom/Fragment Contribution (AFC) method from the work of Meylan and Howard (Meylan, W.M. and P.H. Howard, 1995). The calculated model has a lower bound of -5.0 log units (all predictions lower than this value are set to -5.0). A dataset of compounds with experimental logP values has been built starting from the original dataset provided in EPIsuite™ and, after processing, pruning and cleaning, consists in 9,961 compounds.

The applicability domain of predictions is assessed using an ADI that has values from 0 to 1 (worst to best case). For each index (similarity, accuracy, variability and including the final ADI) three intervals for its values are defined as described for the previous models.

On the pruned training set from EPIsuite™ KowWin module (9,961 compounds), the logP model has the following statistics: Training set: n = 9961; R2 = 0.86; RMSE = 0.76.

The model accepts as input two molecule formats: SDF (multiple MOL file) and SMILES. All molecules found as input are pre-processed before the calculation of molecular descriptors, in order to obtain a standardized representation of compound.

Results given as PDF file consists of a document containing all the information about the prediction. The results are organized in sections and explained in the same way described previously for the other $logK_{ow}$ models.

This version is an updated version of the original model (1.2.0). Improvements are relative to a more robust similarity index, as described in detail in the VEGA website. A further check of structures and experimental data has been performed, resulting in the removal of some compounds from the original dataset (10,005 compounds), which had inconsistent experimental data. This update can influence some calculation, in particular similarity evaluation, so there could be some changes in the applicability domain values produced.

2.2.4 SARpy

SARpy, a free program developed by Politecnico di Milano, is able to automatically identify structural alerts related to a specific property. The software is described in the literature (Ferrari et al., 2013). Briefly, given a training set of molecular structures, with their experimental activity binary labels, SARpy generates every

substructure in the set and mines correlations between the incidence of a particular molecular substructure and the activity of the molecules that contain it. In this way it is possible to extract structural alerts associated to activity or inactivity. In this work it was used to extract structural alerts for different endpoints: the rules of the ready biodegradability model, the structural alerts used for persistence and to identify high or low acute-to-chronic ratio (ACR). Compared to other related programs, SARpy generates fragments with a higher number of atoms, since it starts the fragmentation process from the whole molecule, cutting one atom every time. Thus, it generates fragments, which often are more specific than other programs.

This is done in three steps starting just from the structural SMILES notation:

1. Fragmentation: this novel, recursive algorithm considers every combination of bond breakages working directly on the SMILES string. This fast procedure is capable of computing every substructure of the molecular input set.
3. Evaluation: each substructure is validated as potential SA on the training set. It is a complete match against the training structures, aimed at assessing the predictive power of each fragment.
4. Rule set extraction: from the huge set of substructures collected, a reduced set of rules is extracted in the form: 'IF contains <SA> THEN <apply activity label>'.

The advantage of SARpy is that it is highly flexible, and can be applied to a large variety of datasets. Based on the chemical structure it extracts fragments, which can be related to the effect but also to the lack of effect, which is useful in our case. The threshold to identify the effect can be established by the user, and also this is very useful. Compared to other software extracting fragments, it identifies larger fragments, because the algorithm starts from the largest fragment as possible. This means that the program identifies fragments, which are quite specific.

2.2.5 IstChemFeat

The software istChemFeat 1.0 looks for functional groups of the chemicals in a dataset. It is useful to see if a substance contains a functional group and if this group is active or inactive. A chemical that belongs to an active class has the value above or below the trigger value. The statistics have been described, such as number of components for a certain class, and number of active/inactive compounds. We may obtain classes with chemicals showing property values in a small range, and this is useful, in particular if the number of the chemicals is sufficiently high; conversely, if the range of the values for the chemicals in a class is spread, the class is not useful, and the chemical feature is not related to the property. Thus, the relevance of the class is related on the spread of the range of the values, and on the number of chemicals.

2.2.6 WSKOWWIN v 1.42

WSKOWWIN is one of the standalone models included in the EPIsuite™ v4.1. It estimates water solubility of organic compounds on the basis of the $\log K_{ow}$ and molecular weight (MW). If available, it also uses the experimental melting point to estimate water solubility. $\log K_{ow}$ experimental values, if available, are extracted from an internal database of more than 13200 experimental values, otherwise are calculated through KOWWIN (included in EPIsuite™ v4.1). It contains also experimental water solubility data for 6230 compounds (Meylan, W.M. and P.H. Howard; 1994).

A training set of 1450 compounds with $\log K_{ow}$, water solubility and melting point (Tm) in deg C were used to develop the model based on the following equations:

1. log S (mol/L) = 0.796 - 0.854 log Kow - 0.00728 MW + ΣCorrections
2. log S (mol/L) = 0.693 - 0.96 log Kow - 0.0092(Tm-25) - 0.00314 MW + ΣCorrections

If the melting point is known the second equation is used, otherwise the first one is used.

No automatic evaluation of the applicability domain is available, but compounds outside the range of molecular weight, water solubility and/or log Kow should be considered of low reliability. Moreover, the com-

pounds should not have functional group(s) or other structural features not represented in the training set, and for which no correction factor was developed (US EPA, 2012).

2.2.7 ECOSAR Class Program v 1.11

ECOSAR is one of the standalone models included in the EPIsuite™ v4.1 that is used to predict aquatic toxicity of chemical. It is based on structure similarity using a list of more than 120 chemical classes. The equations are mostly based on the logP (introduced by user, experimental or automatically predicted through KOWWIN v 1.68). The predicted value is compared with the water solubility as calculated by WSKOWWIN v 1.42. For each chemical class equations for LC50 for fish 96 hr, LC50 for daphnid 48 hr, EC50 for algae 72 or 96 hr, Fish ChV (a chronic value calculated as explained below), Daphnid ChV and Algae ChV are available. In some cases also equation for LC50 for fish 96 hr – SW, LC50 for mysid shrimp 96 hr – SW, Fish ChV – SW, Mysid shrimp ChV –SW and LC50 for earthworm 14 d are available. The chronic toxicity values are calculated according to:

$$ChV = 10 ([\log (LOEC*NOEC)]/2$$

There are dedicated models for surfactants and dyes. If no sufficient experimental values are available (excess toxicity classes) equations are derived using a log Kow cut-off in addition to a single experimental toxicity value. If few or no experimental values are available, prediction is done with ACR (marked with a "!") and log Kow cut-off.

No automatic evaluation of the applicability domain is available, but compounds outside the range of molecular weight and logP should be considered of low reliability. Moreover, the user should check each equation used.

In this work ECOSAR was used to estimate acute fish toxicity only (for details see paragraph 2.2.15).

2.2.8 T.E.S.T. v 4.1

The Toxicity Estimation Software Tool (T.E.S.T.) allows a user to estimate 7 toxicity endpoints (i.e. LC50 96h Fathead minnow, LC50 48h *Daphnia magna*, IGC50 48h *Tetrahymena pyriformis*, LC50 oral rat, Bioaccumulation factor, Developmental toxicity and Ames Mutagenicity) and 7 physicochemical properties (i.e. normal boiling point, density, lash point, thermal conductivity, viscosity, surface tension and water solubility) without requiring any external programs. Toxicity can be estimated using one of several advanced QSAR methodologies (i.e. hierarchical method, FDA method, single model method, group contribution method, nearest neighbor method, consensus method and random forest method (only developmental toxicity)). All the descriptors used for the estimation are calculated automatically.

The applicability domain of each model is calculated using a combination of three different methods: model ellipsoid constraint, Rmax constraint and fragment constraint. For the nearest neighbor, to be predicted the compounds must have 3 chemicals with a similarity score (SC) > 0.5. If a method cannot give a prediction inside the applicability domain, the prediction is not available.

2.2.9 VEGA v 1.0.8 – Fathead minnow LC50 96 hr (EPA) v 1.0.6

This is one the models included in the VEGA platform (Benfenati et al., 2013). It estimates the LC50 96 hr for the fathead minnow (*Pimephales promelas*). It is a re-implementation of the original model developed by Todd Martin inside T.E.S.T. software. It is a linear regression based on 21 molecular descriptors, calculated by an in-house software module in which they are implemented as described in: Todeschini and Consonni, 2009.

The applicability domain is automatically calculated through a series of parameters: similar molecules with known experimental value, accuracy (average error) of prediction for similar molecules, concordance with similar molecules (average difference between target compound prediction and experimental values of similar molecules), maximum error of prediction among similar molecules, Atom Centered Fragments similarity check and model descriptors range check. They are summarised in the global AD index. A value greater than

0.85 (max value = 1) means that the prediction is in the applicability domain, between 0.85 and 0.7 that the prediction could be out of domain, and equal or below 0.7 that the prediction is out the applicability domain. (E. Benfenati, A. Manganaro and G. Gini, VEGA-QSAR Workshop, 2013).

2.2.10 VEGA v 1.0.8 – Fish LC50 classification v 1.0.1-DEV

This is the developmental version of a classification model for the acute toxicity for fish (Benfenati et al., 2013). It was developed on a dataset of 568 compounds with experimental toxicity data on fathead minnow (*Pimephales promelas*). The predicted classes are 4: < 1 mg/l, between 1 and 10 mg/l, between 10 and 100 mg/l and > 100 mg/l. The choice of the classes was done considering the requirements of the Classification, Labelling and Packaging regulation. In this work the predictions of the classification model were used only as a confirmation of the continuous value and not alone as a classifier. The test set was a set of 351 compounds with toxicity data on Rainbow trout (*Oncorhynchus mykiss*). The model was developed using SARpy to extract rules for each of the three thresholds.

The applicability domain is automatically calculated through a series of parameters: similar molecules with known experimental value, accuracy (average error) of prediction for similar molecules, concordance with similar molecules (average difference between target compound prediction and experimental values of similar molecules) and Atom Centered Fragments similarity check. They are summarised in the global AD index. A value greater than 0.8 (max value = 1) means that the prediction is in the applicability domain, between 0.8 and 0.65 that the prediction could be out of domain, and equal or below 0.65 that the prediction is out the applicability domain.

2.2.11 Fish Toxicity k-NN/Read-Across model (In-house/VEGA)

It is an in-house model, developed using the VEGA libraries (in particular for the algorithm of similarity calculation) and a novel software tool called istKNN. It is a k-Nearest Neighbour based model, where the prediction is provided using the k most similar compounds. This models searches for the most similar compounds and assigns the property value of the target compound based on the average of the property values of the k most similar chemicals. The number k is found by the software and the developer checks different possibilities. In the particular kNN model we applied, the software is more sophisticated than other kNN models. Indeed, the property values keeps into account how similar are the most similar compounds, and provide weights based on this.

In our case, we used as source of the property values the dataset built from several sources: the database compiled by the MED-Duluth group, the OECD Toolbox, the DEMETRA Project (Rainbow Trout toxicity model) and the work of Su et al. (2014). The dataset contains 973 compounds, where the toxicity is expressed as the mean value of the experimental data on several different species.

Under some conditions, the prediction can be performed on less than k molecules, for instance if some of the similar molecules have a similarity value under a given threshold. Furthermore, when no molecules meet the minimum requirements set for the model, no prediction is provided.

The in-house model has been built with the following settings: number of molecules (k) = 4, similarity threshold = 0.7 (molecules with similarity lower than 0.7 are not considered), similarity threshold for single molecule based prediction = 0.75 (if only one molecule is left for prediction, it is used only if it has a similarity value higher than 0.75), enhance factor = 3 (a value used to increase the relevance of the most similar compounds in the prediction, proportionally to their similarity value) and experimental range = 3.5 (when the k compounds have an experimental range higher than 3.5, prediction is not provided).

The model is validated in a leave-one-out approach: for each molecule, of the training set, its prediction is performed using all the other molecules except the target compound itself (as the model would find a perfect structure match and provide its experimental value instead of a prediction). With such approach, the model yields the following statistics:

- No. of compounds: 973

- Valid predictions: 936
- R2: 0.65
- RMSE: 0.708
- Not predicted compounds: 36 (3.7% of the dataset)

3 RESULTS

3.1 The workflow of the individual properties. General remark.

The conceptual approach to get the PBT assessment is to assess separately P, B and T, and then to merge the results. Here we describe how the separate assessment of P, B and T has been done, below this we will show how the program integrates the results of the separate assessment.

Each separate assessment produces a value, and a figure associated, which is a measurement of the reliability of the assessment.

At the basis of each assessment there are the values, obtained from different sources; experimental and calculated values are used. Experimental values have a higher reliability. Also experimental values are of course associated to uncertainty and variability, but the program only checks inconsistent values resulting from multiple sources, in case of experimental values.

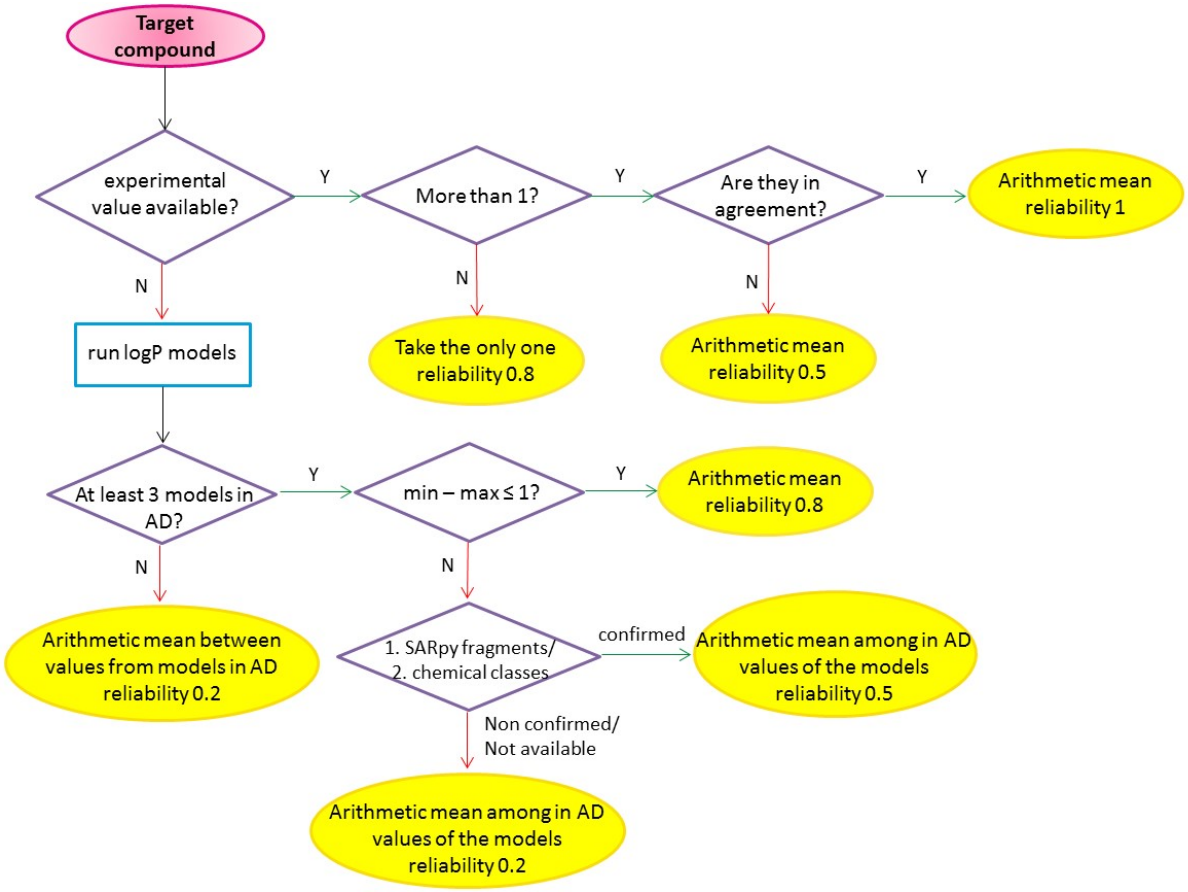
In case of predicted values the uncertainty is associated to the output of the individual but also for the presence of multiple values, which may be consistent or not.

Besides the three P, B and T property we also carefully evaluated the $logK_{ow}$ value, and developed a workflow for this. This value is then used in the B-workflow.

3.1.1 The workflow for $logK_{ow}$

Figure 1 shows the workflow for $logK_{ow}$. The first check is the availability of the experimental data of the target compound.

Figure 1: The LogKow workflow



If more than one experimental value is available and the values are in agreement (threshold 0.35 log units), we calculate the arithmetic mean and obtain the maximum reliability (reliability 1). If they are in disagreement, we calculate the arithmetic mean with medium reliability (reliability 0.5).

If one experimental value is available, we take it with high reliability (reliability 0.8).

In case there are no experimental data, the assessment of the target chemical is based on the predictions with QSAR models, in particular AlogP, MlogP, VEGA KOWWIN, EPA KOWWIN and k-NN.

When the chemical does not fall into the applicability domain of at least three models, we calculate the arithmetic mean between values from the models in which the chemical falls in the applicability domain and a low reliability (reliability 0.2) is assigned.

In case the chemical falls into the applicability domain of at least three models and the predictions are consistent (i.e. the difference between the minimum and the maximum values is lower or equal to 1 log unit), we calculate the arithmetic mean of the predicted values with high reliability (reliability 0.8).

If the condition about the consistence (i.e. the difference between the minimum and the maximum values is lower or equal to 1 log unit) of the predictions is not verified, we use fragments founded with SARpy and chemical classes to see if they confirmed the predictions or not. If the fragments and classes confirm the predictions, we calculate the arithmetic mean among the values of the models in which the compound falls into the applicability domain with a mean reliability (reliability 0.5). In case the fragments or classes are in disagreement with the predictions or are not found, the arithmetic mean is calculated among the values from the models in which the compound is into the applicability domain but with low reliability (reliability 0.2).

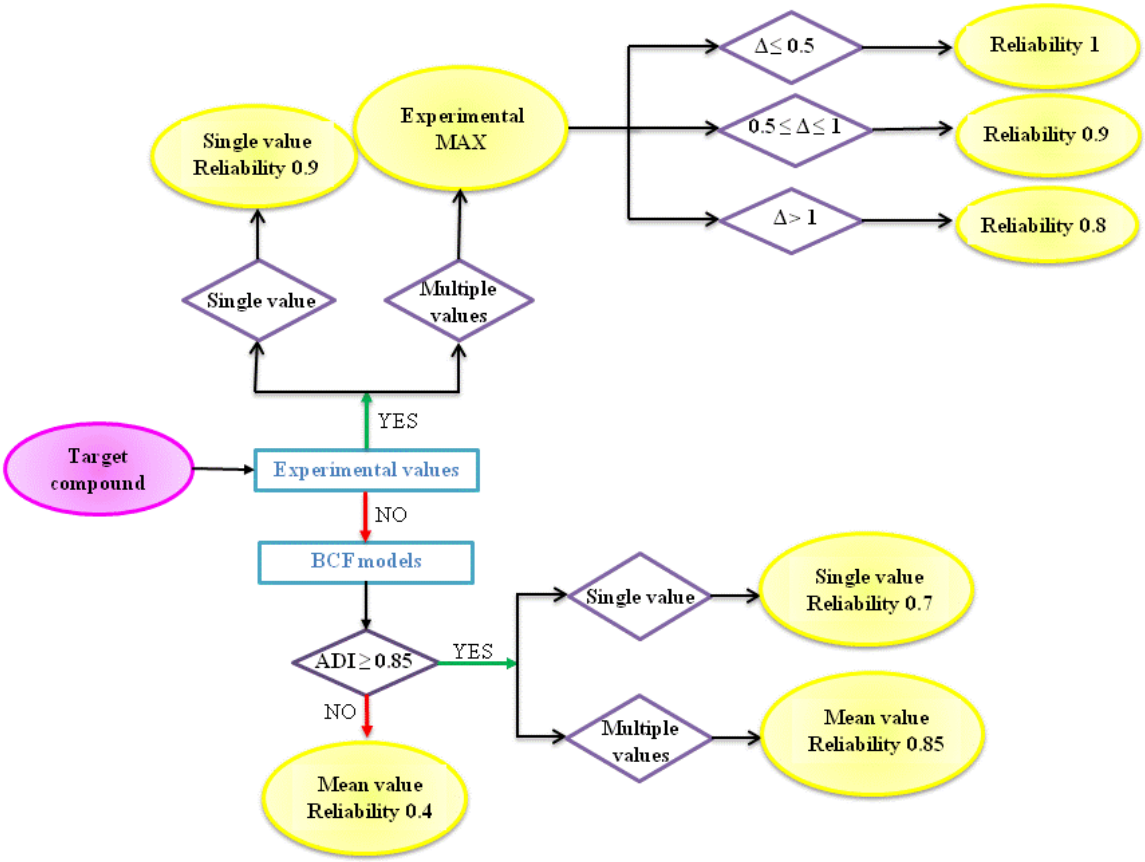
The log Kow estimation is included since it provides a fundamental element for the understanding of the profile of each substance. This estimation is not directly used in the ranking score formula, yet it is strictly related to the BCF results. Indeed, the models used in the workflow to obtain a joint reliable logKow value

are the same that play a relevant role in some models used in the B-workflow. In particular, in the CAESAR BCF model the MLogP value is used as a descriptor (and it plays a key role in the model's algorithm); in the Meylan BCF model, the VEGA LogP model estimation of the logKow is used as a base value for the calculation of the BCF (followed by several correction given from the presence of specific molecular fragments), and also the final reliability is influenced by the reliability of the VEGA LogP model.

3.1.2 The workflow for the bioaccumulation

The assessment of chemicals for bioaccumulation is in this project solely based on the BCF-factor. The workflow is according to the scheme shown in Figure 2.

Figure 2: The BCF workflow



Reliability scores are established to assign the weights differentiating the cases. Table 6 indicates these scores.

First, the system checks whether experimental data are available for the target compound. If there are multiple values and the difference between them is less than 0.5, the higher value of them is reported and the reliability score of 1 is assigned. When the difference between the experimental values goes from 0.5 to 1, the higher value of them is reported and the reliability score of 0.9 is assigned. The same reliability score is assigned when there is a single experimental value. When the difference between the experimental values is greater than 1, the higher value of them is reported and the reliability score of 0.8 is assigned.

Table 6. Reliability scores and criteria to assign

Criteria	Reliability score	
Experimental values with Δ ≤ 0.5	1	High reliability

Experimental values with 0.5 ≤ Δ ≤ 1	0.9	High reliability
Single experimental value	0.9	High reliability
Experimental values with Δ > 1	0.8	Medium reliability
Predicted values with ADI ≥ 0.85	0.85	High reliability
Single value with ADI ≥ 0.85	0.7	Medium reliability
Predicted values with ADI < 0.85	0.4	Low reliability

For the compounds that do not have experimental values, we check the outputs of the BCF models. It is necessary to consider the applicability domain index (ADI). If the target compound has predicted values with ADI values less than 0.85, the arithmetic mean value is calculated and the reliability score of 0.4 is assigned. On the opposite condition, if the predictions for the compound shows ADI values equal or greater than 0.85, the mean value is calculated and the reliability score of 0.85 is assigned. If it happens that there is only one predicted value and ADI is greater than 0.85, this value is reported and the reliability score of 0.7 is assigned.

3.1.3 The workflow for persistence

The assessment of persistence for chemicals was organized as workflow using the different models we developed (Figures 3 and 4).

Since the experimental value is normally considered more reliable compared to the predicted one, the first step is to check if an experimental value is available for the target compound within our datasets. In such a case, it is important to consider if the experimental value is available for all three compartments or not and in case, if they are in agreement. A further criterion to be considered is if the experimental value is nP/P; in this case the reliability of the prediction will be lower because this class has a high rate of intrinsic uncertainty. Thus, if the experimental value is available the assessment is done using the experimental value with a different rate of reliability considering the criteria described above (experimental value is nP/P, experimental value available for each compartment, concordance between experimental values). If the experimental values are available for the three compartment and they indicate the same class, the reliability is maximal (1.0), or medium (0.7) in case of nP/P class. If the experimental values are available for all compartments but not equal, the output is the worst case experimental values with maximum reliability. If experimental values are available for only two compartments and indicate the same class the reliability is high (0.9) (or medium, 0.7, in case of nP/P class), otherwise the output is the worst case experimental value with reliability medium (0.7). In case of only one experimental value available, the reliability is medium (0.7), or low (0.4) in case of nP/P class.

After checking the availability of experimental value, a further assessment is done: if the compound is recognized to be a perfluorinated, the prediction is P/vP with high reliability. Indeed many studies in the literature report that perfluorinated compounds have the capacity to persist in the environment and to be transported far from the emission source (Shoeib et al., 2006; Kim and Kannan, 2007; Fujii et al., 2007; Bao et al., 2011), so we include this rule to our workflow. If the target compound is not perfluorinated, it is evaluated for ready biodegradability property using the model developed by Lombardo et al. (2014). Also in this case the first point is to check if an experimental value is provided or not. In such a case the prediction is nP with maximum reliability (1.0), if not it means that the model may provide a prediction.

As already explained above, ready biodegradability is a screening test for persistence, which means that if a compound is readily biodegradable for sure it is not persistent; in the opposite case when the substance it is non-readily biodegradable, it is not necessarily persistent. Thus, if the compound is predicted by the model as readily biodegradable, the chemical is classified as nP and the reliability of the prediction closely depends on the applicability domain index (ADI) provided by the model itself. The ADI included in VEGA platform is already explained in Cassano et al. (2010) and if it is higher than 0.8 the compound is within the AD of the model (high reliability), if it is between 0.8 and 0.65 the compound could be outside the model AD (medium reliability) and if it is below 0.65 the compound is out of AD model. In the case the prediction is non-readily

biodegradable, possible readily biodegradable, no prediction is provided or if it is predicted readily biodegradable but out of AD, the compound will enter the second part of the workflow.

The following steps have to be done for each compartment (sediment, soil and water) in order to have the final assessment. Primarily the k-NN model is considered. If an experimental value is available within the k-NN dataset models, it will be used for the assessment with maximal reliability 1.0. However, considering the high uncertainly of the nP/P class, the reliability will be lower (high reliability, 0.9) if the experimental value is nP/P compared to the case that the experimental belongs to the other three classes. In case an experimental value is not available, the k-NN predictions will be considered.

The output of the k-NN model can be: nP, nP/P, P/vP, vP or unknown. If the prediction is unknown, it will be checked if SARpy fragments and chemical classes are available for the target compound. Since we extracted SA and chemical classes only for nP and vP compounds, the output of the SARpy fragments and chemical classes' models can be only nP, vP or unknown. If this information is available and is in agreement to each other (i.e. SARpy fragments and chemical classes extracted for nP are available and no fragments or chemical classes for the other category, vP, are provided) the prediction will be nP or vP with low reliability (0.4), otherwise the compound is not predicted. If the prediction of k-NN is available the SARpy fragments and chemical classes will be considered as well.

Depending on a series of conditions, the reliability of the prediction changes. If both SAs and chemical classes are available and in agreement with k-NN prediction the output is the k-NN prediction, with high reliability (0.9) (or medium, 0.7, in case of nP/P class). If only one between SA and chemical classes is available and in agreement with k-NN, the output is the k-NN prediction with medium reliability (0.7), or low (0.4) in case of nP/P class. In case of disagreement among the three predictions or in case is available only the k-NN prediction, the output is the k-NN prediction with low reliability (or not predicted in case of nP/P class). Once the prediction for each compartment has been obtained, the next step is to combine them in order to have a final persistence assessment for the target compound. Since the overall prediction is conservative, the "worse" class always wins and the reliability of the prediction is that of the "worse" class.

For example:

- In water the prediction is vP and reliability is medium (0.7)
- In soil the prediction is nP and reliability is high (0.9)
- In sediment the prediction is P/vP and reliability is low (0.4)
- The final prediction will be vP with medium reliability (0.7)

Another example:

- In water the prediction is vP and reliability is medium (0.7)
- In soil the prediction is vP and reliability is high (0.9)
- In sediment the prediction is P/vP and reliability is low (0.4)
- The final prediction will be vP with high reliability (0.9)

Figure 3: The persistence workflow (first part)

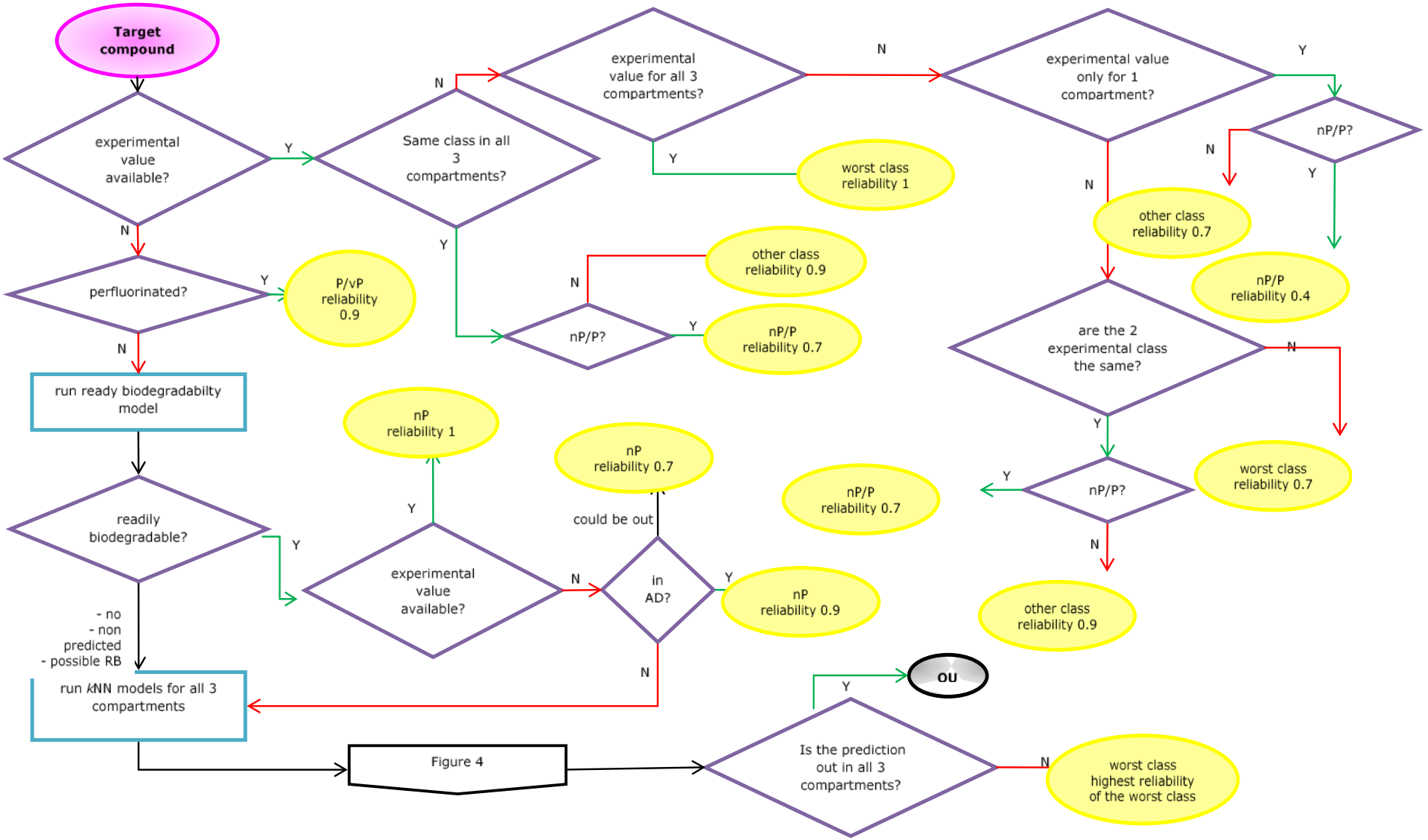
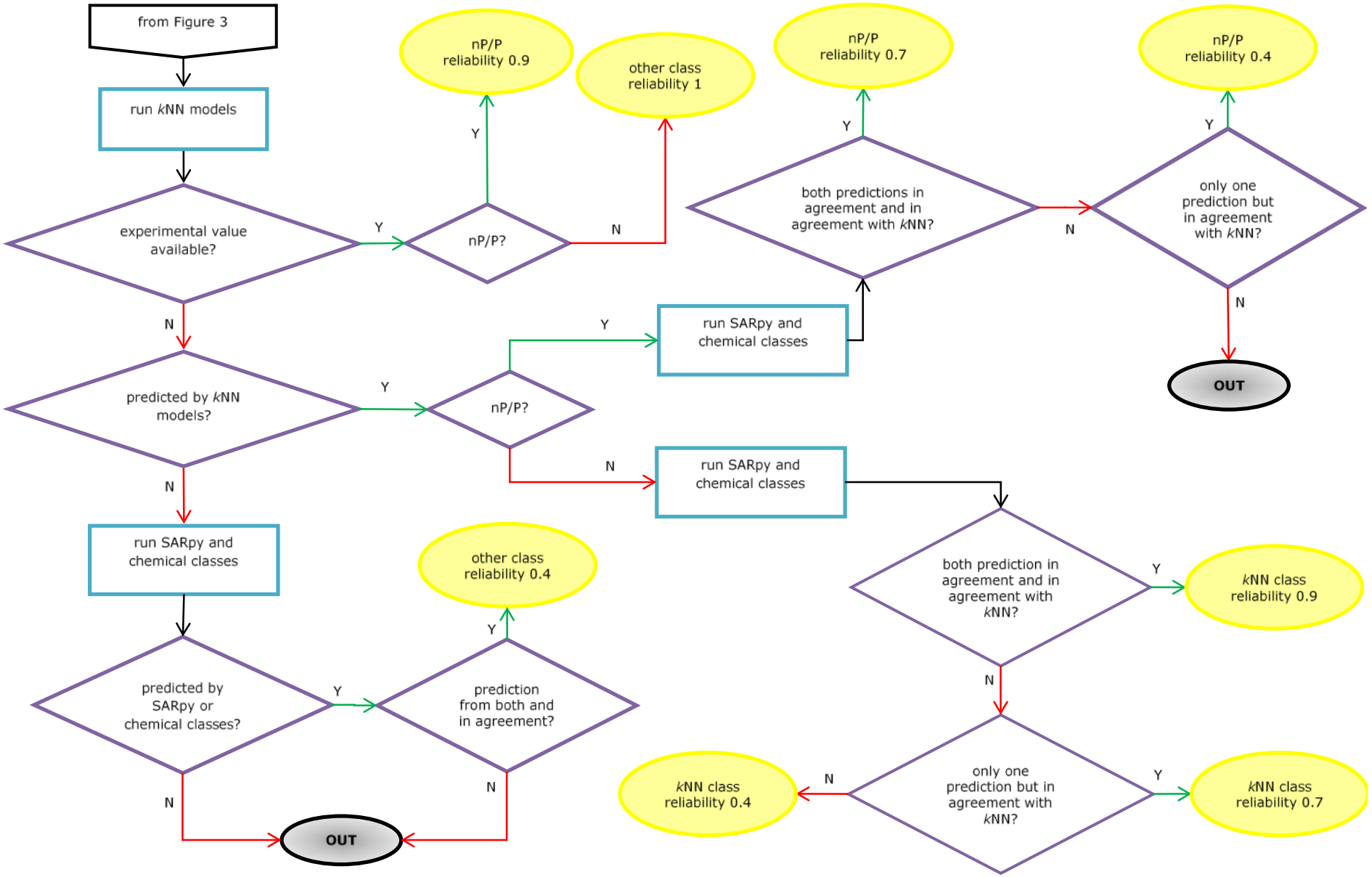


Figure 4: The persistence workflow (second part)



3.1.4 The workflow for toxicity

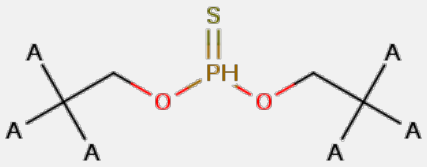
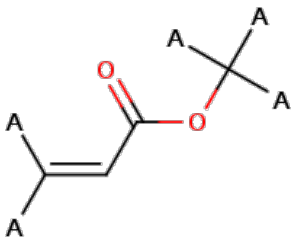
To prioritize substances for the toxicity (T) assessment only fish toxicity was considered, for the lack of reliable in silico models for the other endpoints, as we explained above. The assessment requires ideally also chronic toxicity data. Due to the paucity of these data, we decided to start from acute toxicity and derive the chronic value using the ACR. For this purpose, we built a dataset containing both acute and chronic data, as explained above. Since only few compounds had both the toxicity values we decided to split the compounds in two classes: (1) compounds with high ACR and (2) compounds with low ACR) and then to extract structural alerts for these two classes. To each class the program assigns an ACR that can be used to derive the chronic value from the acute one (either experimental or predicted). It is important to underline that a compound with a high ACR not necessarily is a compound with a high toxicity, but only that it is a compound with a high ratio between the acute and the chronic value. In the same way, a low ACR does not mean that the compound is not toxic, but only that the ratio between acute and chronic values is low.

3.1.4.1 Structural alerts for ACR

Since the number of compounds with experimental chronic toxicity values for fish was too low to develop a QSAR model (91 compounds), our strategy was to use both acute and chronic values to derive the ACR. Then, we verified existing structural alerts and we developed new ones for the estimation of high and low ACR. Existing structural alerts were taken from two sources: Ahlers et al., 2006 and May and Hahn, 2014. In the first study ACRs for fish, daphnids and algae were obtained. In particular for fish Ahlers et al., (2006) found a median ACR of 10.5 that is far below the threshold of 100 of the Technical Guidance Document (TGD) of European Commission on risk assessment (2003). Nevertheless, the threshold of 100 was found not protective for all the chemicals (they reported a maximum ACR for fish of 4250). Considering the distribution of their dataset, they used a threshold of 30 to distinguish between compounds with high and low ACR. On the basis of this threshold they extracted three structural alerts for fish. They also found no correlation between ACR and the log Kow or the mode of action. Another source of ACRs was May and Hahn, 2014. In this work they found a good correlation between acute and chronic values and they confirmed that the threshold of 100 is protective for more than 90 % of the industrial compounds (pesticides were excluded) with a maximum ACR for fish of 1370 and a median of 12.2. Ten structural alerts associated with high ACR (also in this work the threshold considered to discriminate between high and low ACR was of 30) were identified, in addition to the pesticides category. Moreover, the relationship between ACR and logK_{ow} or water solubility was verified and no correlation was found.

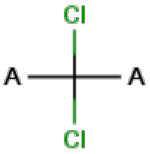
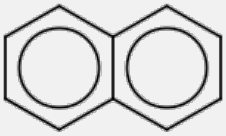
Considering the ACR distribution of our dataset, in which very few compounds had ACR > 30, we used a new threshold of 10 to identify high ACR (i.e. ACR > 10) and low ACR (i.e. ACR < 10). This threshold is more protective. Indeed, 10 is the suggested factor to derive the chronic from the acute toxicity values by Forbes, 2002 that reports the factor suggested by the U.S. Office of Pollution Prevention and Toxics and the Extrapolation (assessment) factors used in the Technical Guidance Documents for existing and new substances legislation within the European Union (CEC 1996). Using SARpy and istChemFeat, new chemical classes/structural alerts were extracted for both high and low ACR. They were examined and compared with the existing ones (Ahlers et al., 2006; May and Hahn, 2014). In some cases, new manually developed structural alerts were added. Finally, 5 structural alerts for high ACR and 10 for low ACR were extracted. The structural alerts were checked using the dataset supplied by UBA with 108 compounds. Some of the previously identified fragments in the literature were confirmed but others not. The not confirmed ones were eliminated: These were 3 structural alerts for high ACR and 4 for low ACR. Final statistics on the 91 compounds are reported in Tab. 7 and 8. In this analysis the two groups of structural alerts were considered separately, with high or low ACR. For this reason the compounds identified by the structural alerts were considered “positive”, and then compounds are labelled depending on which kind of alerts are found. These structural alerts can be used only to derive the chronic toxicity for fish starting from the acute toxicity for fish. To be used for other taxonomic classes new studies would be required.

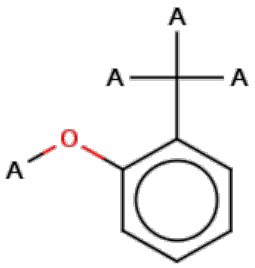
Table 7: Statistics for structural alerts that identify high ACR. Tot Positive represents the number of compounds that contains the structural alert, Tot TP the number of true positive (i.e. the number of compounds identified by the structural alert that have an ACR > 10).

Structural alerts for high ACR		Tot Positive	Tot TP
CCOP(=S)(OCC)		5	5
C(=C)C(=O)OC		3	3
c[NH2]		4	3

A = any atom, including H. (a)C = aromatic carbon. N(H2,A) = aliphatic NH2.

Table 8: Statistics for structural alerts that identify low ACR. Tot Positive represents the number of compounds that contains the structural alert, Tot TP the number of true positive (i.e. the number of compounds identified by the structural alert that have an ACR < 10).

Structural alerts for low ACR		Tot Positive	Tot TP
c1cc(c(cc1Cl))Cl		6	6
C(Cl)Cl		5	5
c1ccc2c(c1)cccc2		4	4

c1cc(c(O)c(c1)C)		9	8

A = any atom, including H. S(A) = aliphatic S.

3.1.4.2 Scheme for the T evaluation (based only on fish toxicity)

Using the ACR structural alerts previously described, the scheme to evaluate the toxicity property for the PBT assessment was built on the basis of fish toxicity only. The scheme is reported in Fig. 5, 6 and 7.

The user of the scheme here described can add experimental acute toxicity values. For each target compound without experimental values, QSAR models (T.E.S.T. v 4.1, VEGA v 1.0.8 – Fathead minnow LC50 96 hr (EPA) v 1.0.6, fish toxicity k-NN/Read-Across model and ECOSAR Class Program v 1.11) should be run. All these models give fish LC50 96h. Some of them allow the user to retrieve the experimental value if available. The k-NN model gives only the experimental value if the target compound is included in the training set.

The scheme discriminates among many situations that may occur:

1. Reliable experimental values inserted by user.
2. No experimental values inserted by user. Two or more experimental values are retrieved by the QSAR models and they are in agreement (the values are considered in agreement if the difference is below 1 log unit).
3. No experimental values inserted by user. Two or more experimental values are retrieved by the QSAR models but they are not in agreement.
4. No experimental values inserted by user. Only one experimental value is available.
5. No experimental values inserted by user or retrieved by QSAR models. Two or more predicted values in the applicability domain (AD) are available and in agreement (as before the values are considered in agreement if the difference is below 1 log unit).
6. No experimental values inserted by user or retrieved by QSAR models. Two or more predicted values in the AD are available but they are not in agreement.
7. No experimental values inserted by user or retrieved by QSAR models. Only one predicted value is available.
8. Both predicted and experimental values are not available.

When more than one reliable experimental value is available and in agreement (situation 1 and 2, Figure 5) the scheme uses the minimum and searches for the presence of the ACR fragments described in the previous paragraph. Depending on the fragments identified different ACR are applied. If only fragments for high ACR are identified the acute toxicity value is divided by a factor of 100. If only fragments for low ACR are identified the acute toxicity value is divided by a factor of 10. In the other cases (no fragments or both fragments for high and low ACR are identified) the factor used is 15. The factor of 100 was chosen because this is considered protective for the majority of the compounds (May and Hahn, 2014). Fragments that identify low ACR identify compounds with an ACR below 10. To be protective, for these compounds an ACR of 10 was chosen. When no fragments or both fragments for high and low ACR are identified, the situation is ambiguous or not defined. For this reason we choose a medium ACR of 15. Indeed, in our dataset the medium ACR is of 19.6, but in other datasets it is lower. For instance, a medium ACR of 12.2 is reported in May and Hahn, 2014 and the median of the ACR in Ahlers et al., 2006 is of 10.5. Since in this last case the ACR may be not

sufficiently protective, the reliability of the prediction (assigned on the basis of three criteria below detailed) is lower than in the other two cases (see Table 9 for the reliability score assignment).

1. The first criterion is the acute value (above or below the threshold of 0.01 mg/l). For the aquatic compartment, compounds with acute values below this threshold can be directly classified as T (EC, 2014). Even if the compound can be classified directly as T, the scheme needs a chronic value for the integrated prioritization. In this case, the fact that the substance is toxic even after a short time (acute toxicity) clearly demonstrates that at least the same level of toxicity can be achieved in chronic conditions. For this reason the reliability of this classification is higher than the one based on acute toxicity values above the threshold.
2. The second criterion is the ACR factor used. If no fragments are present or if both fragments for high and low ACR are identified the reliability of the prediction based on an ACR factor is lower.
3. The third criterion is the source and number of the experimental values. If the experimental value is inserted by the user or there are more experimental values that are in agreement the reliability is higher. When only experimental values retrieved by QSAR models are available and they are not in agreement (situation 3, Figure 5), the scheme uses the minimum (to be protective) and proceeds with the ACR fragments. In this case the reliability (see Table 9) is low due to the low reliability of the experimental values used. In the situation 4 (Figure 6), with only one experimental value retrieved by the QSAR models, the procedure is identical but with reliability slightly lower due to the higher uncertainty of the acute toxicity value.

Table 9: How the reliability score are applied.

	Only fragments for high or only fragments for low		Fragments for both low and high or no fragments	
Multiple exp. values in agreement with LC50 < 0.01	1.0	Very high reliability	0.8	High reliability
Multiple exp. values in agreement with LC50 ≥ 0.01	0.9	Very high reliability	0.7	High reliability
Single exp. value with LC50 < 0.01	0.8	High reliability	0.6	Medium-high reliability
Single exp. value with LC50 ≥ 0.01	0.7	High reliability	0.5	Medium reliability
Multiple exp. values not in agreement	0.4	Medium-low reliability	0.3	Low reliability
Multiple predicted values in agreement	0.6	Medium-high reliability	0.4	Medium-low reliability
Single predicted value confirmed by VEGA-Class	0.5	Medium reliability	0.3	Low reliability
Single predicted value not confirmed by VEGA-Class (not predicted)	0.4	Medium-low reliability	0.3	Low reliability
Multiple predicted values not in agreement confirmed by VEGA-Class	0.3	Low reliability	0.2	Low reliability
Multiple predicted values not in agreement not confirmed by VEGA-Class (not predicted)	0.1	Very low reliability	0.0	Very low reliability

When no acute experimental values are available, the predicted ones should be used considering only the prediction inside the AD of the model. When two or more predicted values are available and in agreement

(situation 5, Figure 7) the scheme uses the minimum and searches for the ACR fragments. The factors are the same described before and the reliability is assigned following the same criteria (reported in Table 9). When the acute predicted values are not in agreement or only one predicted values is available (situation 6, Figure 7 and situation 7, Figure 6 respectively) the VEGA-Class model (VEGA v 1.0.8 – Fish LC50 classification v 1.0.1-DEV) should be run to confirm the prediction. This model classifies the compounds into four classes on the basis of the LC50: < 1 mg/l, 1-10 mg/l, 10-100 mg/l and > 100 mg/l, named respectively T1, T2, T3 and T4. Only the prediction within the AD (i.e. global AD index greater than 0.85) should be considered. If the predicted class is not in agreement with the predicted value (or the minimum predicted value in case of multiple values) the compound cannot be predicted by this scheme. In the other cases the ACR fragments should be searched and the corresponding factors applied. The reliability is assigned following Table 9.

In situation 8 (no experimental and predicted values available, Figure 6) the scheme cannot proceed with the assessment.

Figure 5: The toxicity workflow (first part)

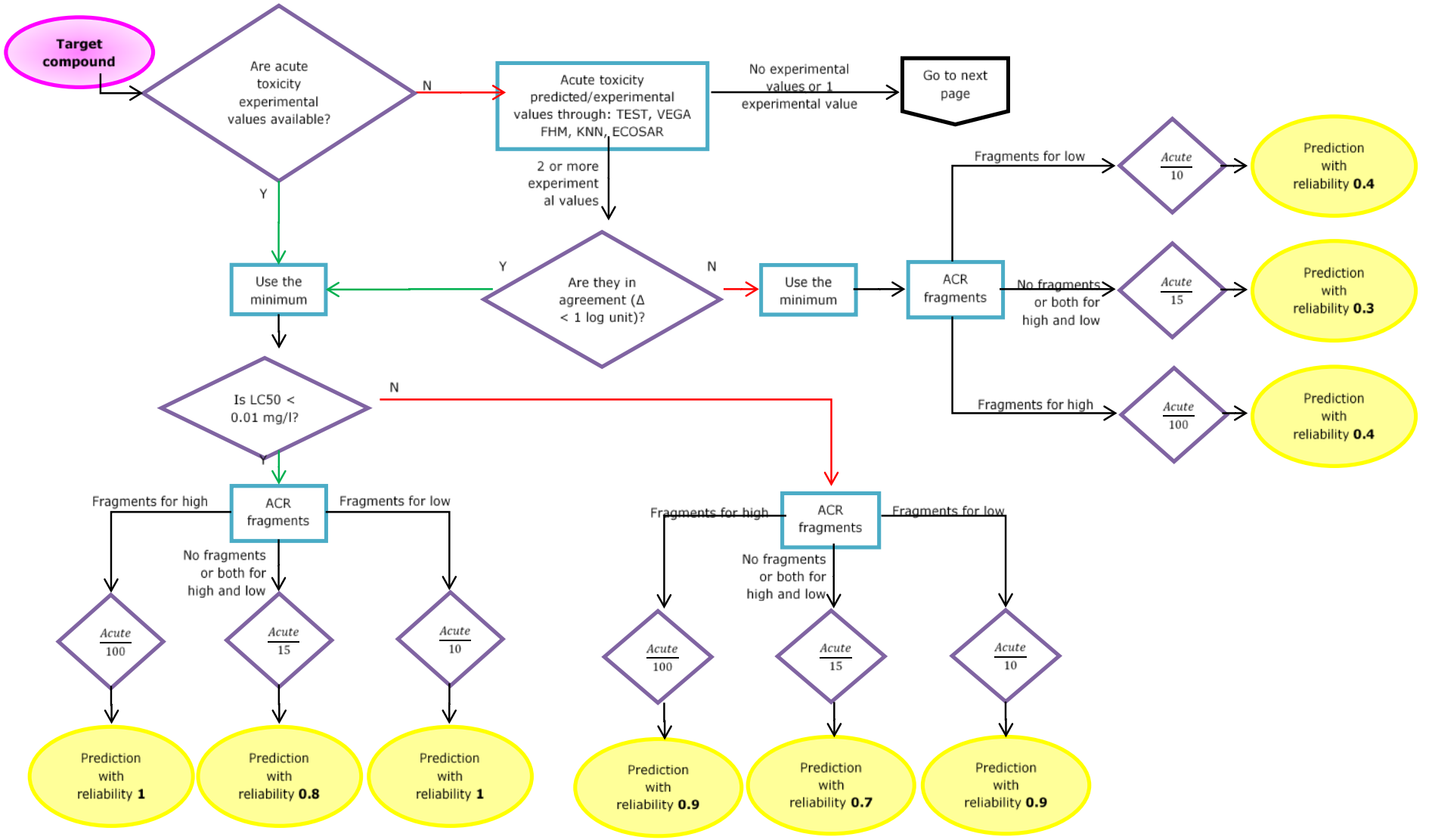


Figure 6: The toxicity workflow (second part)

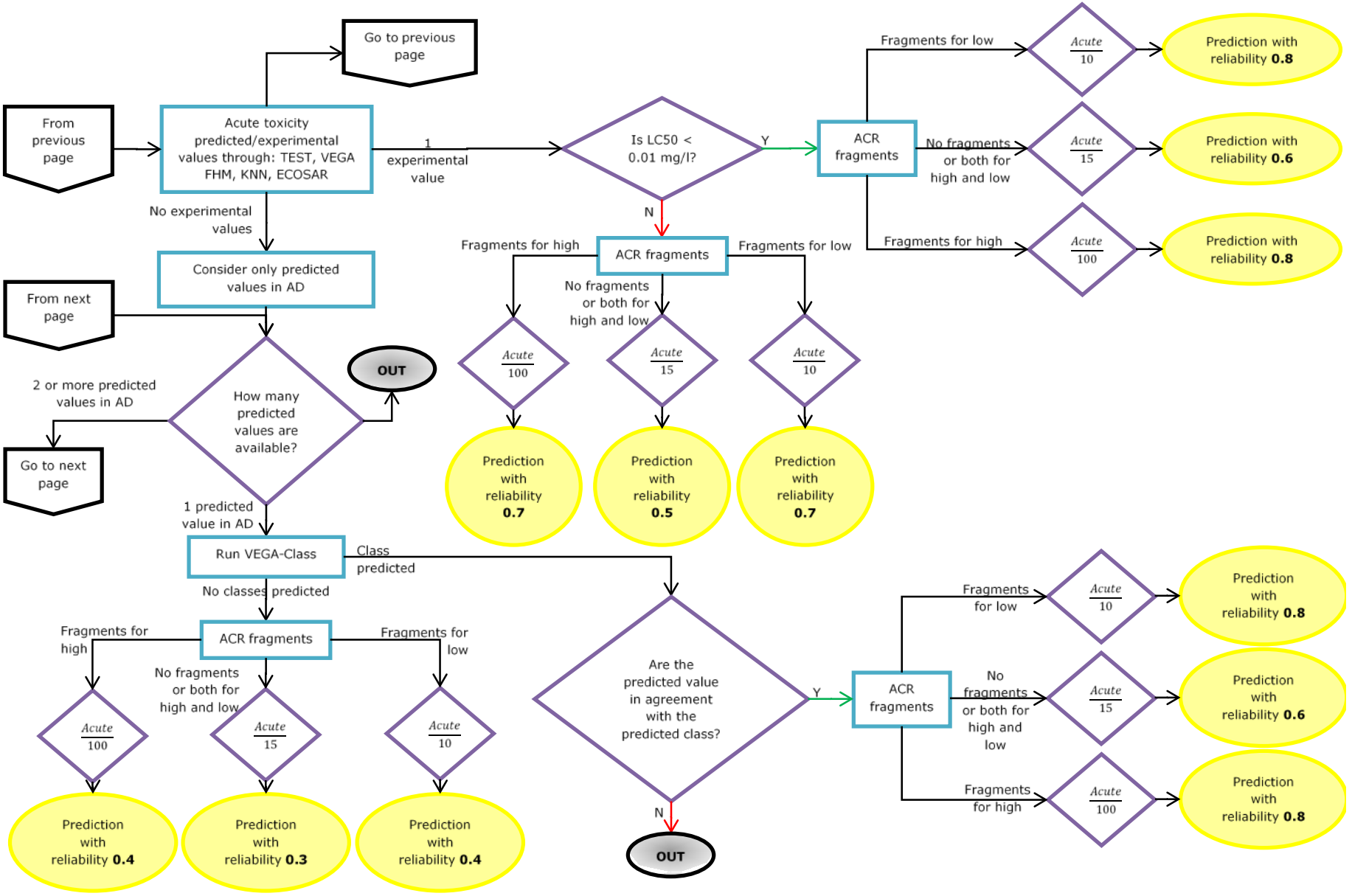
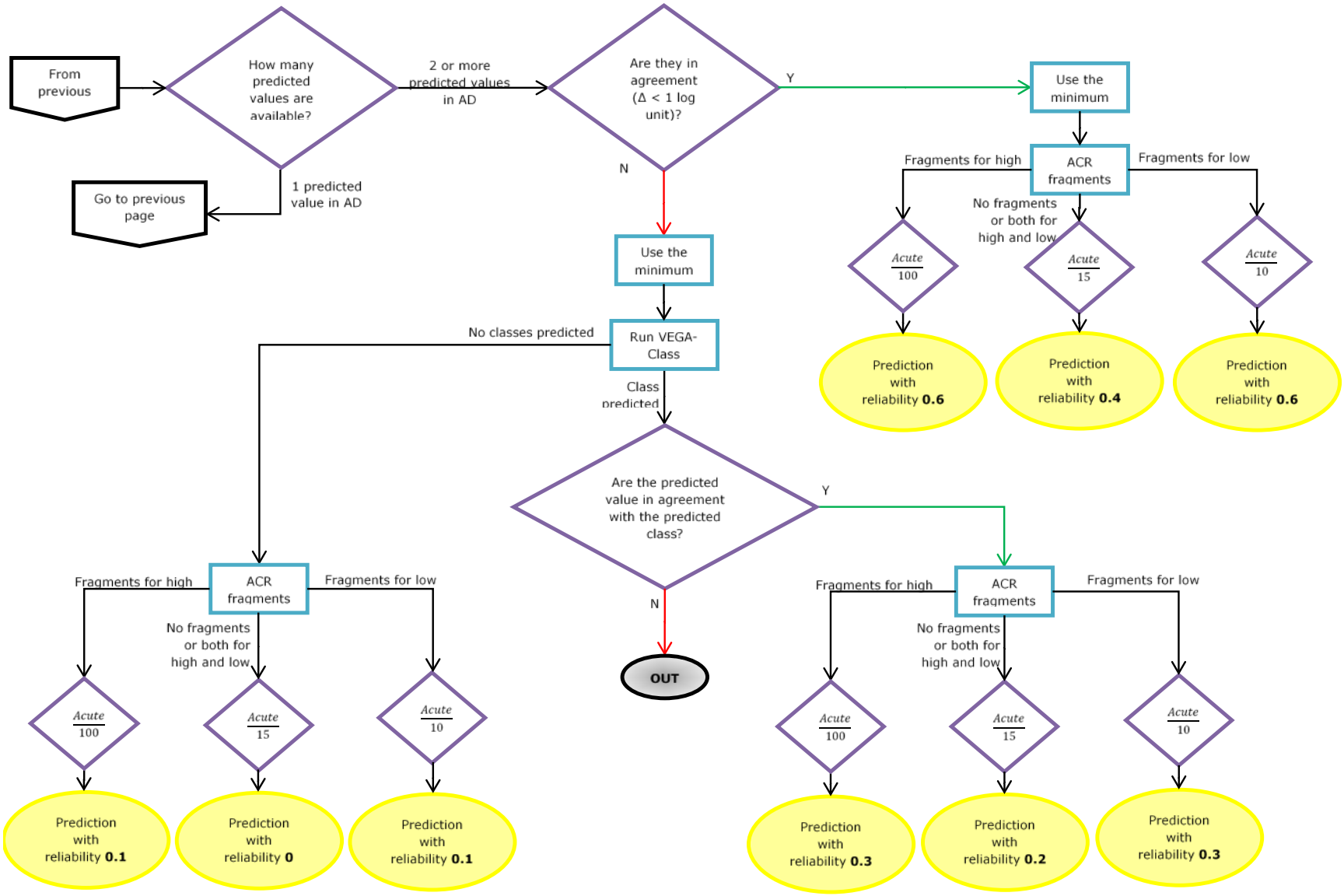


Figure 7: The toxicity workflow (third part)



4 INTEGRATION OF THE WORKFLOWS INTO THE PBT SCORE

This section describes the method used to integrate the predictions (for P, B, and T) of all models used. The validation process, with its results, is described at the end.

4.1 Aggregation process

The starting variables to be used for building the final ranking score are those returned by each of the three workflows, for P, B and T endpoints.

The multiple criteria decision making (MCDM) approach provides methods where a set of input variables are transformed and aggregated. In this work, we decided to add a prior step, in which property value and reliability values for each endpoint are aggregated into a unique value. The overall evaluation for PBT will be based on both these two components: if the substance has been labelled as persistent, bioaccumulative and toxic, but also if the label is “reliable”, on the basis of the uncertainty for each separate label.

Before combining the property value and reliability values, they have to be transformed as required by MCDM in order to obtain a score ranging between 0 and 1, where 1 represents the optimum to be intended as the goal of the ranking approach; indeed in our case it represents the direction of major concern. This is a relevant step especially for the property values, as decisions based on the knowledge of each endpoint and on the considered legal thresholds are part of the choice on how to convert values. For each property we decided to keep the transformed value of 0.5 as an ideal threshold to separate property values going in the direction of major concern (PBT compounds) form those of less concern (non PBT).

Thus, each property has been described with the value from 1 to 0, and the thresholds which are specific for each property have been transferred into the 1 to 0 scale, identifying intermediate values, representing for instance, P, or vP substances. Below we describe for each property these specific property scores.

After these steps, the result consists in three unique scores for P, B and T, which have to be combined using a MCDM technique. After testing different possible approaches, we decided to implement a desirability index, as explained below.

This index represents the global ranking score used to rank compounds. High values (towards 1) are obtained for compounds having a P, B and T assessment performed with good reliability, so they should be the compounds of major concern. On the other side, low values (towards 0) are obtained for compounds with a non-P, non-B and non-T assessment performed with good reliability. Given the algorithm for the calculation of this index, it should be remarked that values falling in the middles of the range (around 0.5) could be obtained for different reasons: indeed, such values could be obtained from (A) compounds for which the properties have reliable predictions values ranging around the thresholds of concerns (for example, B predictions around the 3.3 log units); or (B) compounds with some properties values of concern but with a low reliability.

For an automatic quick evaluation, the 0.5 value of the overall score can be used as a pragmatic threshold: chemicals with a value lower than this threshold could be considered safe.

We notice that the driving factor of the PROMETHEUS project is to tool to prioritize substances, in order to identify substances which are (very) likely to be PBT. For this purpose we tuned to system in this direction. Other purposes may have produced a different strategy. This refers in particular to the fact that substances at the top of the ranking list should very probably have PBT-properties: Wrong results of the screening would result in a loss of time and resources, e.g. by requiring experimental tests on substances which do not have PBT-properties.

Expressed in jargon of the modeller, the software should minimize false positives. We notice that when models are developed for other purposes, they can be optimized in other ways. For instance, it is common

that the in silico model should minimize false negatives (not positives) when used to predict toxic effects of chemicals to be assessed for regulatory purposes.

We also notice that this strategy allows anyhow to introduce a kind of flexibility, using the threshold of the PBT score, and also relatively to the uncertainty of the predictions. Indeed, if the user wants to get a higher number of potential PBT-substances, it can be done proceeding down in the list of the order substances. If the user wants to include also potential PBT-chemicals with higher uncertainty, in order to avoid false negatives, this is also possible by changing the weights assigned to the uncertainty in the PBT-score. We also notice that the uncertainty of the separate P, B and T properties values are reported by the program.

In the future, the software which can be implemented in a user-friendly platform may offer the possibility to tune the output, for instance listing in top the substances also with a high uncertainty, which goes in the direction of a conservative strategy.

Following, the steps used in the program are explained with more details.

4.1.1 Conversion of the Persistence (P) classes into numerical values

The output produced by the P workflow consists of four possible classes (nP, nP/P, P/vP and vP). These classes have been transformed using values between 0 and 1 to separate them without any mathematical function, with the following scheme (table 10):

Table 10: Assignment of numerical values to categories of persistence.

Class	Value
nP	0.3
nP/P	0.6
P/vP	0.8
vP	1.0
unknown	0.5

As mentioned, in this conversion of the property class we decided to keep the value of 0.5 as a threshold to separate concerning and non-concerning assessments.

An additional value has been set for the “unknown”, which is the output of the workflow when no prediction can be provided. For these missing values, the value of 0.5 has been chosen as it represents a situation where it is not possible to decide if the P property should be or not of concern.

4.1.2 Normalization of the Bioaccumulation (B) values

The output produced by the B workflow consists in the predicted value of BCF expressed in log units. For this property, two BCF thresholds are of particular interest: 3.3 (log 2000; threshold for B compounds) and 3.7 (log 5000; threshold for vB compounds).

The BCF values are transformed into a score with a mathematical function based on a composite sigmoid (logistic) function, producing an output which is normalized between o and 1:

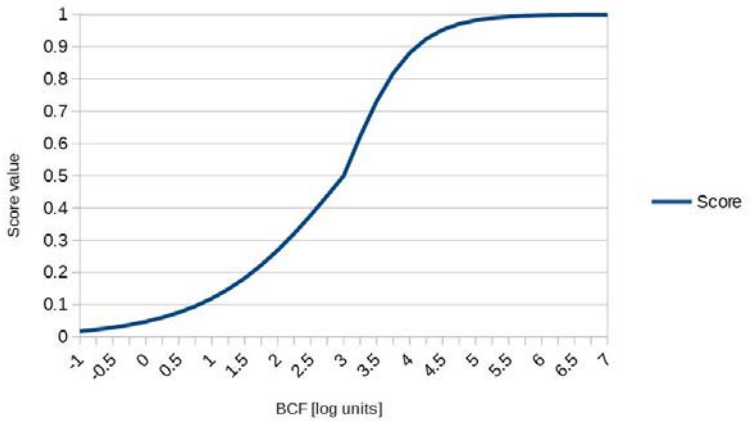
$$Normalized\ BCF\ value = \begin{cases} \dfrac{1}{1+e^{-(BCF-3)}} & for\ BCF < 3.0\log units \\ \dfrac{1}{1+e^{-2(BCF-3)}} & for\ BCF \geq 3.0\log units \end{cases}$$

In this case, the value chosen as a threshold is 3.0 log units (which is transformed to the 0.5 normalized value), also keeping in mind the distribution of the substances in the available collections of values, which indicate a large prevalence of nB substances. Thus, the distribution of the normalized BCF values resulted not too unbalanced towards chemicals with normalized BCF values lower than 0.5.

The mathematical function is composed by the two sigmoid functions above presented, which are used depending if the BCF original value is above the 3 log units or not. We notice that for BCF values higher than the 3.0 threshold the normalized BCF value increases more quickly. This has been done to better discriminate between the two 3.3 and 3.7 BCF thresholds, i.e. between B and vB compounds. Indeed, the score value quickly increases for BCF values between 3.0 and 4.0 log units; for BCF values higher than 4.0, the score goes quickly towards 1, since there is no particular need of discrimination (all compounds are clearly vB).

On the other side, BCF-values lower than 3.0 log units correspond to descending normalized BCF-values. For these substances the normalized BCF value does not decrease as fast as explained before, and under a certain threshold (around 1 log units) all BCF values lower than 1 end with a similar score, with values near 0, as there is no need of particular discrimination (all compounds are clearly nB). Figure 8 shows how this function transforms original BCF-values (with values from -1 to 7 log units) to the corresponding normalized BCF value.

Figure 8: The transformation of the original BCF value into the normalized BCF value for prioritization.



4.1.3 Normalization of the Toxicity (T) values

The output produced by the T workflow consists in the calculated value of fish toxicity expressed in mg/l. This value mainly refers to acute toxicity, but if there is concern for the possible chronic toxicity, the overall toxicity is increased, as presented in the workflow for toxicity.

For this property, the threshold of interest is 0.01 mg/l (where compounds with a toxicity value lower than this threshold are labelled as T).

The toxicity values are transformed into a normalized T value with a sigmoid (logistic) function based:

$$Normalized\ T\ value = 1 - \frac{1}{1 + e^{-(\log(TOX)+1)}}$$

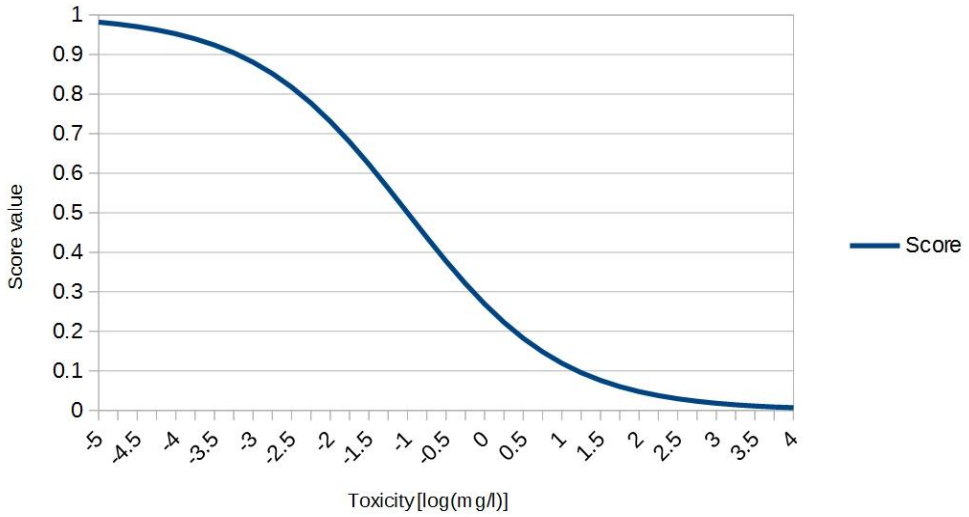
In this case, the value chosen as a threshold is -1 log units, corresponding to 0.1 mg/l (which is transformed to a 0.5 score). In this way, all the values higher than the score value 0.5 represent substances of possible concern, while the threshold at 0.01 corresponds to the score value 0.73. In this case the difference between our -1 log units threshold and the threshold of concern (-2 log units) is slightly more remarked, because compounds with toxicity values around, or lower than, 0.1 mg/l show anyway a toxicity effect, even if they are over the regulatory threshold of 0.01 mg/l.

It should be noted that the transformation function has an inverted optimality: lower values of the property lead to higher score values (as the lower is the mg/l value, the more toxic is the compound). Similarly to the BCF-transformation, the sigmoid function ensures a good discrimination of score values around the threshold of interest, while remarkably higher toxicity values (around or lower 0.001 mg/l) and lower toxicity val-

ues (around or higher than 10 mg/l) are not discriminated anymore and end with similar score values, respectively near 1 and 0.

Figure 9 shows how this function transforms toxicity values (with values from -5 to 4 log units, equal to a range from 0.00001 to 10,000 mg/l) to the corresponding score:

Figure 9: The transformation of the T value into the normalized T value for prioritization.



Like the P-workflow, also the T-workflow can provide some missing values (when unable to perform a prediction). In this case, the default score value of 0.5 is set, as it represents a situation where it is not possible to decide whether the T property should be or not of concern.

4.1.4 Combination of property normalized values with their reliability

As already mentioned, each P, B and T value (obtained with the above explained transformations) is combined with its reliability value to obtain a unique score for each property.

Besides the transformations related to the property values, to get a value ranging from 0 to 1, within each workflow there is the reliability values associated to the property value; also the reliability value ranges from 0 (lowest reliability) to 1 (maximum reliability).

In practice, what we did is something, which can be closely related to the weigh-of-evidence strategy. The weight-of-evidence strategy assesses each value on the basis of the multiple ones, achieving a single value for the final assessment, and the uncertainty of each separate value is taken into account. Typically the weight-of-evidence refers to a single substance and is done manually. The strategy we adopted is very similar, but the main difference is that the assessment is done on many substances together, and in an automatic way, according to the predefined criteria, which are the numerical values above described. This strategy is implemented with the goal of meeting one of the architecture objective, i.e. having compounds reliably predicted as PBT in the top part of the ranking, and those reliably predicted as non-PBT in its bottom part. For this reason, the formula to get the score for a single property is made of two distinct functions:

$$Property\ Score = \begin{Bmatrix} P + (P - 0.5) \cdot (1 - R^{0.5})\ for\ P < 0.5 \\ P - (0.5 - P) \cdot (1 - R^{0.5})\ for\ P \geq 0.5 \end{Bmatrix}$$

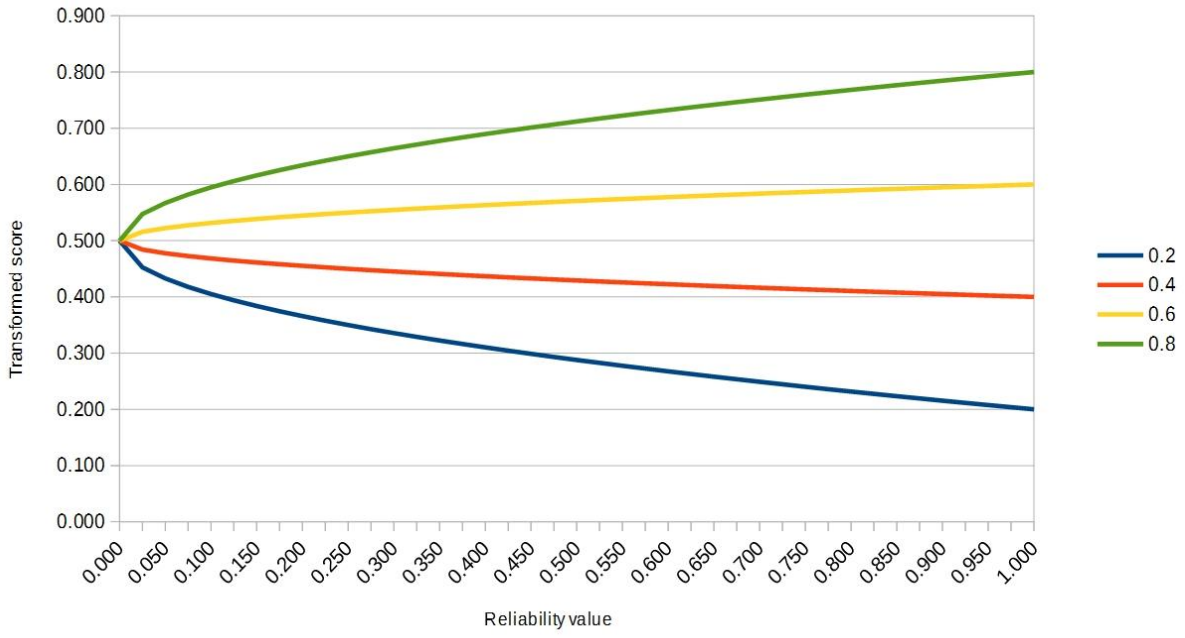
where P is the normalized property score, calculated as explained in the previous sections, and R is the reliability.

This function acts in a different way depending on the starting value of the property's score. Scores above the threshold of 0.5 (thus meaning scores related to compounds of concern) are exponentially lowered when the

related reliability decreases (and thus uncertainty increases). On the other side, scores under 0.5 are exponentially raised when the related reliability decrease. This behaviour is consistent with the target of having high scores for certain PBT compounds, and low scores for certain non-PBT compounds; conversely, all compounds with high level of uncertainty in prediction will tend towards the middle threshold of 0.5. This choice has been done, as we explained before, on the basis of the specific task assigned to the PROMETHUES project by UBA: to minimize false positives, i.e. to be sure as much as possible that the substances in top of the list have PBT-properties. If we allow chemicals with higher uncertainty to merge in the list we will increase the number of false positives. Of course, different choices can be made depending on the specific purpose. In a future platform implementing the program here developed, it is easily feasibility to introduce the possibility for the user to assign no influence to the uncertainty, and thus to avoid the use of the reliability filter. This solution obeys to a different purpose, to identify all possible PBT substances, which is different from the purpose to identify the substances, which "surely" are PBT-substances. We notice that the even the current program of the PBT-score is transparent, and the user of the current program can see separately the value related to each normalized property, and the value of the uncertainty.

Figure 10 reports an example of how the final score (on the Y axis) changes depending on the reliability value (X axis) for four example property's value (0.2, 0.4, 0.6 and 0.8). It is clear that with a low reliability (and thus with high uncertainty) all values tend to the "grey" area in the middle, towards the 0.5 value, while with higher reliability we can distinguish substances on the basis of the property value.

Figure 10: The representation of the overall PBT score depending on the reliability value; four cases are shown.



4.1.5 Calculation of the PBT score

The final step consists of the combination of the three property scores into a unique ranking score to be used for the final ranking. After testing several approaches, coming from MCDM theory, the chosen approach has been what in jargon is call desirability index, and has as general formula:

$$DES = x_1^{w_1} \cdot x_2^{w_2} \cdot \ldots \cdot x_R^{w_R}$$

where x_i represent the i-th input variable, and w_i their weights, i = 1, 2, .., R.

In our architecture, the global desirability has been used with the three global property scores as input, and two weighting scheme has been designed and tested:

$$DES_{PBT} = P^{0.4} \cdot B^{0.4} \cdot T^{0.2}$$

$$DES_{vPvB} = P^{0.5} \cdot B^{0.5}$$

In the first case, the ranking PBT score is calculated taking into account the P, B and T properties with different weights: P and B have an equal weight, while T has a lower weight. This choice follows the indication from UBA, on the basis of the main interest on the P and B properties, also related to the much higher uncertainty associated to the output of the predictive in silico models for aquatic toxicity. Indeed, for aquatic toxicity the best models available are those for fish acute toxicity (compared for instance to those for daphnia), and the robustness of models for chronic aquatic toxicity is even higher (also for the much lower number of experimental data available). But the quality of the best available fish acute toxicity models is still relatively poor. Thus, the T-component in the overall PBT-assessment is relatively more uncertain. Our scheme reflect this assigning a lower weight to the T-component. In the future, if new models would have better results in prediction, the score can be immediately changed.

In the second case, the toxicity prediction is not considered at all, and only P- and B-properties are used, both with an equal weight. This scheme fulfils the request of an alternate ranking system oriented towards vPvB-screening.

The PBT- and vPvB-scores are transparent and thus interpretable, as we explained. There is another useful characteristic to our strategy: It is flexible. What has been developed is the *general* framework and based on this simpler situations can be easily adopted. A specific sub-case which can be preferred by the user is the case of the identification of all the possible PBT-substances. This is a special sub-case of our more general structure. Indeed, it can be easily implemented avoiding the value associated to the reliability. Another possibility is to use specific thresholds for the P-, B- or T-scores as desired by the user, for instance the value 3.3 or 3.7 for bioaccumulation. These are single values, which can be easily implemented as special case of our strategy again: indeed our strategy uses the continuous value (thus all the values) and the specific 3.7 value is a special, limited case in the range of the value. We notice that the opposite would not be possible: having a program working on a few specific thresholds does not allow to switch to a program able to handle all the values. This flexibility allows in the future platform to introduce these special sub-cases, to be chosen by the user through buttons.

4.2 Validation test

After the integration process, it was necessary to carry out a validation test to check the ability of the new platform to differentiate compounds labelled as PBT-compounds from those that are non-PBT.

For this purpose we built up a set of chemicals that contains molecules labelled as PBT and non-PBT obtained from the literature and authorities.

4.2.1 Validation set

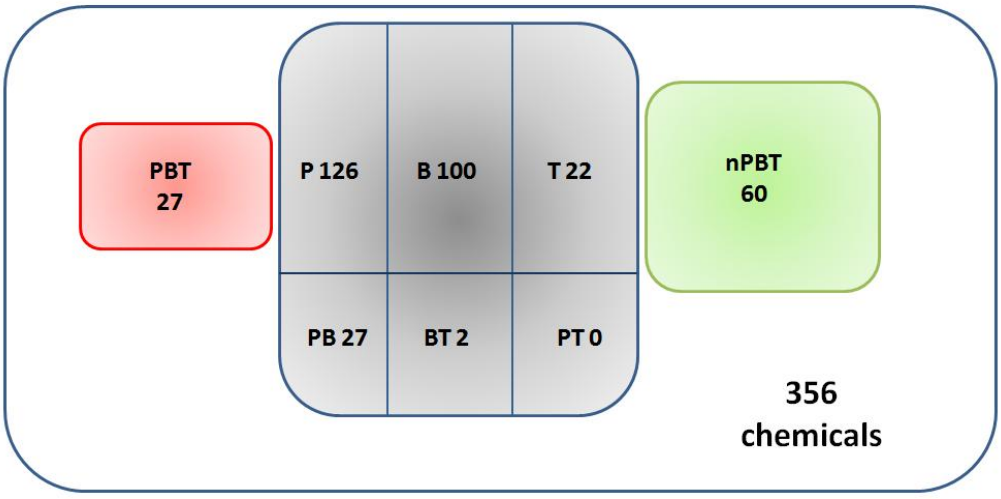
To test the predicting ability of the model for PBT-prioritization we prepared a dataset for validation, extracted from multiple sources of a total of 1875 molecules with experimental data on P, B and / or T. The data about bioaccumulation expressed as log BCF (860 compounds) belong to the Read-Across dataset of VEGA, while the data about the persistence in all the compartments (P_{water} 351, $P_{sediment}$ 297, P_{soil} 568 compounds) are derived from multiple sources available in literature [which consist of Gouin et al., 2004 and Gramatica and Papa, 2007; RIVM (Linders et al., 1994); and USGS (Prioritizing Pesticide Compounds for Analytical Methods Development, 2012)]. Data about fish chronic toxicity (91 compounds) were extracted from the ECOTOX database of the OECD QSAR toolbox. This starting dataset also includes 27 chemicals evaluated as PBT from the REACH Candidate List of ECHA and UBA lists. Table 11 summarizes the information about the data and their sources.

Table 11: Information about sources of the starting dataset

#	Endpoint	Source
27	PBT	Candidate List; UBA list
869	Log BCF	Read-across dataset (VEGA)
351 297 568	P_{water} $P_{sediment}$ P_{soil}	Gouin et al., 2004 and Gramatica and Papa, 2007; RIVM (Linders et al., 1994); and USGS (Prioritizing Pesticide Compounds for Analytical Methods Development, 2012)
91	Fish Chronic Tox	OECD QSAR toolbox (ECOTOX database)

From these 1875 chemicals, we selected a total of 356 molecules in order to obtain a dataset for check of the PBT scheme as described below (Figure 11).

Figure 11: The validation dataset



All the compounds evaluated as PBT in the Candidate List (27) were included in this validation set and we assigned them the label of PBT.

Some chemicals of the starting set are surely not PBT; this is because they are not B with a log BCF<3.3; not P in any compartment (water, soil or sediment), and not T because their chronic toxicity is > 0.01 mg/L. From these compounds, that have a negative feedback for all the endpoints (P, B and T) concurrently, we selected 60 with experimental log BCF <1, labelling them as not PBT.

After the selection because chemicals have experimental values for all P, B and T, we choose chemicals with values of two endpoints only, as described in Table 12.

Table 12: Chemicals with data about two endpoints only

Chemicals	Endpoints
27	P and B
2	B and T
0	P and T

Finally, we also used the chemicals with experimental values related to only one endpoint (P or B or T), as described in Table 13:

Table 13: Chemicals with data about only one endpoint

Chemicals	Endpoints

126	P
100	B
22	T

In details, for the endpoint P, we can also differentiate the chemicals in the set as not P, P/vP and vP.

In conclusion of what has been described above, we summarize the construction of this validation dataset for the model for PBT-prioritization in the Table 14:

Table 14: Construction dataset summary

Chemicals	Features
27	Fulfil PBT criteria
60	Fulfil these conditions simultaneously * nP_{water}; $nP_{sediment}$; nP_{soil} ** log BCF < 1 *** Fish Chronic Tox > 0,01
269	Lack of data for one or more endpoints

4.2.2 Results

The platform is able to assign a ranking close to 1 to all the compounds officially labelled as PBT-substances. Similarly, it is able to assign a ranking close to zero to all those compounds that are eligible to be labelled as "surely" non–PBT. Annex 1 contains all the chemicals, with their values.

Intermediated values of ranking are assigned to those molecules with data lacking for one or will more end-points, defined as "grey-zone".

A qualitative analysis of the results of the ranking from the PBT-score has been performed on the validation set.

Starting from the compounds with highest ranking, it can be seen that the molecules in the top of the list share similar chemical features and belong to congeneric classes of well-known chemicals for which PBT behaviour is documented.

Considering the first 30 molecules, they are all polychlorinated aromatic compounds and several of them are polychloro-biphenyles (PCBs). In the top ten there are well known chlorinated pesticides with condensed rings such as Aldrin, Dieldrin and Chlordane isomers, and with single ring such as hexachlorobenzene (HCB), all included in the list of POPs (Persistent Organic Pollutants – Stockholm Convention 2001). All of them are experimentally vP in all three compartments. Indeed their P prediction always results vP with max-imum reliability. For most of them, the experimental BCF value is available, and they are over the vB-threshold. The T-assessment shows different degrees of reliability, but always towards values of concern, even though no specific experimental data were available for acute aquatic toxicity for these compounds.

Following these molecules, several other compounds are ranked as substances of concerns and share the same chemical features of the above-mentioned chemicals. Examples are CAS number 56348-72-2, belong-ing to the polychlorinated-diphenylethers family, and CAS number 33423-92-6 (1,3,6,8-TCDD) that belongs to the family of tetrachlorodibenzo-p-dioxins. In particular, it is the dioxin with the highest ranking among those in our data set. The other chlorinated dioxins are further down in the list, for their T values and their very low reliability, such as CAS number 35822-46-9 that is predicted as vP and B with high reliability (0.9), but the reliability associated to T value is 0.1.

The first molecule not being polychlorinated is ranked with the 33th position, and it is one of the compounds provided by UBA and labelled as PBT/vPvB. It is a PAH, and it is ranked in this position because, despite

the reliable vP-prediction and the high toxicity prediction, it has a predicted BCF value (of high reliability) of 3.56 and an experimental average value of 3.24. Indeed, while it is surely ranked as a substance of concern, it is ranked after all the other polychlorinated compounds, which clearly show higher BCF-value.

Also, the other compounds provided by UBA and labelled as PBT/vPvB all belong to the PAH class. They are found in different ranking position, even though they all fall in the upper half of the ranking list (score > 0.5). Some of these compounds have nevertheless low BCF-predictions that explain why they are not in higher positions. For example, the lowest BCF (2.24) is found for the chemical with CAS number 218-01-9 (Chrysene). This value is strongly reliable, as it is based on the experimental data found in the VEGA CAESAR and Read-Across models.

Table 15 shows the molecules labelled as PBT by UBA and their experimental BCF-data, as arithmetic mean values, found into the VEGA database, with their related sources, including the EURAS database.

Table 15: BCF experimental values and their sources for PBTs labelled by UBA

CAS no.	LogBCF mean v.	multiple values	Number of. Data	Sources
56-55-3	2.54	2.54	1	Dimitrov*
207-08-9	3.33 (predicted)	2.94 3.70 3.35	3	CAESAR (VEGA) MEYLAN (VEGA) Read-Across (VEGA)
191-24-2	2.71 (predicted)	2.79 4.04 2.64	3	CAESAR (VEGA) MEYLAN (VEGA) Read-Across (VEGA)
50-32-8	2.73	2.68 2.93 2.58 2.73	4	Dimitrov Arnot Arnot Arnot
218-01-9	2.24	2.24	1	Dimitrov
53-70-3	2.80	2.80	1	Dimitrov
206-44-0	3.24	2.71 4.17	2	Dimitrov EURAS
85-01-8	3.40	3.01 3.21 2.85 3.30 3.71 3.62 3.57 3.52 3.50 3.36 3.30 3.18 3.15 3.11	14	Dimitrov Arnot Arnot EURAS EURAS EURAS EURAS EURAS EURAS EURAS EURAS EURAS EURAS EURAS
129-00-0	2.83	2.56 1.72 1.70	8	Dimitrov; Arnot Arnot

		2.99		EURAS
		3.41		EURAS
		3.36		EURAS
		3.15		EURAS
		3.08		EURAS

The compounds that have been added from the EC list of possible PBT or vPvB are also mostly correctly found in the upper half of the ranking list, starting from chemicals CAS number 3846-71-7, 120-12-7 and 1163-19-5. All of them belong to the polybrominated-diphenylethers (PBDEs) family used as flame retardants and well known for their adverse effects to both humans and environment, so restricted under the Stockholm Convention.

Other compounds from this list do not achieve a high total score, for different reasons. For some chemicals, the available models fail to provide a prediction: for CAS number 56-35-9 bis(tributyltin)-oxide (TBTO), because it is an organo-metal, even if it is considered a severe marine pollutant and SVHC (Substance of Very High Concern) for the EU, massively used as marine anti-biofouling agent.

For CAS number 138257-19-9, 678970-15-5, 138257-18-8, 678970-17-7 and 169102-57-2, representing part of the well know flame-retardant family of hexa-bromocyclododecanes (HBCDDs) and also labelled as POPs, the persistence workflow is not able to provide an assessment. These compounds are not predicted by P models since similar molecules are not present in the kNN model training sets. Moreover no SAs or chemical classes were identified for this category of compounds. Only for sediment, one chemical class (for vP) is available, but, following the workflow criteria, it is not enough to generate a final reliable prediction.

Regarding the compounds that achieved a total ranking around the value of 0.5, thus being in the middle of the ranking list, some of them are reported as examples of the meaning of their ranking. Indeed, as already mentioned, this ranking can be achieved in two cases: (1) for compounds that have a reliable set of predictions, with property values for P, B and T below the thresholds of concerns, but not so low to be classified as "safe" compounds, and (2) for compounds that have one or more properties predicted with low reliability, thus they cannot be positioned at the top nor at the bottom of the list due to the uncertainty in their evaluation.

For instance, in the second category there is the compound CAS number 129-43-1, which is 1-hydroxy anthraquinone. It has a total score of 0.515; in this case, the log BCF-value is 2.37. VEGA BCF-models use experimental values, so the value assigned is of the highest reliability. However, it is predicted to be vP but with a very low reliability degree. So, unless a further expert-based assessment is performed to clearly identify its risk of being persistent, this compound falls in the middle of the ranking as it is not possible to determine whether it is a PBT or a surely safe compound.

In the first category, with chemicals with more reliable property values, there is the compound CAS number 292-64-8, an aliphatic cycloalkane. It has a score of 0.505 and its BCF-score and persistence prediction have a reasonably good degree of reliability (respectively 0.85 and 0.7), but the respective values indicate a low hazard (log BCF = 2.69 and a nP/P classification). Thus, while the predictions are reliable, the P and B properties values for this compounds are both far from the thresholds of concern and from the "safe" area, and for this reason this compound falls in the middle of the ranking.

Coming to the bottom of the ranking list, the compounds that achieved lowest scores are as expected those having at least an nP classification or very low BCF values with the highest reliability (which means that some experimental values were available, and that these data were concordant between them and with the models predictions). Indeed, the definition of a non-PBT compound implies that at least it is non-P, non-B or non-T. It should be remarked again that in the final adopted scheme the T-property has a lower weight than P and B, because the assessment of toxicity is more complex, and thus it can be achieved with lower reliability

at the moment, as already explained; thus also the meaning of low total scores is related mainly only to non-P and non-B.

For instance the lowest ranked compound is CAS number 920-66-1 hexafluoro-isopropanol (HFIP), extremely polar and for this reason it is used as a solvent in polymer industry and more generally in organic synthesis. HFIP has one of the lowest BCF-values assessed with reliability equal to 1. This high reliability indicates that the compound has experimental values. In fact, all three BCF models within VEGA (CAESAR, Meylan, Read-Across) have an experimental value for this compound. The Read-Across model has an experimental value equal to log BCF = 0.24 (ADI=0.815). CAESAR and Meylan models have experimental values, log BCF = 0.3 (ADI=1) and log BCF = 0.4 (ADI=1) respectively. Thus, even if the nP classification is provided with low degree of reliability, this substance achieves an overall PBT-score of 0.161.

The second lowest ranked compound, CAS number 141-78-6, ethyl acetate ester, has a slightly higher score of 0.173, and has a certain nP classification but its BCF prediction of log BCF = 0.38 is not so reliable as the one of the previous compound. Esters are widespread in nature and are widely used in industry; their main characteristic is the carboxyl center (RO(C=O)R'), with organic group R in substitution of the hydrogen of carboxyl acids. Their electrophilic functional group reacts with nucleophiles, while the C-H bonds adjacent to the carboxyl are weakly acidic and undergo deprotonation with strong bases. These chemical properties confirm a high reactivity in substitutions, hydrolysis and condensations of esters, which explains their nP and nB classification.

All the compounds that have been added to the validation list with a "non PBT" label (as they had at least on experimentally known property needed for such labelling) are correctly found in the lower half (score < 0.5) of the ranking list.

For instance, the whole class of ketones (CAS number 591-78-6, 107-87-9, 96-22-0, 78-93-3, 67-64-1) are found in the lowest part of the ranking. Ketones are organic compounds that have as characteristic the carbonyl group (C=O), which is polar as a consequence of the fact that the electronegativity of the oxygen is greater than that for carbon. Because the carbonyl group interacts with water by hydrogen bonding, ketones are typically more soluble in water than the related methylene compounds.

As a conclusion, we can say that the integrated model for PBT (vPvB) prioritization separates successfully PBT- vs non-PBT compounds. As expected, the best predictions are achieved with chlorinated aromatic molecules in the top of the list, and in the end of the list with polar and soluble molecules, such as carbonyls, alcohols and esters. These results were obtained thanks to the large number of data with good quality related to these chemical classes of molecules, that we used building the models. The major problem was the lack of available data on fish toxicity and the related low reliability for predicted values. This situation reduces the ranking of same molecules that are possible toxicants.

This overall performance should be seen as a starting point; the platform can be refined with the integration of new models and the implementation within VEGA.

More levels of improvement can be achieved taking into consideration endpoints as endocrine disruption and human toxicity. Models for endocrine disruption and human toxicity (CMR) are already available within VEGA.

In the future, the implementation work could converge in the creation of a program for prioritization of substances to be evaluated for PBT assessment, freely available for the users.

5 References

Ahlers J, Riedhammer C, Vogliano M, Ebert RU, Kühne R, Schüürmann G, 2006. Acute to Chronic Ratios in Aquatic Toxicity-Variation Across Trophic Levels and Relationship with Chemical Structure. Environmental Toxicology and Chemistry, 25: 2937-2945.

Amaury N, Benfenati E, Boriani E, Casalegno M, Chana A, Chaudhry Q, Chretien J R, Cotterill J, Lemke F, Piclin N, Pintore M, Porcelli C, Price N, Roncaglioni A, Toropov A A, 2007. Results of DEMETRA models. in: Benfenati E (Ed.), Quantitative Structure-Activity Relationships (QSAR) for Pesticide Regulatory Purposes, Elsevier Science Ltd, Amsterdam, The Netherlands, 201-281.

ANTARES project: http://www.antares-life.eu/

Arnot JA, Gobas FAPC, 2006. A review of bioconcentration factor (BCF) and bioaccumulation factor (BAF) assessments for organic chemicals in aquatic organisms. Environmental Reviews 14: 257-297.

Aquatic OASIS: http://www.oasis-lmc.org

Aquatic ECETOC: http://www.ecetoc.org/index.php?mact=MCSoap,cntnt01,details,0&cntnt01by_category=5

Aquatic Japan MoE: http://www.safe.nite.go.jp/english/db.html

Bao J, Liu W, Liu L, Jin Y, Dai J, Ran X, Zhang Z, Tsuda S, 2011. Perfluorinated compounds in the environment and the blood of residents living near fluorochemical plants in Fuxin, China. Environmental Science and Technology, 45: 8075-8080.

Benfenati E, Manganaro A, Gini G, 2013. VEGA-QSAR: AI inside a platform for predictive toxicology. Proceedings of the workshop "Populari.ze Artificial Intelligence 2013", December 5th 2013, Turin, Italy, Published on CEUR Workshop Proceedings Vol. 1107

Benfenati E, Pardoe S, Martin T, Gonella Diaza R, Lombardo A, Manganaro A, Gissi A, 2013. Using toxicological evidence from QSAR models in practice. Altex 30: 19-40.

Boriani E, Mariani A, Baderna D, Moretti C, Lodi M, Benfenati E, 2010. ERICA: a multiparametric toxicological risk index for the assessment of environmental healthiness. Environment International 36 : 665-674.

CALEIDOS project: http://www.caleidos-life.eu/pages/project.php

Cappelli CI, Manganelli S, Lombardo A, Gissi A, Benfenati E, 2015. Validation of quantitative structure–activity relationship models to predict water-solubility of organic compounds. Science of The Total Environment 463-464: 781-789.

Cassano A, Manganaro A, Martin T, Young D, Piclin N, Pintore M, Bigoni D, Benfenati E, 2010. CAESAR models for developmental toxicity. Chemistry Central Journal 4: S4.

CEFIC: http://www.cefic-lri.org/lri-toolbox/bcf

CEMC Report No. 200703. 2007. RISK PRIORITIZATION FOR A SUBSET OF DOMESTIC SUBSTANCES LIST CHEMICALS USING THE RAIDAR MODEL FINAL REPORT Prepared by: Jon Arnot and Don Mackay Canadian Environmental Modelling Centre Trent University Prepared for: Environment Canada Scientific authorities: Danaëlle Delage and Don Gutzman.

ChemSpider: http://www.chemspider.com/

Cheng F, Ikenaga Y, Zhou Y, Yu Y, Li W, Shen J, Du Z, Chen L, Xu C, Liu G, Lee PW, Tang Y, 2012. In Silico Assessment of Chemical Biodegradability. Journal of Chemical Information and Modeling 52: 655-669

[CEC] Commission of the European Communities, 1996. Technical guidance documents in support of the Commission Directive 93/67 EEC on risk assessment for new substances and the Commission Regulation (EC) no. 1488/94 on risk assessment for existing substances. Parts I through IV. Brussels: Commission of the European Communities.

[CLP] REGULATION (EC) No 1272/2008 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 16 December 2008 on classification, labelling and packaging of substances and mixtures, amending and repealing Directives 67/548/EEC and 1999/45/EC, and amending Regulation (EC) No 1907/2006

DEMETRA project: http://vega.marionegri.it/wordpress/resources/qsar-in-silico-tools

Dimitrov S, Dimitrova N, Parkerton T, Comber M, Bonnell M, Mekenyan O, 2005. Base-line model for identifying the bioaccumulation potential of chemicals. SAR QSAR in Environmental Research 16: 531-554.

ECHA CHEM: http://www.echemportal.org/

Echem portal: http://www.echemportal.org/

ECOTOX: http://cfpub.epa.gov/ecotox/

EURAS Database: http://www.cefic-lri.org/lri-toolbox/bcf

European Commission, 2012. Guidance on information requirements and chemical safety assessment - European Chemicals Agency. Chapter R.7a: Endpoint specific guidance.

European Commission, 2014. Guidance on information requirements and chemical safety assessment - European Chemicals Agency. Chapter R.11: Endpoint specific guidance.

Ferrari T, Cattaneo D, Gini G, Golbamaki Bakthyari N, Manganaro A, Benfenati E, 2013. Automatic knowledge extraction from chemical structures: the case of mutagenicity prediction. SAR and QSAR in Environmental Research 25: 365-383.

Fujii S, Polprasert C, Tanaka S, Pham Hong Lien N, Qiu Y, 2007. New POPs in the water environment: distribution, bioaccumulation and treatment of perfluorinated compounds-a review paper. Journal of Water Supply: Research and Technology-AQUA 56: 313-326.

Ghose AK, Crippen GM, 1986. Atomic Physicochemical Parameters for Three-Dimensional Structure-Directed Quantitative Structure-Activity Relationships I. Partition Coefficients as a Measure of Hydrophobicity. Journal of Computational Chemistry 7: 565-577.

Ghose AK, Viswanadhan VN, Wendoloski JJ, 1998. Prediction of Hydrophobic (Lipophilic) Properties of Small Organic Molecules Using Fragmental Methods: An Analysis of ALOGP and CLOGP Methods. Journal of Physical Chemistry 102: 3762-3772.

Gouin T, Cousin I, Mackay D, (2004): Comparison of two methods for obtaining degradation half-lives. Chemosphere 56: 531-535.

Gramatica P, Papa E, 2007. Screening and ranking of POPs for Global Half-Life: QSAR approaches for prioritization based on molecular structure. Environmental Science and Technology 41: 2833-2839.

Karrman A, van Bavel B, Jarnberg U, Hardell L, Lindstrom G, 2006. Perfluorinated chemicals in relation to other persistent organic pollutants in human blood. Chemosphere 64: 1582-1591.

Klimisch HJ, Andreae M, Tillmann U, 1997. A systematic approach for evaluating the quality of experimental toxicological and ecotoxicological data. Regulatory Toxicology and Pharamcology 25: 1-5.

Lombardo A, Pizzo F, Benfenati E, Manganaro A, Ferrari T, Gini G, 2014. A new in silico classification model for ready biodegradability, based on molecular fragments. Chemosphere 108: 10-16.

Linders JBHJ, Jansma JW, Mensink BJWG, Otermann K, 1994. Pesticedes: Beneaction or Pandora's box? A synopsis of the environmental aspects or 243 pesticides. RIVM Report 679101014.

Manganaro A, Ballabio D, Consonni V, Mauri A, Pavan M, Todeschini R, 2008 . The DART (decision analysis by ranking techniques) software. in Scientific data ranking methods: theory and applications, ISBN: 978-0-444-53020-2, Elsevier, 193-207.

Maran U, Sild S, Mazzatorta P, Casalegno M, Benfenati E, Romberg M, 2007. Grid Computing for the Estimation of Toxicity: Acute Toxicity on Fathead Minnow (Pimephales promelas). Chapter in: Distributed, High-Performance and Grid Computing in Computational Biology Volume 4360 of the series Lecture Notes in Computer Science pp 60-74.

May and Hahn (2014) Comparison of species sensitivity of Daphnia and fish in acute and chronic testing. Project No. 27448

Meylan, W.M.; Howard, P.H. (1994): Validation of Water Solubility Estimation Methods Using Log Kow for Application in PCGEMS & EPI (Sept 1994, Final Report).

Meylan WM, Howard PH, 1995. Atom/fragment contribution method for estimating octanol-water partition coefficients. Journal of Pharmacolgy Sciences 84: 83-92.

Moody CA, Field JA, 2000. Perfluorinated surfactants and the environmental implications of their use in fire-fighting foams. Environmental Science and Technology 34: 3864-3870.

Moriguchi, I. Hirono S, Nakagome I, Matsushita Y,1992. Simple method of calculating octanol/water partition coefficient. Chemical & Pharmaceutical Bulletin 1: 127-130.

Moriguchi, I. Hirono S, Nakagome I, Hirano H, 1994. Comparison of Reliability of Log P Values for Drugs Calculated by Several Methods. Chemical & Pharmaceutical Bulletin 42: 976-978.

OECD QSAR ToolBox: http://www.oecd.org/chemicalsafety/risk-assessment/theoecdqsartoolbox.htm

OECD Guideline for Testing of Chemicals Test No. 203: Fish, Acute Toxicity Test

OECD Guideline for Testing of Chemicals Test No. 210: Fish Early-lige Stage Toxicity Test

OECD Guideline for Testing of Chemicals Test No. 212: Fish, Short-term Toxicity Test on Embryo and Sac-fry Stages

OECD Guideline for Testing of Chemicals 215: Fish, Juvenile Growth Test

OECD Guideline for Testing of Chemicals Test No. 301: Ready Biodegradability

Pavan M, Worth A, 2008. A set of case studies to illustrate the applicability of DART (Decision Analysis by Ranking Techniques) in the ranking of chemicals. European Commission report EUR 23481 EN, Office for Official Publications of the European Communities, Luxembourg, 2008.

Porcelli C, Roncaglioni A, Chana A, Benfenati E, 2008. A comparison of DEMETRA individual QSARs with an index with an index of evaluation of uncertainty. Chemosphere 71: 1845-1852.

REACH 2006. Registration, Evaluation, Authorisation and restriction of Chemicals (REACH) Regulation (EC) No 1907/2006 of the European Parliament and of the Council of 18 December.

RIVM 2011. Report 601356001/2011. Rorije E, Verbruggen EMJ, Hollander A, Traas TP, Janssen, MPM. Identifying potential POP and PBT substances. Development of a new Persistence/Bioaccumulation-score.

Shoib M, Harner T, Vlahos P, 2006. Perfluorinated Chemicals in the Arctic Atmosphere. Environmental Science and Technology 40: 7577-7583.

Su LM, Liu X, Wang Y, Li JJ, Wang XH, Sheng LX, Zhao YH, 2014. The discrimination of excess toxicity from baseline effect: Effect of bioconcentration. Science of The Total Environment 484: 137-145.

Technical Guidance Document on risk assessment. European Commission, 2003

Todeschini R, Consonni V, 2009. Molecular Descriptors for Chemoinformatics. Methods and Principles in Medicinal Chemistry, Volume 41, Edited by Mannhold, Raimund / Kubinyi, Hugo / Folkers, Gerd. ISBN 978-3-527-31852-0 - Wiley-VCH, Weinheim.

US EPA, 2012. Estimation Programs Interface Suite™ for Microsoft® Windows, v 4.11. United States Environmental Protection Agency, Washington, DC, USA.

VEGA website: http://www.vega-qsar.eu/

Viswanadhan VN, Reddy MR, Bacquet RJ, Erion MD, 1993. Assessment of methods used for predicting lipophilicity: Application to nucleosides and nucleoside bases. Journal Of Computational Chemistry 14: 1019-1026.