UNIVERSITAT ROVIRA I VIRGILI

QUANTITATIVE STRUCTURE FATE RELATIONSHIPS FOR MULTIMEDIA ENVIRONMENTAL ANALYSIS

Izacar Jesús Martínez Brito

ISBN: 978-84-693-4597-9
Dipòsit Legal: T.1010-2010

UNIVERSITAT ROVIRA I VIRGILI
QUANTITATIVE STRUCTURE FATE RELATIONSHIPS FOR MULTIMEDIA ENVIRONMENTAL ANALYSIS
Izacar Jesús Martínez Brito
ISBN:978-84-693-4597-9 /DL:T.1010-2010

Izacar Jesús Martínez Brito

QUANTITATIVE STRUCTURE-FATE RELATIONSHIPS FOR MULTIMEDIA ENVIRONMENTAL ANALYSIS

DOCTORAL THESIS

UNIVERSITAT ROVIRA I VIRGILI

Tarragona, Spain
2010

ISBN:978-84-693-4597-9 /DL:T.1010-2010

Izacar Jesús Martínez Brito

QUANTITATIVE STRUCTURE-FATE RELATIONSHIPS FOR MULTIMEDIA ENVIRONMENTAL ANALYSIS

DOCTORAL THESIS

Supervised by

Dr. Jordi Grifoll i Taverna and Dr. Francesc Giralt i Prat

Department of Chemical Engineering

UNIVERSITAT ROVIRA I VIRGILI

Tarragona, Spain
2010

ISBN:978-84-693-4597-9 /DL:T.1010-2010

UNIVERSITAT ROVIRA I VIRGILI

Universitat Rovira i Virgili
Departament d'Enginyeria Química

Campus Sescelades,Av. Països Catalans,
26, 43007, Tarragona, Spain
Tel.: 977559700
Fax: 977559699

We, Prof. Jordi Grifoll i Taverna and Prof. Francesc Giralt i Prat, members of the Department of Chemical Engineering of the Rovira i Virgili University,

CERTIFY:

That the present study, entitled "QUANTITATIVE STRUCTURE-FATE RELATIONSHIPS FOR MULTIMEDIA ENVIRONMENTAL ANALYSIS" presented by Izacar Jesús Martínez Brito, in partial fulfillment of the requirements for the degree of Doctor, has been carried out under our supervision at the Department of Chemical Engineering of this University.

Tarragona, 19 March 2010

Signatures:

Prof. Jordi Grifoll i Taverna Prof. Francesc Giralt i Prat

ISBN:978-84-693-4597-9 /DL:T.1010-2010

Acknowledgements

I would like to thank Prof. Francesc Giralt and Prof. Jordi Grifoll for their advise; Dr. Gabriela Espinosa, Dr. Robert Rallo and Prof. Yoram Cohen for comments, critics and suggestions; and, the European Union for funding this research work throughout the NOMIRACLE project ("NOvel Methods for Integrated Risk Assessment of CumuLative stressors in Europe", European Commission, FP6 Contract No. 003956).

But, above all, I would like to thank my family and friends. You have always been my source of inspiration.

This study is dedicated to my mother, Eldris Lutecia Brito de Martínez, my father, Izacar del Jesús Martínez Berti, my sister, Rosán Eliza Martínez Brito and my grandmother, Rosa Salazar de Brito. To the loving memory of my grandparents María Visitación Berti de Martínez †, Juan de la Cruz Brito † and Angel Rafael Martínez †.

ISBN:978-84-693-4597-9 /DL:T.1010-2010

Resumen

Excepto para contaminantes químicos comunes considerados de prioridad, las propiedades fisicoquímicas clave de un gran espectro de compuestos tienden a ser desconocidas. Esta falta de datos se vuelve crítica si el número de compuestos a monitorear respecto a su distribución ambiental en múltiples medios se incrementa 10 veces más a causa de la adopción de nuevas regulaciones, como la impulsada por REACH en Europa. Para monitorear estos "nuevos" compuestos y decidir si requieren evaluaciones adicionales, muchas de las propiedades fisicoquímicas necesarias deberán ser estimadas por medio de relaciones cuantitativas de estructura y actividad (QSARs), reglas experimentales que relacionan la estructura molecular de los compuestos con actividad química. Por esta razón, dentro del paquete de trabajo 2.4 del proyecto NOMIRACLE, se ha investigado la posibilidad de analizar la distribución o destino en el ambiente de contaminantes químicos usando información molecular y algoritmos de aprendizaje.

Se sabe que las variables de salida de modelos ambientales de múltiples medios (MEMs) se ven afectados no sólo por las premisas del modelo (procesos ambientales, métodos de cálculo, escalas, etc.) sino también por la incertidumbre en sus variables de entrada. Este estudio analiza la posibilidad de evaluar la distribución ambiental de compuestos, expresada como fracciones másicas adimensionales, directamente a partir de su información molecular en vez de usar MEMs con propiedades fisicoquímicas estimadas por QSARs. Con este fin, se han comparado predicciones de la distribución o destino de compuestos en el ambiente generadas por: a) SimpleBox 3, un MEM de nivel III basado en fugacidades, propagando incertidumbres ya reportadas de propiedades fisicoquímicas por medio de un muestreo estadístico (simulaciones de Monte Carlo); y, b) regresiones de vectores soporte (SVRs) actuando como relaciones cuantitativas de propiedad y destino (QPFRs) o como relaciones cuantitativas de estructura y destino (QSFRs), relacionando fracciones másicas con, respectivamente, propiedades fisicoquímicas relevantes o descriptores moleculares de un juego de compuestos de entrenamiento.

Los análisis de este estudio se refieren a 468 compuestos (incluyendo compuestos prioritarios) emitidos hipotéticamente en aire o agua, en un escenario geográfico fijo representando los Países Bajos (Holanda) como un juego de cinco compartimientos (aire, agua, sedimentos, suelo y vegetación). De los 468 compuestos considerados, 375 se han utilizado como compuestos de trabajo, para entrenar y probar modelos QPFR o QSFR. Los 93 compuestos restantes fueron reservados para la validación externa de los modelos.

Los compuestos de entrenamiento y prueba de cada QPFR ó QSFR fueron seleccionados, por medio del algoritmo de mapas autoorganizativos (SOM), a partir del juego de 375 compuestos de trabajo. El SOM se ha utilizado para establecer mapas de los compuestos en un espacio multidimensional conformado por las variables de entrada (propiedades fisicoquímicas o descriptores moleculares) y salida (fracciones másicas) de cada modelo, agrupando compuestos de trabajo que tienen variables de entradas y salida similares en cada una de las unidades del SOM. En el espacio multidimensional de cada modelo, los compuestos de trabajo más cercanos y

ISBN:978-84-693-4597-9 /DL:T.1010-2010

más alejados a cada unidad del SOM conforman el juego de datos de entrenamiento, mientras que los compuestos de trabajo restantes conforman el juego de datos de prueba. El tamaño de cada SOM se ha ajustado para producir una proporción de compuestos de entrenamiento y de prueba de cerca de, respectivamente, 80 % y 20 % el número de compuestos de trabajo disponibles. El SVR de cada QPFR o QSFR se desarrolló con sus compuestos de entrenamiento, mientras que sus parámetros se ajustaron para predecir de la mejor manera el destino de sus compuestos de prueba. Este paso se ha hecho para garantizar que cada modelo sea capaz de predecir tan bien como sea posible las fracciones másicas de compuestos que, sin ser parte del modelo, comparten ciertas similitudes con los compuestos de trabajo. Finalmente, cada modelo fue validado con los 93 compuestos de validación, no utilizados en ninguna fase del desarrollo de los modelos. El comportamiento de cada modelo con respecto a las predicciones del destino de juegos de compuestos se ha medido en términos del coeficiente cuadrático predictivo (q^2) y de la media de errores absolutos (MAE). QPFRs o QSFRs se han considerado óptimos cuando muestran tanto valores altos de q^2 como valores bajos de MAE, no sólo en los juegos de compuestos de entrenamiento y prueba, sin también en el juego de compuestos de validación.

Aunque varios casos fueron considerados en los reportes del proyecto NOMIRACLE, por simplicidad la mayoría de los análisis descritos aquí se realizaron considerando fracciones másicas ambientales en aire y agua, resultantes de emisiones en agua. En general, compartimientos con fracciones másicas muy bajas mostraron los más altos rangos de variación en estas variables, cuando se propagaba la incertidumbre de propiedades fisicoquímicas a lo largo del MEM de referencia, en algunos casos de hasta 12 unidades logarítmicas (para 468 compuestos: los índices de predicción en aire fueron q^2 = 0.87 y MAE = 0.82; mientras, los índices de predicción para agua fueron q^2 = 0.82 y MAE = 0.18). QPFRs usando propiedades clave, coeficientes de partición y constantes de degradación, produjeron predicciones muy certeras (para 468 compuestos: los índices de predicción en aire fueron q^2 = 0.99 y MAE = 0.10; mientras, los índices de predicción para agua fueron q^2 = 0.99 y MAE = 0.06). Sin embargo, dado que la disponibilidad de datos de partición y degradación se restringe a un número limitado de compuestos, la aplicabilidad del método de análisis ambiental basado en QPFRs se restringe también a tales compuestos.

Los modelos QSFR estiman el destino de contaminantes, no utilizados en el desarrollo de estos modelos, a partir de sus descriptores moleculares y no sus propiedades fisicoquímicas. Una gran ventaja, cuando estas últimas se desconocen. QSFRs se desarrollaron usando uno de dos grupos de descriptores moleculares: el primer grupo comprendía peso molecular (MW) y propiedades moleculares estimadas semiempíricamente con la aproximación PM3 de la teoría de orbitales moleculares, el segundo grupo comprendía MW y el número de constituyentes moleculares en cada compuesto (átomos, enlaces, grupos funcionales y anillos). Las mejores predicciones hechas con QSFRs (para 468 compuestos: los índices de predicción en aire fueron q^2 = 0.78 y MAE = 1.01; mientras, los índices de predicción para agua fueron q^2 = 0.80 y MAE = 0.33) se produjeron a partir del segundo grupo de descriptores (MW y el número de constituyentes moleculares). El algoritmo de SVR pudo estimar el destino en el ambiente de nuevos compuestos (de prueba o validación) con una exactitud aceptable, al comparar compuestos respecto a las secciones de cada molécula en vez de hacerlo respecto a propiedades moleculares promedio.

ISBN:978-84-693-4597-9 /DL:T.1010-2010

Para mejorar las predicciones de QSFRs, se investigó el agrupamiento de compuestos en clases para luego desarrollar QSFR específicos para cada clase. Predicciones mejoradas de fracciones másicas resultaron al agrupar compuestos, no con respecto a su degradación en agua (para 468 compuestos: los índices de predicción en aire fueron q^2 = 0.72 y MAE = 1.13; mientras, los índices de predicción para agua fueron q^2 = 0.57 y MAE = 0.31) sino con respecto a su composición molecular (para 468 compuestos: los índices de predicción en aire fueron q^2 = 0.79 y MAE = 0.84; mientras, los índices de predicción para agua fueron q^2 = 0.86 y MAE = 0.16); porque, de las predicciones de clase en el primer caso se obtuvo una rata de verdaderos positivos del 77.4 % y una rata de falsos positivos del 22.6 %, mientras que de las predicciones de clase en el segundo caso fueron inferiores, con una rata de verdaderos positivos del 100.0 % y una rata de falsos positivos del 0.0 %. Los átomos de un compuesto se pueden calcular fácilmente de su formula molecular, mientras que sus propiedades fisicoquímicas son objeto de variación debido a la incertidumbre en procedimientos tanto experimentales como de estimación. Cualquier falla en la predicción de la clase de un nuevo compuesto lleva a su análisis por medio de un QSFR inapropiado, produciendo resultados extremadamente erróneos. Para tener predicciones correctas del destino de un compuesto, éste debe ser analizado con un QSFR perteneciente a la misma clase química.

Se ha estudiado la predicción de compuestos dentro y fuera de los dominios de aplicabilidad de QSFRs específicos, para clases de compuestos con respecto a su composición, en tres casos: Caso I, basados en el SOM e información sobre constituyentes moleculares; Caso II, basados en el SOM y componentes principales de constituyentes moleculares; y Caso III, la intersección de Casos I y II). Se ha demostrado que las fracciones másicas de nuevos compuestos (de prueba y validación) dentro de dominios de aplicabilidad de QSFRs por cada clase (Caso III: los índices de predicción en aire para 48 compuestos fueron: q^2 = 0.92 y MAE = 0.54; mientras, los índices de predicción en agua para 53 compuestos fueron: q^2 = 0.93 y MAE = 0.16) han sido más precisas que aquellas de compuestos fuera de los DOAs (Caso III: los índices de predicción en aire para 120 compuestos fueron: q^2 = 0.59 y MAE = 1.50; mientras, los índices de predicción en agua para 117 compuestos fueron: q^2 = 0.42 y MAE = 0.35). Extendiendo este estudio a emisiones en aire, tendencias similares se obtuvieron al analizar los mismos compuestos dentro de los DOAs (Caso III: los índices de predicción en aire para 48 compuestos fueron: q^2 = 0.94 y MAE = 0.20; mientras, los índices de predicción en agua para 53 compuestos fueron: q^2 = 0.92 y MAE = 0.27) y fuera de los DOAs (Caso III: los índices de predicción en aire para 120 compuestos fueron: q^2 = 0.53 y MAE = 0.66; mientras, los índices de predicción en agua para 117 compuestos fueron: q^2 = 0.61 y MAE = 0.51).

Adicionalmente, se han comparado los índices de predicción en aire y agua. Se ha observado, al emitir compuestos en uno de estos compartimientos, que los mejores índices de predicción se obtuvieron en un solo compartimiento cuando las emisiones ocurrían en él mismo y no otro en compartimiento. Esto se ha confirmado, tanto para el compartimiento de agua (considerando 53 compuestos dentro de los DOAs en el Caso III: para emisiones en agua, los índices de predicción en agua fueron: q^2 = 0.93 y MAE = 0.16; para emisiones en aire, los índices de predicción en agua fueron: q^2 = 0.92 y MAE = 0.27) como para el compartimiento de aire (considerando 48 compuestos dentro de los DOAs en el Caso III: para emisiones en agua, los índices de

ISBN:978-84-693-4597-9 /DL:T.1010-2010

predicción en aire fueron: $q^2 = 0.92$ y MAE = 0.54; para emisiones en aire, los índices de predicción en aire fueron: $q^2 = 0.94$ y MAE = 0.20).

ISBN:978-84-693-4597-9 /DL:T.1010-2010

Summary

Except for common priority chemical pollutants of current concern, environmental key physicochemical properties tend to be unavailable to a wide spectrum of chemicals. This lack of data becomes critical if the number of chemicals to be screened for multimedia exposure increases over ten-fold due to the adoption of regulatory actions such as REACH in Europe. Most of the properties needed to screen these "new" chemicals and decide if they require further evaluation, will most likely have to be estimated from current Quantitative Structure-Activity Relationship (QSAR) models, understood as a set of experimental rules that relate chemical structure to chemical activity. For this reason, within the work package 2.4 of the NOMIRACLE project, research has been carried out to study the feasibility of assessing the environmental fate of chemical pollutants using molecular information and learning algorithms.

It is known that the outputs of Multimedia Environmental Models (MEMs) are affected by not only the assumptions of the model (environmental processes, calculation methods, scales, etc.) but also by the uncertainty in input parameters. This study analyses the prospect of assessing the environmental distribution of chemicals directly from their molecular information, rather than using MEMs with several physicochemical properties estimated from QSARs. To this end, predictions of the environmental distribution or fate of chemicals, expressed in dimensionless compartmental mass ratios, have been compared between: a) SimpleBox 3, a Level III fugacity MEM, propagating reported uncertainty of key physicochemical properties via statistical sampling (i.e., Monte Carlo simulations); and, b) Support Vector Regressions (SVRs) acting as either Quantitative Property-Fate Relationships (QPFRs) or Quantitative Structure-Fate Relationships (QSFRs), linking mass ratios to, respectively, key physicochemical properties or molecular descriptors of a set of training chemicals.

The assessments of this study were referred to 468 chemicals (including priority chemicals) emitted hypothetically in either air or water, in a fixed geographical scenario representing the Netherlands as a set of five compartments (air, water, sediments, soil and vegetation). Out of the 468 chemicals, 375 were used as work chemicals, for training and testing QPFR or QSFR models. The remaining 93 chemicals were reserved for the external validation of the models.

The training and test chemicals of every QPFR or QSFR model were selected, by means of the Self-Organizing Map (SOM) algorithm, from the set of 375 work chemicals. The SOM mapped the chemicals in a multidimensional chemical space conformed by the input variables (properties or molecular descriptors) and target (mass ratio) variables of each model, clustering work chemicals with similar inputs and targets in each of the SOM units. In the multidimensional space of each model, the closest and farthest work chemicals to each SOM unit conform the training data set, while the remaining chemicals conform the test data set. The size of each SOM was adjusted to yield a proportion of training and test chemicals of about, respectively, 80 % and 20 % the number of available work chemicals. The SVR of every QPFR or QSFR was developed with the training chemicals, while its parameters

ISBN:978-84-693-4597-9 /DL:T.1010-2010

were tuned to predict as well as possible the fate of the test chemicals. This step was meant to guarantee that every model was able to predict as much as possible the mass ratios of chemicals that, without being part of the model, share similarities with the training chemicals. Finally, every model was validated with the 93 validation chemicals, not used in at all in the development of the models. The performance of every model predicting mass ratios with respect to a data set was measured in terms of the square predictive coefficient (q^2) and the mean absolute error (MAE). QPFR or QSFR models were considered optimal when showing both high q^2 values and low MAE values, not only on the training and test data sets, but also on the validation set.

Even when various cases were considered within the NOMIRACLE project, for simplicity most of the assessments described here were carried out considering environmental mass ratios in air and water, resulting from emissions in water. In general, compartments with low mass ratios of chemicals showed the highest ranges of variation in such variables, when propagating the uncertainty of physicochemical properties throughout the reference MEM, in some cases of up to 12 logarithmic units (for 468 chemicals: the performances in air were $q^2 = 0.87$ and MAE = 0.82; while, the performances in water were $q^2 = 0.82$ and MAE = 0.18). QPFRs using key physicochemical properties, partition coefficients and degradation rates, provided very accurate fate predictions (for 468 chemicals: the performances in air were $q^2 = 0.99$ and MAE = 0.10; while, the performances in water were $q^2 = 0.99$ and MAE = 0.06). However, since the availability of partitioning and degradation data is restricted to a limited number of chemicals, the applicability of the QPFR approach is thus restricted to such chemicals.

QSFRs estimate the fate of new chemicals, not used in the development of these models, from their molecular descriptors, not their physicochemical properties. A great advantage, when the latter are unknown. QSFR models were developed using one out of two groups of molecular descriptors, the first group comprised molecular weight (MW) and molecular properties estimated semi-empirically with the PM3 approximation of the Molecular Orbital (MO) theory, the second group comprised MW and counts of molecular constituents (atoms, bonds, functional groups and rings). Best QSFR performances (for 468 chemicals: the performances in air were $q^2 = 0.78$ and MAE = 1.01; while, the performances in water were $q^2 = 0.80$ and MAE = 0.33) resulted when using the second group of descriptors (MW and counts of molecular constituents). The SVR algorithm could estimate the fate of new chemicals (in test or validation data sets) with acceptable accuracy, when comparing chemicals in terms of the sections of every molecule rather than to average molecular properties.

For improving the performance of QSFR models, it was investigated the clustering of chemicals in classes for later developing class-tailored QSFR models. Improved fate predictions resulted when clustering chemicals, not with respect to water degradation (for 468 chemicals: the performances in air were $q^2 = 0.72$ and MAE = 1.13; while, the performances in water were $q^2 = 0.57$ and MAE = 0.31) but with respect to their molecular composition (for 468 chemicals: the performances in air were $q^2 = 0.79$ and MAE = 0.84; while, the performances in water were $q^2 = 0.86$ and MAE = 0.16); because, class predictions in the first case yielded a true positive rate of 77.4 % and a false positive rate of 22.6 %, while class predictions in the second case were much lower than that, a true positive rate of 100.0 % and a false positive rate of 0.0 %. The atoms of a chemical can be easily calculated from its molecular formula, while its

ISBN:978-84-693-4597-9 /DL:T.1010-2010

physicochemical properties are subject to variation due to uncertainties in both experimental and estimation procedures. Any failure in the class prediction of a new chemical leads to its assessment with an inappropriate QSFR model, yielding extremely wrong results. For having the fate of a new chemical well predicted, it must be assessed with a QSFR related to the same chemical class.

The prediction of chemicals in and out the domain of applicability of class tailored-QSFRs, with respect to molecular composition, was studied in three cases: Case I, using the SOM algorithm and information about molecular constituents; Case II, using the SOM algorithm and principal components of molecular constituents; and Case III, the intersection of Cases I and II). It was demonstrated that the environmental mass ratios of new chemicals (test and validation chemicals) within the domains of applicability (DOAs) of class-tailored models (Case III: the performances in air for 48 chemicals were: $q^2 = 0.92$ and MAE = 0.54; while, the performances in water for 53 chemicals were: $q^2 = 0.93$ and MAE = 0.16), were way more accurate than those of outlying chemicals (Case III: the performances in air for 120 chemicals were: $q^2 = 0.59$ and MAE = 1.50; while, the performances in water for 117 chemicals were: $q^2 = 0.42$ and MAE = 0.35). Extending these assessments to emissions in air, similar trends were obtained when analyzing the same chemicals within the DOAs (Case III: the performances in air for 48 chemicals were: $q^2 = 0.94$ and MAE = 0.20; while, the performances in water for 53 chemicals were: $q^2 = 0.92$ and MAE = 0.27) and out of the DOAs (Case III: the performances in air for 120 chemicals were: $q^2 = 0.53$ and MAE = 0.66; while, the performances in water for 117 chemicals were: $q^2 = 0.61$ and MAE = 0.51).

Additionally, comparing the performances of environmental fate predictions in air and water, while emitting chemicals in one of these two compartments, it was observed that best predictive performances were achieved for a single compartment when emissions occur in itself and not in other compartment. This confirmed for both the water compartment (considering 53 chemicals within the DOAs in Case III: for emissions in water, the performances in water were: $q^2 = 0.93$ and MAE = 0.16; for emissions in air, the performances in water were: $q^2 = 0.92$ and MAE = 0.27) and the air compartment (considering 48 chemicals within the DOAs in Case III: for emissions in water, the performances in air were: $q^2 = 0.92$ and MAE = 0.54; for emissions in air, the performances in air were: $q^2 = 0.94$ and MAE = 0.20) of the scenario considered.

ISBN:978-84-693-4597-9 /DL:T.1010-2010

ISBN:978-84-693-4597-9 /DL:T.1010-2010

How to read this thesis book

This manuscript is a doctoral thesis derived from a research work originally prepared for the NOMIRACLE project in form of public deliverables, poster and oral presentations at international conferences and an article at a specialized journal. This document demonstrates how both learning algorithms and molecular information can be used to estimate the environmental fate of chemical pollutants, known sufficient examples of environmental fate for training chemicals. It comprises five chapters. Chapter 1 states the motivation, background, hypothesis, objectives and contributions of this research work. Chapter 2 describes the methods and tools employed, covering relevant technical disciplines: multimedia environmental modeling, statistical sampling, molecular modeling and pattern recognition. Chapter 3 describes the data sets used in the experiments; while Chapter 4 discusses the results of different computerized experiments carried out sequentially following a similar order to that used in the NOMIRACLE project, but applying updated work practices. Chapter 5 states the conclusions of this work, discusses the applicability of the QSFR approach to multimedia environmental analysis, and outlines possible research areas for further developments. Supporting materials are presented in annexes containing: preliminary research works (Annex A), program codes (Annex B), lists of chemicals used in the assessments (Annex C) and data used (Annex D). Since the amount of information contained in some annexes may exceed the capacity of this manuscript, relatively small annexes have been printed and presented with a majuscule letter followed by a dot and number (e.g., like Annexes A.1, B.1, B.2…). Large annexes are only available as standard computer files in the accompanying CD of this manuscript and presented with a majuscule letter followed by a dot, a minuscule letter and a number (e.g., like Annexes A.a1, A.a2, A.a3, A.a4, A.b1, A.b2, A.b3, A.c1…).

The computerized experiments of this thesis work implied the use of techniques and terminology used in very dissimilar disciplines that, when unknown to the reader, might be difficult to understand when studied in a first time. So, an effort has been made to make the information presented in this thesis as clearly as possible to wide audiences. The list of contents of this thesis should be considered as a map, ready to help lost readers find ways to digest this work. Graphs and tables are discussed in every section of the thesis, but they also have extended captions for helping the reader get concise explanations or specific details from an item of interest. Since the accumulated knowledge of each of the involved disciplines is vast, references have been listed at the end of each chapter, helping to associate every section of the thesis work with relevant knowledge or previous research works.

For those interested in having a deep understanding of the findings of this study, the data and models presented in the Annexes can be of great help. They can give the necessary feeling for enhancing the visualization of the trends and results presented in this manuscript.

Please note that the preliminary research works in the Annexes (reports, posters, oral presentations and papers) were edited in series or parallel to the evolution of this study and so their vocabulary, symbols and abbreviations may differ. However, their findings have being used in every step for updating the modeling of QSFRs.

ISBN:978-84-693-4597-9 /DL:T.1010-2010

ISBN:978-84-693-4597-9 /DL:T.1010-2010

Contents

ISBN:978-84-693-4597-9 /DL:T.1010-2010

ISBN:978-84-693-4597-9 /DL:T.1010-2010

ISBN:978-84-693-4597-9 /DL:T.1010-2010

ISBN:978-84-693-4597-9 /DL:T.1010-2010

List of figures

ISBN:978-84-693-4597-9 /DL:T.1010-2010

ISBN:978-84-693-4597-9 /DL:T.1010-2010

List of tables

ISBN:978-84-693-4597-9 /DL:T.1010-2010

ISBN:978-84-693-4597-9 /DL:T.1010-2010

List of abbreviations

0D	0-Dimensional
1D	1-Dimensional
2D	2-Dimensional
3D	3-Dimensional
4D	4-Dimensional
ACS	American Chemical Society
AI	Artificial Intelligence
AIChE	American Institute of Chemical Engineers
AM1	Austin Model 1
ANN	Artificial Neural Network
ASCII	American Standard Code for Information Interchange
BOD	Biological Oxygen Demand
BPN	Backpropagation Network
CAS	Chemical Abstract Service
CERCLA	Comprehensive Environmental Response, Compensation, and Liability Act
CME	Conformation Minimum Energy
DE	Dielectric Energy
DOA	Domain of Applicability
EA	Electron Affinity
ESIS	European Chemical Substances Information System
FN	False Negative
FP	False Positive
GC	Gas Chromatography
GIS	Geographical Information System
HOMO	Highest Occupied Molecular Orbital
HPLC	High Performance Liquid Chromatography
HPV	High Production Volume chemicals
HWIR	Hazardous Waste Identification Rule
InChI	International Chemical Identifier
IP	Ionization Potential
IUPAC	International Union of Pure and Applied Chemistry
KKT	Karush-Kuhn-Tucker conditions
LPV	Low Production Volume chemicals
LRT	Long Range Transport
LRTP	Long Range Transport Potential
LSSVR	Least-Squares Support Vector Regression
LUMO	Lowest Unoccupied Molecular Orbital
MAE	Mean Absolute Error
MEM	Multimedia Environmental Model
MITI	Ministry of International Trade and Industry
MO	Molecular orbital theory
MOPAC	Molecular orbital package, popular software that includes several MO algorithms
MR	Molecular Refractivity
NB	Naive Bayes learning algorithm
NBk	Naive Bayes learning algorithm with kernel estimation
NITE	National Institute of Technology and Evaluation
NOMIRACLE	NOvel Methods for Integrated Risk Assessment of CumuLative stressors in Europe
OECD	Organisation for Economic Co-operation and Development
OpenSMILES	Open source equivalent of the SMILES code
ORATS	Online European Risk Assessment Tracking System
PBT	Persistent Bioaccumulative and Toxic
PCA	Principal Component Analysis
PM3	Parameterized Model 3
PO	Polarizability
POP	Persistent Organic Pollutant

ISBN:978-84-693-4597-9 /DL:T.1010-2010

q^2	Predictive squared coefficient
QPFR	Quantitative Property-Fate Relationship
QSAR	Quantitative Structure-Activity Relationship
QSBR	Quantitative Structure-Biodegradation Relationship
QSFR	Quantitative Structure-Fate Relationship
QSPR	Quantitative Structure-Property Relationship
QSTR	Quantitative Structure-Toxicity Relationship
RBF	Radial Basis Function
ROC	Receiver Operating Characteristic plot
SA	Solvent Accessibility
SE	Steric Energy
SETAC	Society of Environmental Toxicology and Chemistry
SMILES	Simplified Molecular Input Line Entry System
SOM	Self Organizing Map
SRC	Syracuse Research Corporation
SVM	Support Vector Machines
SVR	Support Vector Regression
TN	True Negative
TP	True Positive
UNEP	United Nations Environmental Program

ISBN:978-84-693-4597-9 /DL:T.1010-2010

List of symbols

Physicochemical and fate modeling data

B	Oxygen consumption in a degradation test [mg]
BOD	Biochemical oxygen demand of a chemical in a degradation test [mg]
C	Pollutant concentration [g/m^3, mol/m^3, mol/cm^3]
C_a	Pollutant concentration in medium a [g/m^3, mol/m^3]
C_b	Pollutant concentration in medium b [g/m^3, mol/m^3]
$C_{i,g}$	Concentration of chemical i in compartment g [g/mol]
C_o	Initial pollutant concentration [g/m^3, mol/m^3]
$C_{OH\cdot}$	Concentration of hydroxyl radicals in air [g/m^3, molecule/cm^3]
CORG	Organic carbon content [g/g]
CORG	Average organic carbon content [dimensionless]
D_a	Diffusion coefficient of a chemical in medium a [m^2/s]
D_{air}	Diffusion coefficient of a chemical in air [m^2/s]
deg%	Percentage degradability
Δt	Period of time [yr]
D_{water}	Diffusion coefficient of a chemical in water [m^2/s]
E_i	Emission of chemical i in the system [ton/yr]
G	Number of environmental compartments
g	Compartment g
H	Henry's law constant [Pa·m^3/mol]
J	Pollutant flow by diffusion [g/s, mol/s]
k	Degradation rate [1/s]
K_{ab}	a-b partition coefficient [dimensionless]
k_{air}	Degradation rate constant of a chemical in air [1/s]
K_{aw}	Air-water partition coefficient [dimensionless]
K_{oc}	Organic carbon partition coefficient [L/kg]
$k_{OH\cdot}$	Degradation rate constant of a chemical by hydroxyl radicals in air [m^3/g·s, cm^3/molecule·s]
K_{ow}	Octanol-water partition coefficient [dimensionless]
k_{sed}	Degradation rate constant of a chemical in sediments [1/s]
k_{soil}	Degradation rate constant of a chemical in soil [1/s]
K_{sw}	Solid-water partition coefficient [dimensionless]
k_{water}	Degradation rate constant of a chemical in water [1/s]
MW	Molecular weight [g/mol]
N	Number of chemicals, number of samples
n	Chemical n, sample n
P	Vapor pressure [Pa]
pKa	Acid dissociation constant [dimensionless]
R	Ideal gas constant, 8.314 J/(mol·K)
S_a	Residual mass of a substance at the end of a degradation test [mg]
S_b	Mass of a substance at the beginning of a degradation test [mg]
S_w	Water solubility [mol/m^3, mg/L]
T	Temperature [K]
t	Time [s, min, h, d, yr]
$t_{½}$	Degradation half time [s, h, day, week, month, yr]
T_m	Melting point [ºC, K]
TOD	Theoretical oxygen demand for completing the oxidation of a chemical [mg]
V_g	Volume of compartment g [m^3]
$w_{i,g}$	Mass ratio of chemical i in compartment g [dimensionless]
X	Length [m]
ρ_{soil}	Soil density [kg/L]
ρ_{solid}	Density of solids [kg/L]

Molecular data estimated semi-empirically

CME	Conformation minimum energy [kcal/mole]
DE	Dielectric energy [kcal/mole]
ΔH_f	Heat of formation [kcal/mole]
SE	Steric energy [kcal/mole]
MR	Molar refractivity [m^3/mol]
PO	Polarizability [Å^3]

ISBN:978-84-693-4597-9 /DL:T.1010-2010

SA	Solvent accessibility surface area [$Å^2$]
μ	Dipole moment [debye]
μ_x	Dipole vector X [debye]
μ_y	Dipole vector Y [debye]
μ_z	Dipole vector Z [debye]
EA	Electron affinity [eV]
HOMO	HOMO energy [eV]
IP	Ionization potential [eV]
LUMO	LUMO energy [eV]
${}^0\chi$	Connectivity index (order 0, standard) [dimensionless]
${}^1\chi$	Connectivity index (order 1, standard) [dimensionless]
${}^2\chi$	Connectivity index (order 2, standard) [dimensionless]
${}^1\kappa$	Shape index (basic kappa, order 1) [dimensionless]
${}^2\kappa$	Shape index (basic kappa, order 2) [dimensionless]
${}^3\kappa$	Shape index (basic kappa, order 3) [dimensionless]
${}^0\chi^v$	Valence connectivity index (order 0, standard) [dimensionless]
${}^1\chi^v$	Valence connectivity index (order 1, standard) [dimensionless]
${}^2\chi^v$	Valence connectivity index (order 2, standard) [dimensionless]

Molecular data counting molecular fragments

AC_{all}	Count of all atoms [dimensionless]
$AC_{bromine}$	Count of bromine atoms [dimensionless]
AC_{carbon}	Count of carbon atoms [dimensionless]
$AC_{chlorine}$	Count of chlorine atoms [dimensionless]
$AC_{fluorine}$	Count of fluorine atoms [dimensionless]
$AC_{hydrogen}$	Count of hydrogen atoms [dimensionless]
AC_{iodine}	Count of iodine atoms [dimensionless]
$AC_{nitrogen}$	Count of nitrogen atoms [dimensionless]
AC_{oxygen}	Count of oxygen atoms [dimensionless]
$AC_{phosphorus}$	Count of phosphorus atoms [dimensionless]
$AC_{silicon}$	Count of silicon atoms [dimensionless]
$AC_{sulphur}$	Count of sulphur atoms [dimensionless]
BC_{all}	Count of all bonds [dimensionless]
BC_{single}	Count of single bonds [dimensionless]
BC_{double}	Count of double bonds [dimensionless]
BC_{triple}	Count of triple bonds [dimensionless]
$GC_{aldehyde}$	Count of aldehyde [dimensionless]
GC_{amide}	Count of amide [dimensionless]
GC_{amine}	Count of amine [dimensionless]
$GC_{sec\text{-}amine}$	Count of sec-amine [dimensionless]
$GC_{carbonyl}$	Count of carbonyl [dimensionless]
$GC_{carboxyl}$	Count of carboxyl [dimensionless]
$GC_{carboxylate}$	Count of carboxylate [dimensionless]
GC_{cyano}	Count of cyano [dimensionless]
GC_{ether}	Count of ether [dimensionless]
$GC_{hydroxyl}$	Count of hydroxyl [dimensionless]
GC_{methyl}	Count of methyl [dimensionless]
$GC_{methylene}$	Count of methylene [dimensionless]
GC_{nitro}	Count of nitro [dimensionless]
$GC_{nitroso}$	Count of nitroso [dimensionless]
$GC_{sulfide}$	Count of sulfide [dimensionless]
$GC_{sulfone}$	Count of sulfone [dimensionless]
$GC_{sulfoxide}$	Count of sulfoxide [dimensionless]
GC_{thiol}	Count of thiol [dimensionless]
RC_{all}	Count of all rings [dimensionless]
$RC_{aromatic}$	Count of aromatic rings [dimensionless]
RC_{small}	Count of small rings [dimensionless]
$RC_{5\text{-}m}$	Count of 5 membered rings [dimensionless]
$RC_{a\text{-}5\text{-}m}$	Count of aromatic 5 membered rings [dimensionless]
$RC_{6\text{-}m}$	Count of 6 membered rings [dimensionless]
$RC_{a\text{-}6\text{-}m}$	Count of aromatic 6 membered rings [dimensionless]
$RC_{7\text{-}12\text{-}m}$	Count of 7-12 membered rings [dimensionless]
$RC_{a\text{-}7\text{-}12\text{-}m}$	Count of aromatic 7-12 membered rings [dimensionless]

QPFR and QSFR models

ISBN:978-84-693-4597-9 /DL:T.1010-2010

$\mathbf{C}$	Matrix of environmental fate estimations, of size [N×G]
C_{air}	Concentration of a pollutant in air [g/mol]
C_{sed}	Concentration of a pollutant in sediments [g/mol]
C_{soil}	Concentration of a pollutant in soil [g/mol]
C_{veg}	Concentration of a pollutant in vegetation [g/mol]
C_{water}	Concentration of a pollutant in water [g/mol]
$\mathbf{D}$	Matrix of molecular descriptors, of size [N×D]
D	Number of molecular descriptors [dimensionless]
$\mathbf{E}$	Vector of site-specific parameters, of size [N×J]
f_{MEM}	Function that works as multimedia environmental model
f_{QPFR}	Function that relates physicochemical properties to fate estimations (QPFR)
f_{QSFR}	Function that relates molecular information to fate estimations (QSFR)
G	Number of environmental compartments [dimensionless]
J	Number of environmental compartments in which emissions may occur [dimensionless]
K	Number of physicochemical properties [dimensionless]
K^*	Number of available physicochemical properties, with $K^*<K$ [dimensionless]
$\log_{10}$	Base 10 logarithmic scaling of data
MAE	Mean absolute error
N	Number of chemicals, number of samples [dimensionless]
n	Chemical n, sample n
$N_{[-1,1]}$	Linear normalization of data in the range [-1,1]
$\mathbf{P}$	Matrix of physicochemical properties, of size [N×K]
$\mathbf{P}^*$	Matrix of physicochemical properties, partially incomplete (with missing values),of size [N×K^*]
$\mathbf{P}_{est}$	Matrix of physicochemical properties obtained by experimental methods, of size [N×K]
q^2	Predictive squared coefficient
q^2_{tr}	Predictive squared coefficient that compares predictions of a data set to training targets
$\mathbf{S}$	Vector of site-specific parameters, of size [M×1]
w_{air}	Mass ratio of a pollutant in air [dimensionless]
w_{sed}	Mass ratio of a pollutant in sediments [dimensionless]
w_{soil}	Mass ratio of a pollutant in soil [dimensionless]
w_{veg}	Mass ratio of a pollutant in vegetation [dimensionless]
w_{water}	Mass ratio of a pollutant in water [dimensionless]

ISBN:978-84-693-4597-9 /DL:T.1010-2010

ISBN:978-84-693-4597-9 /DL:T.1010-2010

Chapter 1

Introduction

This thesis study aims to contribute to the environmental modeling of chemical pollutants lacking of partitioning and degradation data. To this end, learning algorithms, widely used in artificial intelligence applications, have been trained to predict the fate of chemicals directly from their molecular structure known representative modeling examples. This introductory chapter explains the motivation, background, hypothesis, objectives and main contributions of this work.

ISBN:978-84-693-4597-9 /DL:T.1010-2010

1.1 Motivation

There is concern about the presence of chemicals in the environment with the capacity to affect ecosystems and human health. For considering how hazardous a chemical can be, it is necessary to evaluate not only its toxicity and reactivity but also its quantity, location and exposure time. The fate of chemical pollutants released in the environment is determined by their tendency or not to persist, bioaccumulate and transport.

Multimedia environmental models are tools used for estimating quantitatively the distribution of pollutants in the environment (Mackay, 2001), which otherwise would be difficult or unpractical to measure in real conditions. These models solve mass balances of pollutants undergoing various environmental processes (e.g., partitioning, transportation, degradation, etc.) in compartments representing different media (e.g., air, water, soil, etc.). Multimedia environmental models estimate concentrations of pollutants in all compartments, which can be subsequently related to toxicity and exposure parameters in standard risk assessments for regulatory and decision making tasks.

Multimedia models require large amounts of data concerning geographic site-specific parameters, emission rates and physicochemical properties of the chemical to assess. Most data are difficult to obtain, site-parameters depend on geographical characterizations (Mackay, 2001) and emissions depend on scarce source data (Breivik et al., 2004; Breivik et al., 2006; Lohmann et al., 2007). In relation to physicochemical properties of chemicals, both experimental data and estimation methods have been compiled for their use in environmental modeling; even so, there is still the need of characterizing not only most existing chemicals but also new chemicals that have yet to be synthesized (Boethling et al., 2004). So, the availability or uncertainty of input data must be taken into account for most multimedia environmental fate assessments (Wania and Mackay, 1999a).

Solely the lack of physicochemical properties of chemicals constitutes a very important issue: they remain unknown until their experimental determination. The large and constantly increasing number of chemicals complicates their complete characterization. The number of chemicals has experienced an exponential growth during the past 200 years (Schummer, 1997a), greatly influenced by the production of new chemicals for varied purposes (Schummer, 1997b). The CAS registry, one of the largest substance registry databases, maintained by the American Chemical Society through the Chemical Abstract Service (CAS) division, reported about 37 million substance and 60 million sequence records by the end of the year 2007 (CAS, 2008). By September 2009, it was reported the 50-millionth unique chemical substance of the CAS registry (Toussant, 2009).

Problems in the experimental determination of physicochemical properties are not only restricted to a matter of costs and time; there are several chemicals for which properties cannot be appropriately measured with current technology, producing noisy values. In the same trend, estimation methods for missing properties may suffer the same kind of limitations as they are usually based on known data of chemicals already characterized. Property estimations may be carried out by a large pool of methods, all

ISBN:978-84-693-4597-9 /DL:T.1010-2010

of them with different levels of uncertainty (Boethling et al., 2004). In general, properties related to partitioning (e.g. melting point, vapor pressure, Henry's law constant, water solubility, etc.) (Boethling et al., 2004; Mackay, 2000) are easier to measure and estimate than properties of degradation processes (Aronson et al., 2006; Howard et al., 1991; Klöpffer and Wagner, 2007; Raymond et al., 2001). Using different experimental or estimation methods, a property may have assigned a wide range of values. With this panorama, the applicability of multimedia environmental models is thus confined to well known chemicals, those for which there are reliable physicochemical data.

Given the gigantic amount of chemicals and little information about them, the attention of regulators have been oriented towards updatable lists of few priority substances. The most known priority lists have been prepared by: the European Commission Community (EEC) (EEC, 1993); the United Nations Economic Commission for Europe (UN-ECE) (UN-ECE, 1998); the United Nations Environment Programme (UNEP) (UNEP, 2001); and, the Comprehensive Environmental Response, Compensation, and Liability Act (CERCLA) (ATSDR and EPA, 2007).

The importance of assessing the multimedia environmental fate of priority chemicals is in no doubt, but chemicals not included in these lists may not be correctly assessed or even considered for risk assessments. So, motivated by the lack of information concerning most commercial chemicals and the risk that they represent for human health and the environment, new regulatory conditions are about to apply in industrialized countries (Tickner et al., 2005). These rules aim to collect information about the characteristics, emission rates and existing volumes of commercial chemicals in order to facilitate decision making tasks regarding the authorization or banning of the latter. The European Union, by means of the REACH regulation (Registration, Evaluation, Authorization and Restriction of Chemicals) (European Commission, 2006), plans to register substances produced in volumes equal or higher than 1 t/year and compile risk assessments for substances produced at rates equal or higher than 10 t/yr. Meanwhile, the United States implements the Inventory Update Rule (US-EPA, 2006) with similar purposes.

Characterizing the massive amount of existing chemicals is a heavy task. Consequently, time may pass before enough and reliable physicochemical data are compiled for assessing the fate of most chemicals. Estimations methods based on molecular structure have proven to be appropriate for predicting chemical activity by means of relationships between analogous chemicals (Hugo, 2002), usually termed Quantitative Structure Relationships (QSARs). They represent an alternative to costly experiments, especially in environmental modeling (Devillers, 2003; Mackay et al., 2003; Mackay and Webster, 2003). QSARs and alike have been widely used for estimating physicochemical properties, environmental parameters, toxicity and health effects of chemical pollutants for regulatory assessments (Cronin et al., 2003; Walker et al., 2002). Moreover, it is expected that newer regulatory initiatives will depend more on QSARs for filling information gaps (Fjodorova et al., 2008; Worth et al., 2007), as databases evolve to contain more and more parameters and assessments of chemicals.

ISBN:978-84-693-4597-9 /DL:T.1010-2010

In most environmental assessments, missing physicochemical properties are usually estimated for their posterior use in multimedia environmental models. However, the more properties are estimated the more uncertain fate estimations may be, collecting the uncertainty propagated by each input parameter. Screening methods are thus required for evaluating the fate of chemical pollutants when their physicochemical properties are incomplete or noisy, situations in which standard multimedia environmental models tend to be highly uncertain.

1.2 Background

1.2.1 Multimedia environmental models

Chemical pollutants may affect organisms at different levels, depending on factors like quantity, exposure time, toxicity and the media in which they are dissolved in. The importance of this matter is as high as its complexity, forcing the need of developing environmental models as simple as possible for describing the fate of pollutants and adding complexity when required (Mackay, 2001).

The environment can be described as a set of homogeneous compartments or phases (typically air, water, sediments, soil, vegetation and biota) with fixed volumes in which gradients of concentration and temperature are negligible. Figure 1-1 shows a representation of the environment as a set of boxes and the pathways that a pollutant may follow as arrows from outside the system and throughout the system, from one compartment to another. Degradation processes in each compartment remove pollutants from the system modifying their structure and generating sub products.

Since transportation and degradation of chemicals may occur in each compartment, mass balances can be set for accounting the rates at which a pollutant i accumulates or disappears in a given phase g of a geographical region:

$$[\text{Accumulation rate}]_{i,g} = [\text{Inflow rate}]_{i,g} - [\text{Outflow rate}]_{i,g} - [\text{Degradation rate}]_{i,g} \quad (1\text{-}1)$$

which allows the evaluation of average concentrations in each compartment once that the quantity of pollutant in each compartment is determined.

Mass balances can be solved assuming the presence or not of steady state conditions, equilibrium or flow. Mackay classified multimedia environmental models in four levels of complexity, according to the assumptions applying in their mass balances (Mackay, 2001). Figure 1-2 shows the assumptions involved in such classification by comparing concentrations of a single pollutant in two-phase systems.

In the absence of perturbation, closed systems (Level I, Figure 1-2a) reach simultaneously both chemical equilibrium and steady state conditions: equilibrium concentrations remain constant with time. In equilibrium, the proportion of solute dissolved in each phase remains constant ($K_{ab} = C_a/C_b$) regardless of the total amount

ISBN:978-84-693-4597-9 /DL:T.1010-2010

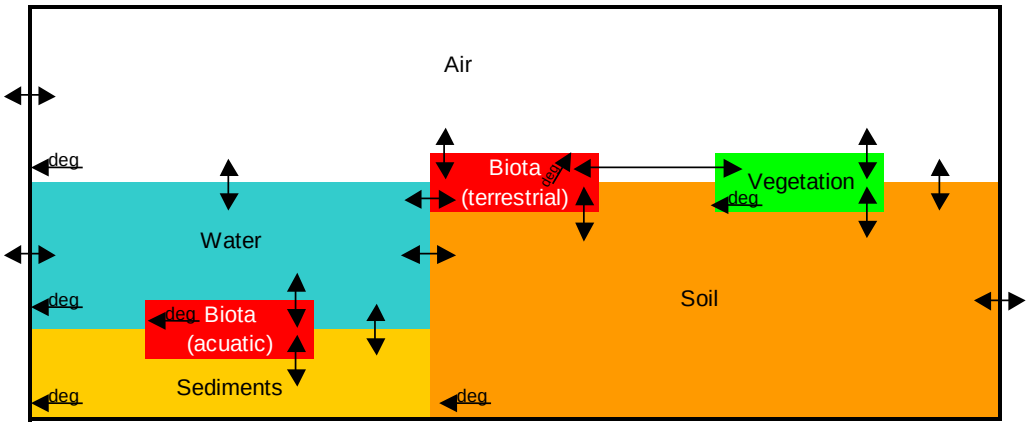


Figure 1-1 Representation of the environment as a set of homogeneous compartments.
In the environment, every medium can be considered a box of homogeneous density and composition that imports and exports chemicals by means of different transport processes. Some chemicals may be partially or totally removed by degradation (deg), other chemicals simply persist. Compartmental multimedia environmental models estimate the tendency of chemicals to distribute through different media, simultaneously, applying mass balances for each media.

of pollutant in the system.

In open systems, the presence of flow allows different sets of conditions: outflows may experience steady state conditions with constant pollutant concentrations that may be equal or not to equilibrium concentrations, depending, respectively, on the availability (Level II, Figure 1-2b) or not (Level III, Figure 1-2c) of time for reaching chemical equilibrium within the system. In the same manner, changing concentrations (Level IV) in outflows may reach (Figure 1-2d) or not (Figure 1-2e) equilibrium.

In some cases, the complexity of a multimedia model can be adjusted for avoiding unnecessary calculation costs. When real conditions change slowly with time, Level III assumptions can be applied for standard multimedia environmental modeling without appreciable inconveniences (e.g., pollutants with the tendency to persist in the environment for relatively long periods of time). Changing conditions are best modeled with Level IV assumptions.

Environmental models have been readily developed for describing individual environmental processes (process models), describing biological uptake (biological uptake models), evaluating the fate of chemicals in generic conditions (evaluative models) and describing the fate of chemicals in real locations at small (regionalized models) and large scales (spatially resolved models) (Wania and Mackay, 1999a).

Process models constitute the core of more specialized models, as the former are added into the latter to account the effect that processes occurring simultaneously exert on the final distribution of chemical pollutants. In the mass balances of multimedia environmental models (Equation 1-1), each term represents a specific environmental process that raises or drops the quantity of chemical pollutants in a medium or more.

ISBN:978-84-693-4597-9 /DL:T.1010-2010

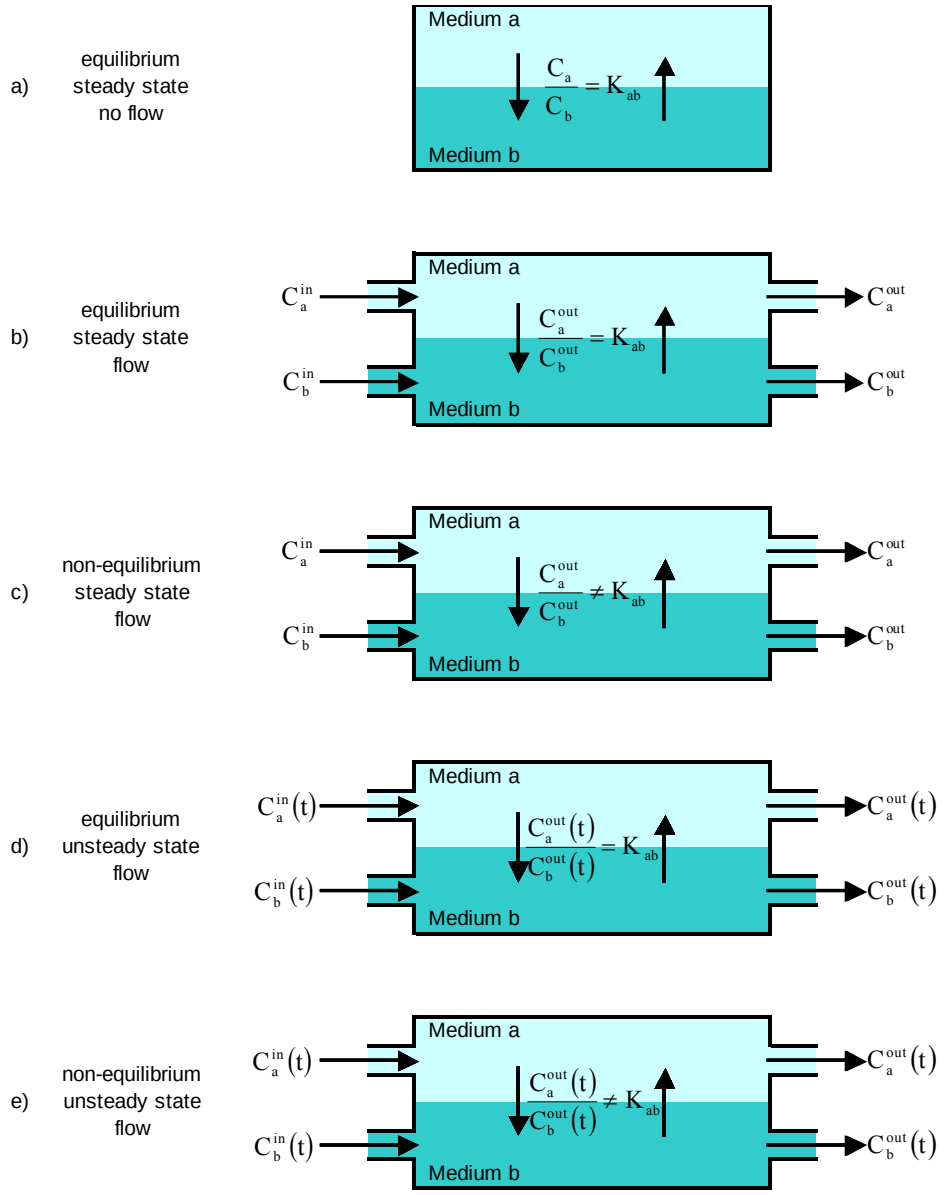


Figure 1-2 Applicable conditions to mass balances of multimedia environmental models.
Compartmental multimedia environmental models are usually classified, according to conditions applying in their mass balances, as Level I (a), Level II (b), Level III (c) or Level IV (d and e). The more complex a model is, the higher its level. (Adapted from Mackay, 2001).

Multimedia models usually represent media like air, water, sediments, soil and biota (vegetation and animals). However, when modeling processes between medium boundaries, other media may also be considered. For example, air may contain aerosols that participate in sorption processes or fall to soil or water compartments by dry and wet deposition processes. Similarly, water may contain suspended sediments that may be deposed, re-suspended and buried. The soil compartment, despite of not being a fluid medium may be affected by the amount of organic matter and water (evaporation, runoff to water, percolation to ground water, etc.) that it contains. All of these compartments may contain biota in different proportions.

ISBN:978-84-693-4597-9 /DL:T.1010-2010

Since the assumption of medium homogeneity is commonly in most multimedia models, average estimations of steady state or time dependent concentrations are obtained from, respectively, Level III and Level IV calculations. Models based on homogeneous compartments may be reasonably used for evaluative and regional assessments. As the volume of the region to assess increases, temporal and spatial variability may be also considered. Environmental media may differ greatly, spatially and temporally, in terms of pressure, temperature, volume, continuity, chemical and physical composition. Spatially resolved models divide each medium into several homogeneous compartments, estimating concentrations for each subdivision of the environment. In some cases, fluid phases may be modeled by 3D differential equations using Eulerian or Lagrangian approaches, i.e., with fixed or moving coordinates, respectively.

Available multimedia environmental models

Environmental models have been used for describing generic portions of the environment in a large variety of configurations, known the properties of the chemicals to assess, emission rates and site-specific parameters. Table 1-1 lists models with very different features, some of them widely used by modelers and regulators. The models contained in Table 1-1 are listed approximately in accordance to their application as evaluative, regionalized or spatially resolved models. General features of these models are also listed: scales, media, inputs and outputs. Specific details can be obtained from their respective references or manuals.

Evaluative models have been widely used for assessing generic conditions. The QWASI model (Mackay et al., 1983) has been originally intended for modeling the fate of pollutants in lakes, in a system composed of air, water, sediments, fish and suspended solids. Some models incorporated soil and vegetation compartments for expanding their use to other locations; good examples are the CEMC models (Level I, Level II, Level III) (Mackay, 1991; Mackay and Paterson, 1991; Mackay et al., 1992a), the EQC model (Mackay et al., 1996a; Mackay et al., 1996b; Mackay et al., 1996c) and ELPOS (Beyer and Matthies, 2001). A modification of the Level III CEMC model (Mackay and Paterson, 1991), the ppLFER model (Breivik and Wania, 2003), incorporated a small set of linear solvation parameters, instead of typical partitioning properties, in an attempt to improve fate estimations for polar organic chemicals. BasinBox (Hollander et al., 2006) is a Level III generic model developed for describing upstream, midstream and downstream sections of rivers.

The fate of pollutants at a global scale has been generically modeled with simple evaluative models like ChemRange (Scheringer, 1996; Scheringer et al., 2004; Scheringer et al., 2002), a one-dimensional homogenous circular system, and CliMoChem (Scheringer et al., 2004; Scheringer et al., 2000), a two dimensional system composed of several latitudinal zones with different volumes and temperatures. Globo-POP (McLachlan et al., 2002; Wania, 2003; Wania and Daly, 2002; Wania and Mackay, 1993; Wania and Mackay, 1995; Wania and Mackay, 1999b; Wania et al., 1999), a zonally averaged multimedia model, divides the atmosphere in 4 layers and describes time dependent processes.

ISBN:978-84-693-4597-9 /DL:T.1010-2010

Table 1-1. Features of available multimedia environmental models.

Model[a,b]	Purpose			Scales[c]				Media[d]					Input[e]					Output[f]		
	Evaluative	Regionalized	Spatially resolved	Local	Regional	Continental	Global	Air	Water	Sediments	Soil	Vegetation	Chemical properties	Emissions	Environmental Parameters	Geophysical Parameters	Meteorological Parameters	Concentrations	Concentration fields	Others
QWASI[+,3-4]	√	-	-	√	-	-	-	√	√[la]	√	-	-	√	√	√	-	-	√	-	√[fu]
CEMC suite[+,1-3]	√	-	-	√	√	-	-	√	√[w,fi,ss]	√	√	-	√	√	√	-	-	√	-	√[fu]
EQC model[+,1-3]	√	-	-	√	√	-	-	√[a,ae]	√[w,ss]	√	√	-	√	√	√	-	-	√	-	√[fu]
ELPOS[+,3]	√	-	-	-	√	-	-	√	√	√	√	√	√	√	√	-	-	√	-	√[op]
ppLFER[+,3]	√	-	-	√	-	-	-	√	√	√	√	-	√[ep]	√	√	-	-	√	-	√[op]
ChemRange[+,3]	√	-	-	-	-	-	√	√	√[suw]	-	√	-	√	√	√	-	-	√	-	√[sr]
CliMoChem[+,4]	√	-	-	-	-	-	√	√	√[suw]	-	√[cs]	-	√	√	√	-	-	√	-	√[sr]
MPI-MBM[+,4]	√	-	-	√	-	-	-	√	√	-	√	√	√	√	√	-	-	√	-	√
MPI-MCTM[3D]	√	-	√	√	√	-	-	√	√[sw,ice]	√	√	√	√	√	√	√	-	√	√	√[de]
GLOBO-POP[+,4]	√	-	-	-	-	-	√	√[la]	√[fw,sw]	√	√[cs,us]	-	√	√	√[ta]	-	-	√	-	√[fu]
SimpleBox[+,n,3-4]	√	-	-	√	√	√	√	√	√[fw,sw]	√[fw,sw]	√[√cs,us]	√	√	√	√	-	-	√	-	√[mf]
EUSES[+,n,3-4]	√	-	-	√	√	√	√	√	√[fw,sw]	√[fw,sw]	√[√cs,us]	√	√	√	√	-	-	√	-	√[ri]
CoZMo-POP[+,3-4]	√	-	-	√	√	√	-	√	√	√	√	√	√	√	√	-	-	√	-	√[mf]
BasinBox[+,3]	√	-	-	√	-	-	-	√	√[ri]	√	√[un,sa]	√	√	√	√	-	-	√	-	√
CHEMGL[+,3]	√	-	-	√[GR]	-	-	-	√[bl,ft,ls]	√[sw]	√	√[sus,vz,gw]	√[pf,pr]	√	√	√	-	-	√	-	√[odp]
ChemFrance[+,3]	√	√	-	√[FR]	√[FR]	-	-	√	√[sw,gw]	√	√[s,gw]	-	√	√	√	-	-	√	-	√[fu]
CalTOX[+,3-4]	√	√	-	√[US]	√[US]	√[US]	-	√	√[suw]	√	√[gs,rs,vz]	√[le,ls]	√	√	√	-	-	√	-	√[ri]
ChemCAN[+3]	√	√	-	-	√[CA]	√[CA]	-	√	√	√	√	√	√	√	√	-	-	√	-	√[fu]
TRIM.FaTE[+,3-4]	√	√	√	-	√[US]	√[US]	-	√	√	√	√	√	√	√	√	√	√	√	-	√[ri]
POPCYCLING-B.[++,4]	-	√	√	-	√[BA]	-	-	√	√[fw,sw]	√	√	√	√	√	√[ta]	-	-	√	-	√
BETR-NA[++,4,GIS]	-	√	√	-	√[US]	√[US]	-	√	√[fw,cw]	√	√	√	√	√	√	√	√	√	√	√
BETR-Europe[++,4,GIS]	-	√	√	-	√[EU]	√[EU]	-	√	√	√	√	√	√	√	√	√	√	-	√	√
BETR-World[++,4,GIS]	-	√	√	-	√[GL]	√[GL]	√[GL]	√	√	√	√	√	√	√	√	√	√	-	√	√[de]
BETR-Global[++,4,GIS]	-	√	√	-	√[GL]	√[GL]	√[GL]	√	√	√	√	√	√	√	√	√	√	-	√	√
IMPACT-2002[++,4,GIS]	-	√	√	-	√[EU]	√[EU]	-	√	√	√	√	-	√	√	√	√	-	-	√	√
EUROS[++,eu]		-	√	-	-	√[EU]	-	√[la,pm]	√	-	√	-	√	√	√	-	√	-	√	√
LOTOS[++,4]	-	-	√	-	-	√[EU]	-	√[at]	-	-	-	-	√	√	√	-	√	-	√	√[de]
LOTOS-EUROS[++,4]	-	√	√	-	-	√[EU]	-	√[la]	√[fw,sw]	-	√[la,cs,us]	-	√	√	√	√	-	-	√	√

ISBN:978-84-693-4597-9 /DL:T.1010-2010

Table 1-1. Features of available multimedia environmental models (Continued).

Model[a,b]	Purpose			Scales[c]				Media[d]					Input[e]					Output[f]		
	Evaluative	Regionalized	Spatially resolved	Local	Regional	Continental	Global	Air	Water	Sediments	Soil	Vegetation	Chemical properties	Emissions	Environmental Parameters	Geophysical Parameters	Meteorological Parameters	Concentrations	Concentration fields	Others
MSCE-POP[++,4]	-	√	√	-	-	-	√	√	√[cr]	√	√	√	√	√	√	√	√	-	√	√[de]
G-CIEMS[++,3-4]	-	-	√	-	√[JP]	-	-	√	√[ri,cw]	-	√	√	√	√	√	√	√	√	√	√
DEHM-POP[3D,eu,4]	-	√	√	-	√[AP]	√[AP]	-	√	√	-	√	-	√	√	√	√	√	-	√	√[de]
FANTOM[3D,eu,4]	-	√	√	-	√[NS]	-	-	√	√[ss]	√	-	-	√	√	√	√	-	-	√	√[de]
GEM/POPs[3D,4]	-	√	-	-	√[CA]	-	-	√	√	-	√	-	√	√	√	-	√	-	√	√
Polair3D-POP[3D,eu,4]	-	√	√	-	√[EU]	√[EU]	-	√	√	-	√	-	√	√	√	-	√	-	√	√

[a] Models (in alphabetical order): BasinBox (Hollander et al., 2006), BETR-Europe (Prevedouros et al., 2004), BETR-Global (MacLeod et al., 2005), BETR-North America (MacLeod et al., 2001), BETR-World (Toose et al., 2004), CalTOX (McKone and Enoch, 2002; McKone et al., 1997; UCLA, 1995), CEMC (Mackay, 1991; Mackay and Paterson, 1991; Mackay et al., 1992a), ChemCAN (Mackay et al., 1991; Webster et al., 2003; Webster et al., 2004), ChemFrance (Devillers et al., 1995), CHEMGL (Zhang et al., 2003), ChemRange (Scheringer, 1996; Scheringer et al., 2004; Scheringer et al., 2002), CliMoChem (Scheringer et al., 2004; Scheringer et al., 2000), CoZMo-POP (Wania et al., 2006), DEMH-POP (Hansen et al., 2004), ELPOS (Beyer and Matthies, 2001), EQC model (Mackay et al., 1996a; Mackay et al., 1996b; Mackay et al., 1996c), EUROS (Leeuw and Rheineck Leyssius, 1990; Matthijsen et al., 2002; Van Loon, 1994; Van Loon, 1995), EUSES (Lijzen and Rikken, 2004; Vermeire et al., 2005; Vermeire et al., 1997), FANTOM (Ilyina et al., 2006), GEM/POPs (Gong et al., 2007; Huang et al., 2007), G-CIEMS (Suzuki et al., 2004), GEM-POPs (Gong et al., 2007; Huang et al., 2007), Globo-POP (McLachlan et al., 2002; Wania, 2003; Wania and Daly, 2002; Wania and Mackay, 1993; Wania and Mackay, 1995; Wania and Mackay, 1999b; Wania et al., 1999), LOTOS (Builtjes, 1992; Schaap et al., 2004), LOTOS-EUROS (Schaap et al., 2005; Schaap et al., 2008), MPI-MBM (Lammel, 2004), MPI-MCTM (Lammel et al., 2001; Semeena et al., 2003), Polair3D-POP (Quéguiner and Musson-Genon, 2008), POPCYCLING-Baltic (Breivik and Wania, 2002; Wania et al., 2000), ppLFER (Breivik and Wania, 2003), QWASI (Mackay et al., 1983), SimpleBox (Brandes et al., 1996; den Hollander and van de Meent, 2004; den Hollander et al., 2004; van de Meent, 1993), TRIM.FaTE (US-EPA, 2002a; US-EPA, 2002b).

[b] Methods for solving mass balances ([+] = few homogeneous compartments, [++] = several homogeneous compartments or grid, [1] = level I, [2] = level II, [3] = Level III, [4] = Level IV, [m-n] = Levels m to n, [n] = nested, [3D] = 3D equations, [eu] = Eulerian, [la] = Lagrangian, [GIS] = retrieval of site-specific data from GIS databases).

[c] Scales covered by the models: local, regional, continental, global. Some models have been developed for real geographic locations ([AP] = Artic Pole, [BA] = Baltic, [CA] = Canada, [FR] = France, [GL] = Global, [GR] = Great Lakes, [EU] = European Union, [JP] = Japan, [NS] = North Sea [US] = United States).

[d] Representation of the environment in standard media: air, water, sediments, soil and vegetation. Some models include other media as well ([ab]= air boundary layer, [at] = atmosphere, [ae] = aerosol, [cr] = cryosphere, [cs] = cultivated soil, [cw] = coastal water, [fi] = fish, [ft] = free troposphere, [fw] = fresh water, [gs] = ground-surface soil, [gw] = groundwater, [ice] = ice, [la] = layers, [le] =plant leaves, [ls] = plant leaf surfaces, [lt] = lower troposphere, [lw] = lake water, [pm] = particulate matter, [rs] = root-zone soil, [rw] = river water, [sa] = saturated soil [ss] = suspended sediments, [sus] =surface soil, [suw] = surface water, [sw] = sea water, [un] = unsaturated soil, [us] = uncultivated soil, [vz] = vadose-zone soil).

[e] Input of models: physicochemical properties ([ep] = solvation energy parameters instead of partitioning properties), emission rates, environmental parameters, geophysical parameters and meteorological parameters.

[f] Output of models: concentrations, concentration fields and others ([LRTP] = long range transport potential, [ov] = overall persistence, [cc] = cold condensation potential, [de] = deposition, [fu] = fugacity, [hi] = history, [m%]= mass percentages, [odp] = ozone depletion potential, [ri] = risk, [sr] = spatial range, [s-r] = source to receptor relationships).

ISBN:978-84-693-4597-9 /DL:T.1010-2010

Some evaluative models have been widely used for regional environmental fate assessments (Table 1-1). In one hand, there are models specifically adapted to regions of concern, like CHEMGL (Zhang et al., 2003), ChemFrance (Devillers et al., 1995), ChemCAN (Mackay et al., 1991; Webster et al., 2003; Webster et al., 2004) and CalTOX (McKone and Enoch, 2002; McKone et al., 1997; UCLA, 1995). On the other hand, there are models designed with several generic compartments for their posterior adaptation to specific regions; this is the case of SimpleBox (Brandes et al., 1996; den Hollander and van de Meent, 2004; den Hollander et al., 2004; van de Meent, 1993), TRIM.FaTE (US-EPA, 2002a; US-EPA, 2002b) and CoZMo-POP (Wania et al., 2006). SimpleBox is a special case, since it is composed of sets of up to 10 compartments nested at different scales (local, regional, continental and global). Some evaluative models linked to exposure and risk models have been widely used for regulatory purposes and risk assessments, CalTOX and TRIM.FaTE are typically used in the United States, while in the European Union it is the case of EUSES (Lijzen and Rikken, 2004; Vermeire et al., 2005; Vermeire et al., 1997), which is based on SimpleBox.

Temporal and spatial variability is a typical feature of more recent multimedia models, requiring more data than the standard evaluative models, usually in form of meteorological and geophysical parameters. Some models retrieve data from large databases supported on the Geographic Information System (GIS). Spatially resolved models offer different resolution levels, depending on how calculations are performed in their mass balances. There are models that divide large regions into smaller interconnected sections (composed of standard homogenous compartments), while other models perform 3D calculations in fluid mediums. Some models use both 3D calculations in fluid media of interest and homogeneous compartments for neighboring media.

Most spatially resolved models are based on several homogeneous compartments. POPCYCLING-Baltic (Breivik and Wania, 2002; Wania et al., 2000), based on 85 homogeneous compartments (4 in the atmosphere, 26 in water, 25 in sediments, 10 in forest canopy, 10 forest soil and 10 agricultural soil boxes), has been used to describe the historical fate of some POPs in the Baltic region. BETR-North America (MacLeod et al., 2001) divides the upper part of the American continent into 24 regions. BETR-Europe (Prevedouros et al., 2004) divides the Europe into 50 regions, while IMPACT-2002 divides the continent into 135 irregular watershed areas (land zones) and 156 separate air zones. BETR-World (Toose et al., 2004) and BETR-Global (MacLeod et al., 2005) divide the terrestrial globe, respectively, into 25 and 288 regions. The model G-CIEMS (Suzuki et al., 2004) represents the air compartment as a grid and rivers and soil as basins, achieving a resolution of up to 5x5 Km^2 in Japan.

Spatially resolved atmospheric models, suitable for volatile pollutants, use either grids or 3D calculations. LOTOS (Builtjes, 1992; Schaap et al., 2004) and EUROS (Leeuw and Rheineck Leyssius, 1990; Matthijsen et al., 2002; Van Loon, 1994; Van Loon, 1995), were developed independently for modeling dispersion and chemical transformation of pollutants in Europe, at the lower troposphere; these models have been merged into LOTOS-EUROS (Schaap et al., 2005; Schaap et al., 2008) to account distribution of pollutants in water and soil compartments as well. There are other atmospheric models, like DEMH-POP (Hansen et al., 2004), MSCE-POP and

ISBN:978-84-693-4597-9 /DL:T.1010-2010

GEM/POPs (Gong et al., 2007; Huang et al., 2007), Polair3D-POP (Quéguiner and Musson-Genon, 2008). A recent model, FANTOM (Ilyina et al., 2006), has been specifically designed for modeling spatial and temporal variability in the ocean, at the North Sea.

Given the large variety of models available, selecting one model in particular depends on the chemicals and region of interest, features in available models (i.e., description of processes, scales, calculation methods, etc.) and data availability. Special care must be taken when selecting parameters for the landscape of interest, since small variations in their values may lead to large variations in chemical fate predictions (Webster et al., 2004).

Environmental models require the management of several variables and assumptions, so it is recommended to use them for the same conditions for which they were developed (Fenner et al., 2005). They produce reasonable results for limited ranges of applicability. Studies comparing the performances of different multimedia models demonstrate how accurate predictions can be and their limitations (Armitage et al., 2007; Hollander et al., 2007; Kawamoto et al., 2001; Lammel et al., 2007; Shatalov et al., 2005; Shatalov et al., 2004), however most comparisons are based on few chemicals, those for which physicochemical properties and emission history are known. All models are expected to undergo further modifications and tests for improving modeling techniques, process descriptions and evaluation of spatial and temporal variability.

Assessments involving spatial and temporal resolution are desirable, but they are limited by the availability of data (accounting temporal and spatial variations) and resources to perform complex calculations. Standard evaluative models with homogenous compartments have proven to give reasonable estimations, making them suitable for screening inexpensively large groups of chemicals without geophysical and meteorological data, as an alternative to spatially resolved models.

Uncertainty in input data

The availability of sufficient and reliable input data is crucial for performing reliable environmental assessments. Multimedia environmental models require large amounts of data. Most measurements and estimations are uncertain, propagating substantial errors throughout the models and affecting the interpretation of analysts, regulators and decision makers. In one hand, field measurements tend to be scarce (records of historical emissions, spatial and temporal variability in media, etc.), forcing the use of average environmental parameters. On the other hand, measurements under laboratory conditions, aside of economical limitations, may be limited technically, leading to scarce or noisy data as well.

Known that difficulties may arise in both field and laboratory measurements, uncertainty analysis is a must in environmental modeling (Wania and Mackay, 1999a). It is known that site parameters (Meyer and Wania, 2007; Webster et al., 2004) and emission rates (Breivik et al., 2004; Breivik et al., 2006; Lohmann et al., 2007) affect the output of environmental models. The same occurs with the properties of chemicals. Physicochemical properties measure the tendency of chemicals to

ISBN:978-84-693-4597-9 /DL:T.1010-2010

participate in different environmental processes and to which extent. It has been observed that errors in data regarding the biodegradation and partitioning of chemicals in the environment may have a significant affect in the output of both standard (Citra, 2004; Eisenberg et al., 1998; Kawamoto et al., 2001; Kühne et al., 1997) and spatially resolved environmental models (Toose et al., 2004).

Despite of neglecting spatial variability, environmental models based on homogeneous compartments may suffer less uncertainty than spatially resolved models, the latter require more data in form of equilibrium or kinetic parameters (Fenner et al., 2004). Since physicochemical properties are present in the terms accounting mass flows in every mass balance (Equation 1-1), their impact on the output of multimedia environmental models depend on the magnitude of the different environmental processes taking place.

1.2.2 Physicochemical properties required in environmental assessments

Partitioning and degradation processes usually influence the most on the distribution of chemicals in the environment, so properties measuring the capacity of chemicals to participate in such processes constitute the major input for multimedia environmental models (Mackay, 2001). The tendency of chemicals to go to one media or another is usually assessed by screening and analyzing the magnitude of these properties (Gouin et al., 2000).

Equilibrium properties

The modeling of chemicals in environmental partitioning processes is typically based on partition coefficients, i.e., the ratio of equilibrium concentrations of two bordering media a and b (heterogeneous equilibrium):

$$K_{ab} = \frac{C_a}{C_b} \tag{1-2}$$

The value of a partition coefficient indicates the proportion in which a chemical distributes in two phases in equilibrium. Partition coefficients are usually determined for systems of air-water (K_{aw}), octanol-water (K_{ow}) and octanol-air (K_{oa}, obtained from the ratio K_{ow}/K_{aw}) since they can be subsequently related to partition coefficients of other systems by means of different correlations.

K_{aw} can be estimated from the ratio of vapor pressure (P_v, Pa) and water solubility (S_w, mol/m^3) or from Henry's law constants (H, Pa.m^3/mol), multiplying in both cases by 1/RT for obtaining dimensionless values (where R is the ideal gas constant 8.314 J/(mol·K) and T is the temperature of the system in K):

ISBN:978-84-693-4597-9 /DL:T.1010-2010

$$K_{aw} = \frac{P_v}{S_w}\left(\frac{1}{RT}\right) \qquad \text{or} \qquad K_{aw} = H\left(\frac{1}{RT}\right) \tag{1-3}$$

K_{aw} values are usually referred to properties experimentally determined. P_v is the pressure exerted by the vapor of a substance in a closed system; S_w is measured accounting the amount of substance dissolved in a given volume of water reaching saturation; and, H can be obtained from saturation concentrations of a substance in air and water. Difficulties arise when measuring P_v for non volatile chemicals or S_w for highly hydrophobic chemicals, improvements on measurement methods are usually required for overcoming these limitations (Mackay et al., 1992b).

Octanol is not present in the environment as a phase, but its similarity, in terms of properties and composition, to different organic phases (like sediments, soil and fat) makes it an ideal substitute of the latter. This facilitates the generation of equilibrium data between water and organic phases from K_{ow}. K_{ow} values are experimentally determined by shaking a closed octanol-water system containing a chemical of interest, measuring equilibrium concentrations of the chemical in both phases and later calculating their ratio (C_o/C_w). Experimental errors in K_{ow} values may result from quantities of emulsified octanol that remain suspended in water during the experiments. K_{ow} values may be also uncertain when determined for highly hydrophobic chemicals. K_{ow} can be used to estimate the organic carbon partition coefficient (K_{oc}, L/kg), since both properties have been found to be somewhat proportional (Karickhoff, 1981):

$$K_{oc} = 0.41 K_{ow} \tag{1-4}$$

The relationship between K_{oc} and K_{ow} may be variable and different correlations have been proposed to estimate K_{oc}, but their reliability tend to be uncertain due to the lack of sufficient experimental K_{oc} values (Gawlik et al., 1997). The soil-water partition coefficient, K_{sw}, may be estimated knowing K_{oc}, the organic carbon content (CORG, g/g) and the density of the solids ($\rho_{solid\text{-}soil}$, kg/L) as follows (Mackay, 2001):

$$K_p = \rho_{solid\text{-}soil}\ CORG\ K_{oc} \tag{1-5}$$

Other partition coefficients can be further derived for accounting partitioning in more specific systems, like lipid-water, fish-water, aerosol-air, vegetation-air, etc. The only restriction is the availability of more basic partition coefficients for multiplying them, invert them or using them in specific correlations (Mackay, 2001).

Kinetic properties

Degradation. Degradation of chemicals is usually expressed in terms of degradation half lives $t_{1/2}$, the time required by a certain amount of a chemical to reach half of its original concentration in a first order reaction controlled by a constant degradation rate k:

ISBN:978-84-693-4597-9 /DL:T.1010-2010

$$\frac{dC}{dt} = -kC \tag{1-6}$$

The equation above, when solved for the time in which the concentration is half of the original concentration, yields:

$$t_{1/2} = \left(\frac{-1}{k}\right)\ln\left(\frac{\frac{1}{2}C_o}{C_o}\right) \tag{1-7}$$

which gives the relation between a half live and its corresponding degradation rate:

$$t_{1/2} = \frac{0.693}{k} \tag{1-8}$$

Measurement of half lives or first order rate constants for most chemicals in the environment represents a challenging problem, environmental degradation processes depend not only on the inherent properties of chemicals but also on the nature of the media they are located (Klöpffer and Wagner, 2007). A common practice is to estimate the mean half live value of a chemical in a medium according to a range of observed half live values. Mackay defined a tabulation of mean half live values for 9 ranges of values (Mackay et al., 1992b), such classification is shown in Table 1-2.

Table 1-2. Mackay's criteria for the classification of chemicals according to their degradation half lives*.

Class	Mean $t_{1/2}$ (h)	Range of $t_{1/2}$ (h)
1	5	<10
2	17 (~ 1 day)	10-30
3	55 (~ 2 days)	30-100
4	170 (~ 1 week)	100-300
5	550 (~ 3 weeks)	300-1000
6	1700 (~ 2 months)	1000-3000
7	5500 (~ 8 months)	3000-10000
8	17000 (~ 2 years)	10000-30000
9	55000 (~ 6 years)	>30000

* (Mackay et al., 1992b).

Diffusion. The transportation of a chemical pollutant can occur macroscopically by advection and microscopically by diffusion. In advection processes, the chemical is moved by a fluid in motion, calculating the pollutant flow (mol/s) from the product of the rate flow of the fluid (m^3/s) and the corresponding pollutant concentration (mol/m^3). Microscopically, the flow of a pollutant i, driven by a concentration gradient in the fluid, occurs towards the region with the lowest concentration. For steady state conditions and one dimension (X), the diffusion process is described by Fick's first law:

$$J = -D_a \frac{dC}{dX} \tag{1-9}$$

ISBN:978-84-693-4597-9 /DL:T.1010-2010

where D_a (m^2/s) is the diffusion coefficient of a chemical in a medium a.

1.2.3 Estimation of physicochemical properties from molecular structure

The need of estimating unavailable physicochemical and toxicity data has raised the demand of quantitative structure-activity relationships (QSARs) (Devillers, 2003; Mackay et al., 2003; Mackay and Webster, 2003). QSARs and alike have been widely used in regulatory assessments (Cronin et al., 2003; Walker et al., 2002). QSARs relate information from the molecular structure of chemicals to a variety of processes, like chemical reactivity, chemical properties or toxicity (Winkler, 2002). When dealing with biological properties, these estimation models are usually referred to as QSARs; but, when used for modeling physicochemical properties, biodegradation or toxicity they are termed, respectively, quantitative structure-property relationships (QSPRs), quantitative structure-biodegradation relationships (QSBRs) or quantitative structure-toxicity relationships (QSTRs).

QSAR hystory. Studies that relate chemical information to a variety of processes have been developed from somewhat independent research lines. The physiological action of substances has been linked to its chemical composition and structure (Crum Brown and Fraser, 1868) and the narcotic potency of a set of organic chemicals was found to be related to their olive oil/water partitioning coefficients (Meyer, 1899; Overton, 1899). Melting points and boiling points were predicted for a series of homologous series of chemicals in a work that is considered to be the first QSPR ever reported (Mills, 1884). The ionization of bases and weak acids was studied, in terms of their molecular structure, under bacteriostatic activity (Albert, 1985; Albert et al., 1945; Bell and Roblin, 1942). Works for the explanation of substituent effects on organic reactions (Hammet, 1935; Hammett, 1970) and the separation of polar, steric and resonance effects (Taft, 1952) were foundations for the posterior development of the QSAR paradigm. Usually, the birth of QSAR models is attributed to works developed, independently, by Hansch and Fujita (1964) in one side and by Free and Wilson (1964) on the other. Structure-activity relationships of plant growth regulators and their dependency on Hammet constants and hydrophobicity were developed and published by 1962 (Hansch et al., 1962). The relative hydrophobicity of a substituent, π, was defined (Fujita et al., 1964) for the partition coefficients of a derivative and the parent molecule, P_X and P_H, respectively:

$$\pi_X = \log(P_X) - \log(P_H) \tag{1-10}$$

These hydrophobic constants were combined with Hammet's electronic constants into the linear Hansch equation and several of its variations (Hansch and Leo, 1995) to describe different types of biological activities:

$$\log\left(\frac{1}{C}\right) = a\sigma + b\pi + c\kappa \tag{1-11}$$

ISBN:978-84-693-4597-9 /DL:T.1010-2010

The equation formulated by Free and Wilson independently considered that all logarithmic biological activity values are the sum of the biological activity of the reference chemical and the group contributions of all substituents attached to different positions of the molecule:

$$\log\left(\frac{1}{C}\right) = \sum_{i=1}^{n} a_i x_i + \mu \qquad (1\text{-}12)$$

Learning algorithms. Research on QSAR has advanced rapidly supported on the introduction of non-linear equations in the models, potent computer based calculations and the definition of thousands of molecular descriptors measuring large varieties of molecular features (Hugo, 2002; Todeschini and Consonni, 2000). Nowadays, the relationship between a chemical process and molecular structure is given by a function, usually unknown and complex, in which parameters of the first (y) are related to a set of molecular descriptors (x_i):

$$\log(y) = f(x_1, x_2, ..., x_n) \qquad (1\text{-}13)$$

with the purpose of predicting the activity of chemicals not used in its development. QSARs must undergo different test stages for assessing their robustness, prediction ability and applicability domain. Thus, the selection of the data sets for training and testing the models constitute an important issue.

QSARs have greatly benefited from the introduction of artificial neural networks (ANNs) (Basheer and Hajmeer, 2000), like backpropagation networks (BPNs) (Hornik et al., 1989; Rumelhart et al., 1986) and radial basis functions (RBF) (Lo, 1998), for fitting data without a prior knowledge of involved functionality. However, finding optimal ANNs is a time consuming problem in which overfitting (Hawkins, 2004) may occur and very different models result from the same training data. Support vector regressions (SVRs) (Drucker et al., 1996), based on support vector machines (SVMs) (Cortes and Vapnik, 1995; Vapnik, 2000) have proven to yield slightly better results and be more robust than classical ANNs, e.g., in classification (Byvatov et al., 2003) and predictive tasks (Bhasin and Raghava, 2004; Hua and Sun, 2001). Additionally, models based on SVRs can be reproduced, i.e., a SVR model can be reconstructed with the same training data used in its development, contrasting ANN based models. SVMs are expected to replace ANNs in QSAR developments as fast as new software packages include SVM-based algorithms in their libraries (Xu et al., 2006). Methods typically used in pattern recognition problems (Jain et al., 2000; Wood, 1996) have also been incorporated in the repertoire of QSAR modeling techniques for manipulating large data sets by clustering, classifying or selecting relevant features (Lavine, 2006).

Uncertainty in QSAR predictions. Assuming that chemicals with similar molecular structure show similar properties, QSARs are meant to estimate chemical activity covering the chemical space as widely as possible (Willighagen et al., 2006). However, small structure differences may lead to large differences in activity (Nikolova and Jaworska, 2003). The predictive performance of QSARs is greatly affected by recurrent factors (Cronin and Schultz, 2003) like the quality of available training data (Stouch et al., 2003), the presence of outliers (Furusjö et al., 2006), the

ISBN:978-84-693-4597-9 /DL:T.1010-2010

selection of input features (Saeys et al., 2007) from large number of descriptors (Bredow and Jug, 2005; Burden et al., 2009; Duca and Hopfinger, 2001; Senese et al., 2004; Todeschini and Consonni, 2000), the selection and tuning of learning algorithms for building relationships (Basheer and Hajmeer, 2000; Xu et al., 2006), the risk of overtraining (Byvatov et al., 2003), the external validation of the models (Golbraikh and Tropsha, 2002; OECD, 2007; Schüürmann et al., 2008) and the definition of applicability domains (Weaver and Gleeson, 2008). The simultaneous optimization of all these elements is a problem that leads to almost infinite posibilites, resulting in a process that forces both modelers and users undergo cycles of optimism and frustration about the benefits of QSAR models (Johnson, 2008).

The development of QSARs is a matter of compromise between understanding, complexity and applicability of the models (Ferenç Darvas et al., 2006). So, QSAR models must not be considered universal and definitive models, but updatable tools that allow data estimations from available resources. Recent developments attempt to generate new types of molecular descriptors (Duca and Hopfinger, 2001; Senese et al., 2004; Todeschini and Consonni, 2000) or simply replace their use with molecular graphs (Goulon et al., 2005; Goulon et al., 2007). However, some time must pass in order to assimilate the applicability of new research trends in molecular modeling to practical applications.

Quantitative structure-property relationships

QSPRs have been developed for estimating basic properties of chemicals, most of them required in standard environmental fate assessments, using different combinations of molecular descriptors and methods (Devillers, 2003). Generally, the accuracy of estimation methods based on QSPRs is no better than that of experimentally determined properties, with some exceptions. Prediction accuracy close to experimental measurements have been achieved in QSPRs restricted for some families of chemicals or QSPRs using sets of test chemicals from their working database, but their performance with independent sets of chemicals have been limited (Taskinen and Yliruusi, 2003). Factors like quality of training data, correlation methods employed and external validation of models has been a matter of debate in the development of QSPRs. For these reason, it is difficult to catalogue any of them as definitive.

When experimentally determined properties can not be obtained from available databases, their estimation from molecular structure is recommended, especially from models accounting a large variety of chemicals and tested for large sets of chemicals not used for their training (Boethling et al., 2004). Among several estimation methods using molecular information, the collection of methods included in the free software package EPI suite^{TM} (SRC, 2008), based mostly on correlations of molecular fragments, has been traditionally recognized to be appropriate for a wide range of chemicals (Boethling et al., 2004).

ISBN:978-84-693-4597-9 /DL:T.1010-2010

Quantitative structure-biodegradation relationships

Known that for certain chemicals degradation or biodegradation processes influence greatly their fate in the environment, several attempts for training QSBRs have been carried out (Raymond et al., 2001). Experimental measurements of degradation are scarce and noisy, so most QSBRs are limited to sets of homologous chemicals. The output from these models may be expressed numerically (e.g., half lives, degradation rates, etc.) or discretely, in which a class is assigned to the chemicals (e.g., persistent or not). The model developed by Boethling et al. (Boethling et al., 1994), based on group contributions, has shown to be better than other models for predictive screenings of a large variety of chemicals, mainly because of the quality and size of its training data set (Raymond et al., 2001). It calculates the probability of a chemical to degrade or not within a range from 1 to 0, respectively. This model has been programmed and named BIOWIN^{TM}, included in EPI Suite^{TM}.

BIOWIN^{TM} has undergone different modifications. BIOWIN^{TM} was originally intended to estimate the probability of a chemical to degrade rapidly or slowly in aerobic conditions (Howard et al., 1992). After a revision of fragments and molecular weight, it was set to estimate the probability of biodegradation from experimental data and estimate primary/ultimate biodegradation times using evaluations of 200 chemicals by 17 experts in the field (Boethling et al., 1994). BIOWIN has included in its fifth and sixth versions (respectively, BIOWIN 5 and BIOWIN 6), 884 chemicals with biodegradation tests from the Japanese Ministry of international Trade and Industry (MITI) (JETOC, 1992), 385 classified as “readily degradable” and 499 classified as “not readily degradable” (Tunkel et al., 2000). The models (linear and non-linear) in this version, based on a total of 42 fragments and MW, have been trained and validated with MITI chemicals selected randomly, predicting correctly 83% of the training chemicals and 81% of the validation chemicals. The MITI experiments are considered to have an appreciable quality because the uniformity in their test conditions (Alikhanidi and Takahashi, 2004).

Despite of the improvements carried out in BIOWIN^{TM}, its use is recommended solely for screening purposes until the availability of more accurate degradation models. In general, its degradation predictions must be considered with caution. Environmental degradation processes are highly variable and correlations to molecular structure are still likely to fail (Aronson et al., 2006). In an attempt to predict degradation half lives of chemicals for their use in multimedia environmental models, a model based on the similarity of molecular structure have been developed and compared to the models of EPI Suite^{TM} (Kühne et al., 2007). The comparison criteria was the capacity of these models to predict correctly the representative classes of 293 chemicals in 4 compartments (air, water, sediments and soil), in accordance to the 9-class scheme proposed by Mackay (Mackay et al., 1992b) and shown in Table 1-2. The model based on structure similarity was said to be superior to the degradation models in EPI Suite^{TM}, according to Kühne et al. (2007). However, the performance of the former with chemicals different than those in the training set has not been tested yet.

ISBN:978-84-693-4597-9 /DL:T.1010-2010

1.2.4. Multimedia environmental modeling from molecular structure

Uncertainty in the input data of multimedia environmental models affects most environmental assessments. Especially, when it is associated to physicochemical properties of chemicals for which reliable experimental data is unavailable or poorly estimated with current methods. This problem, already pointed out by Wania and Mackay (1999a), is very likely to continue in the future as multimedia models rely on properties that must be determined by experimental or estimation procedures, in which uncertainty may be reduced but not completely eliminated.

Standard multimedia models are meant mostly for organic pollutants, but other types of pollutants requires special treatments (e.g., dissociating pollutants, metals, etc.) (Mackay, 2001). The use of partition and degradation properties in most multimedia environmental models limits the application of the latter to chemicals for which such properties can be easily obtained. For these reason, it is required to improve the description of environmental processes and enhance their range of applicability to more chemicals.

Poly-parameter liner free energy relationships

In an attempt to improve multimedia environmental assessments for polar organic chemicals, Breivik and Wania (2003) modified a standard level III model (CEM) by substituting the use of standard partitioning coefficients with solvation parameter models of the form (Abraham, 1993):

$$\log(K_{ab}) = c + rR_2 + s\Pi_2^H + a\sum\alpha_2^H + b\sum\beta_2^H + vV_x \qquad (1\text{-}14)$$

or

$$\log(K_{ab}) = c + rR_2 + s\Pi_2^H + a\sum\alpha_2^H + b\sum\beta_2^H + l\mathrm{Log}(L^{16}) \qquad (1\text{-}15)$$

where partition coefficients (K_{ab}) are related to five solute descriptors: excess molar refraction (R_2), dipolarity/polarizability (π_2^H), overall hydrogen-bond acidity (α_2^H), overall hydrogen bond basicity (β_2^H) and McGowan's characteristic volume (V_x) or the distribution constant of a chemical pollutant in n-hexadecane at 25ºC ($\mathrm{Log}(L^{16})$). The remaining symbols are constants (c, r, s, a, b, v, l).

The model proposed by Breivik and Wania, named ppLFER (poly-parameter linear free energy relationships), establishes a functionality between its output and the characterization of polar chemicals based on both degradation data and solvation parameters (Equations 1-14 and 1-15). After evaluating the model with theoretical solute descriptors for 40 chemicals, Breivik and Wania pointed out the possibility of using chemical structure for describing partitioning behavior and the need of additional research (Breivik and Wania, 2003). In a posterior work, the ppLFER model was assessed with 3 pharmaceuticals, showing that this model may be suitable

ISBN:978-84-693-4597-9 /DL:T.1010-2010

for pharmaceuticals with uncertain K_{ow} values (Zukowska et al., 2006). However, it was also found that half lives of chemicals in water have a major influence on the output of the model and accuracy in such input data is also required.

Structure-fate relationships

Another attempt of linking molecular information to environmental fate assessments involved the use of partial orders and Hasse diagrams to represent structure-fate relationships (Brüggemann et al., 2006) on 19 organic chemicals monitored in the river Main, Germany, for explaining simultaneously four environmental processes, namely, volatilization, sedimentation, persistence and advection. This approach was an effort to derive theoretical relationships between the molecular structure of chemicals and their fate in the environment, but the complexity of the approach makes difficult its extension to a wide number of chemicals as several parameters must be estimated and it is still not clear how to do it properly.

Quantitative structure-fate relationships

The NOMIRACLE project, "NOvel Methods for Integrated Risk Assessment CumuLative stressors in the Environment", studied in its work package 2.4, denominated "Region specific environmental fate", the use of supervised algorithms to estimate the environmental fate of chemicals when key physicochemical properties are unavailable. The output of such study comprised the project deliverables D.2.4.4 (Martínez et al., 2006c; Annex A.1), D.2.4.9 (Martínez et al., 2007b; Annex A.a2), D.4.12 (Martínez et al., 2008d; Annex A.a3) and D.2.4.13 (Martínez et al., 2008a; Annex A.a4).

Considering emission rates in one of various compartments, backpropagation networks were trained to predict level III environmental concentrations of chemicals in five compartments simultaneously (air, water, sediment, soil and vegetation) from reduced sets of properties (Martínez et al., 2006c; Annex A.1), mainly partition coefficients and degradation rate constants.

Since partition and degradation data are usually unavailable for most chemicals, it was proposed the training of supervised learning algorithms to link molecular descriptors to the output of MEMs (Martínez et al., 2007b; Annex A.a2), like standard QSARs linking molecular descriptors to chemical activity (Equation 1-13). The advantage of this approach, here named quantitative structure-fate relationships (QSFRs), with respect to its predecessors (like ppLFER or structure-fate relationships, described above) is that QSFRs are multivariate functions with parameters that can be easily tuned if enough training chemicals are available.

Several experiments were carried out on algorithms using semi-empirical molecular descriptors (Martínez et al., 2007b; Annex A.a2) or descriptors counting molecular constituents (Martínez et al., 2008a; Annex A.a4), yielding better results the latter ones. Other studies considered the extrapolation of scenarios through the use of output sensitivities (Martínez et al., 2007b; Annex A.a2) or the clustering of chemicals for improving fate estimations with class-tailored QSFRs (Martínez et al., 2008d; Annex

ISBN:978-84-693-4597-9 /DL:T.1010-2010

A.a3), but the need of improving the tuning of QSFR models was the main focus of the research within NOMIRACLE (Martínez et al., 2008a; Annex A.a4).

This study, titled "Quantitative structure-fate relationships for multimedia environmental analysis", discusses about the applicability of QSFRs to estimate the fate of chemicals for which physicochemical properties are unavailable. As explained later (Sections 1.3 to 1.5), it builds on experiments (Chapters 2, 3, 4) meant to update the findings within the NOMIRACLE project (Section 5.1), proposing better practices adaptable for future assessments (Sections 5.2 and 5.3).

1.3 Hypothesis

There is a need of estimating the fate of chemicals for which most properties are missing or uncertain, when standard multimedia environmental models are likely to be uncertain as well. Thus, the following hypothesis is formulated:

Since molecular structure is related to partitioning properties and to degradation data, relationships between molecular structure and the output of multimedia environmental models must be expected as well. In addition, such relations may overcome the uncertainty that properties estimated individually usually propagate throughout multimedia environmental models.

1.4 Objectives

The general objective for testing the hypothesis of this work is to relate the molecular structure of chemicals to the output of a standard multimedia model. This implies using machine learning algorithms for establishing quantitative structure-fate relationships (QSFRs) and evaluating the prediction of chemicals not included in the training process. With the general purpose in mind, the following specific objectives have been stated:

1st objective: *Compile data for modeling a reference pollution scenario, to which all analyses will be referred to, which implies: first, compiling input and output data from a standard Level III MEM for two sets of chemicals, one for training and testing learning algorithms (to be used as QPFR, QSFR, classifiers) and the other for their external validation, emitted at hypothetical constant rates in the same geographical scenario; and, second, compiling molecular data for the chemicals to assess.*

2nd objective: *Train learning algorithms to perform environmental fate estimations directly from reduced sets of physicochemical properties, instead of all the properties required by the reference MEM, establishing quantitative property-fate relationships (QPFRs).*

ISBN:978-84-693-4597-9 /DL:T.1010-2010

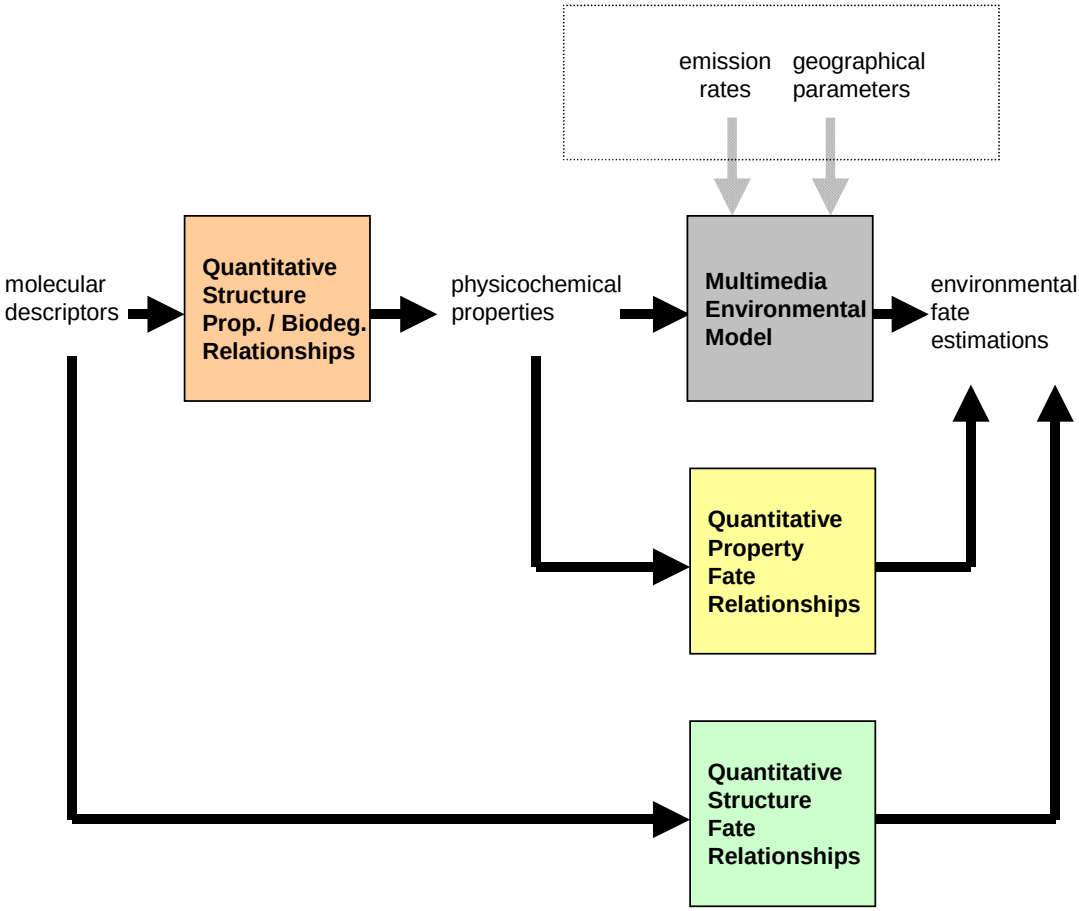


Figure 1-3 Scheme of how QSPRs, QSBRs, QPFRs and QSFRs are used in this work.
This work studies the estimation of environmental fate by means of quantitative property-fate relationships (QPFRs) and quantitative structure-fate relationships (QSFRs). Given constant emission rates and geographical parameters, QPFRs and QSFRs are meant to be alternatives to standard level III multimedia environmental models (MEMs) when large sets of properties must be estimated from either quantitative structure-activity relationships (QSARs) or quantitative structure-biodegradation relationships (QSBRs). The proposed approach is supported on the use of data mining techniques along with multimedia modeling examples, from a parent MEM, for training QPFRs and QSFRs.

3rd objective: *Train learning algorithms to perform environmental fate estimations directly from molecular information, establishing quantitative structure-fate relationships (QSFRs).*

4th objective: *Compare the fate predictions of the reference pollution scenario (1st objective), to fate predictions obtained by alternative paths: a) using the reference MEM with physicochemical properties, previously estimated by publicly available QSPRs and QSBRs; b) using QPFRs with reduced sets of physicochemical properties; c) using QPFRs with reduced sets of physicochemical properties, previously estimated by publicly available QSPRs and QSBRs; and, d) using QSFRs with available molecular information.*

As stated above, for setting a common ground of comparison, all the experiments and analyses of this research work are referred to level III fate predictions of chemicals, emitted hypothetically on a fixed geographical scenario. If others chemicals and

ISBN:978-84-693-4597-9 /DL:T.1010-2010

geographic scenarios were used other results could be obtained, but the applicability of the proposed methodology remains unchanged. The multimedia model selected for reference of QPFRs and QSFRs (1st Objective) is SimpleBox 3 (den Hollander and van de Meent, 2004; den Hollander et al., 2004). The reasons of selecting this model are based on: first, its previous comparison to field data in The Netherlands (Struijs and Peijnenburg, 2002) ; second, its previous comparison to other multimedia models (Armitage et al., 2007; Hollander et al., 2007; Lammel et al., 2007; Shatalov et al., 2005; Shatalov et al., 2004); and third, its extended use within EUSES (Lijzen and Rikken, 2004; Vermeire et al., 2005; Vermeire et al., 1997).

Figure 1-3 represents a scheme of the relationships with multimedia environmental fate estimations used and tested here. The direct inputs and outputs of SimpleBox 3,the reference MEM (1st objective), have been used as reference for the QPFR and QSFR models of this study (considering training, test and validation chemicals). QPFRs have been set to predict the fate of chemicals with reduced sets of physicochemical properties (2nd Objective); while, QSFRs have been set to do so directly from molecular structure, bypassing the use of properties for test and validation chemicals (3rd Objective). The uncertainty analysis on the direct inputs of the MEM (physicochemical properties getting values form statistical distributions), and thus simulating the path of using molecular structure to predict properties and using the latter in the MEM., was meant to compare the resulting fate predictions to those of QPFRs and QSFRs (4th Objective).

1.5 Contributions

This thesis work has been based on research carried out for the work package 2.4 of the project NOMIRACLE (Novel Methods for Integrated Risk Assessment of Cumulative Stressors in Europe), financed by the European Commission (FP6 Contract No. 003956). Table 1-3 lists the research works supporting this manuscript, four reports, three posters, two oral presentations and a paper. During the execution of the NOMIRACLE project, preliminary findings concerning QPFRs and QSFRs were documented in reports (Annexes A.a1, A.a2, A.a3 and A.a4) while results were presented, almost simultaneously, through posters (Annexes A.b1, A.b2 and A.b3) and oral presentations (Annexes A.c1, A.c2). With basis on such preliminary findings, a final paper has been prepared (Annex A.1) for discussing optimal results with QSFRs, considering 375 work chemicals and 80 validation chemicals, demonstrating the capabilities of the QSFR approach to the scientific community and discussing about its application in the assessment of new chemical pollutants.

The preliminary works within NOMIRACLE (Table 1-3) studied extensively QPFRs for emission rates in different compartments, QSFRs with different learning algorithms (like backpropagation networks, radial basis functions and support vector regressions) and sets of molecular descriptors, and class-tailored QSFRs. The present study presents results and discussions for the same geographical scenario used in such works, but implementing a more compact format that shows the evolution of the development of updated QPFRs and QSFRs, considering 375 work chemicals and 93 validation chemicals, aiming to demonstrate updated best practices in the QSFR approach.

ISBN:978-84-693-4597-9 /DL:T.1010-2010

This chapter has stated the problem of assessing the environmental fate of chemicals lacking of reliable properties and a proposal for solving it. The methods and algorithms used in the simulation experiments are described in Chapter 2. Chapter 3 describes both reference multimedia environmental data and molecular data of revised QPFR and QSFR models. Chapter 4 presents results and discussions, while Chapter 5 presents the conclusions of this work, its applicability in multimedia environmental analysis and guidelines for future improvements.

ISBN:978-84-693-4597-9 /DL:T.1010-2010

Table 1-3. List of research works supporting this study.

Item	Date	Title	Type			
			report	poster	oral presentation	paper
1	March 2006	Cognitive neural network-based intelligent system to identify the most important variables for the differences found in partitioning behaviour, transport pathways and exposure routes between chemicals (Martínez et al., 2006c; Annex A.1).	√			
2	May 2006	Modelling chemical multimedia partitioning with neural networks (Martínez et al., 2006a; Annex A.b1).		√		
3	Nov 2006	A Method for Modeling Chemical Multimedia Partitioning with Neural Networks and Classifiers (Martínez et al., 2006b; Annex A.c1).			√	
4	April 2007	Report on the most suitable artificial neural network architectures and molecular descriptors to estimate environmental multimedia behavior, including a sensitivity analysis of the effect of compartment sizes on multimedia concentrations (Martínez et al., 2007b; Annex A.a2).	√			
5	May 2007	Estimation of environmental multimedia partitioning of pollutants from molecular descriptors using artificial neural networks (Martínez et al., 2007a; Annex A.b2).		√		
6	April 2008	Report on the most suitable deterministic and probabilistic algorithms to pre-classify chemicals into families according to their partitioning with the aim of better predicting multimedia concentrations on artificial neural networks for each chemical family (Martínez et al., 2008d; Annex A.a3).	√			
7	April 2008	Estimating fate with neural network models (Martínez et al., 2008c; Annex A.c2).			√	
8	May 2008	Clustering the chemical space to estimate environmental multimedia partitioning of pollutants with kernel methods and molecular information (Martínez et al., 2008b; Annex A.b3).		√		
9	Dec. 2008	Report on the feasibility of predicting multimedia chemical partitioning with artificial neural network models by using functional group counts as input information (Martínez et al., 2008a; Annex A.a4).	√			
10	2010	Multimedia environmental chemical transport and distribution from molecular information (Martínez et al., 2010; Annex A.1)				√

ISBN:978-84-693-4597-9 /DL:T.1010-2010

References

Abraham MH. Scales of solute hydrogen-bonding: their construction and application to physicochemical and biochemical processes. Chem. Soc. Rev. 1993; 22: 73 - 83.

Albert A. Selective Toxicity: The Physicochemical Bases of Therapy. London: Chapman and Hall, 1985.

Albert A, Rubbo S, Goldacre R, Davey M, Stone J. The influence of chemical constitution on antibacterial activity. Part II: A general survey of the acridine series. Br. J. Exp. Pathol. 1945; 26: 160.

Alikhanidi S, Takahashi Y. Pesticide Persistence in the Environment - Collected Data and Structure-Based Analysis. Journal of Computer Chemistry, Japan 2004; 3: 59.

Armitage JM, Cousins IT, Hauck M, Harbers JV, Huijbregts MAJ. Empirical evaluation of spatial and non-spatial European-scale multimedia fate models: results and implications for chemical risk assessment. Journal of Environmental Monitoring 2007; 9: 572-581.

Aronson D, Boethling R, Howard P, Stiteler W. Estimating biodegradation half-lives for use in chemical screening. Chemosphere 2006; 63: 1953.

ATSDR, EPA. 2007 CERCLA priority list of hazardous substances that will be the subject of toxicological profiles and support document. ATSDR, EPA, 2007.

Basheer IA, Hajmeer M. Artificial neural networks: fundamentals, computing, design, and application. Journal of Microbiological Methods 2000; 43: 3.

Bell PH, Roblin RO. Studies in Chemotherapy. VII. A Theory of the Relation of Structure to Activity of Sulfanilamide Type Compounds. J. Am. Chem. Soc. 1942; 64: 2905-2917.

Beyer A, Matthies M. Criteria for Atmospheric Long-Range Transport Potential and Persistence of Pesticides and Industrial chemicals. Institute of Environmental Systems Research (University of Osnabrück), German Environmental Federal Agency, 2001.

Bhasin M, Raghava GPS. Prediction of CTL epitopes using QM, SVM and ANN techniques. Vaccine 2004; 22: 3195.

Boethling RS, Howard PH, Meylan W, Stiteler W, Beauman J, Tirado N. Group contribution method for predicting probability and rate of aerobic biodegradation. Environ. Sci. Technol. 1994; 28: 459-465.

Boethling RS, Howard PH, Meylan WM. Finding and estimating chemical property data for environmental assessment. Environmental Toxicology and Chemistry 2004; 23: 2290-2308.

Brandes LJ, Hollander Hd, Meent. Dvd. SimpleBox 2.0: a nested multimedia fate model for evaluating the environmental fate of chemicals. RIVM, Bilthoven, The Netherlands, 1996, pp. 156.

Bredow T, Jug K. Theory and range of modern semiempirical molecular orbital methods. Theoretical Chemistry Accounts: Theory, Computation, and Modeling (Theoretica Chimica Acta) 2005; 113: 1.

Breivik K, Alcock R, Li Y-F, Bailey RE, Fiedler H, Pacyna JM. Primary sources of selected POPs: regional and global scale emission inventories. Environmental Pollution 2004; 128: 3.

Breivik K, Vestreng V, Rozovskaya O, Pacyna JM. Atmospheric emissions of some POPs in Europe: a discussion of existing inventories and data needs. Environmental Science & Policy 2006; 9: 663.

Breivik K, Wania F. Evaluating a Model of the Historical Behavior of Two Hexachlorocyclohexanes in the Baltic Sea Environment. Environ. Sci. Technol. 2002; 36: 1014-1023.

Breivik K, Wania F. Expanding the Applicability of Multimedia Fate Models to Polar Organic Chemicals. Environmental Science and Technology 2003; 37: 4934.

Brüggemann R, Restrepo G, Voigt K. Structure-Fate Relationships of Organic Chemicals Derived from the Software Packages E4CHEM and WHASSE. Journal of Chemical Information and Modeling 2006; 46: 894.

Builtjes PJH. The LOTOS - Long Term Ozonhe Simulation - project. TNO-MEP, Apeldoorn, The Netherlands,

ISBN:978-84-693-4597-9 /DL:T.1010-2010

1992.

Burden FR, Polley MJ, Winkler DA. Toward Novel Universal Descriptors: Charge Fingerprints. Journal of Chemical Information and Modeling 2009; 49: 710-715.

Byvatov E, Fechner U, Sadowski J, Schneider G. Comparison of Support Vector Machine and Artificial Neural Network Systems for Drug/Nondrug Classification. J. Chem. Inf. Comput. Sci. 2003; 43: 1882-1889.

CAS. CAS Statistical Summary 1907-2007. American Chemical Society, 2008.

Citra MJ. Incorporating Monte Carlo analysis into multimedia environmental fate models. Environmental Toxicology and Chemistry 2004; 23: 1629-1633.

Cortes C, Vapnik V. Support-vector networks. Machine Learning 1995; 20: 273.

Cronin MT, Walker JD, Jaworska JS, Comber MH, Watts CD, Worth AP. Use of QSARs in international decision-making frameworks to predict health effects of chemical substances. Environmental Health Perspectives 2003; 111: 1376-90.

Cronin MTD, Schultz TW. Pitfalls in QSAR. Journal of Molecular Structure: THEOCHEM 2003; 622: 39.

Crum Brown A, Fraser TR. On the connection between chemical constitution and physiologic action. Part 1. On the physiological action of salts of the ammonium bases, derived from strychnia, brucia, thebia, codeia, morphia and nicotia. Trans. Roy. Soc. 1868; 25: pp.151-203.

den Hollander HA, van de Meent D. Appendix to SimpleBox 3.0: A multimedia mass balance model for evaluating the environmental fate of chemicals. RIVM, 2004.

den Hollander HA, van Eijkeren JCH, van de Meent D. SimpleBox 3.0. RIVM, Bilthoven, The Netherlands, 2004.

Devillers J. A decade of research in environmental QSAR. SAR and QSAR in Environmental Research 2003; 14: 1 - 6.

Devillers J, Bintein S, Karcher W. CHEMFRANCE: A regional level iii fugacity model applied to France. Chemosphere 1995; 30: 457.

Drucker H, Burges CJC, Kaufman L, Smola A, Vapnik V. Support Vector Regression Machines. Advances in Neural Information Processing Systems 1996: 155-161.

Duca JS, Hopfinger AJ. Estimation of Molecular Similarity Based on 4D-QSAR Analysis: Formalism and Validation. Journal of Chemical Information and Computer Sciences 2001; 41: 1367-1387.

EEC. COUNCIL REGULATION (EEC) No 793/93 of 23 March 1993 on the evaluation and control of the risks of existing substances. EEC, 1993.

Eisenberg JNS, Bennett DH, McKone TE. Chemical Dynamics of Persistent Organic Pollutants: A Sensitivity Analysis Relating Soil Concentration Levels to Atmospheric Emissions. Environ. Sci. Technol. 1998; 32: 115-123.

European Commission. Regulation (EC) No 1907/2006 of the European Parliament and of the Council of 18 December 2006 concerning the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH), establishing a European Chemicals Agency, amending Directive 1999/45/EC and repealing Council Regulation (EEC) No 793/93 and Commission Regulation (EC) No 1488/94 as well as Council Directive 76/769/EEC and Commission Directives 91/155/EEC, 93/67/EEC, 93/105/EC and 2000/21/EC. European Commission http://ec.europa.eu/environment/chemicals/reach/reach_intro.htm, Brussels, Belgium, 2006.

Fenner K, Scheringer M, Hungerbühler K. Prediction of overall persistence and long-range transport potential with multimedia fate models: robustness and sensitivity of results. Environmental Pollution 2004; 128: 189.

Fenner K, Scheringer M, MacLeod M, Matthies M, McKone T, Stroebe M, et al. Comparing Estimates of Persistence and Long-Range Transport Potential among Multimedia Models. Environ. Sci. Technol. 2005; 39: 1932-1942.

Ferenç Darvas, Oliver Kappe, Gisbert Schneider, Michael Wiese, Kubinyi H. QSAR/QSPR Modelling - Finding Rules in Noisy Data? QSAR & Combinatorial Science 2006; 25: 811-812.

Fjodorova N, Novich M, Vrachko M, Smirnov V, Kharchevnikova N, Zholdakova Z, et al. Directions in QSAR Modeling for Regulatory Uses in OECD Member Countries, EU and in Russia. Journal of Environmental Science

ISBN:978-84-693-4597-9 /DL:T.1010-2010

and Health, Part C 2008; 26: 201 - 236.

Free SM, Wilson JW. A Mathematical Contribution to Structure-Activity Studies. J. Med. Chem. 1964; 7: 395-399.

Fujita T, Iwasa J, Hansch C. A New Substituent Constant, π, Derived from Partition Coefficients. J. Am. Chem. Soc. 1964; 86: 5175-5180.

Furusjö E, Svenson A, Rahmberg M, Andersson M. The importance of outlier detection and training set selection for reliable environmental QSAR predictions. Chemosphere 2006; 63: 99.

Gawlik BM, Sotiriou N, Feicht EA, Schulte-Hostede S, Kettrup A. Alternatives for the determination of the soil adsorption coefficient, KOC, of non-ionicorganic compounds -- a review. Chemosphere 1997; 34: 2525.

Golbraikh A, Tropsha A. Beware of q2! Journal of Molecular Graphics and Modelling 2002; 20: 269.

Gong SL, Huang P, Zhao TL, Sahsuvar L, Barrie LA, Kaminski JW, et al. GEM/POPs: a global 3-D dynamic model for semi-volatile persistent organic pollutants – Part 1: Model description and evaluations of air concentrations. Atmos. Chem. Phys. 2007; 7: 4001.

Gouin T, Mackay D, Webster E, Wania F. Screening Chemicals for Persistence in the Environment. Environ. Sci. Technol. 2000; 34: 881-884.

Goulon A, Duprat A, Dreyfus G. From Hopfield nets to recursive networks to graph machines: Numerical machine learning for structured data. Theoretical Computer Science 2005; 344: 298.

Goulon A, Picot T, Duprat A, Dreyfus G. Predicting activities without computing descriptors: graph machines for QSAR. SAR and QSAR in Environmental Research 2007; 18: 141.

Hammet LP. Chem. Rev. 1935; 17: 125-136.

Hammett LP. Physical Organic Chemistry. New York: McGraw-Hill, 1970.

Hansch C, Fujita T. ρ-σ-π Analysis. A Method for the Correlation of Biological Activity and Chemical Structure. J. Am. Chem. Soc. 1964; 86: 1616-1626.

Hansch C, Leo A. Exploring QSAR. Fundamentals and Applications in Chemistry and Biology. Washington, DC: American Chemical Society, 1995.

Hansch C, Maloney PP, Fujita T, Muir RM. Correlation of Biological Activity of Phenoxyacetic Acids with Hammett Substituent Constants and Partition Coefficients. Nature 1962; 194: 178.

Hansen KM, Christensen JH, Brandt J, Frohn LM, Geels C. Modelling atmospheric transport of Î±-hexachlorocyclohexane in the Northern Hemispherewith a 3-D dynamical model: DEHM-POP. Atmos. Chem. Phys. 2004; 4: 1125.

Hawkins DM. The Problem of Overfitting. J. Chem. Inf. Comput. Sci. 2004; 44: 1-12.

Hollander A, Huijbregts MAJ, Ragas AMJ, Meent D. BasinBox: a generic multimedia fate model for predicting the fate of chemicals in river catchments. Living Rivers: Trends and Challenges in Science and Management, 2006, pp. 21.

Hollander A, Sauter F, den Hollander H, Huijbregts M, Ragas A, van de Meent D. Spatial variance in multimedia mass balance models: Comparison of LOTOS-EUROS and SimpleBox for PCB-153. Chemosphere 2007; 68: 1318.

Hornik K, Stinchcombe M, White H. Multilayer feedforward networks are universal approximators. Neural Networks 1989; 2: 359.

Howard PH, Boethling RS, Stiteler WM, Meylan WM, Hueber AE, Beauman JA, et al. Predictive model for aerobic biodegradability developed from a file of evaluated biodegradation data. Environmental Toxicology and Chemistry 1992; 11: 593-603.

Howard PS, Boethling RS, Jarvis WF, Meylan WM, Michalenko EM. Handbook of environmental degradation rates. Chelsea, MI: Lewis Publications, 1991.

Hua S, Sun Z. A novel method of protein secondary structure prediction with high segment overlap measure:

ISBN:978-84-693-4597-9 /DL:T.1010-2010

support vector machine approach. Journal of Molecular Biology 2001; 308: 397.

Huang P, Gong SL, Zhao TL, Neary L, Barrie LA. GEM/POPs: a global 3-D dynamic model for semi-volatile persistent organic pollutants; Part 2: Global transports and budgets of PCBs. Atmos. Chem. Phys. 2007; 7: 4015.

Hugo K. From Narcosis to Hyperspace: The History of QSAR. Quantitative Structure-Activity Relationships 2002; 21: 348-356.

Ilyina T, Pohlmann T, Lammel G, Sündermann J. A fate and transport ocean model for persistent organic pollutants and its application to the North Sea. Journal of Marine Systems 2006; 63: 1.

Jain AK, Duin RPW, Jianchang M. Statistical pattern recognition: a review. Pattern Analysis and Machine Intelligence, IEEE Transactions on 2000; 22: 4.

JETOC. Biodegradation and Bioaccumulation Data of Existing Chemicals Based on the Chemical Substances Control Law (CSCL Japan). Japan Chemical Industry Ecology-Toxicology & Information Center (JETOC), Tokyo, 1992.

Johnson SR. The Trouble with QSAR (or How I Learned To Stop Worrying and Embrace Fallacy). Journal of Chemical Information and Modeling 2008; 48: 25-26.

Karickhoff SW. Semi-empirical estimation of sorption of hydrophobic pollutants on natural sediments and soils. Chemosphere 1981; 10: 833.

Kawamoto K, MacLeod M, Mackay D. Evaluation and comparison of multimedia mass balance models of chemical fate: application of EUSES and ChemCAN to 68 chemicals in Japan. Chemosphere 2001; 44: 599.

Klöpffer W, Wagner B. Persistence revisited. Environmental Science and Pollution Research 2007; 14: 141.

Kühne R, Breitkopf C, Schüürmann G. Error propagation in fugacity level-III models in the case of uncertain physicochemical properties. Environmental Toxicology and Chemistry 1997; 16: 2067-2069.

Kühne R, Ebert R-U, Schüürmann G. Estimation of Compartmental Half-lives of Organic Compounds - Structural Similarity versus EPI-Suite. QSAR & Combinatorial Science 2007; 26: 542-549.

Lammel G. Effects of time-averaging climate parameters on predicted multicompartmental fate of pesticides and POPs. Environmental Pollution 2004; 128: 291.

Lammel G, Feichter J, Leip A. Long-range transport and global distribution of semivolatile organic compounds: A case study on two modern agrochemicals. Max Planck Institute for Meteorology, Hamburg, Germany, 2001, pp. 44 pp.

Lammel G, Klöpffer W, Semeena V, Schmidt E, Leip A. Multicompartmental fate of persistent substances. Environmental Science and Pollution Research 2007; 14: 153.

Lavine BK. Pattern Recognition. Critical Reviews in Analytical Chemistry 2006; 36: 153 - 161.

Leeuw FAAM, Rheineck Leyssius HJ. Modeling study of SOx and NOx transport during the January 1985 SMOG episode. Water, Air, & Soil Pollution 1990; 51: 357.

Lijzen JPA, Rikken MGJ. European Union System for the Evaluation of Substances 2.0 (EUSES 2.0); background report. RIVM, Bilthoven, the Netherlands., 2004, pp. 454.

Lo JT-H. Multilayer perceptrons and radial basis functions are universal robust approximators. IEEE International Conference on Neural Networks - Conference Proceedings. 2, 1998, pp. 1311.

Lohmann R, Breivik K, Dachs J, Muir D. Global fate of POPs: Current and future research directions. Environmental Pollution 2007; 150: 150.

Mackay D. Multimedia Environmental Models - The Fugacity Approach. Chelsea, MI: Lewis Publishers, 1991.

Mackay D. Multimedia Environmental Models - The Fugacity Approach. Boca Ratón: Lewis Publishers, 2001.

Mackay D, Di Guardo A, Paterson S, Cowan CE. Evaluating the environmental fate of a variety of types of chemicals using the EQC model. Environmental Toxicology and Chemistry 1996a; 15: 1627.

Mackay D, Di Guardo A, Paterson S, Kicsi G, Cowan CE. Assessing the fate of new and existing chemicals: a

ISBN:978-84-693-4597-9 /DL:T.1010-2010

five-stage process. Environmental Toxicology and Chemistry 1996b; 15: 1618-1626.

Mackay D, Di Guardo A, Paterson S, Kicsi G, Cowan CE, Kane DM. Assessment of chemical fate in the environment using evaluative, regional and local scale models: Illustrative application to chlorobenzene and linear alkylbenzene sulfonates. Environmental Toxicology and Chemistry 1996c; 15: 1638-1648.

Mackay D, Hubbarde J, Webster E. The role of QSARs and fate models in chemical hazard and risk assessment. QSAR & Combinatorial Science 2003; 22: 106-112.

Mackay D, Joy M, Paterson S. A quantitative water, air, sediment interaction (QWASI) fugacity model for describing the fate of chemicals in lakes. Chemosphere 1983; 12: 981.

Mackay D, Paterson S. Evaluating the multimedia fate of organic chemicals: a level III fugacity model. Environ. Sci. Technol. 1991; 25: 427-436.

Mackay D, Paterson S, Shiu WY. Generic models for evaluating the regional fate of chemicals. Chemosphere 1992a; 24: 695.

Mackay D, Paterson S, Tam DD. Assessments of chemical fate in Canada: continued development of a fugacity model., 1991.

Mackay D, Shiu W-Y, Ma KC. Illustrated Handbook of Physical-Chemical Properties and Environmental Fate for Organic Chemicals: Lewis Publishers Inc., 1992b.

Mackay D, Wan-Yiu Shiu. Physical-Chemical Properties and Environmental Fate Handbook on CD-ROM. Chapman & Hall / CDCnetBASE, 2000.

Mackay D, Webster E. A perspective on environmental models and QSARs. SAR and QSAR in Environmental Research 2003; 14: 7.

MacLeod M, Riley WJ, McKone TE. Assessing the Influence of Climate Variability on Atmospheric Concentrations of Polychlorinated Biphenyls Using a Global-Scale Mass Balance Model (BETR-Global). Environ. Sci. Technol. 2005; 39: 6749-6756.

MacLeod M, Woodfine DG, Mackay D, McKone TE, Bennett DH, Maddalena RL. BETR North America: A regionally segmented multimedia contaminant fate model for North America. Environmental Science & Pollution Research; Journal Volume: 8; Journal Issue: 3; Other Information: Journal Publication Date: 2001; PBD: 1 Mar 2001 2001.

Martínez I, Espinosa G, Grifoll J, Cohen Y, Giralt F. Modelling chemical multimedia partitioning with neural networks. SETAC Europe 16th Annual Meeting, The Hague, The Netherlands, 2006a.

Martínez I, Espinosa G, Rallo R, Grifoll J, Cohen Y, Giralt F. A Method for Modeling Chemical Multimedia Partitioning with Neural Networks and Classifiers. AICHE Annual Meeting, San Francisco, United States, 2006b.

Martínez I, Espinosa G, Rallo R, Grifoll J, Cohen Y, Giralt. F. Estimation of environmental multimedia partitioning of pollutants from molecular descriptors using artificial neural networks. SETAC Europe 17th Annual Meeting, Oporto, Portugal, 2007a.

Martínez I, Grifoll J, Giralt F, Rallo R, Espinosa G. Report on the feasibility of predicting multimedia chemical partitioning with artificial neural network models by using functional group counts as input information. Universitat Rovira i Virgili, Tarragona, Spain, 2008a. NOMIRACLE Report 2.4.13

Martínez I, Grifoll J, Giralt J, Rallo R and Cohen Y. Multimedia environmental chemical transport and distribution from molecular information. STOTEN. To be submitted in 2010.

Martínez I, Grifoll J, Giralt F, Rallo R, Espinosa G, Cohen Y. Clustering the chemical space to estimate environmental multimedia partitioning of pollutants with Kernel methods and molecular descriptors. SETAC Europe 18th Annual Meeting, Warsawa, Poland, 2008b.

Martínez I, Grifoll J, Rallo R. Cognitive neural network-based intelligent system to identify the most important variables for the differences found in partitioning behaviour, transport pathways and exposure routes between chemicals. Universitat Rovira i Virgili, Tarragona, Spain, 2006c. NOMIRACLE Report 2.4.4

Martínez I, Grifoll J, Rallo R, Espinosa G, Giralt F. Estimating fate with Neural network models. NoMiracle Workshop on Chemical Exposure, UFZ Leipzig, Germany, 2008c.

ISBN:978-84-693-4597-9 /DL:T.1010-2010

Martínez I, Grifoll J, Rallo R, Giralt F. Report on the most suitable artificial neural network architectures and molecular descriptors to estimate environmental multimedia behavior, including a sensitivity analysis of the effect of compartment sizes on multimedia concentrations. Universitat Rovira i Virgili, Tarragona, Spain, 2007b. NOMIRACLE Report 2.4.9

Martínez I, Grifoll J, Rallo R, Giralt F. Report on the most suitable deterministic and probabilistic algorithms to pre-classify chemicals into families according to their partitioning with the aim of better predicting multimedia concentrations on artificial neural networks for each chemical family. Universitat Rovira i Virgili, Tarragona, Spain, 2008d. NOMIRACLE Report 2.4.12

Matthijsen Jg, Sauter F, De Waal ES. Modelling of particulate matter on a European scale. In: Keller J, Andreani-Aksojoglu S, editors. GLOREAM Symposium, Wengen, Switzerland, 2002.

McKone TE, Enoch KG. CalTOX™, A Multimedia Total Exposure Model Spreadsheet User's Guide Version 4.0 (Beta). Environmental Energy Technologies Division, Lawrence Berkeley National Laboratory., Berkeley, California, United States., 2002.

McKone TE, Hall D, Kastenberg. WE. CalTOX Version 2.3 Description of Modifications and Revisions. Human and Ecological Risk Division Department of Toxic Substances, Control California Environmental Protection Agency, Sacramento, California, US., 1997.

McLachlan MS, Czub G, Wania F. The Influence of Vertical Sorbed Phase Transport on the Fate of Organic Chemicals in Surface Soils. Environ. Sci. Technol. 2002; 36: 4860-4867.

Meyer H. Zur Theorie der Alkoholnarkose. Erste Mittheilung. Welche Eigenschaft der Anästhetica bedingt ihre narkotische Wirkung? Arch. Exp. Pathol. Pharmakol. 1899; 42: pp. 109–118.

Meyer T, Wania F. What environmental fate processes have the strongest influence on a completely persistent organic chemical's accumulation in the Arctic? Atmospheric Environment 2007; 41: 2757.

Mills E. On melting-point and boiling-point as related to chemical composition. Philosophical Magazine 1884; 17: pp. 173-187.

Nikolova N, Jaworska J. Approaches to Measure Chemical Similarity - a Review. QSAR & Combinatorial Science 2003; 22: 1006-1026.

OECD. Guidance Document on the Validation of (Quantitative) Structure-Activity Relationship [(Q)SAR] Models. OECD Series on Testing and Assessment 69., 2007.

Overton E. Studien über die Narkose. 1899; 4: pp. 88-135.

Prevedouros K, MacLeod M, Jones KC, Sweetman AJ. Modelling the fate of persistent organic pollutants in Europe: parameterisation of a gridded distribution model. Environmental Pollution 2004; 128: 251.

Quéguiner S, Musson-Genon L. Modelling of Atmospheric Transport of POPs at the European Scale with a 3D Dynamical Model Polair3D-POP. Air Pollution Modeling and Its Application XIX, 2008, pp. 669.

Raymond JW, Rogers TN, Shonnard DR, Kline AA. A review of structure-based biodegradation estimation methods. Journal of Hazardous Materials 2001; 84: 189.

Rumelhart DE, Hinton GE, Williams. RJ. Learning representations by back propagating errors. Nature 1986; 323: 533-536.

Saeys Y, Inza I, Larranaga P. A review of feature selection techniques in bioinformatics. Bioinformatics 2007; 23: 2507-2517.

Schaap M, Roemer M, Sauter F, Boersen GAC, Timmermans RMA, Builtjes PJH. LOTOS-EUROS: documentation. TNO, Apeldoorn, The Netherlands, 2005.

Schaap M, Timmermans RMA, Roemer M, Boersen GAC, Builtjes PJH, Sauter FJ, et al. The LOTOS EUROS model: description, validation and latest developments. International Journal of Environment and Pollution 2008; 32: 270.

Schaap M, van Loon M, ten Brink HM, Dentener FJ, Builtjes PJH. Secondary inorganic aerosol simulations for Europe with special attention to nitrate. Atmos. Chem. Phys. 2004; 4: 857.

Scheringer M. Persistence and Spatial Range as Endpoints of an Exposure-Based Assessment of Organic

ISBN:978-84-693-4597-9 /DL:T.1010-2010

Chemicals. Environ. Sci. Technol. 1996; 30: 1652-1659.

Scheringer M, Salzmann M, Stroebe M, Wegmann F, Fenner K, Hungerbühler K. Long-range transport and global fractionation of POPs: insights from multimedia modeling studies. Environmental Pollution 2004; 128: 177.

Scheringer M, Stroebe M, Held H. Chemrange 2.1—A Multimedia Transport Model for Calculating Persistence and Spatial Range of Organic Chemicals. Swiss Federal Institute of Technology Zürich, Potsdam Institute for Climate Impact Research, 2002.

Scheringer M, Wegmann F, Fenner K, Hungerbuhler K. Investigation of the Cold Condensation of Persistent Organic Pollutants with a Global Multimedia Fate Model. Environ. Sci. Technol. 2000; 34: 1842-1850.

Schummer J. Scientometric studies on chemistry I: The exponential growth of chemical substances, 1800–1995. Scientometrics 1997a; 39: 107.

Schummer J. Scientometric studies on chemistry II: Aims and methods of producing new chemical substances. Scientometrics 1997b; 39: 125.

Schüürmann G, Ebert R-U, Chen J, Wang B, Kühne R. External Validation and Prediction Employing the Predictive Squared Correlation Coefficient - Test Set Activity Mean vs Training Set Activity Mean. Journal of Chemical Information and Modeling 2008; 48: 2140-2145.

Semeena VS, Feichter J, Lammel G. Effects of various scenarios upon entry of DDT and γ-HCH into the global environmental on their fate as predicted by a multicompartment chemistry-transport model. Fresenius Environmental Bulletin 2003; 12: 925-939.

Senese CL, Duca J, Pan D, Hopfinger AJ, Tseng YJ. 4D-Fingerprints, Universal QSAR and QSPR Descriptors. Journal of Chemical Information and Computer Sciences 2004; 44: 1526-1539.

Shatalov V, Mantseva E, Baart A, Bartlett P, Breivik K, Christensen J, et al. POP Model Intercomparison Study. Stage II. Comparison of mass balance estimates and sensitivity studies, 2005.

Shatalov V, Mantseva E, Baart A, Bartlett P, Breivik K, Christensen J, et al. POP Model Intercomparison Study - Stage I. Comparison of descriptions of main processes determining POP behaviour in various environmental compartments. In: East MSC-, editor, 2004.

Syracuse Research Corporation. EPI Suite v4.00. SRC, 2008.

Stouch TR, Kenyon JR, Johnson SR, Chen X-Q, Doweyko A, Li Y. In silico ADME/Tox: why models fail. Journal of Computer-Aided Molecular Design 2003; 17: 83.

Struijs J, W.J.G.M.Peijnenburg. Predictions by the multimedia environmental fate model SimpleBox compared to field data: Intermedia concentration ratios of two phthalate esters. RIVM, Bilthoven, 2002, pp. 62.

Suzuki N, Murasawa K, Sakurai T, Nansai K, Matsuhashi K, Moriguchi Y, et al. Geo-Referenced Multimedia Environmental Fate Model (G-CIEMS): Model Formulation and Comparison to the Generic Model and Monitoring Approaches. Environ. Sci. Technol. 2004; 38: 5682-5693.

Taft RW. Polar and Steric Substituent Constants for Aliphatic and o-Benzoate Groups from Rates of Esterification and Hydrolysis of Esters. J. Am. Chem. Soc. 1952; 74: 3120-3128.

Taskinen J, Yliruusi J. Prediction of physicochemical properties based on neural network modelling. Advanced Drug Delivery Reviews 2003; 55: 1163.

Tickner J, Geiser K, Coffin M. The U.S. Experience in Promoting Sustainable Chemistry (9 pp). Environmental Science and Pollution Research 2005; 12: 115.

Todeschini R, Consonni V. Handbook of Molecular Descriptors: Wiley-VCH, 2000.

Toose L, Woodfine DG, MacLeod M, Mackay D, Gouin J. BETR-World: a geographically explicit model of chemical fate: application to transport of [alpha]-HCH to the Arctic. Environmental Pollution 2004; 128: 223.

Toussant M. A scientific milestone. Chemical & Engineering News 2009; 87: 3.

Tunkel J, Howard PH, Boethling RS, Stiteler W, Loonen H. Predicting ready biodegradability in the Japanese Ministry of International Trade and Industry test. Environmental Toxicology and Chemistry 2000; 19: 2478-2485.

ISBN:978-84-693-4597-9 /DL:T.1010-2010

UCLA. CalTOX™, A Multimedia Total Exposure Model For Hazardous-Waste Sites Spreadsheet User's Guide Version 1.5. 1995.

UN-ECE. Protocol to the 1979 convention on long range transboundary air pollution on persistent organic pollutants and executive body decision 1998/2 on information to be submitted and the procedure for adding substances to annexes I, II or III to the protocol on persistent organic pollutants. UN-ECE, United Nations, New York and Geneva, 1998.

UNEP. Final act of the conference of plenipotentiaries on the Stockholm convention on persistent organic pollutants. UNEP, United Nations, 2001.

US-EPA. TRIM-FaTE Technical Support Document. Volume I: Description of Module. US Environmental Protection Agency, North Carolina, 2002a.

US-EPA. TRIM-FaTE Technical Support Document. Volume II: Description of Chemical Transport and Transformation Algorithms. US Environmental Protection Agency, North Carolina, 2002b.

US-EPA. Inventory Update Rule. Office of Pollution Prevention and Toxics, Environmental Protection Agency http://www.epa.gov/oppt/iur/, Washington, 2006.

van de Meent D. SIMPLEBOX: a generic multimedia fate evaluation model. RIVM, Bilthoven, The Netherlands, 1993.

Van Loon M. Numerical smog prediction, I: the physical and chemical model. CWI, Amsterdam, The Netherlands, 1994.

Van Loon M. Numerical smog prediction, II: grid refinement and its application to the dutch smog prediction model. CWI, Amsterdam, The Netherlands, 1995.

Vapnik VN. The Nature of Statistical Learning Theory: Springer-Verlag New York, Inc., 2000.

Vermeire T, Rikken M, Attias L, Boccardi P, Boeije G, Brooke D, et al. European union system for the evaluation of substances: the second version. Chemosphere 2005; 59: 473.

Vermeire TG, Jager DT, Bussian B, Devillers J, den Haan K, Hansen B, et al. European Union System for the Evaluation of Substances (EUSES). Principles and structure. Chemosphere 1997; 34: 1823.

Walker JD, Carlsen L, Hulzebos E, Simon-Hettich B. Global Government applications of analogues, SARs and QSARs to predict aquatic toxicity, chemical or physical properties, environmental fate parameters and health effects of organic chemicals. SAR and QSAR in Environmental Research 2002; 13: 607.

Wania F. Assessing the Potential of Persistent Organic Chemicals for Long-Range Transport and Accumulation in Polar Regions. Environ. Sci. Technol. 2003; 37: 1344-1351.

Wania F, Breivik K, Persson NJ, McLachlan MS. CoZMo-POP 2 - A fugacity-based dynamic multi-compartmental mass balance model of the fate of persistent organic pollutants. Environmental Modelling & Software 2006; 21: 868.
Wania F, Daly GL. Estimating the contribution of degradation in air and deposition to the deep sea to the global loss of PCBs. Atmospheric Environment 2002; 36: 5581.

Wania F, Mackay D. Modelling the global distribution of toxaphene: A discussion of feasibility and desirability. Chemosphere 1993; 27: 2079.

Wania F, Mackay D. A global distribution model for persistent organic chemicals. Science of The Total Environment 1995; 160-161: 211.

Wania F, Mackay D. The evolution of mass balance models of persistent organic pollutant fate in the environment. Environmental Pollution 1999a; 100: 223.

Wania F, Mackay D. Global chemical fate of α-hexachlorocyclohexane. 2. Use of a global distribution model for mass balancing, source apportionment, and trend prediction. Environmental Toxicology and Chemistry 1999b; 18: 1400-1407.

Wania F, Mackay D, Li Y-F, Bidleman TF, Strand A. Global chemical fate of α-hexachlorocyclohexane. 1. Evaluation of a global distribution model. Environmental Toxicology and Chemistry 1999; 18: 1390-1399.

Wania F, Persson J, Di Guardo A, McLachlan MS. The POPCYCLING-Baltic Model. A Non-Steady-State

ISBN:978-84-693-4597-9 /DL:T.1010-2010

Multicompartment Mass Balance Model of the Fate of Persistent Organic Pollutants in the Baltic Sea Environment. Norwegian Institute for Air Research, Kjeller, Norway, 2000.

Weaver S, Gleeson MP. The importance of the domain of applicability in QSAR modeling. Journal of Molecular Graphics and Modelling 2008; 26: 1315.

Webster E, Hubbarde J, Mackay D. Upgrading the ChemCAN Model: Version 4.95 to 6.00. Canadian Environmental Modelling Centre, Trent University., Peterborough, Ontario K9J 7B8, CANADA., 2003.

Webster E, Mackay D, Di Guardo A, Kane D, Woodfine D. Regional differences in chemical fate model outcome. Chemosphere 2004; 55: 1361.

Willighagen EL, Wehrens R, Buydens LMC. Molecular Chemometrics. Critical Reviews in Analytical Chemistry 2006; 36: 189 - 198.

Winkler DA. The role of quantitative structure - activity relationships (QSAR) in biomolecular discovery. Brief Bioinform 2002; 3: 73-86.

Wood J. Invariant pattern recognition: A review. Pattern Recognition 1996; 29: 1.

Worth AP, Bassan A, De Bruijn J, Saliner AG, Netzeva T, Patlewicz G, et al. The role of the European Chemicals Bureau in promoting the regulatory use of (Q)SAR methods. SAR and QSAR in Environmental Research 2007; 18: 111.

Xu Y, Zomer S, Brereton RG. Support Vector Machines: A Recent Method for Classification in Chemometrics. Critical Reviews in Analytical Chemistry 2006; 36: 177 - 188.

Zhang Q, Crittenden JC, Shonnard D, Mihelcic JR. Development and evaluation of an environmental multimedia fate model CHEMGL for the Great Lakes region. Chemosphere 2003; 50: 1377.

Zukowska B, Breivik K, Wania F. Evaluating the environmental fate of pharmaceuticals using a level III model based on poly-parameter linear free energy relationships. Science of The Total Environment 2006; 359: 177.

ISBN:978-84-693-4597-9 /DL:T.1010-2010

Chapter 2

Methods

The environmental assessment of chemical pollutants by means of QPFR or QSFR models is founded on techniques of multimedia environmental modeling, molecular modeling and pattern recognition. This chapter describes briefly these techniques and explains how they can be blended into a methodology that uses computerized supervised learning algorithms, common in artificial intelligence applications, for relating available examples of multimedia environmental modeling data to key physicochemical properties or molecular information in the case of, respectively, QPFRs or QSFRs.

ISBN:978-84-693-4597-9 /DL:T.1010-2010

2.1 Multimedia environmental modeling

MEMs estimate the distribution of chemical pollutants in the environment known physicochemical properties of pollutants, emission rates and site-specific data (Mackay, 2001). Due to computation and data costs, these models are required to be as simple as possible, without sacrificing the mathematical description of processes taking place in the region of interest. Level III MEMs (Mackay and Paterson, 1991), assuming steady state and non-equilibrium conditions, are usually recommended because involve a reasonable compromise between computational complexity and standard environmental processes (Mackay et al., 1992).

SimpleBox

SimpleBox (Brandes et al., 1996; den Hollander and van de Meent, 2004; den Hollander et al., 2004; van de Meent, 1993) is a nested multimedia fate model fashioned to Mackay's style of describing the environment: a set of compartments representing homogeneous media with mass balance and transport equations at different levels of complexity (Mackay, 2001). SimpleBox may perform Level III and Level IV calculations.

Earlier versions of SimpleBox have been used as foundations for the European Union System for the Evaluation of Substances (EUSES) (Lijzen and Rikken, 2004; Vermeire et al., 2005; Vermeire et al., 1997), designed to not only estimate the fate of chemicals but also to evaluate the risk of chemicals to humans and the environment, according to requirements from the European Union (Directive 92/32/EC, EC Council Regulation (EC) 793/93 and EC Directive 98/8/EC).

SimpleBox 3.0, as its previous versions (Brandes et al., 1996), is a nested multimedia model. It consists of four scales (den Hollander et al., 2004): local, regional, continental and global. The local scale is contained within the regional scale, which is contained within the continental scale and so on. Additionally, the global scale contains not only the continental scale but also a moderate zone, a tropic zone and an artic zone that work as background for the continental and regional scales (Brandes et al., 1996). Both the regional and continental scales are divided in 10 compartments representing different media: air, fresh water, sea water, fresh water sediments, sea water sediments, natural soil, agricultural soil, other soil, natural vegetation and agricultural vegetation. The local scale is divided in 8 compartments representing the same media as those contained within the regional and continental scale except sea water and sea water sediments. The zones at the global scale contain solely 4 compartments: air, water, sediments and soil. Default values for parameters in all the compartments are already included in SimpleBox 3.0 (den Hollander and van de Meent, 2004), but they can be modified by the user of the model according to his/her needs.

SimpleBox 3.0 requires as input, physicochemical properties of pollutants, emission rates and geographical parameters. The required physicochemical properties are the following: molecular weight (MW, g/mol); melting point (T_m); vapor pressure (P_v,

ISBN:978-84-693-4597-9 /DL:T.1010-2010

Pa); water solubility (S_w, mg/L); diffusion coefficients in air (D_{air}, m^2/s) and water (D_{water}, m^2/s); dimensionless partition coefficients for air-water (K_{aw}), solids-water (K_{sw}) and octanol-water systems (K_{ow}); and, degradation rates in air (k_{air}, 1/s), water (k_{water}, 1/s), sediments (k_{sed}, 1/s) and soil (k_{soil}, 1/s). If a property is not given by the user (1st option), it is estimated from other properties available (2nd option) or a default value is assigned (3rd option) (den Hollander and van de Meent, 2004). Internally, SimpleBox 3.0 adjusts temperature dependent properties to temperatures in the different scales and zones.

Environmental fate estimations from SimpleBox 3 are mainly expressed in form of average concentrations and mass fluxes for Level III calculations and time dependant concentrations for Level IV calculations. They are obtained from non-equilibrium computations over a set of J mass balance equations (as many as compartments in the model), for a given chemical i and several compartments j, with the form (Brandes et al., 1996):

$$V_j \frac{dC_{i,j}}{dt} = EMIS_{i,j} + IMP_{i,j} - EXP_{i,j} - DEGRD_{i,j} - LCH_{i,j} - BRL_{i,j} + ADV_{i,j \to j^*} + DIFF_{i,j \to j^*} \quad (2\text{-}1)$$

Terms in the equation above (Equation 2-1) have a first order dependency on the concentration of chemical i in box j ($C_{i,j}$). The linearity of the model can be verified by checking that concentrations in each compartment, j, are directly proportional to the emission rate.

2.2 Statistical sampling

Statistical sampling, also referred to as Monte Carlo simulations (Metropolis, 1987), is based on the generation of random values for evaluating how numerical models respond to several input variables. This methodology is usually employed in models for which analytical assessments are quite complex. Given the number of repetitive evaluations required, simulations for statistical sampling are best for computer based calculations.

For studying the response of a model to different situations, it is essential to define which variables remain fixed (deterministic values) and which are affected by random values (stochastic values). Subsequently, a planning for studying the response of the model is also required. Once that the variables affecting the model have been listed, it is necessary to describe how random values occur in each of the stochastic variables. This is done by selecting a probability density function, which relates the magnitude of possible random values with their probability to occur.

Historical data is required to determine a statistical distribution that best fits the variability of such data. When such information is available, it is possible to determine the parameters of the selected distribution as well, which will be used later in the generation of random values. In some cases, the values of a variable are discrete and have the same probability of occurring: for n possible values of x, the probability of occurring each value is 1/n (uniform discrete distribution). For continuous

ISBN:978-84-693-4597-9 /DL:T.1010-2010

variables, functions may be defined for describing equal probabilities in values of a definite range (uniform continuous distribution), producing a rough estimate when data is scarce (triangular distribution), describing symmetrical probabilities around a mean value (normal distribution) or when the logarithms of a variable have a symmetrical distribution (log-normal distribution).

The most widely used statistical distribution is the normal distribution (also called Gaussian), it accounts very well variability in parameters undergoing a symmetrical probability of acquiring values around a mean value. Normal distributions are very easy to use, known the mean ($\bar{x}$) and the standard deviation (SD) of the data, with no skewness or kurtosis in the probability distribution of values. $\bar{x}$ and SD are generally defined, for a set of observations, as follows:

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{n} x_i \tag{2-2}$$

and

$$SD = \sqrt{\frac{1}{N}\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{2-3}$$

It may occur that the normal distribution fails to fit the continuous values of a process, making necessary the evaluation of other statistical functions until fitting correctly the parameters to simulate. A typical case in which the normal distribution tends to fail is that of parameters resulting from multiplicative effects or defined to be positive and close to 0: symmetry is not present in the original scale of such parameters, but chances are high that their logarithms fit a normal distribution. For these situations, the log-normal distribution can be valid, it fits well the probability distribution of small positive values and values that tend to vary several orders of magnitude (Limpert et al., 2001). The log-normal distribution may be used to describe typical biological and environmental processes, where parameters may experience very different orders of magnitude.

Applying statistical sampling to a mathematical model requires a standard set of tasks (Doane, 2004).When simulating a process with statistical sampling, it is important to identify its inputs and outputs, assigning statistical functions to the input variables in which variability is supposed to occur. The output variables of the model vary in accordance to the variability generated by each input variable, within the probability domain of its statistical function, as propagated throughout the mathematical model of the process. Both the inputs and outputs of the model should be stored for later evaluating their uncertainty. The whole simulation procedure must be carried out by a computer program with the capacity to generate random values, from the selected statistical distributions, and evaluate the equations of the model in every realization.

2.3 Molecular modeling

ISBN:978-84-693-4597-9 /DL:T.1010-2010

Molecular modeling is the emulation or explanation of the behavior of molecules, applied in a wide number of disciplines (chemistry, physics, biology, engineering, etc.). Molecular modeling is fundamental for chemometrics (Brown et al., 1994), the application of mathematical or statistical methods to chemical data, which is also the core of QSARs and alike (Hugo, 2002). Generally, QSAR models rely on functions relating chemical activity to molecular descriptors (Equation 1-15), previously calculated by different theoretical methods (Willighagen et al., 2006).

Molecular orbital theory. The molecular orbital theory (MO theory) is a methodology that aims to determine molecular structure treating electrons as moving elements influenced by the nuclei of the entire molecule, not assigned to the bonds between atoms. The development of the MO theory has been based on several theoretical developments during the 20th century. One important step in its evolution is marked by the Hartree-Fock (HF) method of molecules; which, based on atoms, defined molecular orbitals (ψ_i) as eigenfunctions of the self-consistent Hamiltonian field (H), leading to coupled differential equations of difficult resolution (Pople, 1999).

The HF model was later refined with the work of Roothaan (Roothaan, 1951), producing a major advance (Zerner, 2000) by assuming that the molecular orbital wave function in a molecule (ψ_i) is equivalent to the linear combination of its N constituent atomic orbitals (χ_i):

$$\psi_i = \sum_{i=1}^{N} c_{ij}\chi_i \qquad (2\text{-}4)$$

where the coefficients c_{ij} can be determined by placing the equation above into the Schrödinger equation and applying the variational principle. With this approach, the determination of molecular structure is linked to solving a set of equations, some of them with integrals of still difficult resolution.

Semi-empirical applications of the molecular orbital theory. Determining molecular structure with calculations based on the Hartree-Fock method is unfeasible, even for small molecule systems. In result, approximations have been introduced in the HF theory for allowing the resolution of the equations involved, giving birth to methods classified as semi-empirical. The number of semi-empirical methods is vast and each of them has inherent advantages and disadvantages when modeling molecular structure (Bredow and Jug, 2005).

The most widely used semi-empirical methods are the Austin Model 1 (AM1) (Dewar et al., 1985) and the Parameterized Model 3 (PM3) (James, 1989). They are included in most standard software packages for computerized molecular modeling. AM1 based calculations, based on the iterative search of parameters in involved equations, are reasonably fast and robust. PM3 calculations, based on a more sophisticated optimization algorithm, gives acceptable results for molecules resembling those used in the training of the algorithm, but may yield strange predictions when such condition is not met. Since AM1 and PM3 are usually the semi-empirical methods of choice, their performances have been subject of comparison for different modeling problems (Bredow and Jug, 2005).

ISBN:978-84-693-4597-9 /DL:T.1010-2010

Molecular descriptors. Molecular descriptors encode information from molecular structure onto numerical parameters, measuring different characteristics of molecules for their posterior use in numerical models, comparisons and analysis (Todeschini and Consonni, 2000). These parameters are calculated with basis on different theories and methods. The variety of possible molecular structures is large; so, the number of possible descriptors for measuring their characteristics is large as well.

There are different types of molecular descriptors, depending on the information source used in their determination. Molecular descriptors can be classified as 0D, 1D, 2D, 3D or 4D. 0D descriptors are those derived from the molecular formula, the simplest chemical representation: molecular weight, number and type of atoms. 1D descriptors are derived from substructure list representations: functional groups, rings, bonds, etc. 2D descriptors measure topological information, describing how atoms are bonded in a molecule (considering also types of bonding and specific atomic interactions). 3D descriptors are obtained from geometrical representations of molecules (three-dimensional models) and may be referred to electronic, steric and shape features.

Research on molecular descriptors is still a very active field of research, so new descriptors are still being developed for measuring more and more features, especially for applications requiring highly complex molecules (pharmaceutics, genetics, polymer applications, etc.). Recent works have led to the development of 4D descriptors; however, there is still no agreement on which definition should represent this new category: one definition is based on the interaction field of molecules (Todeschini and Consonni, 2000) while another is based on their different conformations (Duca and Hopfinger, 2001; Senese et al., 2004).

In general, the calculation of molecular descriptors is grounded on the generation of molecular models for representing the structure of the chemicals to analyze. Traditional molecular models are based on 2D or 3D schemes indicating, respectively, how the atoms of a molecule are distributed on it, or, the exact location of atoms in space (considering atom size, bonds, angles, etc.).

With the aim to allow the modeling of molecules with computers, methods have been developed to describe molecular structure in simple ways, for its easy interpretation by both humans and computers. One remarkable example is the Simplified Molecular Input Line Entry System, known as SMILES^{TM} (Anderson et al., 1987; Weininger, 1988; Weininger et al., 1989), a simple code introduced in the 1980s and currently under development by Daylight Chemical Information Systems Inc. that allows, in a single string line, the characterization of most molecules with ASCII characters.

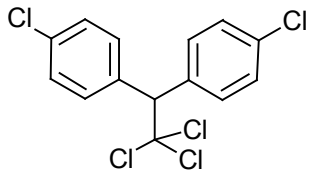
Newer codes are also aiming to describe molecular structure with their own syntax, but still they have not achieved the privileged position of the SMILES^{TM} notation, which is included in most molecular software packages. Examples of recent molecular notation schemes are InChl^{TM} (McNaught, 2006) and OpenSMILES, introduced and maintained, respectively, by the International Union of Pure and Applied Chemistry (IUPAC) and The Blue Obelisk (Guha et al., 2006). Figure 2-1 shows an example of how the molecule of 1,1,1-trichloro-2,2-bis-(4-chlorophenyl)ethane (CAS: 50-29-3) can be represented by means of its corresponding molecular formula (Figure 2-1a), its

ISBN:978-84-693-4597-9 /DL:T.1010-2010

a) Molecular formula: $C_{14}H_9Cl_5$

b) SMILES code: Clc1ccc(cc1)C(c2ccc(Cl)cc2)C(Cl)(Cl)Cl

c) 2D model



d) 3D model

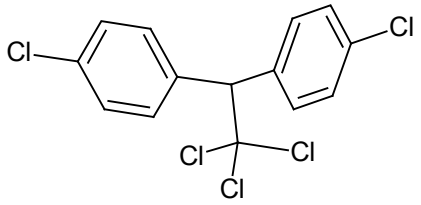


Figure 2-1. Standard schemes for representing molecular structures.

The molecular structure of a chemical can be represented, at different levels of complexity, by its molecular formula (a), its SMILES^{TM} code (b), a 2D model (c) and a 3D model (d). Note that all representations in this figure (a-d) are referred to the molecule of 1,1,1-trichloro-2,2-bis-(4-chlorophenyl)ethane (CAS: 50-29-3).

SMILES^{TM} code (Figure 2-1b), a 2D model (Figure 2-1c) and a 3D model (Figure 2-1d).

Given a molecular structure, energy-based descriptors are usually estimated through a steepest descent algorithm until a conformational minimum energy (CME) is achieved, when the structure of interest achieves the most stable geometry. If the conformation of a molecule is recalculated several times, different CME values may be obtained, as the algorithm encounters different minima during the optimization process. Selecting an optimal molecular conformation implies choosing the conformation with the lowest energy and discarding all those with higher energy values.

Descriptors measuring energy parameters may be easier to interpret than descriptors measuring other molecular features by means of abstract indexes. Some of the most widely used parameters are the heat of formation (ΔH_f), the highest occupied molecular orbital (HOMO), the lowest occupied molecular orbital (LUMO), the dipole moment (μ), among others. They are usually preferred because their theoretical definitions are easy to interpret.

ΔH_f represents the change of enthalpy accompanying the formation of 1 mole of a substance in its standard state from its conforming elements in their standard states. ΔH_f gives an indication of how stable a molecule is: the more stable a molecule, the lower its ΔH_f value. Negative ΔH_f values are associated to exotermic reactions of formation. The difference between HOMO and LUMO, termed the band gap, is an indicator of the excitability of a molecule: the smaller the band gap, the more

ISBN:978-84-693-4597-9 /DL:T.1010-2010

excitable a molecule is. μ indicates the capacity of a molecule to behave as a dipole, and so the capacity of a molecule to be soluble in polar or non-polar phases.

Thousands of descriptors have been defined and others are still under development. When optimizing QSAR based models, the molecular descriptors to implement in a QSAR may be selected heuristically, by means of mathematical algorithms, or according to their theoretical contribution to the understanding of a chemical process (Willighagen et al., 2006). Some descriptors measuring very specific molecular features may be hard to interpret, and their use in a QSAR model may not necessarily improve its performance and utility. In general, a QSAR model should be as simple as possible and ready to be employed by users not having the exact same tools of its developers.

2.4 Pattern recognition

Practically, plenty of information can be retrieved from any process or element. Excess of information does not necessarily imply better understanding; but, it does imply the presence of both useful and pointless information, mixed. The need of finding relations from vast amounts of data has been recognized in both quotidian activities and specialized fields long time ago (telecommunications, business administration, marketing, science and engineering, etc.), leading to a discipline that today is termed data mining (Fayyad et al., 1996; Witten and Frank, 2005) and that keeps evolving along with computer developments.

Data mining implies the retrieval of potentially useful information from large data sets, usually by means of computer algorithms with capacity to identify patterns in data. Pattern recognition is referred to information retrieval relying in learning algorithms, procedures that allow machines to extract and process information with different purposes. Most learning algorithms have their roots on developments for machine learning and artificial intelligence (Winston, 1992), which attempt to emulate intellectual behavior.

Data preprocessing. The observations of a data set may be called samples, examples or instances; the variables or characteristics describing each data point may be called features or attributes. When training algorithms for specific processes, most part of the work must be dedicated to the preparation of the working data set. Data must be collected, cleaned, analyzed and preprocessed prior to its utilization for training algorithms and later performing predictions with already developed models.

Errors are most likely to occur during the manipulation of data: collecting data from different sources, copying, etc. For this reason, careful selection of data samples is required. Once that the data has been collected and cleaned, it must be analyzed to determine if it is suitable or not for its processing with the available algorithms. For numerical data, it is important to use data with both a wide range of values and a homogeneous density of data points all over the numerical space of interest. If these conditions are not met, the available data must be transformed (preprocessed) for reaching such conditions.

ISBN:978-84-693-4597-9 /DL:T.1010-2010

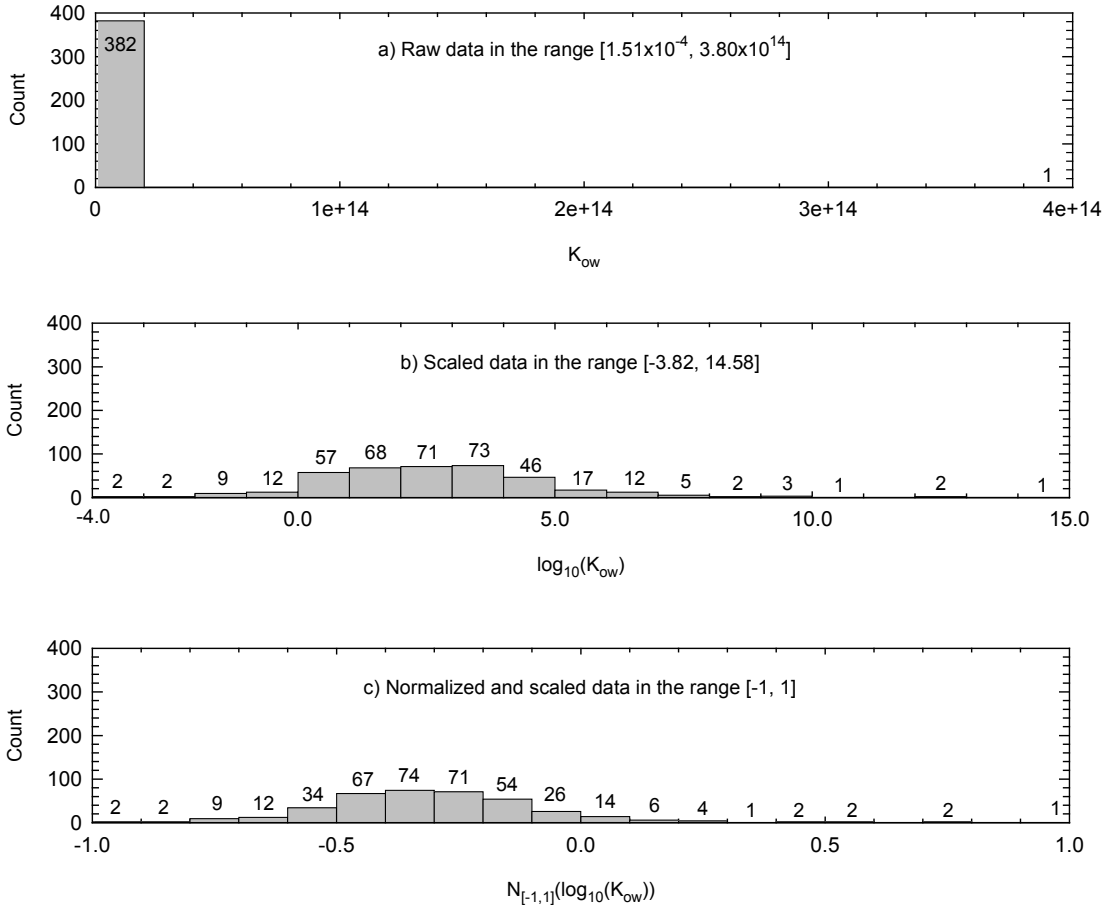


Figure 2-2. Common pre-processing data techniques.
In some cases, highly skewed data require transformations for producing smooth data distributions and helping learning algorithms to process them. Logarithmic scaling and linear normalization are among the most common pre-processing techniques. In this figure, raw K_{ow} values of 383 chemicals (a) are compared to K_{ow} values scaled logarithmically (b) and to K_{ow} values scaled logarithmically and later normalizad in the range [-1, 1] (c). Note that the presence of Kow values with several orders of magnitude difficults the discrimination of data points in such data set (a), the logarithmic scaling of the raw values (b) produces a smoother distribution while a posterior normalization redefines the scale (c).

Most learning algorithms may suffer serious efficiency reductions because of insufficient computer resources, noisy data or inconsistent data points (outliers). Usually, it must be analyzed if within the collected data there are redundant or useless variables and data samples. If any of these elements are encountered, they should be eliminated from the data set of interest. Direct data observation may be used for doing so; however, when the size of the data set is prohibitive, clustering and classifying algorithms may be of great help.

Figure 2-2 shows an example of how a variable with different orders of magnitude can be adapted to its analysis, the histograms in this figure are referred to K_{ow} values for 383 chemicals. K_{ow} can get very small values for highly lipophobic chemicals and extremely large values for highly lipophilic chemicals. Figure 2-2a shows that 382 out of 383 chemicals have small K_{ow} values, while there is 1 chemical with a marked tendency to dissolve in lipids. It is possible to determine that 1 chemical is highly lipophilic and that the remaining 382 chemicals are extremely less lipophilic.

ISBN:978-84-693-4597-9 /DL:T.1010-2010

However, it is not clear to which degree the latter are lipophobic, or lipophilic, among themselves. The original scale in which raw K_{ow} values are expressed may affect significantly all posterior calculations and analysis, making very difficult the discrimination of chemicals with respect to K_{ow}.

Clearly, the distribution of chemicals in Figure 2-2a is highly skewed and transformations are required for producing a smoother distribution. Figure 2-2b shows how chemicals are distributed in a logarithmic scale. The distribution of chemicals with respect to logarithmic K_{ow} values (Figure 2-2b) is smoother than that based simply on raw K_{ow} values (Figure 2-2a). Now, a more clear discrimination can be performed, different degrees of octanol-water partitioning can be identified in the set of selected chemicals: there are 25 lipophobic chemicals with K_{ow} values below 1.00 (0 in the logarithmic scale), 26 highly lipophobic chemicals with K_{ow} values above 1.00×10^6 (6 in the logarithmic scale) and 332 chemical with Kow values between 1.00 and 1.00×10^6.

When there are several variables to analyze, their range values may differ greatly and some variables may eclipse others. A convenient data transformation technique is the normalization of data, which sets uniform weights for the sets of variables to analyze. This may be done by forcing all variables of interest to be in the same scale. A linear transformation may be used for setting the maximum and minimum of each scale to be [0,1] or [-1,1], etc.

For normalizing every data point y_n of N samples referred to a given variable (y) in the range [-1,1], the following expression is used:

$$N_{[-1,1]}(y_n) = 2\left(\frac{y_n - y_{min}}{y_{max} - y_{min}}\right) - 1 \qquad (2\text{-}5)$$

where y_{min} and y_{max} are, respectively, the minimum and maximum values that can be found in the data set, with respect to all the data points and the variable to normalize. Figure 2-2c shows a histogram for normalized logarithmic K_{ow} values in the range [-1,1]. It can be observed that Figure 2-2c tends to preserve the distribution of Figure 2-2b, but setting a new scale of values. If more physicochemical properties are to be analyzed, they should be preprocessed as done with K_{ow} for the example.

There are different approaches that can be applied in the preparation of a data set prior to its analysis either manually or by means of computerized learning algorithms. However, it all depends on two important factors. First, unnecessary variables and data points must be removed; and, second, both past and new data samples must be in the same scale.

Artificial neural networks. Artificial neural networks (ANNs) are mathematical models of biological neurons, originally developed with the purpose of imitating brain activities. With time, the applicability of ANNs has evolved towards the solution of a large variety of mathematical problems, especially those in which data are noisy or incomplete (Basheer and Hajmeer, 2000).

ISBN:978-84-693-4597-9 /DL:T.1010-2010

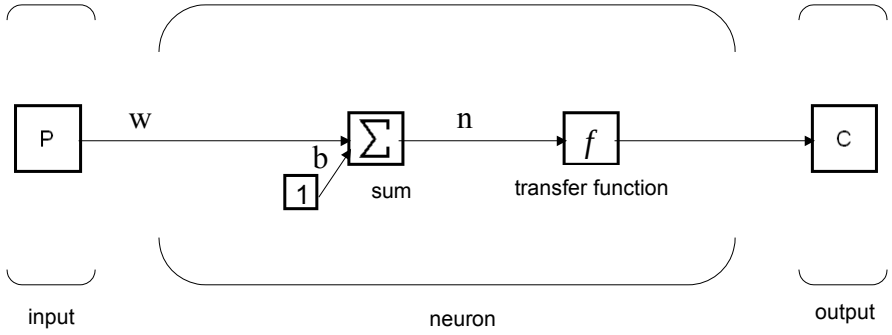


Figure 2-3. Information flow in a single artificial neuron.
Mathematically, an artificial neuron emulates the functionality (f) of a real biological neuron by generating a signal (c) in response to the stimulus provoked by an incoming signal (p) of varied strength (weight, w) and a given bias (b).

ANNs are based on the same elements that constitute biological neural networks. Figure 2-3 shows an scheme of the elements that constitute a single artificial neuron. An artificial neuron is conformed by its cell body (transfer function, f), different synapses of variable strength (weight, w) to receive signals from other neurons (input, p); and, an axon to send signals to other neurons as well (output, c). A signal coming out of a neuron follows the same equation:

$$c = f(wp + b) \tag{2-6}$$

For a single layer of S neurons in parallel, the inputs are: a vector **p** of size R×1 a weight matrix **W** of size S×R and a bias vector **b** of size S×1 (**n** = **Wp** + **b**). The output of a layer of neurons is a vector **c** of size S×1 calculated as follows:

$$\mathbf{c} = \mathbf{f}(\mathbf{Wp} + \mathbf{b}) \tag{2-7}$$

Choosing an ANN architecture requires considering the complexity of the problem to solve, the computation capacity available, the stability of the ANN system to use and its training algorithm (supervised, unsupervised, etc.). The inputs and outputs of a problem correspond to those of a network, while the transfer functions at the output layer of the network correspond to the specification of the outputs in the problem. ANNs can be configured for processing data in a wide variety of ways, being the most common classification and function approximation tasks.

Supervised and unsupervised learning. The way a learning algorithm works defines the type of learning it performs, it may be based on ANNs, rules, data vectors, etc. Most common algorithms may be supervised and unsupervised. A supervised algorithm generates a function for mapping the input of a process to its outputs, usually termed targets (the desired output). Unlike supervised algorithms, an unsupervised algorithm does not require data labeled with the output of a process, it simply processes the input data without external influence. There are other machine learning schemes, like transduction (similar to supervised learning, but without creating functions), semi-supervised learning (combining labeled and unlabeled data)

ISBN:978-84-693-4597-9 /DL:T.1010-2010

and reinforcement learning (correcting what has been learned while interacting with a guide).

Based on the capacity of learning algorithms to detect patterns in previous data, these tools can be used for explaining past observations or predicting future trends or events. The process in which a learning algorithm adjusts its internal parameters to fit a data set (under any learning scheme) is usually referred to as training; in this stage, algorithms "learn" from data. When using a trained algorithm for predicting trends in new data, the algorithm compares the patterns in the new data set with the patterns it has learned from past data, for later producing a response.

Training. Training is the process in which the inner parameters of a learning algorithm are adjusted (for example, the weights and biases of ANNs), with basis on a set of data samples, the training data set. A trained algorithm should reproduce what it has learned in order to explain past data or perform forecastings. The training data provide the required information to do so, but the manner in which they are processed by the algorithm affects its own predictive power. For this reason, it is important to select training data as diverse as possible and give some freedom to the algorithm to fit them. When the algorithms fits very well the training data problems may occur: overtrained algorithms, set to identify high standards in the training data, can not find similarities in new data and may produce highly erroneous predictions.

Test and validation. The prediction power of any learning algorithm is affected by its training; however, for determining how predictive a trained algorithm is, it is required to evaluate the algorithm with an independent data set, not used in its training. In this manner, it can be determined if a trained algorithm can generalize well or if, on the contrary, fails to predict trends in new data. It is usual to evaluate the predictive performance of a trained algorithm on test data until optimal training settings are achieved, for later evaluating solely its performance on validation data. When the data of a process are scarce, selecting data sample for training and validating the models becomes an additional problem that can be tackled with n-fold cross-validation (n-fold CV): the working data set is divided iteratively into n subsets, for training and testing n models with, respectively, (n-1)/n and 1/n, fractions of the original working data set. The leave one out validation (LOO) procedure is a variation of the n-fold CV in which all data vectors except one are used for training while the remaining vector is used for validation, n is then equal to the number of data vectors available. The performance of an algorithm trained and tested with the n-fold CV or LOO procedures is reported by averaging the performance indicators in each independent set.

Data mining techniques. There is a large variety of processing techniques based on learning algorithms (McClean and Robert, 2001). When there is no prior knowledge referred to a high dimensional data set, tasks based on supervised algorithms are appropriate for finding subgroups of data with common attributes (clustering) and visualizing all data points. When the input and output of process are known, it may be of interest using past data for predicting the outputs of a process without using it explicitly, because of its involved costs or poor performance; in this case, supervised learning algorithms can be used for classifying data points and approximating complex functions.

ISBN:978-84-693-4597-9 /DL:T.1010-2010

2.4.1 Visualization and clustering algorithms

Large datasets may be very difficult to analyze at any stage of a data mining project. Reducing their complexity becomes an important step for gaining an understanding of hidden relationships or improving the performance of computerized algorithms. There is a great variety of algorithms (Jain et al., 1999) that can be applied, under different learning schemes, for visualizing and clustering complex datasets.

Principal Component Analysis

Principal component analysis (PCA), also known as Karhunen-Loève transform, is an unsupervised algorithm widely used for reducing the number of dimensions in data sets, extracting features and generating simple data visualizations (Jolliffe, 2002). It is defined as a linear projection that minimizes the average projection cost (Pearson, 1901), expressed as the mean squared distance between data points and their projections.

It may also be defined as the orthogonal projection of data onto a linear space characterized with lower dimensions, in such way that the variance of the projected data is maximized (Hotelling, 1936). Figure 2-4 shows how data points characterized by two independent variables (x, y) are projected into an orthogonal space (u_1), the principal subspace.

For demonstrating how raw data is projected onto orthogonal space, it is assumed that a set of N observations $\{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N\}$, an Euclidean variable ($\mathbf{x}_n$) with D dimensions and the M dimensions of the orthogonal projection are known. Assuming the case for one single dimension (M = 1) in the orthogonal space, the direction of such space can be represented by a unit vector with D dimensions $\mathbf{u}_1$. The projection of each data point $\mathbf{x}_n$ is given by the scalar $\mathbf{u}_1^T \mathbf{x}_n$. The variance of the projected data is given by:

$$\frac{1}{N}\sum_{n=1}^{N}\left\{\mathbf{u}_1^T\mathbf{x}_n - \mathbf{u}_1^T\bar{\mathbf{x}}\right\}^2 = \mathbf{u}_1^T\mathbf{S}\mathbf{u}_1 \qquad (2\text{-}8)$$

where $\bar{\mathbf{x}}$ is the sample set mean and $\mathbf{u}_1^T\bar{\mathbf{x}}$ is the mean of the projected data. **S** is the data covariance matrix, defined as:

$$\mathbf{S} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n - \bar{\mathbf{x}})(\mathbf{x}_n - \bar{\mathbf{x}})^T \qquad (2\text{-}9)$$

For maximizing the projected variance, $\mathbf{u}_1^T\mathbf{S}\mathbf{u}_1$, with respect to the unit vector, $\mathbf{u}_1$, an unconstrained maximization is applied, imposing the normalization condition $\mathbf{u}_1^T\mathbf{u}_1 = 1$ by means of the Lagrange multiplier λ_1:

ISBN:978-84-693-4597-9 /DL:T.1010-2010

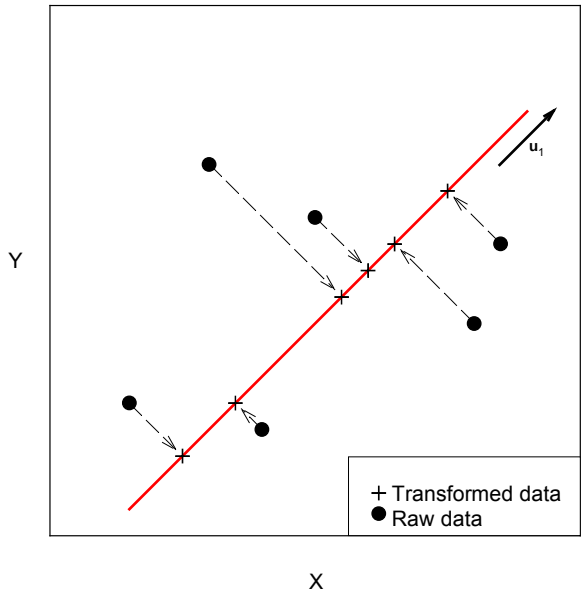


Figure 2-4. Data projections based on the principal component analysis.

In a PCA projection, original data points are projected onto an orthogonal low-dimensional space that minimizes the average projection cost.

$$\mathbf{u}_1^T\mathbf{S}\mathbf{u}_1 + \lambda_1\left(1 - \mathbf{u}_1^T\mathbf{u}_1\right) = 0 \tag{2-10}$$

setting the derivative with respect to $\mathbf{u}_1$ equal to zero, a stationary point is achieved when:

$$\mathbf{S}\mathbf{u}_1 = \lambda_1\mathbf{u}_1 \tag{2-11}$$

indicating that $\mathbf{u}_1$ is an eigenvector of $\mathbf{S}$. Multiplying by $\mathbf{u}_1^T$ from the left side and noticing that $\mathbf{u}_1^T\mathbf{u}_1 = 1$, the expression above takes the form:

$$\mathbf{u}_1^T\mathbf{S}\mathbf{u}_1 = \lambda_1 \tag{2-12}$$

from which the first principal component is obtained, the eigenvector $\mathbf{u}_1$. This component has associated the largest eigenvalue, λ_1, and so, a maximum variance.

Other principal components can be defined repeating the procedures used for obtaining the first principal component. Every new direction must be chosen in a way that the projected variance is maximized, in the middle of all possible directions orthogonal to those already found. For M dimensions, this is reduced to the definition of M eigenvectors $\mathbf{u}_1,\mathbf{u}_2,\ldots,\mathbf{u}_M$ of the data covariance matrix S, with the corresponding eigenvalues $\lambda_1,\lambda_2,\ldots,\lambda_M$.

K-means

ISBN:978-84-693-4597-9 /DL:T.1010-2010

The K-means algorithm (MacQueen, 1967) is a supervised algorithm that clusters N data points into K partitions, known a value K lower than N (K<N). It minimizes a distortion measure, given by:

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \left\| \mathbf{x}_n - \boldsymbol{\mu}_K \right\|^2 \tag{2-13}$$

this definition gives the sum of the squares of the distances of each data point, $\mathbf{x}_n$, to its closest prototype vector, $\boldsymbol{\mu}_k$ (a vector representing the kth cluster). r_{nk} is a binary indicator variable (for k = 1,…,K), equal to 1 when a data point is assigned to a cluster k ($r_{nk} = 1$, for a cluster k) and equal to 0 when not ($r_{nk} = 0$, for a cluster $j \neq k$). The assignment of r_{nk} values to the different data points is usually known as the 1-of-K coding.

The goal of the K-means algorithm is to find a set of values r_{nk} and a set of prototype vectors $\boldsymbol{\mu}_k$ for minimizing J in an iterative procedure. First, r_{nk} values are estimated to be 1 or 0 according to a set of initial prototype vectors $\boldsymbol{\mu}_k$. Secondly, fixed a set of r_{nk} values, all prototype vectors are optimized by setting to zero the derivative of the distortion measure (Equation 2-13) with respect to $\boldsymbol{\mu}_k$:

$$2\sum_{n=1}^{N} r_{nk} \left(\mathbf{x}_n - \boldsymbol{\mu}_K \right) = 0 \tag{2-14}$$

and solving for $\boldsymbol{\mu}_k$:

$$\boldsymbol{\mu}_k = \frac{\sum_{i=1}^{N} r_{nk} \mathbf{x}_n}{\sum_{n=1}^{N} r_{nk}} \tag{2-15}$$

for estimating again the 1-of-K coding (set of r_{nk} values) and another set of prototype vectors ($\boldsymbol{\mu}_k$) for the data set, until reaching a minimum J. The K-means algorithm may be slow or imprecise for some cases, so other clustering algorithms may be more suitable instead (Lance et al., 2004). However, its simplicity and functionality make it appropriate for exploring the clustering of unknown datasets, prior to further assessments.

Self Organizing Maps

The Self-Organizing Map (SOM) (Kohonen et al., 1996), also known as Kohonen map, is based on ANNs applying both vector quantization and projection algorithms in unsupervised conditions. It is widely used for clustering and visualizing high-dimensional data sets.

The neurons of a SOM are arranged on a lattice of any regular shape (either 2D or 3D: rectangular, hexagonal, cylindrical, etc.) in which each neuron is represented by a weight vector of dimensions d, where d is the dimension of the SOM input vectors.

ISBN:978-84-693-4597-9 /DL:T.1010-2010

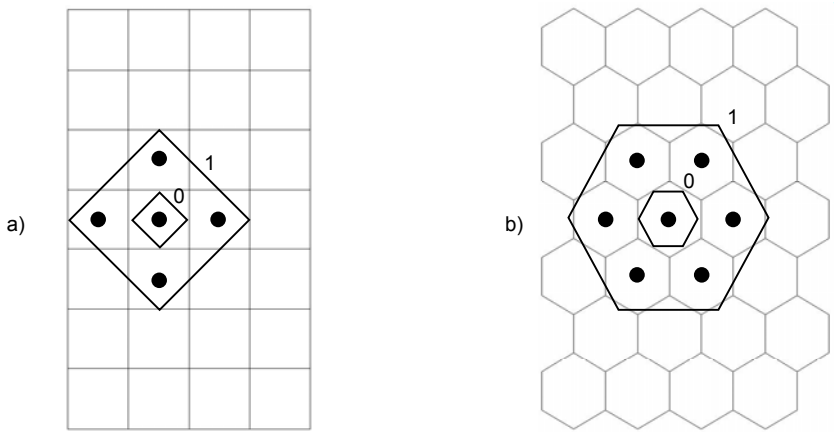


Figure 2-5 Distribution of artificial neurons in self organizing maps.
The neurons of SOMs, also called SOM units, are disposed in lattices of varied spatial configurations. This figure shows two SOMs, one with a rectangular lattice and 5 clustered data points (a) and another with a hexagonal lattice 7 clustered data points (b). Data points clustered in SOM units may have neighbor data points with, until some extent, similar characteristics. In this figure, points marked with 0, are surrounded by other data points (marked with 1), 4 in the rectangular lattice (a) and 6 in the hexagonal lattice (b).

Figure 2-5 shows the lattices of two SOMs with the same number of units, 28 neurons (organized in lattices of 7x4 neurons), the first SOM has a rectangular lattice (Figure 2-5a) while the second one has a hexagonal lattice (Figure 2-5b).

SOMs are trained in an iterative manner. For each epoch in the training of a SOM, one sample vector **x** from the input data is chosen at random and compared with all the weight vectors of the SOM via a similarity measure. The distance (or Euclidean distance) of a weight vector, **m**, close to an input vector **x** is calculated as:

$$\left\|\mathbf{x}-\mathbf{m}_{\mathrm{BMU}}\right\| = \min_{i}\left\{\left\|\mathbf{x}-\mathbf{m}_{i}\right\|\right\} \tag{2-16}$$

where the best matching unit (BMU) is the neuron whose weight vector has the greatest similarity or shortest distance with the input sample **x**. The equation above is modified for accounting the contribution of different elements in the selection of BMUs: missing data values do not contribute at all ($\left\|\mathbf{x}-\mathbf{m}_{i}\right\| = 0$); and, every variable may contribute or not, depending on its associated mask (with values between 0 or 1).

The distance measure used in the selection of BMUs has the form:

$$\left\|\mathbf{x}-\mathbf{m}\right\|^{2} = \sum_{k\in K} w_{k}\left(x_{k}-m_{k}\right)^{2} \tag{2-17}$$

where K is the set of available variables of the samples vector **x**. x_k, m_k and w_k are, respectively, the kth component of the sample, the kth component of the weight vector and the kth mask value.

The goodness of a SOM is assessed in terms of the mean quantization error ($\bar{q}_{error}$) and the mean topological error ($\bar{t}_{error}$) (Uriarte and Martín, 2005):

ISBN:978-84-693-4597-9 /DL:T.1010-2010

$$\bar{q}_{error} = \frac{1}{N}\sum\left\|\mathbf{x}_i - \mathbf{m}_{\mathbf{x}_i}\right\| \tag{2-18}$$

and

$$\bar{t}_{error} = \frac{1}{N}\sum_{n=1}^{N} u(\mathbf{x}_i) \tag{2-19}$$

where: N is the number of data vectors; $\mathbf{m}_{\mathbf{x}_i}$ is the best matching unit (BMU, also called SOM unit or SOM prototype) of the corresponding data vector $\mathbf{x}_i$; and, $u(\mathbf{x}_i)$ is a function that yields 1 if the first and second BMUs of $\mathbf{x}_i$ are adjacent and, 0 otherwise.

After finding the BMUs, the weights of the SOM are updated. After training, each neuron of the SOM represents the vectors of the input space that have been classified in the cell and its neighborhoods. The rule for updating the weights of each unit i of the SOM is given by:

$$\mathbf{m}_i(t+1) = \mathbf{m}_i(t) + h_{BMU,i}(t)\left[\mathbf{x}(t) - \mathbf{m}_i(t)\right] \tag{2-20}$$

The neighborhood kernel, $h_{BMU,i}(t)$, formed by a neighborhood function and a learning rate function, is a non-increasing function of time and of the distance between unit i and the BMU, defining the region of influence that the input sample **x** has on the SOM:

$$h_{BMU,i}(t) = h\left(\left\|\mathbf{r}_{BMU} - \mathbf{r}_i\right\|, t\right)\alpha(t) \tag{2-21}$$

Figure 2-6 and Figure 2-7 are referred to a practical example illustrating how a SOM can be used to visualize and cluster data. The example has been prepared as follows: First, a dataset has been normalized in the range [-1,1] prior to the SOM training, the dataset is composed of 383 chemicals characterized by logarithmic values of their vapor pressure (P_v), water solubility (S_w) and air-water partition coefficient (K_{aw}, Equation 1-3). Second, a SOM has been set to have, approximately, as many units as vectors in the dataset (24x16 units) in a hexagonal lattice. Third the dataset is presented to the SOM. During the training phase of the SOM, the prototypes of each neuron were adjusted automatically by the SOM itself to fit the dataset as much as possible.

Figure 2-6a shows a 2-D PCA projection that confirms that, after the training phase, the SOM prototypes have been located very close to most data vectors, fitting well the dataset of interest. In Figure 2-6b, it can be observed that the example SOM has clustered the 383 chemicals in its neurons (or SOM units): some neurons are empty, but there are others clustering 1, 2, 3 or 4 chemicals. Figure 2-6c shows the component planes of the SOM, mapping the values assigned to every SOM prototype in every of the three logarithmic properties, which are somewhat comparable to the values of the fitted dataset and give an insight of the distribution of data vectors in the

ISBN:978-84-693-4597-9 /DL:T.1010-2010

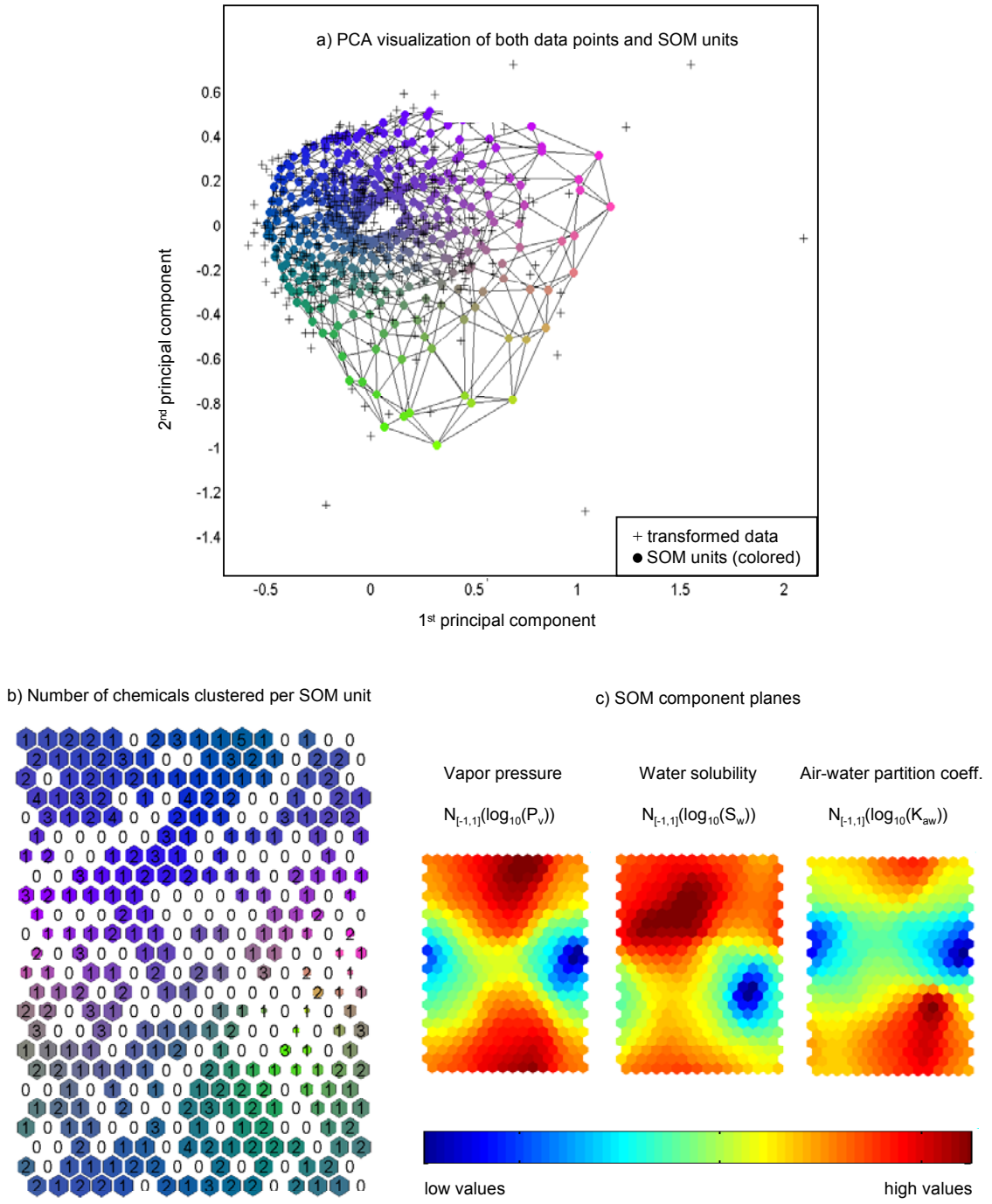


Figure 2-6. Possible data visualization schemes on self organizing maps.
This figure shows how a SOM, with 24x16 units disposed in a hexagonal lattice, allows the visualization of a set of 383 chemicals characterized by logarithmic values of P_v, S_w and K_{aw}: making a PCA projection of both data points and SOM units (b), counting the number of chemicals clustered in every SOM unit (b); and, by means of SOM component planes (c).

SOM. At first sight, any of the visualization schemes of Figure 2-6 (a, b or c) may seem difficult to interpret for the user lacking of experience with SOMs. However, it is important to remember that the presented visualization schemes are equivalent. They are simply different points of view for the same problem, fitting a dataset with the neurons of a SOM.

ISBN:978-84-693-4597-9 /DL:T.1010-2010

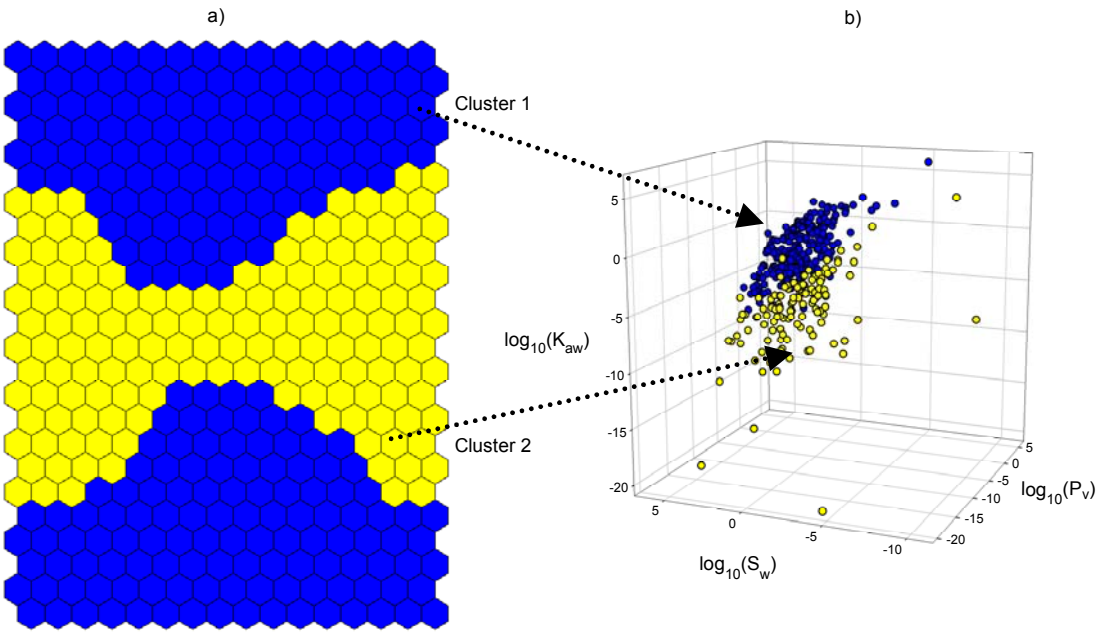


Figure 2-7. Clustering of self organizing maps.
This figure shows how the SOM of Figure 2.7 can be clustered with basis on its component planes (a), extending the clustering to the 383 chemicals already used in its training (b). In this example, cluster 1 and cluster 2 are associated to chemicals with, respectively, high K_{aw} and low K_{aw} values.

The SOM algorithm clusters data points in its neurons. However, further clustering is possible. Figure 2-7a shows how the SOM of Figure 2-6 has been clustered with the Davies-Bouldin algorithm, very similar to K-means. The number of clusters was set to 2 in the clustering algorithm, so the SOM was divided into two regions: one referred to high K_{aw} values and another with low K_{aw} values. The resulting two clusters coincide with the high and low-value regions of the K_{aw} component (Figure 2-6c). In the three-dimensional coordinate system of partitioning properties (P_v, S_w, K_{aw}), chemicals with similar partitioning behavior are located close to each other. The example SOM has learned to map the partitioning properties of 383 chemicals (Figure 2-6) and has clustered them with respect to their tendency to partition to air or water (Figure 2-7).

Using the SOM for datasets composed of data points with 1 to 3 dimensions may be redundant. However, when the number of data points and dimensions are high, the applicability of the SOM algorithm may be of great help. Several parameters can be adjusted to help a SOM fits its training data: shape, lattice, number of units (neurons), learning function, neighborhood kernel, etc. It must be noticed that, in general, datasets with several dimensions and few data points may lead to poor fitting. So it is up to the user to test different SOM settings when processing his/her working data for finding an optimal model.

2.4.2 Classifiers

ISBN:978-84-693-4597-9 /DL:T.1010-2010

Some machine learning algorithms can be used for data classification, known a set of previously labeled items of a training data set. Typically, classification problems require finding a classifier that best maps the characteristics of new data to their real classes. There is no single learning algorithm that works best on all classification problems, the performance of each classifier depends on the features of the data to be classified. For this reason, it is common practice to test various algorithms for the same classification problem and compare their predictions for the test data.

The outcomes of a classifier can be described as true positive (TP), false positive (FP), true negative (TN) and false negative (FN). Regarding one class, elements classified as members of such class are TP when their classification is correct and FP when incorrect; meanwhile, elements classified as member of other classes are TN or FN, when correctly or incorrectly classified, respectively. The performance of a classifier on a dataset is estimated calculating the rates of true positive (TP) and false positive (FP) predictions:

$$TP_{rate} = \left(\frac{TP}{TP + FN} \right) 100\% \tag{2-22}$$

$$FP_{rate} = \left(\frac{FP}{FP + TN} \right) 100\% \tag{2-23}$$

and comparing such values in a two-dimensional plot, in which high values of TP_{rate} and low values of FP_{rate} for a test data set indicate acceptable predictions. Another measurement of the performance of a classifier may be obtained using the F-measure:

$$F\text{-measure} = \left(\frac{2TP}{2TP + FP + FN} \right) \tag{2-24}$$

Naive Bayes

The Naive Bayes classifier (George and Langley, 1995), supported on Bayes's rule of conditional probability (Barnard and Bayes, 1958) and widely used in supervised learning tasks, works given two assumptions: the predictive attributes are conditionally independent given the class; and, no hidden or latent attributes influence the prediction process. It says that for a given hypothesis H and evidence X that bears on that hypothesis, the probability of the hypothesis conditional on the evidence is as follows:

$$p[H = h \mid X = x] = \frac{p[H = h]p[X = x \mid H = h]}{p[X = x]} \tag{2-25}$$

and that, for N pieces of evidence, the term p[X = x | H = h] is given by:

$$p[X = x \mid H = h] = \prod_{i}^{N} p[X_i = x_i \mid H = h] \tag{2-26}$$

ISBN:978-84-693-4597-9 /DL:T.1010-2010

p[X = x], the denominator in the equations above (Equation 2-24 and Equation 2-25), is not estimated and disappears when normalizing so that the sum of p[H = h | X = x] is 1. When processing numeric attributes, the classifier is assumed to have a Gaussian probability distribution:

$$p[X = x \mid H = h] = \frac{1}{\sqrt{2\pi\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad \text{when x lies in } [x-\varepsilon/2,\ x+\varepsilon/2] \qquad (2\text{-}27)$$

Variations on the Naïve Bayes classifier replace the probability distribution by other density estimation methods that may reduce prediction errors on natural and artificial data sets. When it is difficult to know the probability distribution to use, a Kernel density function may yield good results (George and Langley, 1995):

$$p[X = x \mid H = h] = \frac{1}{nh} \sum_j K\left(\frac{x - \mu_i}{h}\right) \quad \text{where } h = \sigma \text{ and } K = g(x,0,1) \qquad (2\text{-}28)$$

Decision trees

Decision trees are models that characterize the conditions of an event to occur, classifying data in every branch of a tree-like graph according to different conditioning features. Classification or regressions tasks can be performed with decision tree algorithms when processing, respectively, categorical or numerical data. Since the outputs of a process are required when training decision tree algorithms, these algorithms work under supervised learning conditions.

One of the most simplest tree-based algorithms is the classification and regression trees (CART) algorithm (Breiman et al., 1984). Consider a data set with N vectors, in which every vector is characterized by a set of D input features, $\{i_1,\ldots, i_D\}$ and a target feature {t}. When the partitioning of the input space is known and the associated error function is minimized (based on a sum of squares), the optimal value for a predictive variable in any given region is given by the average values of t_n running over the data points falling in that region.

When determining the structure of a decision tree, the optimization process of minimizing the error function resulting from fitting the algorithm to the training data may become infeasible. This process implies the selection of input features and thresholds for each branch that, for large multivariate and large datasets, may have associated high computational costs. An alternative greedy optimization is usually applied, which creates a single-node tree covering the entire input space and adds nodes to the tree, one at a time. In every step, a selection of one of the D input variables and the associated threshold is carried out by exhaustive search until an optimal selection is found, characterized by the local average of the predictive variable. The whole greedy optimization process is repeated for all possible input variables, selecting the variable with the lowest associated error.

A problem associated to the greedy optimization of tree-based algorithms is when to stop the addition of tree nodes during the training process. The simplest approach is to

ISBN:978-84-693-4597-9 /DL:T.1010-2010

stop the training process when an error threshold is achieved; but, every available split reduces slightly the error of the algorithm, resulting in several splits when reaching the error threshold. To overcome this problem, usually large trees are extended until reaching a limit based on the number of data points associated to the leaf nodes, for later pruning back the original tree.

The pruning process balances the residual error of the tree against a parameter measuring its complexity. The pruning of a tree, Tr_0, is carried out by collapsing internal nodes and merging the corresponding regions and generating a subtree of T_0, T, that complies the condition $T \subset T_0$. The leaf nodes of a tree are indexed by $\tau = 1, \ldots, |T|$, where $|T|$ is the maximum number of leaf nodes. Every leaf node τ has associated a region of the input space, R_τ, and a set of data points, N_τ. The optimal prediction for a region R_τ is given by:

$$y_\tau = \frac{1}{N_\tau} \sum_{x_n \in R_\tau} t_n \tag{2-29}$$

with the associated residual sum of squares:

$$Q_\tau(T) = \sum_{x_n \in R_\tau} \{t_n - y_\tau\}^2 \tag{2-30}$$

The pruning criterion for regression tasks is expressed as:

$$C(T) = \sum_{\tau=1} Q_\tau(T) + \lambda |T| \tag{2-31}$$

where λ is a regularization parameter that controls the exchange between the overall residual sum, $Q_\tau(T)$, and the number of leaf nodes $|T|$. λ is usually determined by cross-validation.

For classification tasks, the pruning criterion is based on performance measures different than those based on errors. Two common performance measurements used in the pruning of tree-based classifiers are the Gini index:

$$Q_\tau(T) = \sum_{k=1}^{K} p_{\tau k}(1 - p_{\tau k}) \tag{2-32}$$

and the cross-entropy:

$$Q_\tau(T) = \sum_{k=1}^{K} p_{\tau k} \ln(p_{\tau k}) \tag{2-33}$$

where $p_{\tau k}$ is the proportion of data points laying in the region R_τ of the class k, given a set of K classes (k = 1,…, K). Note that these two performance measurements achieve a maximum at $p_{\tau k} = 0.5$ and become zero when $p_{\tau k} = 0$ and $p_{\tau k} = 1$, helping the formation of regions in which a large number of data points are clustered in a class.

ISBN:978-84-693-4597-9 /DL:T.1010-2010

In general, decision trees tend to be very sensitive in relation to variations in their training data. Different data splits may be obtained for slightly different training data sets. However, the graphical representation generated by these algorithms makes them suitable for getting an intuitive understanding of the composition of large multivariable data sets. There are several algorithms modeling decision trees with different variations, some of the most known are: the ID3 algorithm (Quinlan, 1986), a decision tree based on the entropy performance measurement meant to produce small trees rather than large trees; the C4.5 algorithm (Quinlan, 1993), an extension of ID3 that examines the normalized information gain resulting from choosing an attribute for splitting the data; the J4.8 algorithm (Quinlan, 1993), meant for generating pruned or unpruned C4.5 decision trees; and, the Random Forest algorithm (Breiman, 2001), meant for constructing a forest of random trees.

Support Vector Machines

Support vector machines (SVMs) are algorithms that build mathematical structures from data vectors selected during the training phase (Cortes and Vapnik, 1995). SVMs perform well for classification problems involving two classes, when data are either linearly separable or not.

When the available data vectors of a two-class classification problem are linearly separable, one may choose from a variety of solutions, depending on how a standard classifier algorithm optimizes its errors. Figure 2-8a makes a graphical representation of such situation. Instead, SVMs search an optimal solution by establishing a line for which the distance, or margin, between itself and vectors lying on the boundaries is maximum, as in Figure 2-8b. The solution given by SVMs is unique and "supported" on vectors on the boundaries, regardless of any other elements in the data set.

When the data to classify are non-linearly separable, SVMs search for an optimal hyper-plane different from the original data space (or input space), a higher dimensional space, using projections of the original data by means of a feature function $\varphi(\mathbf{x})$. The feature function meets a necessary condition: the product of the feature function $\varphi(\mathbf{x})$ evaluated on two generic training vectors $\mathbf{x}_i$ and $\mathbf{x}_j$ must have an equivalent in the input space where the kernel $K(\mathbf{x}_i, \mathbf{x}_j)$ operates.

Algorithms depending on the number of support vectors rather than on the dimensionality of the feature space have decision functions, non-linear in the input space and based on the convolution of the inner product, with the form:

$$f(x) = \sum_{i=1}^{N} t_i \alpha_i K(\mathbf{x}_i, \mathbf{x}) + b \qquad (2\text{-}34)$$

equivalent to linear decision functions in the high-dimensional feature space $\psi_1(x), \ldots, \psi_N(x)$; where $K(\mathbf{x}_i, \mathbf{x})$ is a convolution of the inner product in the feature space. For finding the coefficients in either the separable case or the non-separable case it is enough to find the maximum of the function:

ISBN:978-84-693-4597-9 /DL:T.1010-2010

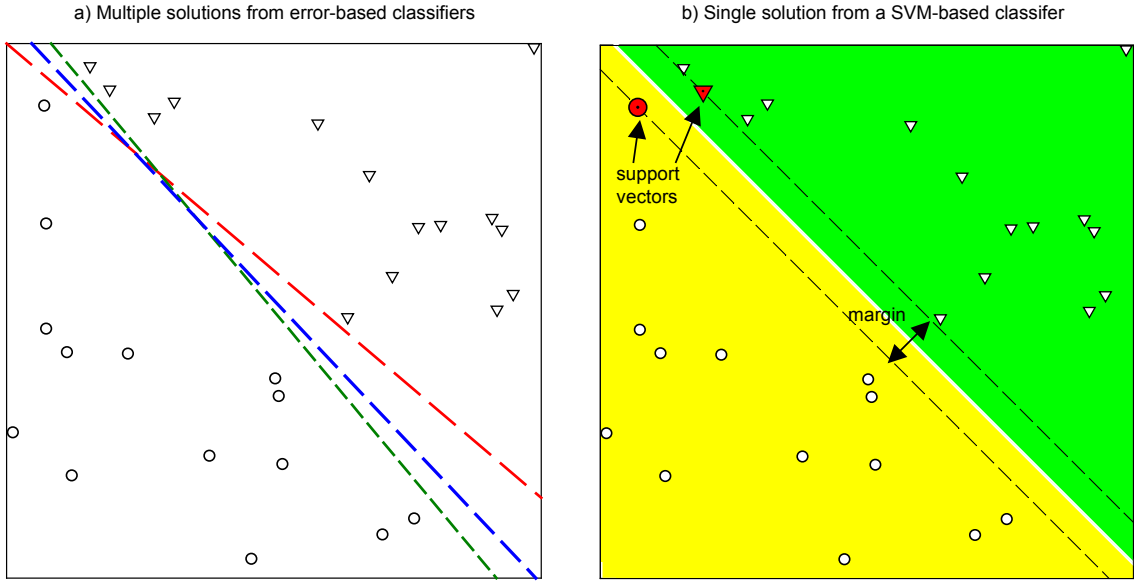


Figure 2-8. Decision boundaries of classifiers.
This figure shows how data points of a two-class linearly separable dataset are separated by standard error-based classifiers (a) and a SVM-based classifier (b). Multiple solutions can be derived from error-based classifiers as they find local minima during their training (a). Instead, support vectors provide an unique solution supported on selected data points (b).

$$W(\alpha) = \sum_{i=1}^{L} \alpha_i - \frac{1}{2} \sum_{i,j}^{L} \alpha_i \alpha_j t_i t_j K(\mathbf{x}_i, \mathbf{x}_j) \qquad (2\text{-}35)$$

subject to these restrictions:

$$\alpha \geq 0 \text{ and } \sum_{i=1}^{L} a_i y_i = 0 \qquad \text{for } i = 1, 2, \ldots, L. \qquad (2\text{-}36)$$

Using different functions for the convolution of inner products, $K(\mathbf{x}, \mathbf{x}_i)$, different types of nonlinear decision surfaces can be obtained. The most common are based on polynomials, radial basis functions and backpropagation networks:

for a polynomial learning machine: $$K(\mathbf{x}, \mathbf{x}_i) = [(\mathbf{x} \cdot \mathbf{x}_i) + 1]^d \qquad (2\text{-}37)$$

for a radial basis function machine: $$K(|\mathbf{x} - \mathbf{x}_i|) = \exp\left\{ \frac{-|\mathbf{x} - \mathbf{x}_i|^2}{2\sigma^2} \right\} \qquad (2\text{-}38)$$

for a two-layer neural network: $$K(\mathbf{x}, \mathbf{x}_i) = S[v(\mathbf{x} \cdot \mathbf{x}_i) + c] \qquad (2\text{-}39)$$

SVMs may perform classifications with considerable robustness and efficiency when compared to algorithms based on standard ANNs. A SVM-based model can be always reproduced if its training data remains unaltered. This is not the case of models based

ISBN:978-84-693-4597-9 /DL:T.1010-2010

on ANNs, which optimize their inner parameters (weights, biases) searching for a minimum error, yielding dissimilar models when finding different minima (a discussion for backpropagation ANNs is given later in this chapter).

2.4.3 Multivariate function approximators

Learning algorithms can be used for performing regressions involving several input and output variables. Their training process typically requires adjusting the parameters of a complex functional structure until closely matching a target multivariate function. This involves the training of algorithms under supervised learning conditions.

Multivariate regressions are required for predicting the outputs of a process when its operation is unpractical, known the inputs and outputs for a set of known cases. Several property estimation methods rely on QSARs (Section 1.2.3), grounded on multivariate regressions relating molecular descriptors linked to chemical activity (Equation 1-15). With the aim of guiding the development and validation of QSAR models, the Organisation for Economic Co-operation and Development (OECD), in the 37th OECD's Joint Meeting of the Chemicals Committee and the Working Party on Chemicals, Pesticides and Biotechnology (Joint Meeting) agreed on the "OECD Principles for the Validation, for Regulatory Purposes, of (Q)SAR Models" (OECD, 2007):

"To facilitate the consideration of a (Q)SAR model for regulatory purposes, it should be associated with the following information:

1. a defined endpoint;

2. an unambiguous algorithm;

3. a defined domain of applicability;

4. appropriate measures of goodness-of-fit, robustness and predictivity;

5. a mechanistic interpretation, if possible."

which, coupled to a large lists of methods, for the proper integration of QSARs into regulatory/decision-making frameworks, should be taken with some flexibility, depending on the needs and constraints of specific regulatory authorities.

A simple way of assessing a multivariate function approximator is calculating the mean absolute error (MAE) running over an entire dataset:

$$\mathrm{MAE} = \frac{\sum_{g=1}^{G}\sum_{n=1}^{N}\left|t_{n,g} - p_{n,g}\right|}{NG} \qquad \text{for N samples and G output variables} \qquad (2\text{-}40)$$

ISBN:978-84-693-4597-9 /DL:T.1010-2010

where the differences between target values ($t_{n,g}$) and the predictions of the algorithm ($p_{n,g}$) are averaged for all the samples of a data set (n = 1,…,N) and all the output variables (g = 1,…,G) in the process. Low MAE values indicate, in average, good predictions.

Since the objective of developing multivariate regressions is predicting new trends rather than describing known observations, emphasis is made on assessing the predictive capacity of trained models. The predictive squared correlation coefficient (q^2) is meant to indicate how well a single output variable g is individually predicted for all the elements of a dataset:

$$q^2 = 1 - \frac{\sum_{n=1}^{N}\left(p_n - t_n\right)^2}{\sum_{n=1}^{N}\left(t_n - \bar{t}_{dataset}\right)^2} \quad \text{for a dataset and a single output variable g} \qquad (2\text{-}41)$$

where, in average, the difference between predictions (p_n) and targets (t_n) is compared to the difference between the targets (t_n) and the average target value in the dataset ($\bar{t}_{dataset}$). q^2 yields 1 when optimal, 0 when predictions are as good as the average values and negative values when the averages are better estimators than the actual estimations. It has been suggested that trained algorithms should yield $q^2 > 0.5$ when evaluated on an external dataset, not used in the training of the models, for ensuring their predictive capacity.

However, it has been argued that the q^2 coefficient by itself is not enough for assessing the predictive capacity of a model, so a set of measurements has been proposed by Golbraikh and Tropsha (2002) and Tropsha et al. (2003). These authors suggest that a model, when evaluated on an external dataset, should comply with the following conditions:

$$q^2_{tr} > 0.5 \qquad (2\text{-}42)$$

$$R^2 > 0.6 \qquad (2\text{-}43)$$

$$0.85 \le k \le 1.15 \quad \text{or} \quad 0.85 \le k' \le 1.15 \qquad (2\text{-}44)$$

$$\frac{\left(R^2 - R_0^2\right)}{R^2} < 0.1 \quad \text{or} \quad \frac{\left(R^2 - R'^2_0\right)}{R^2} < 0.1 \qquad (2\text{-}45)$$

where:

$$q^2\big|_{tr} = 1 - \frac{\sum_{n=1}^{N}\left(t_n - p_n\right)^2}{\sum_{n=1}^{N}\left(t_n - \bar{t}_{tr}\right)^2} \qquad (2\text{-}46)$$

ISBN:978-84-693-4597-9 /DL:T.1010-2010

$$R^2 = \frac{\left(\sum_{n=1}^{N}(t_n - \bar{t}_{dataset})(p_n - \bar{p}_{dataset})\right)^2}{\left(\sum_{n=1}^{N}(t_n - \bar{t}_{dataset})^2\right)\left(\sum_{n=1}^{N}(p_n - \bar{p}_{dataset})^2\right)} \tag{2-47}$$

$$k = \frac{\sum_{n=1}^{N} t_n p_n}{p_n^2} \tag{2-48}$$

$$k' = \frac{\sum_{n=1}^{N} t_n p_n}{t_n^2} \tag{2-49}$$

$$R_0^2 = 1 - \frac{\sum_{n=1}^{N}(p_n - kp_n)^2}{\sum_{n=1}^{N}(p_n - \bar{p}_{dataset})^2} \tag{2-50}$$

$$R'^2_0 = 1 - \frac{\sum_{n=1}^{N}(t_n - k't_n)^2}{\sum_{n=1}^{N}(t_n - \bar{t}_{dataset})^2} \tag{2-51}$$

with all sums running over the elements of the external dataset and evaluating the average of the targets and predictions of the set, respectively, $\bar{t}_{set}$ and $\bar{p}_{set}$. Please note that $q^2|_{tr}$ (Equation 2-46) especially compares, in average, the difference between targets (t_n) and predictions (p_n) with respect to the difference between targets (t_n) and the average target value of the training set ($\bar{t}_{tr}$), not the average target value in the dataset on evaluation ($\bar{t}_{dataset}$, as in Equation 2-41).

Recently, it has been noted that the definition of the squared correlation coefficient with respect to the training set ($q^2|_{tr}$) may overestimate systematically the prediction capability of a model, yielding values higher than q^2 or R^2 when evaluated (Schüürmann et al., 2008). So Schüürmann and coworkers have suggested that q^2 should be used instead of q^2_{tr}; and, that the OECD guidelines for the validation of QSARs (OECD, 2007) should be modified to replace $q^2|_{tr}$ by q^2.

There is still a lot of controversy on how QSAR models should be validated. However, aside of the different kinds of performance measurements that can be applied to QSAR models, methods involving the visual inspection of targets and predictions generally constitute an important factor for assessing the goodness of QSARs. QSARs are dynamic models in the sense that they can be updated, as long as

ISBN:978-84-693-4597-9 /DL:T.1010-2010

new optimal conditions are found (training data, core algorithms, internal and external parameters, etc.).

Backpropagation Networks

Backpropagation networks (BPNs) are multilayer feed-forward neural networks that work under supervised learning, in which both the inputs and outputs of a problem are presented to the network. BPNs can work as function approximators with architectures based on at least one hidden layer of neurons with sigmoid transfer functions (Cybenko, 1989; Funahashi, 1989; Hornik et al., 1989). They are trained with the backpropagation algorithm (Parker, 1985; Rumelhart et al., 1986), which minimizes squared errors, using the chain rule for calculating derivatives of the squared error with respect to the weight and biases from the last to the first layer of the network.

The backpropagation algorithm is explained in the following lines for a BPN of M layers. Figure 2-9 shows a concise scheme of how the information is propagated through BPNs during their training. First, an input vector (**p**) is propagated forward through the network for calculating the outputs of each layer ($\mathbf{c}^m$) with the expression:

$$\mathbf{c}^m = \mathbf{f}^m(\mathbf{W}^m\mathbf{c}^{m-1} + \mathbf{b}^m) \text{ for m = 1, …, M, where } \mathbf{c}^0 = \mathbf{p} \quad (2\text{-}52)$$

Second, the sensitivities ($\mathbf{s}^m$) are calculated and propagated backward through the network:

$$\mathbf{s}^M = -2\dot{\mathbf{F}}^M(\mathbf{n}^M)(\mathbf{t} - \mathbf{a}^M) \text{ for the output layer (the last layer, m = M)} \quad (2\text{-}53)$$

$$\mathbf{s}^m = \dot{\mathbf{F}}^m(\mathbf{n}^m)(\mathbf{W}^{m+1})^T s^{m+1} \text{ for the hidden layers m = M-1, …, 2, 1.} \quad (2\text{-}54)$$

where:

$$\dot{\mathbf{F}}(\mathbf{n}^m) = \begin{bmatrix} \dot{f}^m(n_1^m) & 0 & \dots & 0 & 0 \\ 0 & \dot{f}^m(n_2^m) & & & 0 \\ \vdots & & . & & \vdots \\ 0 & & & . & 0 \\ 0 & 0 & \dots & 0 & \dot{f}^m(n_{s^m}^m) \end{bmatrix} \quad (2\text{-}55)$$

$$\dot{f}^m(n_j^m) = \frac{\partial f^m(n_j^m)}{\partial n_j^m} \quad (2\text{-}56)$$

ISBN:978-84-693-4597-9 /DL:T.1010-2010

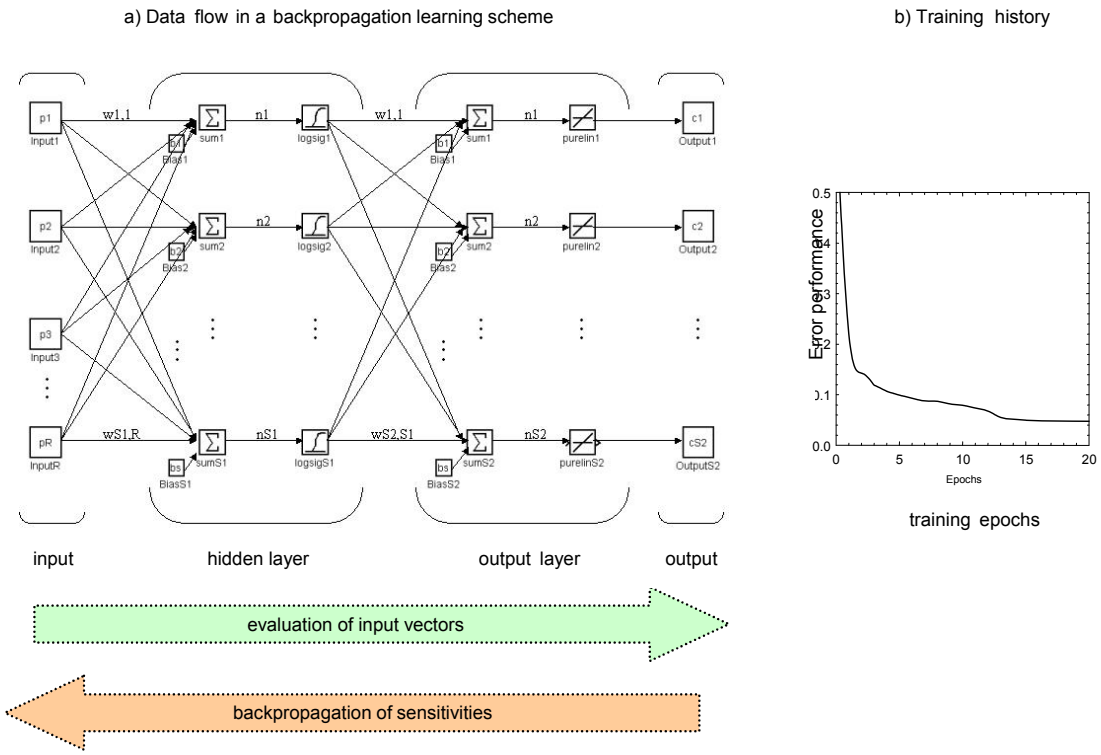


Figure 2-9. Training of backpropagation networks.
Backpropagation networks adjust their internal parameters with basis on the comparison they perform between target and predicted values for a given training data set. This figure shows how data flows through a two-layer feedforward network trained by the backpropagation algorithm (a), with performance errors decreasing for every training epoch (b). In every epoch (a), the input data is presented to the network, the predictions are compared to their corresponding target values and the inner parameters (weights, biases) of the network are updated with the backpropagation of sensititivites until a minimum performance error is achieved (b).

$$n_i^m = \sum_{j=1}^{S^{m-1}} w_{i,j}^m a_j^{m-1} + b_i^m \tag{2-57}$$

Third, the weights and biases of all the layers in the network are updated by means of the approximate steepest descent rule:

$$\mathbf{W}^m(k+1) = \mathbf{W}^m(k) - \alpha \mathbf{s}^m (\mathbf{a}^{m-1})^T \tag{2-58}$$

$$\mathbf{b}^m(k+1) = \mathbf{b}^m(k) - \alpha \mathbf{s}^m \tag{2-59}$$

Fourth, the whole procedure is repeated until minimum errors in the outputs of the network are obtained.

The backpropagation algorithm has two major shortcomings, requires long computational times and may become unstable for high dimensional data. Additionally, there is always the possibility of overtraining when the repetitions of the algorithm are not stopped on time; for these reason, an error threshold must be specified for stopping the training process when an optimal solution has been obtained. There are several variations of the backpropagation algorithm to accelerate

ISBN:978-84-693-4597-9 /DL:T.1010-2010

its convergence. One of the fastest methods is the Levenberg-Marquadt optimization algorithm (Hagan and Menhaj, 1994).

Radial Basin Functions

A radial basis function (RBF) is a two-layer neural network that contains basis functions in its hidden layer, usually Gaussian bell functions, and linear functions in its output layer (Lo, 1998). RBFs require the determination of the mean and standard deviation from the input data to calculate the output from each output neuron, given by:

$$g_i(x) = \exp\left(\frac{-|x - c_i|^2}{\sigma^2}\right) \tag{2-60}$$

where for a given neuron i, c_i is its centre, σ_i is its radius (also called spread) and $|x-c_i|$ is the Euclidean distance between the input vectors and the ith centre. With the prediction of the network calculated according to the expression:

$$y(x) = \sum_{i=0}^{n-1} w_{ij} \cdot g_i(x) + w_{0j} \tag{2-61}$$

Support Vector Regressions

SVMs can also be used as multivariate function approximators (Drucker et al., 1996), usually referred to as Support Vector Regressions (SVRs). The original SVM algorithm is altered with the application of loss functions, usually called ε-insensitive loss functions, required for making models not only robust but also sparse. These functions are very important for estimating dependencies for large numbers of data vectors; the magnitude of ε is inversely proportional to the amount of support vectors included in a model.

Given a training dataset $\{(\mathbf{x}_1,\mathbf{t}_1),\ldots,(\mathbf{x}_N,\mathbf{t}_N)\}$ with N points composed of inputs ($\mathbf{x}_i$) of dimension D and targets of dimension 1 (t_i), the goal is to establish a regression function f(x), as flat as possible, with at most a deviation of magnitude ε for all the targets (t_i) in the dataset. Errors are accepted only if they are lower than ε, rejecting deviations larger than this. For linear functions, f(x) takes the form:

$$f(x) = \langle \mathbf{w}, \mathbf{x} \rangle + b \qquad \text{with:} \quad \mathbf{w}, \mathbf{x} \in R^D,\ b \in R \tag{2-62}$$

in a convex optimization problem in which the norm ($\|\mathbf{w}\|^2 = \langle \mathbf{w}, \mathbf{w} \rangle$) is minimized:

$$\frac{1}{2}\|\mathbf{w}\|^2 \tag{2-63}$$

ISBN:978-84-693-4597-9 /DL:T.1010-2010

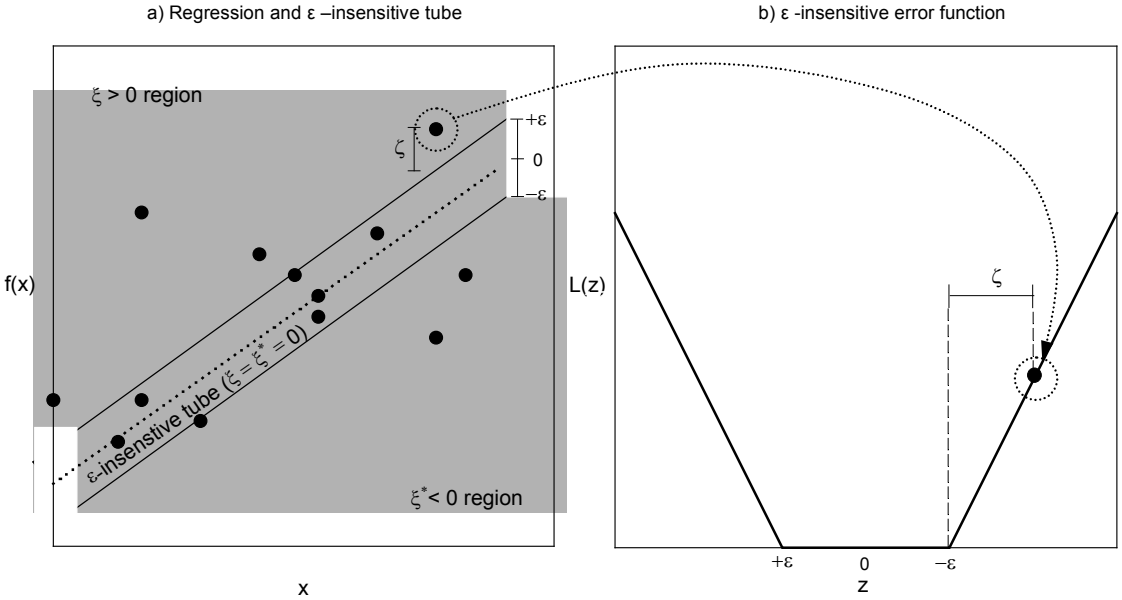


Figure 2-10. Training of support vector regressions.
The support vector regression is surrounded by the ε-insensitive tube (a), where $\xi = \xi^* = 0$. Data points outside the ε-insensitive tube contribute to the cost, identified by the ε-insensitive error function (b).

subject to:

$$t_i - \langle \mathbf{w}, \mathbf{x}_i \rangle - b \leq \varepsilon \tag{2-64}$$

$$\langle \mathbf{w}, \mathbf{x}_i \rangle + b - t_i \leq \varepsilon \tag{2-65}$$

The minimization problem above (Equations 2-61, 2-62 and 2-63) assumes tacitly that a function f(x) exists and that it approximates all pairs ($\mathbf{x}_i$,t_i) with ε precision. Slack variables (ξ, ξ^*) can be introduced to deal with infeasible constrains of this problem to reformulate it as the minimization of the function:

$$\frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{N}\left(\xi_i + \xi_i^*\right) \tag{2-66}$$

subject to:

$$t_i - \langle \mathbf{w}, \mathbf{x}_i \rangle - b \leq \varepsilon + \xi_i \tag{2-67}$$

$$\langle \mathbf{w}, \mathbf{x}_i \rangle + b - t_i \leq \varepsilon + \xi_i^* \tag{2-68}$$

$$\xi_i, \xi_i^* \geq 0 \tag{2-69}$$

ISBN:978-84-693-4597-9 /DL:T.1010-2010

where C>0, the regularization parameter, determines the swapping between the flatness of f(x) and the amount up to which deviations larger than ε are tolerated. This leads to the establishment of the ε-insensitive error function, which, when linear, has the form:

$$|\xi|_\varepsilon = \begin{cases} 0 & \text{if} |\xi| \le \varepsilon \\ |\xi| - \varepsilon & \text{otherwise} \end{cases} \tag{2-70}$$

selecting points that lay outside the e-insensitive tube, the region around the regression, as a contribution to the cost insofar, (deviations are penalized linearly). Points are assigned $\xi > 0$ or $\xi^* > 0$, when laying, respectively, above or below the ε-insensitive tube. Point within the tube have $\xi = \xi^* = 0$.

The new optimization problem (Equations 2-64, 2-65, 2-66, 2-67 and 2-68) can be solved by further applying a dual formulation, that allows the extension of the algorithm to non-linear functions by the application of Lagrange multipliers, yielding the function:

$$w = \sum_{i=1}^{N} (\alpha i - \alpha_i^*) x_i \; ; \quad \text{thus} \quad f(x) = \sum_{i=1}^{N} (\alpha i - \alpha_i^*) \langle \mathbf{x}_i, \mathbf{x} \rangle + b \tag{2-71}$$

a linear combination of the training patterns (x_i), independent from the dimensionality of the input space (D), dependent solely on the number of support vectors. B can be computed with the application of the Karush-Kuhn-Tucker (KKT) conditions (Karush, 1939; Kuhn and Tucker, 1950).

The support vector regression (Equation 2-69) can become non-linear, preprocessing the input patterns (x_i) in the feature space *F* of a kernel function Φ: $R^D \rightarrow F$.

$$w = \sum_{i=1}^{N} (\alpha i - \alpha_i^*) \Phi(x_i); \quad \text{thus} \quad f(x) = \sum_{i=1}^{N} (\alpha i - \alpha_i^*) k \langle \mathbf{x}_i, \mathbf{x} \rangle + b \tag{2-72}$$

leaving the optimization problem as the search of the flattest function in feature space, instead of the original input space.

2.5 Multimedia environmental modeling from pattern recognition

2.5.1 Philosophy of QPFRs and QSFRs

MEMs estimate the distribution of chemical pollutants in the environment from data describing specific geographical locations and the physicochemical properties and emission rates of pollutants of concern (Section 1.2.1). Given the large variety of

ISBN:978-84-693-4597-9 /DL:T.1010-2010

input and output parameters involved, any MEM can be considered as a multivariate function that, in matrix form, can be defined as:

$$\mathbf{C} = f_{\mathrm{MEM}}(\mathbf{P}, \mathbf{E}, \mathbf{S}) \tag{2-73}$$

where **C** is a matrix of environmental fate estimations (in terms of concentrations, mass fractions, fugacity values, etc.), **P** a matrix of physicochemical properties, **E** a matrix of emission rates and **S** a vector of site-specific parameters. These terms can be subsequently defined, for N chemical pollutants characterized by K physicochemical properties and emitted on J out of G environmental compartments, as follows:

$$\mathbf{C} = \begin{bmatrix} C_{1,1} & . & . & . & C_{1,G} \\ . & . & & & . \\ . & & C_{n,g} & & . \\ . & & & . & . \\ C_{N,1} & . & . & . & C_{N,G} \end{bmatrix} \tag{2-74}$$

$$\mathbf{P} = \begin{bmatrix} P_{1,1} & . & . & . & P_{1,K} \\ . & . & & & . \\ . & & P_{n,k} & & . \\ . & & & . & . \\ P_{N,1} & . & . & . & P_{N,K} \end{bmatrix} \tag{2-75}$$

$$\mathbf{E} = \begin{bmatrix} E_{1,1} & . & . & . & E_{1,J} \\ . & . & & & . \\ . & & E_{n,j} & & . \\ . & & & . & . \\ E_{N,1} & . & . & . & E_{N,J} \end{bmatrix} \tag{2-76}$$

$$\mathbf{S} = \begin{bmatrix} S_{1,1} \\ . \\ S_{m,1} \\ . \\ S_{M,1} \end{bmatrix} \tag{2-77}$$

When the emission rates of a set of chemicals remain constant on a fixed geographical scenario (**E** and **S** constants), it is possible to consider a MEM as a multivariate function that relates the physicochemical properties (**P**) of pollutants to environmental fate estimations (**C**):

$$\mathbf{C} = f_{\mathrm{MEM}}(\mathbf{P}) \qquad \text{if } \mathbf{E} \text{ and } \mathbf{S} \text{ remain constant} \tag{2-78}$$

simplifying the original multimedia environmental modeling approach (Equation 2-73) and focusing the environmental assessment of chemicals solely on their physicochemical properties.

ISBN:978-84-693-4597-9 /DL:T.1010-2010

Uncertainty in standard environmental assessments. For assessing the environmental fate of chemicals for which physicochemical properties are missing, it is common practice to estimate every missing property from available QSPR and QSBR methods (Boethling et al., 2004) and evaluate a MEM as usual (Mackay, 2001), known emission rates and a geographic scenario (Figure 1-4). In result, different levels of uncertainty must be expected at the output of the MEM. If $\mathbf{P}^{est}$ is a matrix of physicochemical properties totally or partially estimated, the outcome of a MEM using $\mathbf{P}^{est}$ as input can be either reasonably approximated to the outcome of the same model (**C**) using a set of reference properties **P** (Equation 2-78):

$$\mathbf{C} \approx f_{\mathrm{MEM}}(\mathbf{P}^{\mathrm{est}}) \qquad \text{if } \mathbf{P} \approx \mathbf{P}^{\mathrm{est}} \tag{2-79}$$

or, in the worst cases, wrongly estimated:

$$\mathbf{C} \neq f_{\mathrm{MEM}}(\mathbf{P}^{\mathrm{est}}) \qquad \text{if } \mathbf{P} \neq \mathbf{P}^{\mathrm{est}} \tag{2-80}$$

depending on the amount of estimated properties, the uncertainty associated to each value and their role in the equations of the MEM.

Quantitative property-fate relationships.

When some physicochemical properties are unavailable for chemicals of concern, alternative environmental fate predictions can be obtained from available physicochemical data, using QPFRs (Figure 1-4). Supervised learning algorithms (like the ones described in Section 2.4.3) can be used, given a set of training chemicals, to establish relationships between reduced set of properties to the outputs of a MEM (Martínez et al., 2006a; Martínez et al., 2006b):

$$\mathbf{C} \approx f_{\mathrm{QPFR}}(\mathbf{P}^{*}) \qquad \text{if } \mathbf{P} \text{ exists} \tag{2-81}$$

where **C** is a matrix of fate predictions generated by a reference MEM for a set of training chemicals (Equation 2-78), **P** is a matrix with all the K properties required by the reference MEM and $\mathbf{P}^{*}$ is a matrix with a reduced number of properties K^{*} ($K^{*} < K$). The environmental assessment of new chemicals for which some properties are available can be performed with QPFRs, as long as the former have the exact set of K^{*} available properties required by the latter as input.

Quantitative structure-fate relationships

When key physicochemical properties are either unavailable or extremely noisy for chemicals of concern, alternative environmental fate predictions can be obtained from molecular information, using QSFRs (Figure 1-4). Supervised learning algorithms (Section 2.4.3) can be used, given a set of training chemicals, to establish relationships between available molecular information to the outputs of a MEM (Martínez et al., 2007a; Martínez et al., 2007b):

ISBN:978-84-693-4597-9 /DL:T.1010-2010

$$\mathbf{C} \approx f_{\mathrm{QSFR}}(\mathbf{D}) \qquad \text{if } \mathbf{P} \text{ exists} \tag{2-82}$$

where **C** is a matrix of fate predictions generated by a reference MEM for a set of training chemicals (Equation 2-78), **P** is a matrix with all the K properties required by the reference MEM and **D** is a matrix with L molecular descriptors. Known the molecular structure of a new chemical of concern, it is possible to calculate any type of molecular descriptors for later assessing its environmental fate through QSFRs.

2.5.2 Training supervised learning algorithms to emulate MEMs as QPFRs or QSFRs

In this thesis, QPFR and QSFR models have been developed for estimating level III mass ratios (Equation 3-1) in compartments of the reference pollution scenario to be described in Chapter 3. Every model presented and discussed in Chapter 4 predicts a mass ratio w_g from either a set of available properties $p_1,...,p_{K*}$, in the case of QPFRs; or, a set of molecular descriptors $d_1,...,d_L$, in the case of QSFRs. These models, based on supervised learning algorithms, require the same considerations that apply in the development of standard property estimation methods relying on the QSAR approach (Section 1.2.3):

- Compiling training data with the highest possible quality (Stouch et al., 2003)
- Avoiding the presence of outliers (Furusjö et al., 2006).
- Selecting appropiate input features (Saeys et al., 2007) from large number of descriptors (Bredow and Jug, 2005; Burden et al., 2009; Duca and Hopfinger, 2001; Senese et al., 2004; Todeschini and Consonni, 2000)
- Selecting and tuning the learning algorithms for building the models (Basheer and Hajmeer, 2000; Xu et al., 2006).
- Overcoming the risk of overtraining in the models (Byvatov et al., 2003).
- Validating externally the models (Golbraikh and Tropsha, 2002; OECD, 2007; Schüürmann et al., 2008).
- Assessing the domain of applicability of the models (Weaver and Gleeson, 2008).

The simultaneous optimization of all these factors is a problem that leads to almost infinite hypothesis (Johnson, 2008). So, taking this in mind, such factors have been adapted and merged into a methodology that builds emulators of any given MEM for available well-known chemicals, as described in Table 2-1. For allowing the tuning of algorithms in their training phase, available work chemicals must be characterized by a set of attributes (physicochemical properties or molecular descriptors) and fate estimations (the outputs of a MEM); to be precise, the inputs and targets of the algorithms, respectively. For assessing the fate of new chemicals, solely selected attributes are needed.

ISBN:978-84-693-4597-9 /DL:T.1010-2010

Table 2-1. Methodology used for training, testing and validating QPFRs and QSFRs in this study.

Step	Action	Description
1st	Pre-processing work data	Work data, conformed by both the input and target variables of a QSFR for a set of available chemicals, are pre-processed, per variable, by base 10 logarithmic scaling (if having values spanning more than two orders of magnitude) and normalization in the range [-1, 1] (according to Equation 2-5).
2nd	Selecting the input variables of a model	The input variables to use in a model are selected by either expert criteria (supported on the literature, assumptions and practical conditions) or empirical data filtering by the CFS algorithm (Hall, 1999), depending on feasibility and generalization capabilities of the algorithm.
3rd	Building the training and test data sets of a model	For every model, training and test data sets are derived from available work chemicals in the reference scenario by means of the SOM algorithm: about 80 % of available work chemicals are dedicated to training the model, while the rest of work chemicals are reserved for testing its performance while tuning its parameters (3rd step). The SOM algorithm, based on the implementation of the SOM toolbox 5 for Matlab (Vesanto et al., 2000), has been used to force the diversity of the training data set and the representation of the test data set in the former as follows (Annex B.1; coupled to Annexes B.2, B.3 where pertinent): First, SOMs of different sizes are trained to fit all available work chemicals in the input-target space of every model. The SOMs are set to have toroidal shapes and hexagonal lattices, for diminishing their respective mean quantization errors (Equation 2-18) and mean topological errors (Equation 2-19) as much as possible, while all chemicals are characterized by the normalized inputs and target variables of the model to train. Second, from each resulting SOM, work chemicals are included into a candidate training data set when showing the lowest or highest quantization error with respect to the closest SOM unit, having extreme values (the lowest or highest values in the whole work data set) in target variables or, in the case of QPFRs, in physicochemical properties as well. All work chemicals not following these characteristics are moved to the corresponding candidate test data set instead. Third, pairs of candidate training and test data sets are considered for the development of models when the number of training chemicals is about 80 % (±5 %) the total number of work chemicals. That is, with a relation of training-test chemicals of about 4:1 in which the training chemicals tend to surround the test chemicals in a PCA space (Pearson, 1901) conformed by the inputs and targets of the model.
4th	Pre-processing the validation data set	New data, conformed by the inputs and target variables of a model for a set of chemicals not used at all in the development of the models, are preprocessed under the same conditions in which the work data was preprocessed, applying base 10 logarithms if applicable and normalizing in the range [-1, 1] (Equation 2-5) according to the minimum and maximum values contained in the work data set.
5th	Training, testing and validating a model	Any model, based on supervised learning, emulates a reference MEM with a form resembling that of standard QSARs (Equation 1-13). QPFRs (Equation 2-81) have the form: $$N_{[-1,1]}\left(\log_{10}\left(w_g\right)\right) = f_{QPFR}\left(N_{[-1,1]}\left(\log_{10}\left(p_1\right)\right),\ldots,N_{[-1,1]}\left(\log_{10}\left(p_{K*}\right)\right)\right) \quad (2\text{-}83)$$ while QSFRs (Equation 2-82) have the form: $$N_{[-1,1]}\left(\log_{10}\left(w_g\right)\right) = f_{QSFR}\left(N_{[-1,1]}\left(d_1\right),\ldots,N_{[-1,1]}\left(d_L\right)\right) \quad (2\text{-}84)$$ where a w_g is the dimensionless mass ratio of a compartment g (the target variable, defined by the Equation 3-1) and *f* is the function resulting from the training of a supervised learning algorithm as QPFR or QSFR. …

ISBN:978-84-693-4597-9 /DL:T.1010-2010

Table 2-1. Methodology used for training, testing and validating QPFRs and QSFRs in this study (continued).

Step	Action	Description
5th	Training, testing and validating a model	… The models presented in this work are based on the SVR algorithm with RBF kernel functions. The ε-SVR implementation in the software package RapidMiner 4.4 (Mierswa et al., 2006) has been used to build the QPFRs and QSFRs of Chapter 4, per compartment g, with basis on a candidate training data set that contains about 80% of available work chemicals (selected with a SOM, as explained in the 3rd step). For every compartment and set of input features considered, an iterative evaluation of 4000 models is implemented for tuning the parameters of an optimal SVR model (Annex B.4): C, γ, ε and p. For every combination of parameters, a SVR is developed with the training data set and evaluated on the test and validation data sets. An optimal SVR model is selected when having the lowest mean absolute error (MAE) on the test data set among the SVRs with the 10 highest squared correlation (R^2) values on the test data set. This criteria aims to select a model with optimal generalization capabilities based on chemicals not included in the training set, but somehow represented in it. The MAE and R^2 measurements are calculated (Annex B.5), respectively, with Equation 2-40 and Equation 2-47, per compartment (G = 1) and over the normalized logarithmic mass ratios of all the chemicals of a given data set (tr = training, te = test or val = validation). Having selected a SVR model for an optimal set of parameters, its accuracy is estimated by means of both a 10-fold cross validation (CV) and a leave one out (LOO) validation procedure running over all the available work chemicals (Annexes B.6 and B.7, respectively). In both cases, the MAE and R^2 values of all subsets are averaged. Note that so far the outputs of the SVRs are normalized logarithms of mass ratios.
6th	Post-processing of fate predictions	Initially, predictions of normalized logarithmic mass ratios for all the data sets (training, test and validation sets) are obtained by evaluating them in a QPFR or QSFR model, respectively, Equation 2-83 or Equation 2-84. Later, they are denormalized using Equation 2-5 backwards, solving y_n from $N_{[-1,1]}(y_n)$, where $y_n = \log_{10}(w_g)$, yielding logarithmic mass ratios.
7th	Measuring the performance of a model	For measuring the performance of a compartmental QPFR or QSFR model with respect to a single data set, its predictions are compared with respect to the target values, i.e., the reference mass ratios originally generated by the reference MEM for the pollution scenario considered. The differences between targets and predictions are estimated, in average, calculating a mean absolute error over logarithmically scaled predictions as follows: $$\mathrm{MAE} = \frac{1}{N}\sum_{n=1}^{N}\left\lvert \log_{10}\left(w_n^{\text{target}}\right) - \log_{10}\left(w_n^{\text{predicted}}\right)\right\rvert \quad (2\text{-}85)$$ the lower the MAE of a data set, the lower the differences between the targets and predictions of all chemicals in the set. The predictive performance of a model is assessed in terms of the predictive squared coefficient suggested by Schüürmann et al. (2008), q^2, as follows: $$q^2 = 1 - \frac{\sum_{n=1}^{N}\left(\log_{10}\left(w_{n,g}^{\text{predicted}}\right) - \log_{10}\left(w_{n,g}^{\text{target}}\right)\right)^2}{\sum_{n=1}^{N}\left(\log_{10}\left(w_{n,g}^{\text{target}}\right) - \frac{1}{N}\sum_{n=1}^{N}\log_{10}\left(w_{n,g}^{\text{target}}\right)\right)^2} \quad (2\text{-}86)$$ with the q^2 coefficient varying in the range $(-\infty,1]$. Models with q^2 values closer to 1 have a high predictive performance, but when having q^2 values equal or lower than zero their predictions are worst than simply averaging all targets.

ISBN:978-84-693-4597-9 /DL:T.1010-2010

References

Anderson E, Veith GD, Weininger D. SMILES: A line notation and computerized interpreter for chemical structures. U.S. Environmental Protection Agency, Environmental Research Laboratory-Duluth, Duluth, MN 55804, 1987.

Barnard GA, Bayes T. Studies in the History of Probability and Statistics: IX. Thomas Bayes's Essay Towards Solving a Problem in the Doctrine of Chances. Biometrika 1958; 45: 293-315.

Basheer IA, Hajmeer M. Artificial neural networks: fundamentals, computing, design, and application. Journal of Microbiological Methods 2000; 43: 3.

Boethling RS, Howard PH, Meylan WM. Finding and estimating chemical property data for environmental assessment. Environmental Toxicology and Chemistry 2004; 23: 2290-2308.

Brandes LJ, den Hollander H, van de Meent D. SimpleBox 2.0: a nested multimedia fate model for evaluating the environmental fate of chemicals. RIVM, Bilthoven, The Netherlands, 1996, pp. 156.

Bredow T, Jug K. Theory and range of modern semiempirical molecular orbital methods. Theoretical Chemistry Accounts: Theory, Computation, and Modeling (Theoretica Chimica Acta) 2005; 113: 1.

Breiman L. Random Forests. Machine Learning 2001; 45: 5.

Breiman L, Friedman J, Olshen RA, Stone CJ. Classification and Regression Trees: Chapman & Hall/CRC, 1984.

Brown SD, Blank TB, Sum ST, Weyer LG. Chemometrics. Analytical Chemistry 1994; 66: 315-359.

Cortes C, Vapnik V. Support-vector networks. Machine Learning 1995; 20: 273.

Cybenko G. Approximation by superpositions of a sigmoidal function. Mathematics of Control, Signals and Systems. 1989; 2: 303-314.

den Hollander HA, van de Meent D. Appendix to SimpleBox 3.0: A multimedia mass balance model for evaluating the environmental fate of chemicals. RIVM, 2004.

den Hollander HA, van Eijkeren JCH, van de Meent D. SimpleBox 3.0. RIVM, Bilthoven, The Netherlands, 2004.

Dewar MJS, Zoebisch EG, Healy EF, Stewart JJP. Development and use of quantum mechanical molecular models. 76. AM1: a new general purpose quantum mechanical molecular model. Journal of the American Chemical Society 1985; 107: 3902-3909.

Doane DP. Using Simulation to Teach Distributions. Journal of Statistics Education 2004; 12.

Drucker H, Burges CJC, Kaufman L, Smola A, Vapnik V. Support Vector Regression Machines. Advances in Neural Information Processing Systems 1996: 155-161.

Duca JS, Hopfinger AJ. Estimation of Molecular Similarity Based on 4D-QSAR Analysis: Formalism and Validation. Journal of Chemical Information and Computer Sciences 2001; 41: 1367-1387.

Fayyad UM, Piatetsky-Shapiro G, Smyth P, Uthurusamy R. Advances in knowledge discovery and data mining: The MIT Press, 1996.

Funahashi K. On the approximate realization of continous mappings by neural networks. Neural Netwoks. 1989; 2: 183-192.

George HJ, Langley P. Estimating Continuous Distributions in Bayesian Classifiers. Eleventh Conference on Uncertainty in Artificial Intelligence. Morgan Kaufmann, San Mateo, 1995, pp. pp. 338-345.

Golbraikh A, Tropsha A. Beware of q2! Journal of Molecular Graphics and Modelling 2002; 20: 269.

Guha R, Howard MT, Hutchison GR, Murray-Rust P, Rzepa H, Steinbeck C, et al. The Blue Obelisk Interoperability in Chemical Informatics. Journal of Chemical Information and Modeling 2006; 46: 991-998.

Hagan MH, Menhaj MB. Training feedforward networks with the Marquardt algorithm. IEEE Transactions on

ISBN:978-84-693-4597-9 /DL:T.1010-2010

Neural Networks 1994; 5: 989-993.

Hall MA. Correlation-based Feature Selection for Machine Learning. Department of Computer Science. Ph.D. thesis. The University of Waikato, Hamilton, New Zealand, 1999.

Hornik K, Stinchcombe M, White H. Multilayer feedforward networks are universal approximators. Neural Networks 1989; 2: 359.

Hotelling H. Relations between two sets of variates. Biometrika 1936; 28: 321-377.
Hugo K. From Narcosis to Hyperspace: The History of QSAR. Quantitative Structure-Activity Relationships 2002; 21: 348-356.

Jain AK, Murty MN, Flynn PJ. Data clustering: a review. ACM Comput. Surv. 1999; 31: 264-323.

James JPS. Optimization of parameters for semiempirical methods II. Applications. Journal of Computational Chemistry 1989; 10: 221-264.

Jolliffe I. Principal Component Analysis: Springer, 2002.

Karush W. Minima of functions of several variables with inequalities as side constraints. Dept. of Mathematics. Master's Thesis. University of Chicago, 1939.

Kohonen T, Oja E, Simula O, Visa A, Kangas J. Engineering applications of the self-organizing map. Proceedings of the IEEE 1996; 84: 1358.

Kuhn HW, Tucker AW. Nonlinear Programming. Proceedings of 2nd Berkeley Symposium on Mathematical Statistics and Probabilistic. University of California Press, Statistical Laboratory of the University of California, Berkeley, 1950, pp. 481-492.

Lance P, Ehtesham H, Huan L. Subspace clustering for high dimensional data: a review. SIGKDD Explor. Newsl. 2004; 6: 90-105.

Lijzen JPA, Rikken MGJ. European Union System for the Evaluation of Substances 2.0 (EUSES 2.0); background report. RIVM, Bilthoven, the Netherlands., 2004, pp. 454.

Limpert E, Stahel WA, Abbt M. Log-normal Distributions across the Sciences: Keys and Clues. BioScience 2001; 51: pp. 341–352.

Lo JT-H. Multilayer perceptrons and radial basis functions are universal robust approximators. IEEE International Conference on Neural Networks - Conference Proceedings. 2, 1998, pp. 1311.

Mackay D. Multimedia Environmental Models - The Fugacity Approach. Boca Ratón: Lewis Publishers, 2001.

Mackay D, Paterson S. Evaluating the multimedia fate of organic chemicals: a level III fugacity model. Environ. Sci. Technol. 1991; 25: 427-436.

Mackay D, Paterson S, Shiu WY. Generic models for evaluating the regional fate of chemicals. Chemosphere 1992; 24: 695.

MacQueen J. Some methods for classification and analysis of multivariate observations. In: Neyman LMLCaJ, editor. Fifth Berkeley Symposium on Mathematical Statistics and Probability. University of California Press, Statistical Laboratory of the University of California, Berkeley, 1967, pp. 281-297.

Martínez I, Espinosa G, Grifoll J, Cohen Y, Giralt F. Modelling chemical multimedia partitioning with neural networks. SETAC Europe 16th Annual Meeting, The Hague, The Netherlands, 2006a.

Martínez I, Espinosa G, Rallo R, Grifoll J, Cohen Y, Giralt. F. Estimation of environmental multimedia partitioning of pollutants from molecular descriptors using artificial neural networks. SETAC Europe 17th Annual Meeting, Oporto, Portugal, 2007a.

Martínez I, Grifoll J, Rallo R. Cognitive neural network-based intelligent system to identify the most important variables for the differences found in partitioning behaviour, transport pathways and exposure routes between chemicals. Universitat Rovira i Virgili, Tarragona, Spain, 2006b.

Martínez I, Grifoll J, Rallo R, Giralt F. Report on the most suitable artificial neural network architectures and molecular descriptors to estimate environmental multimedia behavior, including a sensitivity analysis of the effect of compartment sizes on multimedia concentrations. Universitat Rovira i Virgili, Tarragona, Spain, 2007b.

ISBN:978-84-693-4597-9 /DL:T.1010-2010

McClean SI, Robert AM. Data Mining and Knowledge Discovery. Encyclopedia of Physical Science and Technology. Academic Press, New York, 2001, pp. 229.

McNaught A. The IUPAC International Chemical Identifier: InChl—A New Standard for Molecular Informatics. CHEMISTRY International 2006; 28: 12-15.

Metropolis N. The beginning of the Monte Carlo method. Los Alamos Science 1987.

Mierswa I, Wurst M, Klinkenberg R, Scholz M, Euler T. YALE: Rapid Prototyping for Complex Data Mining Tasks. In: Ungar L, Craven M, Gunopulos D, Eliassi-Rad T, editors. Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-06). ACM, Philadelphia, PA, USA, 2006, pp. 935-940.

OECD. Guidance Document on the Validation of (Quantitative) Structure-Activity Relationship [(Q)SAR] Models. OECD Series on Testing and Assessment 69. 2007. OECD Document ENV/JM/MONO(2007)

Parker DB. Learning-logic: Casting the cortex of the human brain in silicon. Center for Computational Research in Economics and Management science, MIT, Cambridge, MA., 1985.

Pearson K. On lines and planes of closest fit to systems of points in space. The London, Edinburgh and Dublin Philosophical Magazine and Journal of Science, sixth series 1901; 2: 559-572.

Pople JA. Nobel Lecture: Quantum chemical models. Reviews of Modern Physics 1999; 71: 1267.

Quinlan JR. Induction of decision trees. Machine Learning 1986; 1: 81.

Quinlan R. C4.5: Programs for Machine Learning. San Mateo, Ca.: Morgan Kaufmann Publishers, 1993.

Roothaan CCJ. New Developments in Molecular Orbital Theory. Reviews of Modern Physics 1951; 23: 69.

Rumelhart DE, Hinton GE, Williams. RJ. Learning representations by back propagating errors. Nature 1986; 323: 533-536.

Schüürmann G, Ebert R-U, Chen J, Wang B, Ku□hne R. External Validation and Prediction Employing the Predictive Squared Correlation Coefficient - Test Set Activity Mean vs Training Set Activity Mean. Journal of Chemical Information and Modeling 2008; 48: 2140-2145.

Senese CL, Duca J, Pan D, Hopfinger AJ, Tseng YJ. 4D-Fingerprints, Universal QSAR and QSPR Descriptors. Journal of Chemical Information and Computer Sciences 2004; 44: 1526-1539.

Todeschini R, Consonni V. Handbook of Molecular Descriptors: Wiley-VCH, 2000.

Tropsha A, Gramatica P, Gombar VK. The Importance of Being Earnest: Validation is the Absolute Essential for Successful Application and Interpretation of QSPR Models. QSAR & Combinatorial Science 2003; 22: 69-77.

Uriarte EA, Martín FD. Topology Preservation in SOM. International Journal of Applied Mathematics and Computer Sciences 2005; 1: 19.

van de Meent D. SIMPLEBOX: a generic multimedia fate evaluation model. RIVM, Bilthoven, The Netherlands, 1993.

Vermeire T, Rikken M, Attias L, Boccardi P, Boeije G, Brooke D, et al. European union system for the evaluation of substances: the second version. Chemosphere 2005; 59: 473.

Vermeire TG, Jager DT, Bussian B, Devillers J, den Haan K, Hansen B, et al. European Union System for the Evaluation of Substances (EUSES). Principles and structure. Chemosphere 1997; 34: 1823.

Vesanto J, Himberg J, Alhoniemi E, Parhankangas J. SOM Toolbox for Matlab 5, 2000.

Weininger D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. Journal of Chemical Information and Computer Sciences 1988; 28: 31-36.

Weininger D, Weininger A, Weininger JL. SMILES. 2. Algorithm for generation of unique SMILES notation. Journal of Chemical Information and Computer Sciences 1989; 29: 97-101.

Willighagen EL, Wehrens R, Buydens LMC. Molecular Chemometrics. Critical Reviews in Analytical Chemistry

ISBN:978-84-693-4597-9 /DL:T.1010-2010

2006; 36: 189 - 198.

Winston PH. Artificial Intelligence: Addison Wesley, 1992.

Witten IH, Frank E. Data Mining: Practical machine learning tools and techniques. San Francisco, U.S.: Morgan Kaufmann, 2005.

Zerner MC. Perspective on "New developments in molecular orbital theory". Theoretical Chemistry Accounts: Theory, Computation, and Modeling (Theoretica Chimica Acta) 2000; 103: 217.

ISBN:978-84-693-4597-9 /DL:T.1010-2010

ISBN:978-84-693-4597-9 /DL:T.1010-2010

Chapter 3

Reference pollution scenario

For predicting the environmental fate of new chemicals in the absence of key physicochemical properties, it is necessary to have enough examples of the environmental distribution of well known chemicals. Such data constitute a reference pollution scenario that may allow learning algorithms find relationships between environmental fate and either few physicochemical properties or molecular descriptors, creating QPFRs or QSFRs, respectively. This chapter describes the reference scenario of the algorithms employed and discussed later in Chapter 4.

ISBN:978-84-693-4597-9 /DL:T.1010-2010

3.1 General description

For establishing relationships between key properties or molecular information and fate, the inputs and outputs of a multimedia environmental model for several chemicals are required. This implies the use of available data for generating fate modeling examples for conditions of concern. For developing the models to be shown and discussed in Chapter 4, a hypothetical reference pollution scenario has been considered: Level III fate estimations for 468 chemical pollutants, 375 work chemicals (Annex C.1) and 93 validation chemicals (Annex C.2), emitted at a constant rate of 1 ton/yr in either water or air at The Netherlands. This scenario is based on chemicals of concern, for which biodegradation in water have been thoroughly tested (JETOC, 1992), and a Level III multimedia model originally developed and tested for the Netherlands, SimpleBox (Brandes et al., 1996; den Hollander and van de Meent, 2004; den Hollander et al., 2004; van de Meent, 1993).

Preliminary versions of the reference pollution scenario, also referred to The Netherlands but considering diverse sources of degradation data and emissions in various compartments, were used in the preliminary reports of the NOMIRACLE project (Martínez et al., 2008a; Martínez et al., 2006; Martínez et al., 2007; Martínez et al., 2008b), contained in Annexes A.a1 to A.a4. Since degradation data in water is usually a critical input for most multimedia fate models (Aronson et al., 2006), the final version of the reference scenario considers chemicals for which degradation rates in water have been derived from MITI-I degradability tests (NITE, 2002), as explained later in this chapter.

The Netherlands. The Netherlands has been modeled with SimpleBox 3 as a set of 5 homogeneous compartments (air, water, sediments, soil and vegetation), taking as reference an original modeling of the region with SimpleBox (Struijs and Peijnenburg, 2002) as an area of 40000 km^2 (divided in 1200 km^2 of fresh water, 10800 km^2 of natural soil, 24000 km^2 of agricultural soil and 4000 km^2 of other soil) next to 40000 km^2 of sea water. The height of the air compartment is 300 m, the wind speed is 1.5 m/s and the temperature of the system is 12 ºC. With the exception of the landscape parameters specified by Struijs and Peijnenburg (2002), all SimpleBox 3 default parameters (den Hollander and van de Meent, 2004; den Hollander et al., 2004) have been left unchanged. These parameters are listed in Table 3-1.

As discussed in section 2.1, SimpleBox 3 describes the environment as a set of homogenous compartments at different geographic scales (local, regional, continental and global). In the reference pollution scenario, the Netherlands is modeled with the regional scale of the SimpleBox 3 model, which comprises originally 10 homogeneous compartments: air, fresh water, sea water, fresh water sediments, sea water sediments, natural soil, agricultural soil, other soil, natural vegetation and agricultural vegetation. For simplifying all subsequent analysis, similar compartments have been merged, as shown in Table 3-2, into 5 general compartments: air, water, sediments, soil and vegetation. Note that SimpleBox 3 models the depth of soil compartments in terms of an effective depth for each pollutant that varies according to the degradation, diffusion and advection suffered of every chemical pollutant in soil.

ISBN:978-84-693-4597-9 /DL:T.1010-2010

Table 3-1. Landscape parameters used in SimpleBox 3 for modeling The Netherlands.

Nº	Parameter*	Symbol	Units	Value
1	Area of sea water +	AREAsea.R	m^2	4.00×10^{10}
2	Area of land+	AREAland.R	m^2	4.00×10^{10}
3	Total area in the regional system	SYSTEMAREA.R	m^2	8.00×10^{10}
4	Total area in the local system++	SYSTEMAREA.L	m^2	1.00×10^{-6}
5	Area fraction of fresh water+	AREAFRAC.w1R	-	1.50×10^{-2}
6	Area fraction of sea water +	AREAFRAC.w2R	-	5.00×10^{-1}
7	Area fraction of natural soil +	AREAFRAC.s1R	-	1.35×10^{-1}
8	Area fraction of agricultural soil +	AREAFRAC.s2R	-	3.00×10^{-1}
9	Area fraction of other soil +	AREAFRAC.s3R	-	5.00×10^{-2}
10	Height of air compartment +	HEIGHT.aR	m	3.00×10^{2}
11	Annual precipitation	RAINrate.R	m/s	2.22×10^{-8}
12	Water run off from natural soil	WATERrun.s1R	m^3/s	5.99×10^{1}
13	Water run off from agricultural soil	WATERrun.s2R	m^3/s	1.33×10^{2}
14	Water run off from other soil	WATERrun.s3R	m^3/s	2.22×10^{1}
15	Dry aerosol deposition rate	DRYDEPaerosol.R	m/s	6.68×10^{-7}
16	Standard mass fraction of organic carbon soil/sed.	CORG	-	2.00×10^{-2}
17	Mass fraction of organic carbon in natural soil	CORG.s1R	-	2.00×10^{-2}
18	Mass fraction of organic carbon in agr. soil	CORG.s2R	-	2.00×10^{-2}
19	Mass fraction of organic carbon in other soil	CORG.s3R	-	2.00×10^{-2}
20	Vegetation mass on natural soil	VEGmass.v1R	kg/m^2	1.20×10^{0}
21	Vegetation mass on agricultural soil	VEGmass.v2R	kg/m^2	1.80×10^{0}
22	Leaf area index of natural vegetation	LAI.v1R	-	3.90×10^{0}
23	Leaf area index of agricultural vegetation	LAI.v2R	-	2.70×10^{0}
24	Interception of wet aerosol deposition by nat. veg.	IFWETaerosol.v1R	-	5.00×10^{-2}
25	Interception of wet aerosol deposition by agr. veg.	IFWETaerosol.v2R	-	2.50×10^{-2}
26	Wet density of natural vegetation	RHO.v1R	kg/m^3	9.00×10^{2}
27	Wet density of agricultural vegetation	RHO.v2R	kg/m^3	9.00×10^{2}
28	Effective depth of natural soil~	PENdepth.s1R	m	3.00×10^{-2} to 1.00×10^{0}
29	Effective depth of agricultural soil~	PENdepth.s2R	m	3.00×10^{-2} to 1.00×10^{0}
30	Effective depth of other soil~	PENdepth.s3R	m	3.00×10^{-2} to 1.00×10^{0}
31	Mineral density of sediments and soil	RHOsolid	kg/m^3	2.50×10^{3}
32	Mixed depth of fresh water sediments	DEPTH.sd1R	m	3.00×10^{-2}
33	Mixed depth of sea water sediments	DEPTH.sd2R	m	3.00×10^{-2}
34	Volume fraction of water in natural soil	FRACw.s1R	-	2.00×10^{-1}
35	Volume fraction of water in agricultural soil	FRACw.s2R	-	2.00×10^{-1}
36	Volume fraction of water in other soil	FRACw.s3R	-	2.00×10^{-1}
37	Volume fraction of air in natural soil	FRACa.s1R	-	2.00×10^{-1}
38	Volume fraction of air in agricultural soil	FRACa.s2R	-	2.00×10^{-1}
39	Volume fraction of air in other soil	FRACa.s3R	-	2.00×10^{-1}
40	Suspended matter in fresh water	SUSP.wR	kg/m^3	1.50×10^{-2}
41	Suspended matter in sea water	SUSP.wR	kg/m^3	3.00×10^{-3}
42	Mixed depth of fresh water	DEPTH.wR	m	3.00×10^{0}
43	Mixed depth of sea water	DEPTH.wR	m	1.00×10^{1}
44	Net sediment accumulation rate from fresh water	NETsedrate.wR	m/s	8.69×10^{-11}
45	Net sediment accumulation rate from sea water	NETsedrate.w2R	m/s	5.33×10^{-13}
46	Regional temperature+	TEMP.R	K	2.85×10^{2}
47	Mass fraction of organic carbon in f. w. sediments	CORG.sdR	-	5.00×10^{-2}
48	Mass fraction of organic carbon in s. w. sediments	CORG.sdR	-	5.00×10^{-2}
49	Regional wind speed+	WINDspeed.R	m/s	1.50×10^{0}

* All parameters in this table have been assigned SimpleBox 3 default values (den Hollander and van de Meent, 2004; den Hollander et al., 2004), except when noted: + = Values obtained from the report of Struijs and Peijnenburg (Struijs and Peijnenburg, 2002), ++ = Values assigned for removing the local scale, ~ Variable values resulting from degradation, diffusion and advection processes in soil per chemical (den Hollander and van de Meent, 2004; den Hollander et al., 2004). All other default parameters not included in this table can be found in the original documentation of SimpleBox 3 (den Hollander and van de Meent, 2004; den Hollander et al., 2004).

ISBN:978-84-693-4597-9 /DL:T.1010-2010

Table 3-2. Compartments considered in the reference pollution scenario.

Compartments in the regional scale of SimpleBox 3	Compartments in the reference pollution scenario
Air	Air
Fresh water Sea water	Water
Fresh water sediment Sea water sediment	Sediment
Natural soil~ Agricultural soil~ Other soil~	Soil
Natural vegetation Agricultural vegetation	Vegetation

~ The depths of soil compartments in SimpleBox 3 vary as functions of degradation, diffusion and advection processes in soil.

Chemicals of concern. In total, 468 chemicals of concern have been considered, those for which degradability in water, determined by measuring the biological oxygen demand (BOD), agrees in up to 10 % with degradability estimated with total organic carbon (TOC) methods (NITE, 2002). There is a high degree of heterogeneity in the molecular structures of the selected chemicals. These chemicals have been divided randomly in two sets: a first set with 375 work chemicals (Annex C.1) reserved for creating training and test data sets for QPFR and QSFR models; and, a second set, with 93 chemicals (Annex C.2), reserved for the external validation of the models. The only limitation imposed to the validation chemicals is to have each fate properties and molecular descriptors within the ranges that characterize the work chemicals.

The diversity of the selected chemicals is also manifest in their production volumes. Out of the whole set of 468 chemicals, 243 (51.9 %) and 114 (24.4 %) are classified, respectively, as High Production Volume (HPV) chemicals and Low Production Volume (LPV) chemicals, while the remaining 111 chemicals (23.7 %) are not classified neither HPV nor LPV. Currently, the European chemical Substances Information System (ESIS) lists 2782 HPV chemicals and 7829 LPV chemicals (Allanou, 2005). Figure 3-1 shows, through pie charts, the number of working and validation chemicals in the reference pollution scenario that are classified as HPV or LPV chemicals

It must be pointed out that some of the 468 selected chemicals appear in various priority lists: 57 (12.2 %) are listed in the 2007 CERCLA priority list, 44 (9.2 %) are listed in the Online European Risk Assessment Tracking System (ORATS) and 6 (1.3 %) are listed in the 12-chemical priority list of the United Nations Environmental Program (UNEP).

For obtaining molecular descriptors of the structures of the chemicals considered in this study, it was required the availability of both SMILES codes and 3D models of all the molecules involved. Both the SMILES codes and 3D molecular structures are shown for both the 375 work and 93 validation chemicals in, respectively, Annex C.1 and Annex C.2. For facilitating the visualizations of both data sets, the chemicals of each data set are ordered according to their MW.

ISBN:978-84-693-4597-9 /DL:T.1010-2010

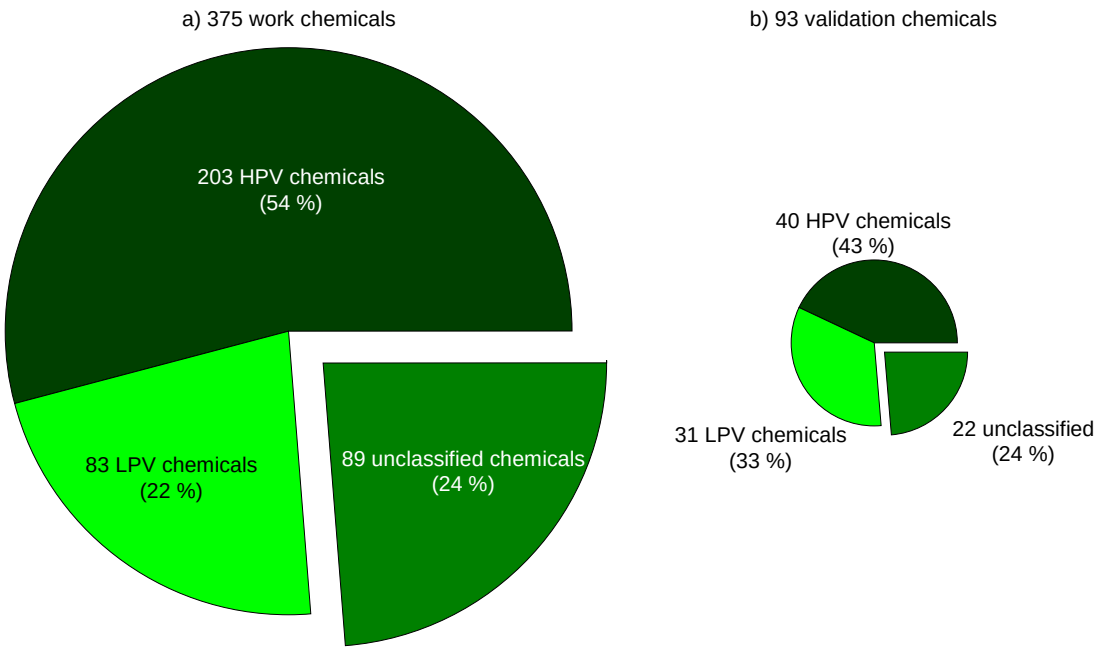


Figure 3-1 Production volume of chemicals used in the reference pollution scenario.
In total, 468 chemicals of concern have been compiled. They have been divided in two sets, a work set of 375 chemicals (a) and a validation set of 93 chemicals (b). The majority of these chemicals have been classified as either high production volume (HPV) chemicals or low production chemicals (LPV) by the European chemical Substances Information System (ESIS) (Allanou, 2005).

3.2 Target and input variables for QPFRs and QSFRs

When building QPFR and QSFR models with basis on supervised learning algorithms, both the input and output variables of a reference MEM referred to training and test chemicals are required for, respectively, building a model and tuning its parameters (Table 2.1). The same variables are also required for the validation chemicals, but only for measuring the performance of a model resulting from the training and test phases with chemicals not used in its development .

As mentioned in Section 2.5.2, the target and input variables of the learning algorithms are, respectively, the outputs of the reference MEM and attributes of the chemicals of concern (physicochemical properties, in the case of QPFRs; or, molecular descriptors, in the case of QSFRs). Note that fate estimations for all the 468 chemicals of concern, in each of the compartments of the reference scenario (air, water, sediments, soil and vegetation), were retrieved from SimpleBox 3 (as explained in Section 3.1). In the following lines of this section, the inputs and targets of the QSPR and QSFR models of this study are described in detail. Which are presented in Annex C.a1, in the CD accompanying this manuscript.

ISBN:978-84-693-4597-9 /DL:T.1010-2010

3.2.1 Target variables of QPFRs & QSFRs: Level III environmental mass ratios

The output of the reference multimedia model, SimpleBox 3, has been expressed in terms of Level III dimensionless mass ratios in air, water, sediments, soil and vegetation (all merged compartments listed in Table 3-1). The mass ratio of a pollutant in each compartment is calculated as follows:

$$w_{n,g} = \frac{C_{n,g} V_g}{E_n \Delta t} \tag{3-1}$$

where $C_{n,g}$ (g/m^3) is the steady state concentration of a pollutant n as estimated by a multimedia model for a compartment g of volume V_g (m^3), during a period of time Δt = 1yr for a total emission rate in the system of magnitude E_n (ton/yr). Note that compartmental concentrations estimated by SimpleBox 3 are directly proportional to the emission rate in the system (den Hollander and van de Meent, 2004; den Hollander et al., 2004). This is the reason why unitary emissions have been considered in all the simulation experiments of this reference scenario (Section 3.1).

The targets of QPFRs and QSFRs, mass ratios estimated according to Equation 3-1 for the 468 chemicals of this study after evaluating their physicochemical properties in SimpleBox 3, are presented in Annex C.a1 for emissions in water and air. Table 3-3 and Table 3-4 list the value ranges of such mass ratios for, respectively, emissions in water and air. These value ranges are delimited by the minimum and maximum values resulting for the work and validation data sets.

Table 3-3. Value ranges of dimensionless level III mass ratios estimated by SimpleBox 3 for the work and validation chemicals, considering emissions in water.

Mass ratio	Symbol	Units	Work data set		Validation data set	
			min	max	min	max
Dimensionless mass ratio in air	w_{air}	-	2.67×10^{-25}	1.11×10^{-2}	5.00×10^{-18}	5.09×10^{-3}
Dimensionless mass ratio in water	w_{water}	-	4.85×10^{-9}	6.28×10^{-1}	1.01×10^{-6}	6.28×10^{-1}
Dimensionless mass ratio in sediments	w_{sed}	-	3.36×10^{-11}	7.96×10^{-3}	5.19×10^{-9}	7.57×10^{-3}
Dimensionless mass ratio in soil	w_{soil}	-	3.01×10^{-22}	4.84×10^{-2}	1.10×10^{-14}	4.74×10^{-2}
Dimensionless mass ratio in vegetation	w_{veg}	-	8.05×10^{-25}	1.37×10^{-2}	1.40×10^{-12}	9.57×10^{-3}

Table 3-4. Value ranges of dimensionless level III mass ratios estimated by SimpleBox 3 for the work and validation chemicals, considering emissions in air.

Mass ratio	Symbol	Units	Work data set		Validation data set	
			min	max	min	max
Dimensionless mass ratio in air	w_{air}	-	7.63×10^{-12}	1.22×10^{-2}	6.47×10^{-10}	5.22×10^{-3}
Dimensionless mass ratio in water	w_{water}	-	1.56×10^{-9}	2.37×10^{-1}	9.23×10^{-7}	2.34×10^{-1}
Dimensionless mass ratio in sediments	w_{sed}	-	3.70×10^{-12}	1.46×10^{-3}	1.07×10^{-9}	1.44×10^{-3}
Dimensionless mass ratio in soil	w_{soil}	-	3.31×10^{-12}	1.29×10^{0}	4.53×10^{-10}	1.30×10^{0}
Dimensionless mass ratio in vegetation	w_{veg}	-	1.08×10^{-9}	7.95×10^{-2}	2.30×10^{-7}	5.72×10^{-2}

ISBN:978-84-693-4597-9 /DL:T.1010-2010

3.2.2 Input variables of QPFRs: Physicochemical properties

Data of physicochemical properties, at 25 °C, have been collected for the chemicals of the reference pollution scenario, giving priority to experimental values whenever possible; otherwise, estimations have been used instead. Experimental and estimated values for MW, T_m, S_w, P_v, K_{ow}, H and degradation hydroxyl rate constants ($k_{OH\cdot}$, $cm^3/(mol\cdot s)$) have been retrieved from PHYSPROP (SRC), while experimental results of ready biodegradability tests in water (MITI-I) have been retrieved from the Japanese National Institute of Technology and Evaluation (NITE, 2002). Some of these data have been processed further for their use in SimpleBox 3, as explained later in this section, the final collection of properties is presented in Annex C.a1. Table 3-5 lists value ranges of the physicochemical properties compiled for the chemicals in the work and validation data sets.

Partitioning coefficients. Dimensionless K_{aw} values were determined directly, from either experimental or estimated H values, using the equation 1-3. Dimensionless K_{sw} partition coefficients were estimated, from either experimental or estimated K_{ow} values, using the correlation included in SimpleBox 3:

$$K_{sw} = (1.26K_{ow}^{0.81})\frac{(CORG\ \rho_{solid})}{1000} \qquad (3\text{-}2)$$

for an average organic carbon content of 2% and solid soil density of 2.5 kg/L (den Hollander and van de Meent, 2004; den Hollander et al., 2004).

Degradation rates. Degradation rates constants in air (k_{air}, 1/s) have been directly calculated from degradation rates of chemicals exposed to hydroxyl radicals ($k_{OH\cdot}$), assuming pseudo first order reactions in air. This reaction occurs under a second order

Table 3-5. Value ranges of physicochemical properties entered in SimpleBox 3 for the work and validation chemicals.

Physicochemical properties*	Symbol	Units	Work data set		Validation data set	
			min	max	min	max
Molecular weight[P]	MW	g/mol	4.41×10^{1}	9.59×10^{2}	6.01×10^{1}	4.31×10^{2}
Melting point[P]	T_m	ºC	-1.60×10^{2}	3.90×10^{2}	-9.50×10^{1}	3.12×10^{2}
Solubility in water[P]	S_w	mg/L	9.48×10^{-4}	4.07×10^{10}	1.19×10^{-3}	1.72×10^{8}
Vapor pressure[P]	P_v	Pa	2.24×10^{-21}	2.28×10^{6}	1.47×10^{-13}	1.45×10^{0}
Octanol-water partition coefficient[P]	K_{ow}	-	1.21×10^{-19}	9.67×10^{5}	5.93×10^{-11}	4.28×10^{3}
Air-water part. coefficient (Henry's law)[P]	K_{aw}	-	1.25×10^{-11}	1.00×10^{6}	1.44×10^{-6}	1.00×10^{6}
Solid-water partition coefficient[S1]	K_{sw}	-	5.62×10^{-3}	3.80×10^{14}	7.41×10^{-3}	4.47×10^{11}
Degradation rate in air[P]	k_{air}	1/s	5.96×10^{-12}	3.59×10^{-4}	1.42×10^{-7}	3.20×10^{-4}
Degradation rate in water[M]	k_{water}	1/s	4.15×10^{-9}	3.81×10^{-6}	4.15×10^{-9}	3.23×10^{-6}
Degradation rate in sediments[CF]	k_{sed}	1/s	1.19×10^{-9}	1.09×10^{-6}	1.19×10^{-9}	9.24×10^{-7}
Degradation rate in soil[CF]	k_{soil}	1/s	4.15×10^{-9}	3.81×10^{-6}	4.15×10^{-9}	3.23×10^{-6}
Diffusion coefficient in air[S2]	D_{air}	m^2/s	1.11×10^{-7}	5.20×10^{-7}	1.66×10^{-7}	4.45×10^{-7}
Diffusion coefficient in water[S2]	D_{water}	m^2/s	1.16×10^{-11}	5.39×10^{-11}	1.72×10^{-11}	4.61×10^{-11}

* Some properties have been retrieved or converted from: [P] = PHYSPROP (SRC); [M] = MITI-I biodegradability tests (NITE, 2002). While, other properties have been estimated from: [S1] = K_{ow} based correlations (den Hollander and van de Meent, 2004; den Hollander et al., 2004); [S2] = MW based correlations (den Hollander and van de Meent, 2004; den Hollander et al., 2004); or, [CF] = reported conversion factors (Aronson and Howard, 1999).

ISBN:978-84-693-4597-9 /DL:T.1010-2010

reaction scheme:

$$OH\cdot + Chemical \longrightarrow Products \tag{3-3}$$

with the following degradation rate:

$$r_{n,air} = k_{OH\cdot} C_{OH\cdot} C_{n,air} \tag{3-4}$$

where r_{air} ($g/m^3 \cdot s$) is the degradation rate in air, $k_{OH\cdot}$ is the second-order reaction constant ($m^3/g.s$) (SRC, 2008) and $C_{OH\cdot}$ (g/m^3) is the concentration of hydroxyl radicals in air. Considering a global average concentration of hydroxyl radicals of $C_{OH\cdot} = 2.66x10^{-11}$ g/m^3 (Prinn et al., 2001), pseudo first-order degradation rates have the form:

$$r_{air} = k_{air} C \qquad r_{n,air} = k_{n,air} C_{n,air} \tag{3-5}$$

where the pseudo first degradation rate constant is:

$$k_{air} = k_{OH\cdot} C_{OH\cdot} \tag{3-6}$$

Degradation rates in water (k_{water}, 1/s) have been calculated from MITI-I biodegradability tests (NITE, 2002). These tests have been originally reported to measure the degradability (deg%) of a substance, previously incubated in presence of activated sludge, by either direct and indirect methods. The direct methods used in the MITI-I tests included total organic carbon (TOC), high performance liquid chromatography (HPLC) and gas chromatography (GC). Indirect tests measured the biological oxygen demand (BOD) of the samples. The degradability has been determined in the direct methods as follows:

$$deg\% = \left(\frac{S_b - S_a}{S_b} \right) 100 \tag{3-7}$$

where S_b (mg) is the residual mass of the test substance at the end of the test and S_a (mg) the mass of substance in a blank test with water only. For indirect measurement methods the degradability has been measured with the following equation:

$$deg\% = \left(\frac{BOD - B}{TOD} \right) 100 \tag{3-8}$$

where BOD (mg) is the biochemical oxygen demand of the test substance, B (mg) is the oxygen consumption in the basic culture medium inoculated with the activated sludge and TOD (mg) is the theoretical oxygen demand required for complete oxidation of the test substance.

Correlations for degradability values determined from BOD and TOC have been satisfactory; but, this has not been the case of correlations of BOD and TOC with chromatographic techniques, which have shown to be worse (Sedykh and Klopman, 2007). In this study, k_{water} has been estimated from the percentage of degradation

ISBN:978-84-693-4597-9 /DL:T.1010-2010

(deg%) and the corresponding period of time (t, weeks), determined for BOD tests in agreement to TOC methods in up to 10%, as follows (using the equation 1-8):

$$k_{water} = \left(\frac{-1}{t}\right)\ln\left(1-\frac{deg\%}{100}\right)\left(\frac{1}{604800}\right) \tag{3-9}$$

Please note that degradability values reported to be higher than 99 % or lower than 1 % have been set to be equal to, respectively, 99 % or 1%. Due to error measurements, some degradability values have been originally reported to be higher than 100% or negative (NITE, 2002).

Since data for degradation in sediments and soil are usually scarce, conversion factors have been used to estimate degradation rates in these two media. It has been reported that degradation half lives in water are similar to those in soil, while degradation rates in soil tend to be 3 to 4 times faster that degradation rates in flooded soil (Aronson and Howard, 1999). With such information, it is assumed that degradation rates in soil are equal to those in water and that degradation rates in sediments are 3.5 times slower than those in soil (considering that sediments behave as flooded soil):

$$k_{soil} = k_{water} \tag{3-10}$$

and

$$k_{sed} = \left(\frac{2}{7}\right)k_{soil} \tag{3-11}$$

Diffusion coefficients. Diffusion coefficients in air (D_{air}, m^2/s) and water (D_{water}, m^2/s) have been estimated from MW values according to the following correlations (den Hollander and van de Meent, 2004; den Hollander et al., 2004):

$$D_{air} = 2.57 \cdot 10^5 \sqrt{\frac{18}{1000MW}} \tag{3-12}$$

and

$$D_{water} = 2 \cdot 10^9 \sqrt{\frac{32}{1000MW}} \tag{3-13}$$

where the molecular weight unit is g/mol.

3.2.3 Input variables of QSFRs: Molecular descriptors

Molecular descriptors were compiled, from either SMILES codes or 3D molecular representations of the chemicals considered in this work (Annexes C.1 and C.2), using the CACHE software (Fujitsu, 2004). Such descriptors are presented in Annex C.a1, like the physicochemical properties and mass ratios discussed below.

ISBN:978-84-693-4597-9 /DL:T.1010-2010

Table 3-6. Value ranges of theoretical molecular properties of the work and validation chemicals.

Descriptor*	Symbol	Units	Work data set		Validation data set	
			min	max	min	max
Heat of Formation	ΔH_f	kcal/mole	-1341.59	145.89	-374.49	45.22
Molar Refractivity	MR	m^3/mol	11.31	169.07	17.07	123.36
Polarizability	PO	$Å^3$	4.43	68.96	7.00	49.54
Total hybridization dipole moment	μ_{hyb}	debye	0.00	2.60	0.00	2.45
Total point charge dipole moment	μ_{pc}	debye	0.00	10.14	0.01	8.73
Total sum dipole moment	μ	debye	0.00	11.24	0.00	9.67
Area	Area	$Å^2$	77.05	622.79	106.50	567.43
Volume	Vol	$Å^3$	59.03	709.73	89.13	592.79
Number of filled levels	NFL	-	7.00	121.00	13.00	82.00
HOMO energy	HOMO	eV	-12.97	-7.97	-12.16	-8.39
LUMO energy	LUMO	eV	-3.15	3.48	-2.52	2.97
Ionization potencial	IP	eV	7.97	12.97	8.39	12.16
Electron affinity	EA	eV	-3.48	3.15	-2.97	2.52
Connectivity index (order 0, standard)	${}^0\chi$	-	2.00	33.58	3.41	21.42
Connectivity index (order 1, standard)	${}^1\chi$	-	1.00	21.12	1.91	13.41
Connectivity index (order 2, standard)	${}^2\chi$	-	0.00	22.10	1.00	11.48
Valence connectivity index (order 0, standard)	${}^0\chi^v$	-	1.99	26.05	2.57	19.13
Valence connectivity index (order 1, standard)	${}^1\chi^v$	-	0.81	14.41	1.32	13.02
Valence connectivity index (order 2, standard)	${}^2\chi^v$	-	0.00	13.07	0.58	9.32
Shape index (kappa alpha, order 1)	${}^1\kappa$	-	2.21	38.07	3.77	26.96
Shape index (kappa alpha, order 2)	${}^2\kappa$	-	0.46	28.00	1.27	25.96
Shape index (kappa alpha, order 3)	${}^3\kappa$	-	0.00	28.00	0.77	25.96

* All descriptors were estimated semi-empirically with the CACHE software (Fujitsu, 2004).

A first group of descriptors, derived from 3D molecular representations, included 22 diverse theoretical molecular attributes: heat of formation (ΔH_f), molar refractivity (MR), polarizability (PO), total hybridization dipole moment (μ_{hyb}), total point charge dipole moment (μ_{pc}), total sum dipole moment (μ), area (Area), volume (Vol), number of filled levels (NFL), highest occupied molecular orbital energy (HOMO), lowest occupied molecular orbital energy (LUMO), ionization potential (IP), electron affinity (EA), connectivity indexes (${}^0\chi$, ${}^1\chi$, ${}^2\chi$), valence connectivity indexes (${}^0\chi^v$, ${}^1\chi^v$, ${}^2\chi^v$) and kappa alpha shape indexes (${}^1\kappa$, ${}^2\kappa$, ${}^3\kappa$). ΔH_f, EA, IP, HOMO, LUMO, μ_{hyb}, μ_{pc}, and μ were calculated at the minimum energy geometry determined by optimization with MOPAC and parameters from the Parameterized Model 3 (PM3) (Stewart, 1989). MR was calculated using the atom typing scheme of Ghose et al. (1988). The indexes ${}^0\chi$, ${}^1\chi$, ${}^2\chi$, ${}^0\chi^v$, ${}^1\chi^v$ and ${}^2\chi^v$ were determined from the atoms and bonds in chemical samples at the time of evaluation (Kier and Hall, 1986), while indexes ${}^1\kappa$, ${}^2\kappa$, ${}^3\kappa$ were derived from counts of one-bond, two-bond and three-bond fragments, each count being made relative to fragment counts in reference structures which possess a maximum and minimum value for that number of atoms (Hall and Kier, 1992). Table 3-6 lists the value ranges of these 22 theoretical descriptors for the chemicals in the work and validation data sets.

Another selection of descriptors, derived from SMILES codes, included 43 counts of molecular constituents: atoms, bonds, functional groups and rings. The calculation of these descriptors calculation is very simple, simple sums of the constituents of every molecular model. Table 3-7 lists the value ranges of these 43 simple descriptors for the chemicals in the work and validation data sets.

Notice that different types of information are associated to the sets of descriptors listed in Tables 3-6 and 3-7. The descriptors in the former set (Table 3-6) provide

ISBN:978-84-693-4597-9 /DL:T.1010-2010

Table 3-7. Value ranges of molecular constituent counts of the work and validation chemicals.

Descriptor*	Symbol	Work data set		Validation data set	
		min	max	min	max
Atom Count (all atoms)	AC_{all}	5	89	10	81
Atom Count (bromine)	$AC_{bromine}$	0	10	0	3
Atom Count (carbon)	AC_{carbon}	1	32	2	26
Atom Count (chlorine)	$AC_{chlorine}$	0	8	0	6
Atom Count (fluorine)	$AC_{fluorine}$	0	27	0	3
Atom Count (hydrogen)	$AC_{hydrogen}$	0	60	3	54
Atom Count (iodine)	AC_{iodine}	0	0	0	0
Atom Count (nitrogen)	$AC_{nitrogen}$	0	6	0	3
Atom Count (oxygen)	AC_{oxygen}	0	8	0	8
Atom Count (phosphorus)	$AC_{phosphorus}$	0	1	0	1
Atom Count (silicon)	$AC_{silicon}$	0	0	0	0
Atom Count (sulphur)	$AC_{sulphur}$	0	4	0	2
Bond Count (all bonds)	BC_{all}	4	88	10	80
Bond Count (single bonds)	BC_{single}	4	88	9	80
Bond Count (double bonds)	BC_{double}	0	18	0	8
Bond Count (triple bonds)	BC_{triple}	0	2	0	2
Group Count (aldehyde)	$GC_{aldehyde}$	0	1	0	1
Group Count (amide)	GC_{amide}	0	2	0	2
Group Count (amine)	GC_{amine}	0	2	0	2
Group Count (sec-amine)	$GC_{sec\text{-}amine}$	0	2	0	2
Group Count (carbonyl)	$GC_{carbonyl}$	0	2	0	2
Group Count (carboxyl)	$GC_{carboxyl}$	0	2	0	2
Group Count (carboxylate)	$GC_{carboxylate}$	0	0	0	0
Group Count (cyano)	GC_{cyano}	0	2	0	2
Group Count (ether)	GC_{ether}	0	4	0	3
Group Count (hydroxyl)	$GC_{hydroxyl}$	0	4	0	3
Group Count (methyl)	GC_{methyl}	0	9	0	9
Group Count (methylene)	$GC_{methylene}$	0	3	0	0
Group Count (nitro)	GC_{nitro}	0	3	0	1
Group Count (nitroso)	$GC_{nitroso}$	0	1	0	0
Group Count (sulfide)	$GC_{sulfide}$	0	4	0	2
Group Count (sulfone)	$GC_{sulfone}$	0	1	0	1
Group Count (sulfoxide)	$GC_{sulfoxide}$	0	0	0	0
Group Count (thiol)	GC_{thiol}	0	1	0	1
Ring Count (all rings)	RC_{all}	0	12	0	2
Ring Count (aromatic rings)	$RC_{aromatic}$	0	4	0	2
Ring Count (small rings)	RC_{small}	0	7	0	0
Ring Count (5 membered)	$RC_{5\text{-}m}$	0	4	0	1
Ring Count (aromatic 5 membered)	$RC_{a\text{-}5\text{-}m}$	0	2	0	0
Ring Count (6 membered)	$RC_{6\text{-}m}$	0	4	0	2
Ring Count (aromatic 6 membered)	$RC_{a\text{-}6\text{-}m}$	0	4	0	2
Ring Count (7-12 membered)	$RC_{7\text{-}12\text{-}m}$	0	2	0	1
Ring Count (aromatic 7-12 membered)	$RC_{a\text{-}7\text{-}12\text{-}m}$	0	0	0	0

* All descriptors have been calculated with the CACHE software (Fujitsu, 2004).

information about the overall behavior of molecules, while the descriptors in the latter set (Table 3-7) simply provide information about the number of constituents of the molecules. Depending on a given problem and available data, different types and numbers of descriptors may be more appropriate than others. Theoretical descriptors, like those listed in Table 3-6, have been used for predicting some physicochemical properties (Devillers, 2003; Raymond et al., 2001; Taskinen and Yliruusi, 2003), while descriptors identifying molecular fragments or constituents, similar to those in Table 3-7, have been widely recommended for predicting both physicochemical properties (Boethling et al., 2004) and degradation data (Raymond et al., 2001).

ISBN:978-84-693-4597-9 /DL:T.1010-2010

References

Allanou R. European chemical Substances Information System (ESIS), Version 4.20, 2005. http://www.chem.unep.ch/irptc/sids/oecdsids/

Aronson D, Boethling R, Howard P, Stiteler W. Estimating biodegradation half-lives for use in chemical screening. Chemosphere 2006; 63: 1953.

Aronson D, Howard PH. Evaluating Potential POP/PBT Compounds for Environmental Persistence. SRC, North Syracuse, United states, 1999.

Boethling RS, Howard PH, Meylan WM. Finding and estimating chemical property data for environmental assessment. Environmental Toxicology and Chemistry 2004; 23: 2290-2308.

Brandes LJ, den Hollander H, van de Meent D. SimpleBox 2.0: a nested multimedia fate model for evaluating the environmental fate of chemicals. RIVM, Bilthoven, The Netherlands, 1996, pp. 156.

den Hollander HA, van de Meent D. Appendix to SimpleBox 3.0: A multimedia mass balance model for evaluating the environmental fate of chemicals. RIVM, 2004.

den Hollander HA, van Eijkeren JCH, van de Meent D. SimpleBox 3.0. RIVM, Bilthoven, The Netherlands, 2004.

Devillers J. A decade of research in environmental QSAR. SAR and QSAR in Environmental Research 2003; 14: 1-6.

Fujitsu BGo. CAChe Software. BioSciences Group, Fujitsu Computer Systems, Beaverton. 2004.

Ghose AK, Pritchett A, Crippen GM. Atomic physicochemical parameters for three dimensional structure directed quantitative structure-activity relationships III: Modeling hydrophobic interactions. Journal of Computational Chemistry 1988; 9: 80-90.

Hall LH, Kier LB. Reviews in Computational Chemistry, Ch. 9: ed. K.B. Lipkowitz and D.B. Boyd, 1992.

JETOC. Biodegradation and Bioaccumulation Data of Existing Chemicals Based on the Chemical Substances Control Law (CSCL Japan). Japan Chemical Industry Ecology-Toxicology & Information Center (JETOC), Tokyo, 1992.

Kier LB, Hall LH. Molecular Connectivity in Structure-Activity Analysis. New York: John Wiley & Sons Inc., 1986.

Martínez I, Grifoll J, Giralt F, Rallo R, Espinosa G. Report on the feasibility of predicting multimedia chemical partitioning with artificial neural network models by using functional group counts as input information. Universitat Rovira i Virgili, Tarragona, Spain, 2008a.

Martínez I, Grifoll J, Rallo R. Cognitive neural network-based intelligent system to identify the most important variables for the differences found in partitioning behaviour, transport pathways and exposure routes between chemicals. Universitat Rovira i Virgili, Tarragona, Spain, 2006.

Martínez I, Grifoll J, Rallo R, Giralt F. Report on the most suitable artificial neural network architectures and molecular descriptors to estimate environmental multimedia behavior, including a sensitivity analysis of the effect of compartment sizes on multimedia concentrations. Universitat Rovira i Virgili, Tarragona, Spain, 2007.

Martínez I, Grifoll J, Rallo R, Giralt F. Report on the most suitable deterministic and probabilistic algorithms to pre-classify chemicals into families according to their partitioning with the aim of better predicting multimedia concentrations on artificial neural networks for each chemical family. Universitat Rovira i Virgili, Tarragona, Spain, 2008b.

NITE. Chemical Risk Information Platform (CHRIP). National Institute of Technology and Evaluation. 2002. Accesed on 2006.

Raymond JW, Rogers TN, Shonnard DR, Kline AA. A review of structure-based biodegradation estimation methods. Journal of Hazardous Materials 2001; 84: 189.

Sedykh A, Klopman G. Data analysis and alternative modelling of MITI-I aerobic biodegradation. SAR and QSAR

ISBN:978-84-693-4597-9 /DL:T.1010-2010

in Environmental Research 2007; 18: 693 - 709.

SRC. Interactive PhysProp Database Demo. Syracuse Research Corporation.

Stewart JJP. Optimization of parameters for semiempirical methods II. Applications. Journal of Computational Chemistry 1989; 10: 221-264.

Struijs J, Peijnenburg WJGM. Predictions by the multimedia environmental fate model SimpleBox compared to field data: Intermedia concentration ratios of two phthalate esters. RIVM, Bilthoven, 2002, pp. 62.

Taskinen J, Yliruusi J. Prediction of physicochemical properties based on neural network modelling. Advanced Drug Delivery Reviews 2003; 55: 1163.

van de Meent D. SIMPLEBOX: a generic multimedia fate evaluation model. RIVM, Bilthoven, The Netherlands, 1993.

ISBN:978-84-693-4597-9 /DL:T.1010-2010

ISBN:978-84-693-4597-9 /DL:T.1010-2010

Chapter 4

Quantitative structure-fate relationships

Multimedia environmental models perform reasonable estimations of the fate of chemicals, if there are known physicochemical properties of the chemicals to assess, emission rates and site-specific parameters. The assessment of a large number of chemicals gets complicated when their physicochemical properties are unknown, making necessary the use of a large pool of property estimation methods that, depending on the assumptions and techniques involved, provide values that may differ considerably from experimental values. Here, it is discussed the use of learning algorithms to predict, from molecular information, the fate of chemicals for which key physicochemical properties are unavailable.

ISBN:978-84-693-4597-9 /DL:T.1010-2010

4.1 Screening chemicals in level III conditions

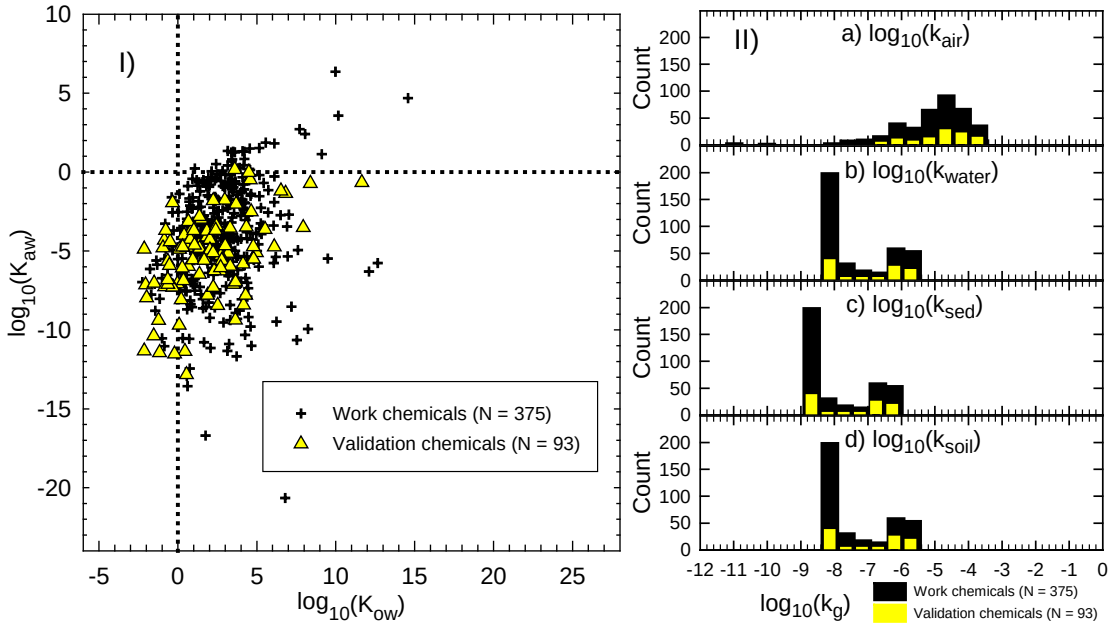
The environmental screening of chemicals can be roughly estimated by means of their partition coefficients, using 2D graphs per pair of independent coefficients for suggesting the final distributions in three contiguous media with mass balances at level II conditions (Gouin et al., 2000). The advantage of such method is that chemicals with extreme partition coefficients may not require some compartmental degradation data, especially in the compartments in which their presence is estimated to be minimal. However, such approach is no longer valid for level III conditions, in which the system is considered to have non-equilibrium and steady state conditions.

Figure 4-1 shows a screening of partition coefficients, degradation data and fate estimations for emissions in water (Annex C.a1) referred to the chemicals (Annexes C.1 and C.2) in the reference pollution scenario already described in Chapter 3. They are displayed in a fashion somewhat similar to that proposed by Gouin et al., but level II mass balances are not applied because the reference scenario is not in equilibrium and fate estimations based solely on partitioning coefficients would markedly differ from those estimated by the reference level III MEM. However, we can still have a preliminary view of the functionality between physicochemical properties and fate estimations (Equation 2-78) in the reference scenario by inspecting the distribution of chemicals in each of the subplots of Figure 4-1. The value ranges of the partition coefficients of Figure 4-1I and degradation rates of Figure 4-1II produce, when used simultaneously in the reference MEM, the value ranges of mass ratios shown in Figure 4-1III. The spaces occupied by the work chemicals in each of the subplots of Figure 4-1 give an insight of the DOA of the work data set, referred to the functionality of the MEM. The validation chemicals, selected to have properties within the ranges reported for the work chemicals, are clearly within the DOA of the latter.

The visual screening of available chemicals helps to identify the existence of regions with low density of examples, in which few chemicals may behave as outliers with respect to the rest. Inspecting Figure 4-1I, it can be noticed that only few of the chemicals considered in the reference scenario have partition coefficients markedly different than the majority: some are highly hydrophobic (very high K_{ow} values), some have a strong tendency to volatilize (very high K_{aw} values) and others are simply non-volatile (very low K_{aw} values). The degradation data of the reference scenario, represented as histograms in Figure 4-1II, indicate roughly that most of the chemicals of the scenario tend to undergo degradation faster in air than in water, sediments or soil. The level III mass ratios resulting for emissions in water (Annex C.a1) in the reference scenario (Chapter 3), represented in histograms in Figure 4-1III, indicate that the highest mass ratios occur in the water compartment, where emissions take place. Of course, though partitioning properties and degradation rates contain relevant information about the tendency of chemicals to behave in the environment, it is clear that the final distribution of chemicals is affected by the attributes of the geographic scenario selected for the assessment.

Graphically, simple plots of solely partitioning or degradation data can not offer a

ISBN:978-84-693-4597-9 /DL:T.1010-2010



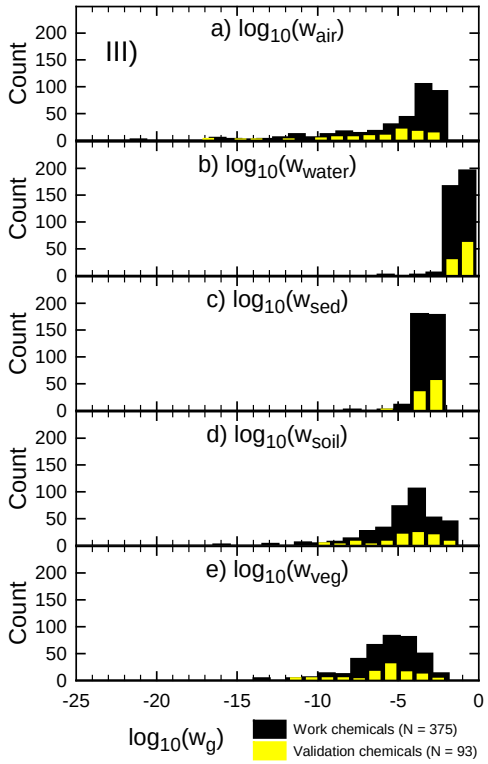


Figure 4-1. Simple graphical screening of the key inputs and outputs of a MEM, for emissions in water.
This figure shows for 375 work and 93 validation chemicals: I) a 2D plot of key partitioning coefficients (K_{ow} and K_{aw}); II) histograms of key degradation rate constants (k_{air}, k_{water}, k_{sed} and k_{soil}); and, III) histograms of level III fate estimations (w_{air}, w_{water}, w_{sed}, w_{soil} and w_{veg}) resulting from a reference MEM for emissions of 1 ton/yr in water at The Netherlands.

clear view of the environmental distribution of chemicals of concern in level III conditions, a more realistic assessment implies the use of a level III MEM. This is a multivariate problem that can be tackled with unsupervised learning algorithms for data visualization.

A level III graphic screening of the environmental distribution of pollutants can be performed, in a somewhat similar manner to the level II method proposed by Gouin et al., by processing all the inputs and outputs of a reference MEM for a population of

ISBN:978-84-693-4597-9 /DL:T.1010-2010

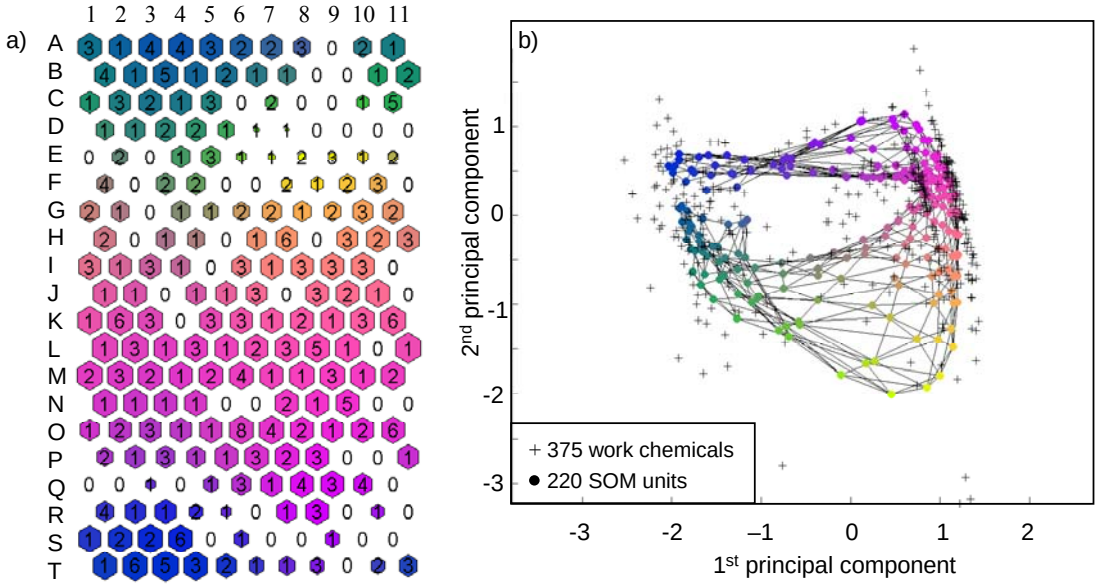


Figure 4-2. Fitting of work chemicals, characterized by all the inputs and outputs of a MEM for emissions in water, with a SOM.
This figure shows: a) the number of work chemicals clustered in each of the units of a SOM; and, b) a visualization of the work chemicals and SOM units in simplified 2D visualization of the original 18D multivariate space, characterized by the first two principal components of the data with a cumulative variance of 79 %.

known chemicals with a multivariate learning algorithm like the SOM (Section 2.4). This helps to summarize in one picture the effects that different combinations of properties have on the final distribution of chemicals.

Figure 4-2 shows a SOM (Annex D.a1) adjusting the 375 work chemicals (Annex C.1) of the reference scenario in a 18D multivariate space conformed by the normalized logarithms of all the inputs (MW, T_m, S_w, P_v, K_{ow}, K_{aw}, K_{sw}, k_{air}, k_{water}, k_{sed}, k_{soil}, D_{air}, D_{water}) and outputs (w_{air}, w_{water}, w_{sed}, w_{soil}, w_{veg}) of the reference MEM. The SOM, with 20x11 units, minimizes the Euclidean distances between the data points and the SOM units in the multivariate space.

Figure 4-2a shows the number of chemicals clustered in every SOM unit. Figure 4-2b shows a visualization of the 375 work chemicals and the SOM units (identifiable in Figure 4-2a through color codes) in a 2D space, characterized by the first two principal components of the data (Equations 4-1 and 4-2): PC1 and PC2, respectively, the 1st and 2nd principal components of the work data set. The cumulative variance of these two components is 79 %, indicating that this pair of variables inherited a great deal of relevant information from the original input and output variables of the MEM, providing a reasonable 2D approximation of the original 18D space.

The PCA analysis of the 375 work chemicals offers an orthogonal visualization of their attributes in low dimensions, in this case, the first and second principal components of their properties and environmental mass ratios. Meanwhile, the SOM analysis of the same data offers both clustering and visualization of the data.

ISBN:978-84-693-4597-9 /DL:T.1010-2010

$$
\begin{aligned}
PC1 = \; & 0.092\, N_{[-1,1]}(\log_{10}(MW)) - 0.377\, N_{[-1,1]}(\log_{10}(T_m)) \\
& +0.096\, N_{[-1,1]}(\log_{10}(K_{sw})) + 0.190\, N_{[-1,1]}(\log_{10}(K_{aw})) \\
& -0.037\, N_{[-1,1]}(\log_{10}(P_v)) - 0.298\, N_{[-1,1]}(\log_{10}(S_w)) \\
& -0.194\, N_{[-1,1]}(\log_{10}(K_{ow})) - 0.094\, N_{[-1,1]}(\log_{10}(k_{air})) \\
& -0.034\, N_{[-1,1]}(\log_{10}(k_{water})) + 0.006\, N_{[-1,1]}(\log_{10}(k_{sed})) \\
& +0.019\, N_{[-1,1]}(\log_{10}(k_{soil})) - 0.001\, N_{[-1,1]}(\log_{10}(D_{air})) \\
& -0.014\, N_{[-1,1]}(\log_{10}(D_{water})) + 0.815\, N_{[-1,1]}(\log_{10}(w_{air})) \\
& -0.040\, N_{[-1,1]}(\log_{10}(w_{water})) - 0.021\, N_{[-1,1]}(\log_{10}(w_{sed})) \\
& -0.000\, N_{[-1,1]}(\log_{10}(w_{soil})) + 0.000\, N_{[-1,1]}(\log_{10}(w_{veg}))
\end{aligned}
\tag{4-1}
$$

$$
\begin{aligned}
PC2 = \; & 0.069\, N_{[-1,1]}(\log_{10}(MW)) - 0.288\, N_{[-1,1]}(\log_{10}(T_m)) \\
& -0.321\, N_{[-1,1]}(\log_{10}(K_{sw})) + 0.346\, N_{[-1,1]}(\log_{10}(K_{aw})) \\
& +0.245\, N_{[-1,1]}(\log_{10}(P_v)) + 0.704\, N_{[-1,1]}(\log_{10}(S_w)) \\
& -0.273\, N_{[-1,1]}(\log_{10}(K_{ow})) - 0.200\, N_{[-1,1]}(\log_{10}(k_{air})) \\
& +0.027\, N_{[-1,1]}(\log_{10}(k_{water})) - 0.130\, N_{[-1,1]}(\log_{10}(k_{sed})) \\
& +0.026\, N_{[-1,1]}(\log_{10}(k_{soil})) + 0.034\, N_{[-1,1]}(\log_{10}(D_{air})) \\
& -0.007\, N_{[-1,1]}(\log_{10}(D_{water})) - 0.002\, N_{[-1,1]}(\log_{10}(w_{air})) \\
& -0.000\, N_{[-1,1]}(\log_{10}(w_{water})) + 0.000\, N_{[-1,1]}(\log_{10}(w_{sed})) \\
& +0.000\, N_{[-1,1]}(\log_{10}(w_{soil})) - 0.000\, N_{[-1,1]}(\log_{10}(w_{veg}))
\end{aligned}
\tag{4-2}
$$

Compared to Figure 4-1, Figure 4-3 provides an enhanced visualization of each of the 18 dimensions that constitute the chemical space of the 375 work chemicals emitted in water (Annex A.c1) in the reference scenario. Every dimension, is represented as a SOM plane with colors that indicate the magnitude of each SOM unit (previously identified as colored hexagons in Figure 4-1a and colored circles in Figure 4-2a). Figure 4-3 shows an enhanced graphical representation of the simple graphs of Figure 4-1 for the 375 work chemicals in the reference pollution scenario, providing a more understandable screening of the inputs and outputs of the used MEM. It can be verified the dependency of some properties to those that generated them, as the similarity of their respective component planes indicates. Both D_{air} and D_{water} are inversely proportional to MW, K_{sw} is proportional to K_{ow}, and both k_{sed} and k_{soil} are proportional to k_{water}. In an analogous manner, relationships between the mass ratios and the independent properties can be screened through the SOM planes, but giving special attention to specific zones of the latter.

Observing the SOM-based visualization in Figure 4-3 and remembering in which compartment the chemicals are emitted (in water, in this case) it is possible to analyze how they move to other compartments with respect to their properties. The planes of w_{water} and w_{sed} are inversely proportional to K_{ow}, and, because of the magnitude of different colored areas, it can be noticed that the majority of the work chemicals are hydrophilic while few ones are highly hydrophobic (those chemicals located in the center-right zone of the upper part of the SOM have very high K_{ow} values). The

ISBN:978-84-693-4597-9 /DL:T.1010-2010

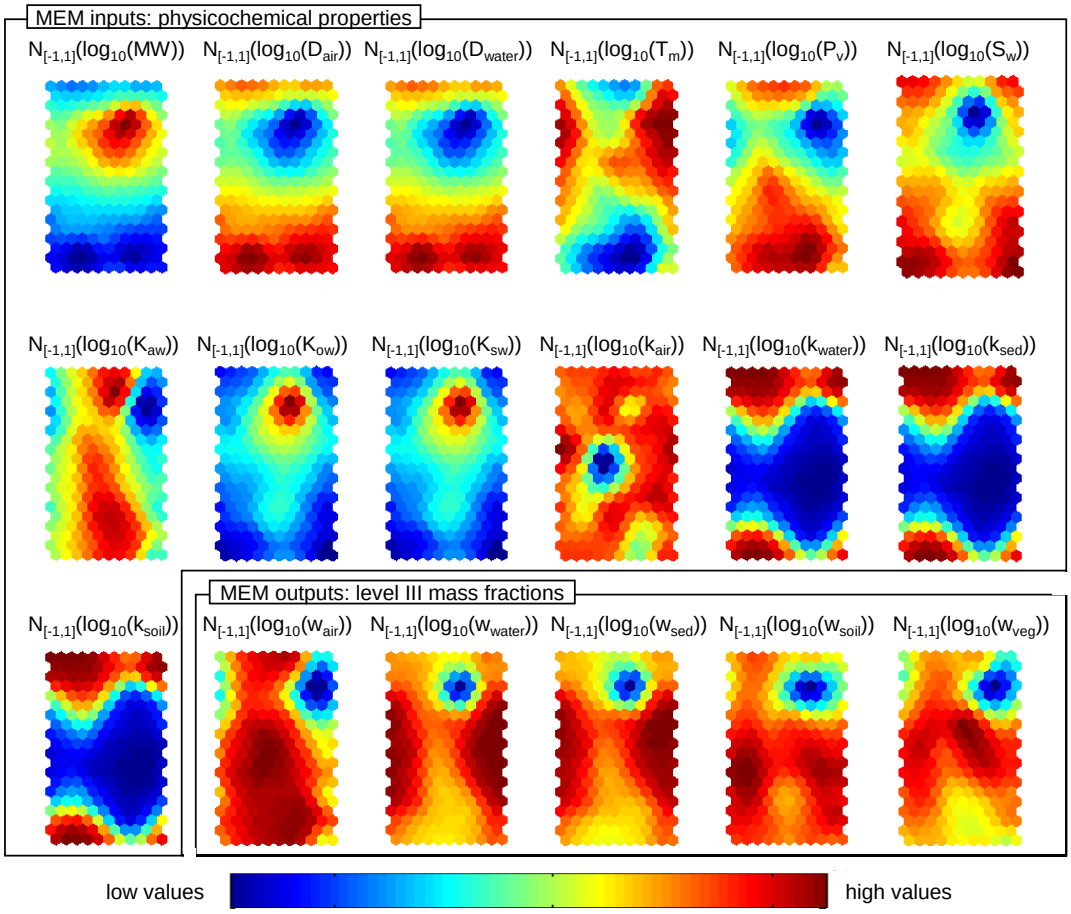


Figure 4-3. Multivariate screening, through SOM planes, of the inputs and outputs of a MEM for emissions in water.
This figure shows the values of the SOM prototypes in each of the dimensions conforming the chemical space of the 375 work chemicals of the reference pollution scenario: MW, T_m, S_w, P_v, K_{ow}, K_{aw}, K_{sw}, k_{air}, k_{water}, k_{sed}, k_{soil}, D_{air}, D_{water}, w_{air}, w_{water}, w_{sed}, w_{soil}, w_{veg}). The work chemicals are characterized by logarithmic properties and mass ratios normalized in the range [-1,1].

chemicals in the mid-level part of the SOM are persistent, as the plane of k_{water} indicates very low values for that zone, corroborated with the corresponding medium to high w_{water} values indicated in the w_{water} plane. Some of these persistent chemicals evaporate easily. The persistent chemicals in the extreme left and right zones of the mid-part of the SOM remain in water because they have medium to low K_{aw} values, while those in the center of the mid-part of the SOM go to air, reaching later the soil and vegetation compartments. Similar analyses can be performed at a more detailed level by focusing attention on every SOM unit independently.

For a general overview of the entire set of work chemicals, the SOM can be clustered into somewhat big portions as Figure 4-4 indicates. The SOM introduced in Figures 4-2 and 4-3, was divided into two sections applying the K-means algorithm: one section contains chemicals with high water degradability (with half lives, $\tau_{½}$, between 2 days and 5.5 months) and another with low water degradability (with half-lives between 3.3 months and 5.3 years). It can be verified that the clustering of the SOM in Figure

ISBN:978-84-693-4597-9 /DL:T.1010-2010

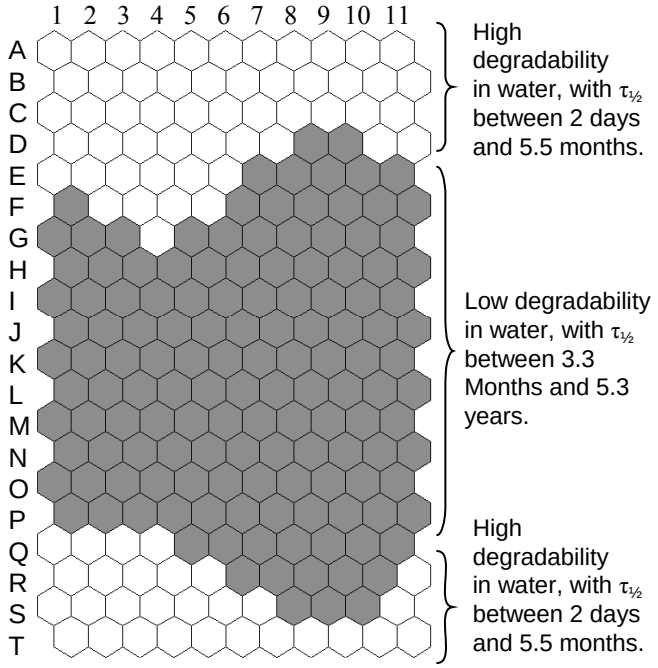


Figure 4-4. SOM clustering in search of relationships between key variables.

This figure shows how the SOM referred to the 375 work chemicals of the reference scenario can be divided with basis on their water degradability. The two sections were identified by the application of the K-means and Davies-Bouldin algorithm.

4-4 is greatly influenced by the k_{water} values of the work chemicals by tracing a line that cuts the center of Figure 4-2b from its upper-left side to its down-right side, taking as guide the color code in Figure 4-2a and the SOM plane of k_{water} in Figure 4-3. For the current reference scenario (Chapter 3), in which constant emissions take place in water, the degradability in this compartment has a great influence on the final partitioning of chemicals to other compartments.

For a general overview of the entire set of work chemicals, the SOM can be clustered into somewhat big portions as Figure 4-4 indicates. The SOM introduced in Figures 4-2 and 4-3, was divided into two sections applying the K-means algorithm: one section contains chemicals with high water degradability (with half lives, $\tau_{½}$, between 2 days and 5.5 months) and another with low water degradability (with half-lives between 3.3 months and 5.3 years). It can be verified that the clustering of the SOM in Figure 4-4 is greatly influenced by the k_{water} values of the work chemicals by tracing a line that cuts the center of Figure 4-2b from its upper-left side to its down-right side, taking as guide the color code in Figure 4-2a and the SOM plane of k_{water} in Figure 4-3. For the current reference scenario (Chapter 3), in which constant emissions take place in water, the degradability in this compartment has a great influence on the final partitioning of chemicals to other compartments.

The SOM algorithm can be reasonably applied to the graphical screening of multimedia environmental modeling data in level III conditions, giving to the modeler an insight of how a known set of chemicals of concern can be environmentally distributed according to their physicochemical properties. The SOM offers an approximate representation of each of the dimensions of a data set that, when referred to the attributes of chemicals and the output of a level III MEM, allows the analysis of the involved variables from different points of view (one by one, in groups, in map sections, etc.) and a clearer understanding of the mechanisms taking place.

ISBN:978-84-693-4597-9 /DL:T.1010-2010

4.2 Variability in the outputs of MEMs from properties estimated with QSPRs and QSBRs

The complexity of the factors required for predicting chemical activity from molecular information usually limits the accuracy of QSAR-based estimation methodologies (Johnson, 2008). Properties of chemicals not used in the development of QSPR or QSBR models are known to be estimated with substantial errors (Taskinen and Yliruusi, 2003) and so there is consensus on the use of QSARs validated externally for large sets of chemicals (OECD, 2007). Under such premise, several methods have been largely recommended for predicting partitioning (Boethling et al., 2004) and degradation (Raymond et al., 2001) data for a wide range of chemicals. Generally, quantitative estimations can be performed for partitioning properties (Boethling et al., 2004). In contrast, estimations for degradation data are rather qualitative (Aronson et al., 2006). Issues associated to the experimental determination of environmental degradation are still difficult to characterize (Klöpffer and Wagner, 2007), limiting the availability of reliable training data (Aronson et al., 2006) for developing QSBR models and, ultimately, limiting key inputs of standard MEMs (Kühne et al., 2007).

The uncertainty associated to key physicochemical properties, like partitioning and, more notably, degradation data, has been recognized to exert a great influence on the outputs of standard MEMs (Citra, 2004; Eisenberg et al., 1998; Kawamoto et al., 2001; Kühne et al., 1997; Toose et al., 2004). In such cases, wide uncertainties in the inputs of a MEM may cause as well wide uncertainties in its outputs (Equation 2-80). The QSPR and QSBR methods compiled in EPIsuite (SRC, 2008), developed and validated for a wide number of chemicals, are among the most widely recommended methods for estimating partitioning properties (Boethling et al., 2004) and the degradability or not of chemicals (Raymond et al., 2001) from molecular structure. However, the accuracy of its degradability estimation methods are not accurate enough to provide numerical degradability measures, limited solely to discrete degradation estimates for general purpose environmental screenings (Aronson et al., 2006). For this reason, the need of methods capable of providing degradation estimates ready to use in standard MEMs remains intact (Kühne et al., 2007). The large variety of factors affecting the degradability of chemicals in the environment is such that there is still a lot of work to be done for measuring and modeling such process (Klöpffer and Wagner, 2007).

For simulating the effect that uncertainty in physicochemical properties estimated from QSPRs or QSBRs have on standard level III fate estimations, 1000 combinations of random property values have been propagated for each chemical throughout the reference MEM of the reference pollution scenario (Chapter 3), simultaneously, for all independent properties affecting the estimations of the model: T_m, P_v, H, K_{ow}, k_{air} and k_{water}. Uncertainty in all the remaining properties were not considered because of their dependency (D_{air}, D_{water}, K_{sw}, k_{sed} and k_{soil}), negligible uncertainty (MW) or no direct intervention in the model (by definition, S_w has already been considered in the ratio $H = P_v/S_w$). Figure 4-5 shows the resulting cause-effect relationships between all properties (inputs) and mass ratios (outputs). Note that the uncertainty analysis is

ISBN:978-84-693-4597-9 /DL:T.1010-2010

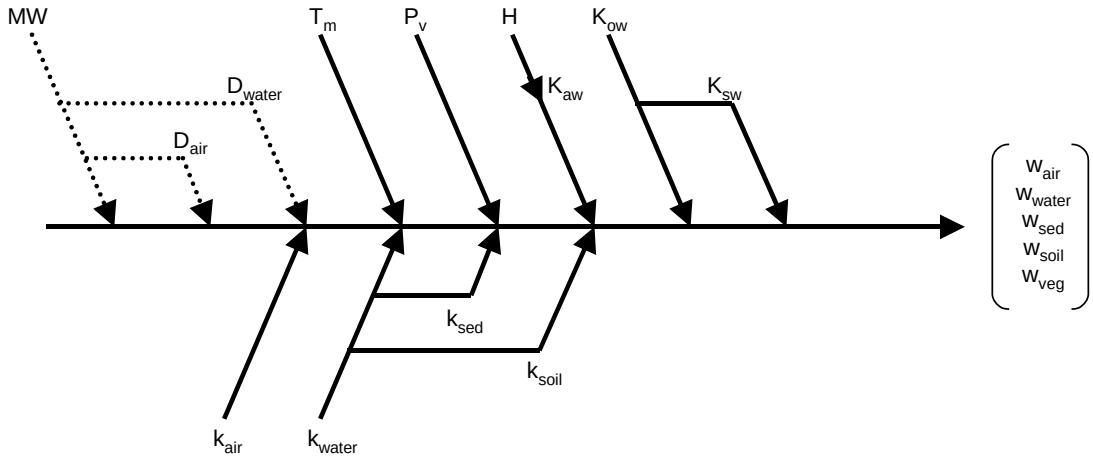


Figure 4-5. Main sources of uncertainty considered on the MEM of the reference pollution scenario.
This fishbone diagram shows a cause-effect relationship between independent properties and level III mass ratios. The most important sources of uncertainty are shown with solid arrows, while negligible sources of uncertainty are shown with dotted arrows.

Table 4-1. Statistical distributions assigned to independent properties affecting the reference pollution scenario.

Property	Assumed distribution for simulations	Statistics reported for recommended QSPRs and QSBRs			
		Data set	Statistic parameters*	Units	Source
T_m	Normal	validation	SD = 58.00	K	(Boethling et al., 2004)
P_v	Log-normal	validation	SD = 0.717	mmHg	(Boethling et al., 2004)
H	Log-normal	training	SD = 0.440*,T,3	$\log_{10}$(mg/L)	(Boethling et al., 2004)
K_{ow}	Log-normal	validation	SD = 0.427*,V,4	$\log_{10}$(atm·m^3/mol)	(Boethling et al., 2004)
k_{air}	Discrete	training	P(0) = 0.48, P(±1) = 0.37, P(±2) = 0.13, P(±>2) = 0.02	-	(Kühne et al., 2007)
k_{water}	Discrete	training	P(0) = 0.52, P(±1) = 0.35, P(±2) = 0.08, P(±>2) = 0.05	-	(Kühne et al., 2007)

* For QSPRs, the parameters have been reported in standard deviations, SD, in logarithmic values when noted; for QSBRs, the reported parameters are probabilities, P(C), that indicate if a chemical has been classified as member of a degradation class C (0 = correct class, ±1 = neighbor category predicted, ±2 = two categories differing and ±>2 = more than two categories differing) in the 9-class scale proposed by Mackay et al. (1992).

referred to all the 468 chemicals of the reference pollution scenario (375 work and 93 validation chemicals) and that the random values have been generated by statistical distributions of widely recommended QSPRs (Boethling et al., 2004) and prototype QSBRs (Kühne et al., 2007), with statistic parameters listed in Table 4-1.

With the standard deviations (SD) given in Table 1, continuous distributions have been assigned for T_m, P_v, S_w, H and K_{ow}. It has been assumed that a variable follows a normal distribution if the standard deviation given by Boethling et al. (2004) is in unit variables. When the standard deviation is given in logarithmic units, normal-logarithmic distributions (Limpert et al., 2001) have been considered. Although the

ISBN:978-84-693-4597-9 /DL:T.1010-2010

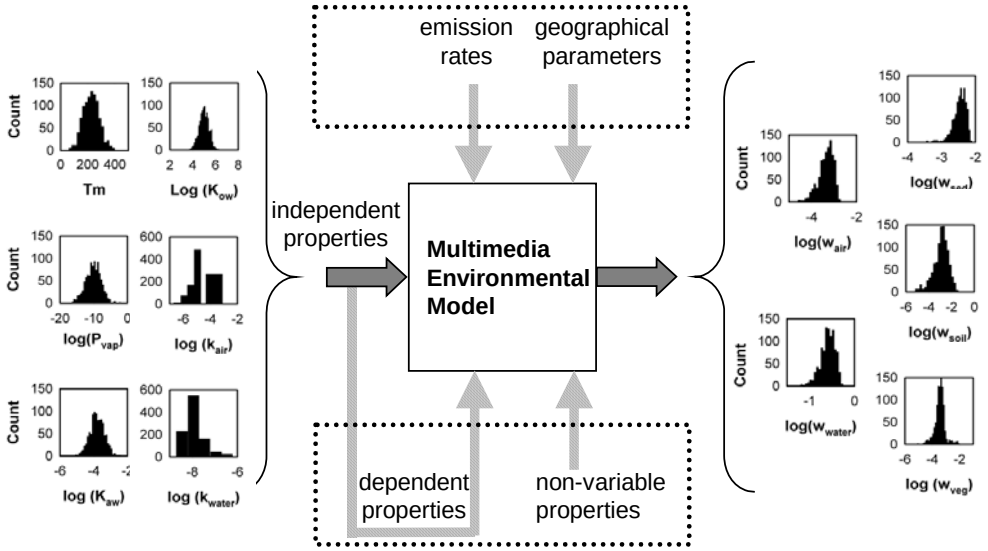


Figure 4-6. Ranges of variation in the mass ratios estimated by the MEM of the reference pollution scenario for Endrin, resulting from a statistical sampling of key independent properties in 1000 iterations, for emissions in water.
This diagram shows statistical distributions for both key input and output variables of the MEM for a single pollutant.

standard deviation of P_v is given in terms of mmHg, a lognormal distribution has been used to avoid negative values in chemicals with very low P_v values. Since degradation data is usually predicted in term of classes, non-uniform discrete distributions have been chosen for k_{air} and k_{water}, assuming that a correct prediction for the degradation class of a chemical has a probability equal to P(0) and that probabilities for incorrect classes below and above the correct class are symmetrical and equal to half the probability corresponding to the number of differing categories, as reported in Table 1 from QSBRs using structural similarity through atom centered fragments (Kühne et al., 2007) for predicting degradation classes as listed in Table 1-2.

A graphical representation of the statistical distributions in both input and output variables of the MEM used in the reference pollution scenario is given in Figure 4-6 for Endrin (CAS: 72-20-8), a very persistent organic pollutant. For emissions in water, the target logarithmic mass ratios of this chemical in the reference scenario are: $\log_{10}(w_{air})$ = -3.11, $\log_{10}(w_{water})$ = -0.65, $\log_{10}(w_{sed})$ = -2.47, $\log_{10}(w_{soil})$ = -3.35; and, $\log_{10}(w_{veg})$ = -2.05. However, when only six independent properties are varied according to the statistics reported in Table 4-1 the logarithmic variation ranges in such outputs are, respectively: 2.18, 1.16, 1.33, 3.92 and 3.02. The lowest variation range (1.16) occurs in the water compartment, where emissions take place; while, higher variation ranges resulted for neighboring compartments.

As discussed in Section 4.1 for emissions in the water compartment, the 468 chemicals move from such compartment to its immediate neighboring compartments: air and sediment; and, from air, the chemicals go directly to soil and vegetation

ISBN:978-84-693-4597-9 /DL:T.1010-2010

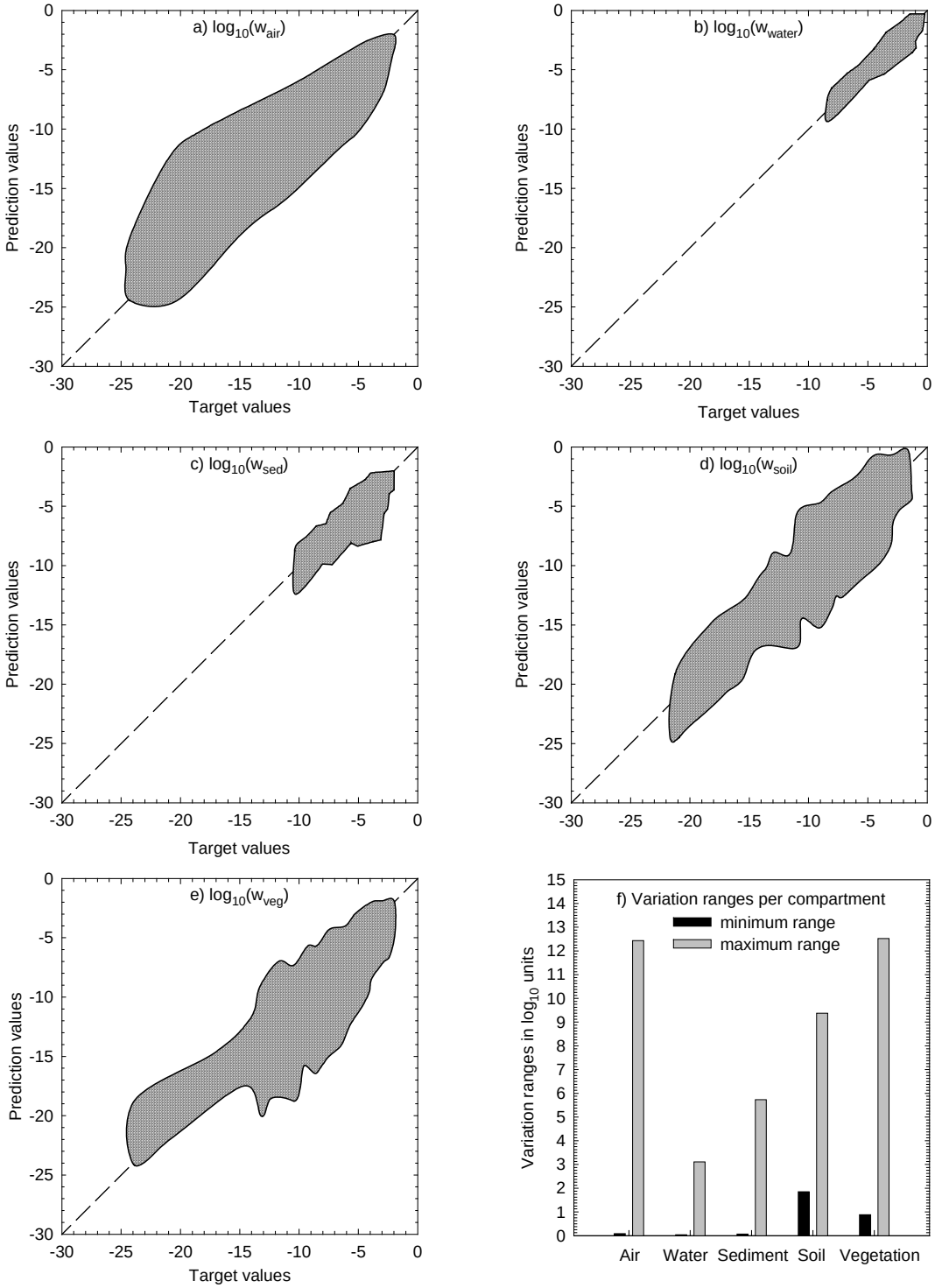


Figure 4-7. Ranges of variation in the mass ratios estimated by the MEM of the reference pollution scenario for 468 chemicals emitted in water, from a statistical sampling of key independent properties in 1000 iterations.

Range comparisons among all chemicals, distributed in air (a), water (b), sediments (c), soil (d) and vegetation (e). An additional comparison is referred to the minimum and maximum ranges of variation reported for all compartments (f). The variations ranges are listed for all 468 chemicals in Annex D.c1.

ISBN:978-84-693-4597-9 /DL:T.1010-2010

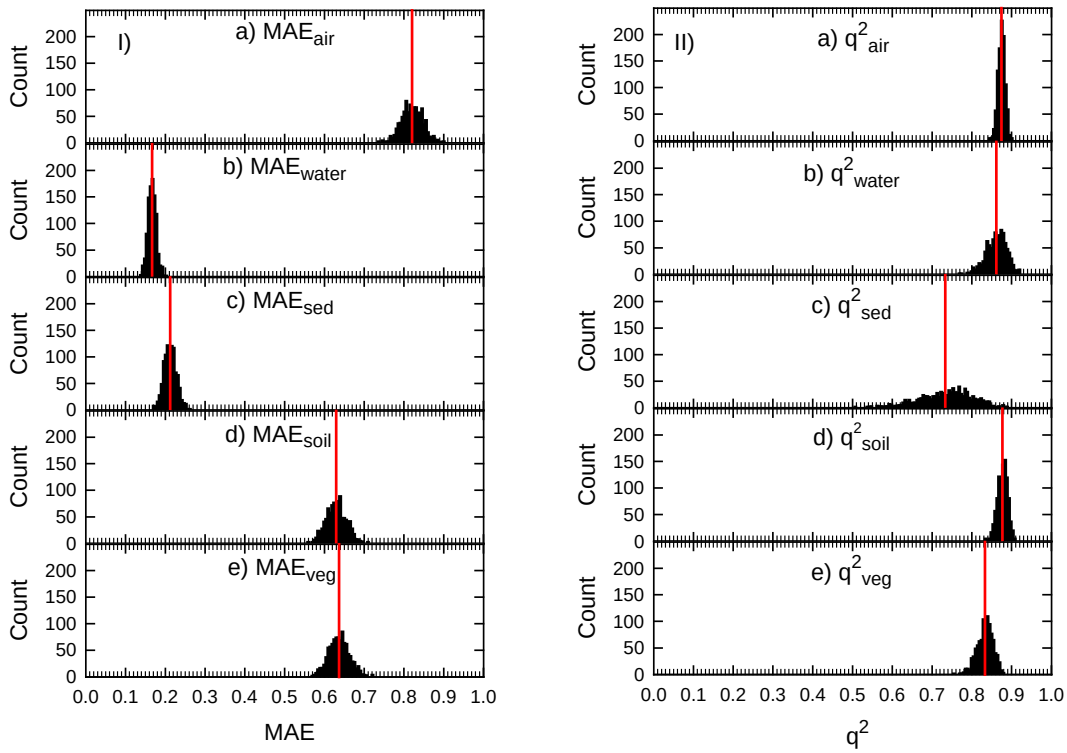


Figure 4-8 Measurement of the predictive capacity of the MEM in the reference scenario in terms of MAE and q^2 over all 468 chemicals emitted in water, resulting from a statistical sampling of independent properties in 1000 iterations.

This figure shows how MAE measurements running over all the 468 chemicals of the reference pollution scenario are low for the water and sediment compartments and high for the other compartments (I), despite of that fact that high q^2 measurements seem to indicate that fate predictions in all compartments seem to be good (II).

compartments. When random property values are propagated throughout the reference MEM using the statistical distributions of Table 4-1 for all chemicals, the lower variation ranges occur in the mass ratios estimated for the water compartment (where emissions take place); while, much higher variations occur in all the other compartments, especially in the air, soil and vegetation compartments. Such tendencies can be observed in the subplots a to e of Figure 4-7, which compare the mass ratios estimated by the MEM when affected by random properties (prediction values) to its reference estimations (target values), originally resulting from the reference properties described in Chapter 3.

Since chemical emissions in the water compartment are being analyzed, the water compartment concentrates more than 1 % of the mass emission in the system (as logarithmic mass ratios above –2 indicate, see subplot b of Figure 4-1III) for most chemicals considered. With these ideas in mind, we can see that as chemicals move from the water compartment (Figure 4-7b) to immediately neighboring compartments, air (Figure 4-7a) and sediment (Figure 4-7c), mass ratios in these compartments become smaller and show wider variability with respect to their target values. As the considered chemicals reach, from air, the soil (Figure 4-7d) and vegetation (Figure 4-7e) compartments, the logarithmic mass ratios in the latter inherit the variability suffered by the mass ratios in the former. Figure 4-7f shows the minimum and

ISBN:978-84-693-4597-9 /DL:T.1010-2010

maximum variability of mass ratios reported among all the 468 chemicals; it confirms that for the reference pollution scenario, the narrower and wider variations in the output of the MEM take place, respectively, in water and air.

When measuring the predictive performance of the reference MEM, interesting tendencies can be identified. The evaluation of all the 468 chemicals of the scenario in terms of compartmental MAE values (Equation 2-85) reinforces what has been already observed in Figure 4.7: differences between target and predicted values tend to be minimal in the water compartment, where emissions take place in the scenario, as MAE values per iterative simulations on the entire set of 468 chemicals show (subplot I of Figure 4.8). However, when the chemicals are assessed in terms of q^2 values (Equation 2-86), the goodness of all the predictions can be overestimated, with the exception of those in the sediment compartment, as all compartmental q^2 values are extremely high (subplot II of Figure 4-8).

As pointed out in the previous section, k_{water} plays an important role in the final environmental distribution of all the chemicals in the scenario; the discrete variability of k_{water} (Table 4-1) distorts the predicted amount of chemicals in water, from which partitioning to other compartments takes place. This result is somewhat analogous to a previous work, in which the statistical sampling of herbicides emitted in soil was shown to affect the estimated overall persistence time in the system, primarily because of the variability in soil degradation half-lives (Citra, 2004). Depending on the "real" mean reference property values of a chemical, the random property values generated by statistical distributions of standard property estimation methods (Table 4-1) produced variations in the outputs of the reference MEM (Figure 4-7), that were, in the worst cases, of several orders of magnitude in logarithmic units. Annex D.c1 lists variation ranges for emissions in water and emissions in air. In the same manner, it can be inferred that, depending on the domain of applicability of available QSPRs and QSBRs, the output of standard MEMs should undergo a similar variability.

4.3 Fate predictions from QPFRs

Chemicals lacking of some physicochemical properties cannot be assessed with MEMs because the functionality property-fate cannot be evaluated (Equation 2-78), unless every property is individually estimated via standard QSPR or QSBR methods (Figure 1-4). Alternatively, the capacity of supervised learning algorithms (Witten and Frank, 2005), like ANNs (Basheer and Hajmeer, 2000) or SVRs (Smola and Schölkopf, 2004), to recognize patterns from noisy or incomplete data (Jain et al., 2000) can be used to estimate quantitatively the fate of new chemicals from few available properties, simply evaluating available properties in QPFRs (Equation 2-81).

In preliminary experiments, it was found that BPNs working as QPFRs could emulate accurately the property-fate functionality of a reference MEM, predicting level III logarithmic concentrations in five compartments simultaneously for The Netherlands, from solely partitioning and degradation data of chemicals emitted in one out of

ISBN:978-84-693-4597-9 /DL:T.1010-2010

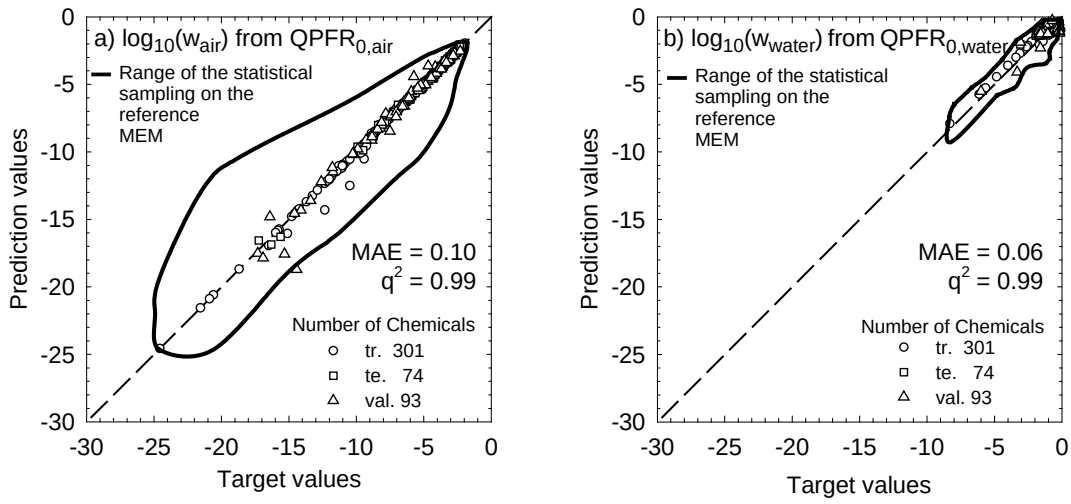


Figure 4-9. Predictions from QPFRs, based on SVRs using independent but key properties as input (Set 0), for air (a) and water compartments (b), considering emissions in water.

Few partitioning and degradation properties are enough for emulating a reference MEM with supervised learning algorithms, producing fate estimations with high accuracy.

various compartments (Martínez et al., 2006c; Annex A.a1). Since partitioning and degradation properties have a direct influence in the mass balances of the MEM used for generating the target values, the information they provide is enough for emulating a MEM straightforward.

Considering emissions in water like in Sections 4.1 and 4.2, let's analyze the air and water compartments of the reference pollution scenario, in which the highest and lowest fate estimations take place (Figure 4-7f). SVR-based QPFRs were tuned (Annex D.b1) and used to predict level III logarithmic mass ratios in these two compartments (Annex D.c2) from a set of few but meaningful inputs, independent partitioning and degradation properties (Set 0): $\log_{10}(K_{aw})$, $\log_{10}(K_{ow})$, $\log_{10}(k_{air})$, $\log_{10}(k_{water})$. QPFRs have been trained, tested and validated following the procedure listed in Table 2-3: The 18D SOM already presented in Section 4.2 (Figures 4-2, 4-3 and 4-4) and compiled in Annex D.a1 was used for building the training and test data sets with, respectively, 301 and 74 chemicals from the original set of 375 work chemicals. Note that both the inputs and target of every QPFR model are, respectively, normalized logarithmic properties (Set 0) and a normalized logarithmic compartmental mass ratio as shown in Equation 2-83.

Figure 4-9 shows predictions for air and water from two models, respectively, $QPFR_{0,air}$ (Figure 4-9a) and $QPFR_{0,water}$ (Figure 4-9b). It can be observed that, as most chemicals lie in the diagonal (down-left to up-right) of each subplot of Figure 4-9, prediction values are very close to their corresponding target values; also indicated by very low MAE and high q^2 measurements. It can be observed as well that such QPFRs outperform when compared to the statistical sampling of the reference MEM already studied in Section 4.2 (Figures 4-7a and 4-7b). Solely with the pair of partition coefficients and the pair of degradation rates contained in Set 0, very accurate fate predictions were obtained for chemicals not used in the development of the QPFR

ISBN:978-84-693-4597-9 /DL:T.1010-2010

models: the test and validation chemicals. However, this was possible because the key properties of such "new" chemicals were known by the time of the assessment.

It is known that for most chemicals the availability of partitioning and degradation data is precisely limited, specially for the latter (Klöpffer and Wagner, 2007). So, the applicability of QPFRs is restricted to new chemicals of concern for which accurate key physicochemical properties are already available, from either accurate measurements or existing estimation methods (QSPRs and/or QSBRs). When physicochemical properties are unavailable, fate predictions from molecular information might be an alternative for the environmental assessment of chemical pollutants, as detailed in Section 4.4, below.

4.4 Fate predictions from QSFRs

Known the shortcomings of assessing the fate of chemicals with either MEMs relying on a wide number of estimated physicochemical properties (Section 4.2) or with simple QPFRs (Section 4.3), the availability of another fate estimation methodology would be of great interest to environmental modelers. The possibility of estimating the fate of new chemicals, bypassing the explicit use of their physicochemical properties, with QSFRs (Equation 2-82) implies solely the use of molecular information (Figure 1-4).

Several property estimation methods (Devillers, 2003; Raymond et al., 2001; Taskinen and Yliruusi, 2003) rely on multivariate correlations using as input a wide variety of molecular descriptors (topological, electronic, geometric, etc.) derived from semi-empirical approximations of the molecular orbital (MO) theory (Bredow and Jug, 2005). Other estimation methods, relating activity to fragment contributions derived from the SMILES notation (Weininger, 1988; Weininger et al., 1989), have been widely recommended for predicting partitioning data (Boethling et al., 2004) and degradation data (Raymond et al., 2001) for a wide range of chemicals, which is the case of the models traditionally included in EPI suite^{TM} (SRC, 2008). So, it seems plausible the direct prediction of environmental multimedia fate from molecular information via QSFRs (Equation 4-11), grounded on either basic theoretical descriptors (derived from semi-empirical MO models) or counts of molecule constituents (atoms, bonds, functional groups and rings).

Within the NOMIRACLE project, several experiments were carried out for studying QSFRs and identifying best practices for their application in standard multimedia environmental modeling (Table 1-3). In one experiment, both QPFR and QSFR models were evaluated on the same scenario using, respectively, key physicochemical properties (partitioning and degradation data) and semi-empirical molecular descriptors for estimating the fate of chemicals emitted in one our of various compartments (Martínez et al., 2006b; Annex A.c1): fate predictions from QPFRs were more accurate than those from QSFRs; but QSFRs could show rough but meaningful fate trends, solely from molecular information. This experiment demonstrated that molecular information could be linked to chemical fate, but that

ISBN:978-84-693-4597-9 /DL:T.1010-2010

special adjustments would be required. For this reason, subsequent experiments within the project studied: the use of different supervised learning algorithms, in Annexes A.a2 (Martínez et al., 2007b), A.b2 (Martínez et al., 2007a) and A.a4 (Martínez et al., 2008a); and, the use of different sets of molecular descriptors, in Annexes A.a2 (Martínez et al., 2007b), A.b2 (Martínez et al., 2007a), A.a4 (Martínez et al., 2008a) and A.b3 (Martínez et al., 2008b).

The selection of supervised learning algorithms for building QSFRs is simply a matter of compromise between computational feasibility and applicability (e.g., ANNs may be appropriate for QSFRs with several outputs but are very likely to suffer overtraining; while, SVRs yield reproducible QSFR, but solely for single outputs). In this section, it will be discussed solely the influence of different sets of molecular descriptors in the performance of QSFRs. The implementation of QSFRs for chemicals belonging to specific chemical classes will be discussed in detail separately in section 4.5.

Table 4-2 compares the features and performances of the QPFRs presented in Figure 4-9 to those of QSFR prototypes, also optimized (Annex D.b1) with the procedure of Table 2-3, when modeling fate in the air and water compartments of the reference pollution scenario from different sets of descriptors: a) few theoretical descriptors selected empirically by the CFS filtering algorithm (Hall, 1999) for each compartment from a starting set of 23 descriptors (MW and the 22 semi-empirically estimated descriptors): 4 descriptors for air (set i-a: μ_{hyb}, μ_{pc}, ${}^{0}\chi$, ${}^{1}\chi$) and 6 descriptors for water (set i-b: MW, μ_{hyb}, μ_{pc}, HOMO, ${}^{2}\kappa$, ${}^{3}\kappa$); b) A unique set of 23 theoretical descriptors including MW and the 22 semi-empirically estimated descriptors (set ii: MW, ΔH_f, MR, PO, μ_{hyb}, μ_{pc}, μ, Area, Vol, NFL, HOMO, LUMO, IP, EA, ${}^{0}\chi$, ${}^{1}\chi$, ${}^{2}\chi$, ${}^{0}\chi^{v}$, ${}^{1}\chi^{v}$, ${}^{2}\chi^{v}$, ${}^{1}\kappa$, ${}^{2}\kappa$, ${}^{3}\kappa$); and, c) A unique set of descriptors based on MW and 38 non-zero counts of molecular constituents (set iii: MW, 10 atom counts (all atoms, bromine, carbon, chlorine, fluorine, hydrogen, nitrogen, oxygen, phosphorus and sulphur), 4 bond counts (all bonds, single bonds, double bonds and triple bonds), 16 functional group counts (aldehyde, amide, amine, sec-amine, carbonyl, carboxyl, cyano, ether, hydroxyl, methyl, methylene, nitro, nitroso, sulfide, sulfone and thiol) and 8 ring counts (all rings, aromatic rings, small rings, 5 membered, aromatic 5 membered, 6 membered, aromatic 6 membered and 7-12 membered)).

Per each QSFR model (Annex D.c2), referred to a set of descriptors (sets i-a, i-b, ii and iii) and a compartment, a SOM was trained with the 375 work chemicals for generating a pair of optimal training and test data sets (Annexes Da.2 to D.a7). QSFR models using few descriptors performed poorly (for w_{air}: $QSFR_{ia,air}$; for w_{water}: $QSFR_{ib,water}$) when compared to those using more descriptors (for w_{air}: $QSFR_{ii,air}$, $QSFR_{iii,air}$; for w_{water}: $QSFR_{ii,water}$, $QSFR_{iii,water}$) as their R^2_{te} values indicate, also confirmed with the average performances on the 10-fold CV and the LOO procedures: R^2_{10CV} and R^2_{LOO}, respectively. Such tendencies are also applicable for the validation phase, as R^2_{val} values indicate. Figure 4-10 compares the MAE performances on denormalized fate predictions already reported in Table 4-2, i.e, logarithmic mass ratios. The lowest errors have been achieved with the QSFRs referred to the compartment in which emissions take place, the water compartment, while the highest errors resulted from the QSFR models referred to the air compartment. Following

ISBN:978-84-693-4597-9 /DL:T.1010-2010

Table 4-2. SVR prototypes of QPFRs and QSFRs for the air and water compartments of the reference pollution scenario, considering emissions in water.

		Air compartment				Water compartment			
		$QPFR_{0,air}$	$QSFR_{ia,air}$	$QSFR_{ii,air}$	$QSFR_{iii,air}$	$QPFR_{0,water}$	$QSFR_{ib,water}$	$QSFR_{ii,water}$	$QSFR_{iii,water}$
Attributes	set	0	ia	ii	iii	0	ib	ii	iii
	total	4	4	23	39	4	6	23	39
	type*	PP	TD	TD	CC	PP	TD	TD	CC
Number of available chemicals	training+	301[a1]	288[a2]	297[a4]	300[a6]	301[a1]	300[a3]	307[a5]	299[a7]
	test+	74[a1]	87[a2]	78[a4]	75[a6]	74[a1]	75[a3]	68[a5]	76[a7]
	validation	93	93	93	93	93	93	93	93
SVR parameters	C	150	1	5	300	150	150	10	25
	γ	1	10	1	0	1	0	1	1
	ε	$1.0x10^{-6}$	$1.0x10^{-1}$	$2.5x10^{-1}$	$1.0x10^{-2}$	$1.0x10^{-5}$	$2.5x10^{-1}$	$2.5x10^{-1}$	$1.0x10^{-6}$
	p	$1.0x10^{-6}$	$1.0x10^{-5}$	$1.0x10^{-3}$	$1.0x10^{-2}$	$1.0x10^{-6}$	$1.0x10^{-5}$	$1.0x10^{-5}$	$1.0x10^{-1}$
Support vectors	total	301	271	232	259	301	233	212	124
Prediction performances on normalized data‡	R^2_{tr}	1.00	0.66	0.93	0.85	1.00	0.19	0.94	0.89
	R^2_{te}	1.00	0.57	0.70	0.86	1.00	0.51	0.70	0.75
	R^2_{val}	0.98	0.05	0.25	0.46	0.98	0.30	0.51	0.67
	MAE_{tr}	0.01	0.10	0.05	0.07	0.01	0.27	0.04	0.07
	MAE_{te}	0.01	0.11	0.10	0.07	0.01	0.08	0.08	0.08
	MAE_{val}	0.02	0.27	0.22	0.16	0.02	0.11	0.11	0.10
10-fold CV on normalized data	R^2_{10CV}	0.95	0.28	0.50	0.76	0.87	0.20	0.40	0.44
	MAE_{10CV}	0.02	0.18	0.16	0.10	0.04	0.12	0.10	0.11
LOO on normalized data	R^2_{LOO}	0.92	0.27	0.48	0.77	0.84	0.15	0.39	0.47
	MAE^2_{LOO}	0.02	0.18	0.16	0.10	0.03	0.12	0.10	0.10
Prediction performances on denormalized data	q^2_{tr}	1.00	0.64	0.93	0.85	0.99	0.20	0.93	0.86
	q^2_{te}	1.00	0.53	0.69	0.86	0.95	0.48	0.65	0.60
	q^2_{val}	0.97	-0.24	0.13	0.42	0.96	0.26	0.50	0.63
	MAE_{tr}	0.06	1.18	0.61	0.81	0.04	0.46	0.17	0.30
	MAE_{te}	0.11	1.28	1.12	0.81	0.08	0.34	0.33	0.34
	MAE_{val}	0.24	3.04	2.54	1.83	0.10	0.47	0.45	0.42

* Type of input variables: PP = physicochemical properties; MO = semi-empirical MO descriptors; and, CC = MW and simple counts of molecular constituents.
+ Chemicals selected with specific SOMs, presented in: [a1] = Annex D.a1, [a2] = Annex D.a2, [a3] = Annex D.a3, [a4] = Annex D.a4, [a5] = Annex D.a5, [a6] = Annex D.a6 and [a7] = Annex D.a7.
‡ Prediction performances obtained during the tuning of the SVR algorithm in each case, presented in Annex D.b1.

trends already identified when applying statistical sampling on the reference MEM (Figure 4-7).

In standard QSPRs and QSBRs, it is common practice to use as less descriptors as possible (Mager and Mager, 1992; Wold, 1992); but, since QSFRs attempt to emulate MEMs, in which diverse environmental processes are simulated simultaneously, few descriptors seem to offer little information to predict the fate of test chemicals and even less information for validation chemicals. This is especially true for the air compartment of the reference scenario (Figure 4-10a), in which the variation with respect to the reference MEM tends to be higher in any case.

ISBN:978-84-693-4597-9 /DL:T.1010-2010

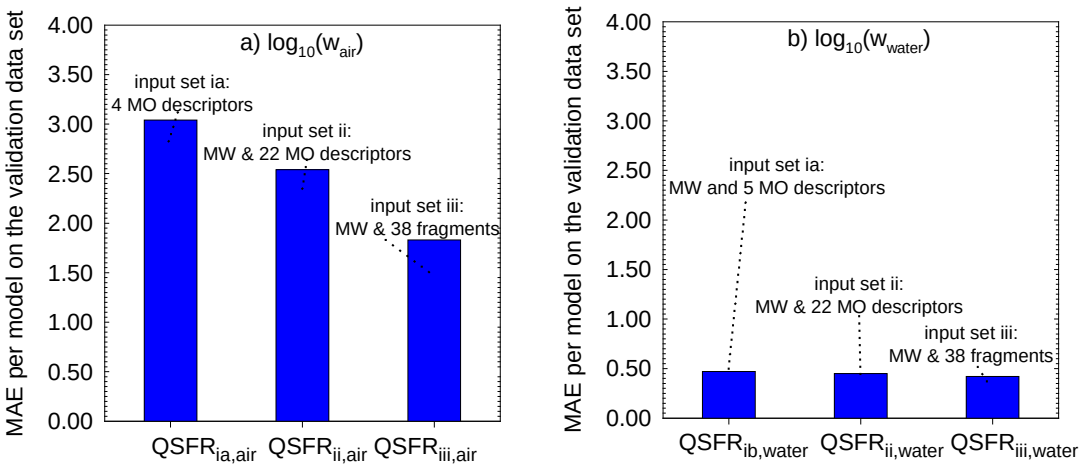


Figure 4-10. MAE errors of logarithmic mass ratios in air (a) and water (b), predicted for the 93 validation chemicals by QSFR models using different sets of molecular descriptors.

Per compartment, best environmental fate estimations are obtained from QSFRs using as input MW and counts of molecular fragments (atoms, bonds, groups and rings).

QSFRs with a diverse number of semi-empirically determined descriptors can provide good performances on test chemicals, selected to be somewhat similar to the training chemicals (Table 2-3), as it occurs in the models $QSFR_{ii,air}$ and $QSFR_{ii,water}$ (Table 4-2). But, it must be noted that poor fate predictions can also be obtained when assessing independent chemicals, not used at all in the optimization of the algorithms: $QSFR_{ii,air}$ predicted poorly the fate of the 93 validation chemicals in air, with MAE and q^2 over logarithmic mass ratios of, respectively, 2.54 and 0.13. It must be also noted that the prediction accuracy of QSFRs can diminish, if the descriptors of new chemicals to assess are estimated with a semi-empirical MO method different than that used in the development of the QSFR models, there are marked differences between existing semi-empirical MO methods (Bredow and Jug, 2005).

QSFRs using as inputs counts of molecular constituents (atoms, bonds, functional groups and rings) have provided the best fate estimations for both test and validation chemicals. The fate predictions resulting from $QSFR_{iii,air}$ and $QSFR_{iii,water}$ have been superior to all the other QSFR models listed in Table 4-2. Figure 4-11 compares the fate predictions of $QSFR_{iii,air}$ (Figure 4-11a) and $QSFR_{iii,water}$ (Figure 4-11b) to the variations ranges resulting from the statistical sampling of the reference MEM (Figures 4-7a and 4-7b). All predictions values for $\log_{10}(w_{water})$ are within the variation ranges; while, most predictions values for $\log_{10}(w_{air})$ also lie within the variation ranges, with some exceptions. After checking the structure of each of the 468 chemicals of the reference scenario (Annexes A.C1 and A.C2), it has been noted that their structures were extremely diverse, not only with respect to the presence of rings, but also with respect to their composition (with very dissimilar atom types, like bromine, chlorine, fluorine, nitrogen, oxygen, phosphorus and sulphur). This gives an insight of why QSFRs using few semi-empirical MO descriptors were poor fate predictors: these simply could not provide enough information for discriminating chemicals, where constituent counts do that more efficiently.

ISBN:978-84-693-4597-9 /DL:T.1010-2010

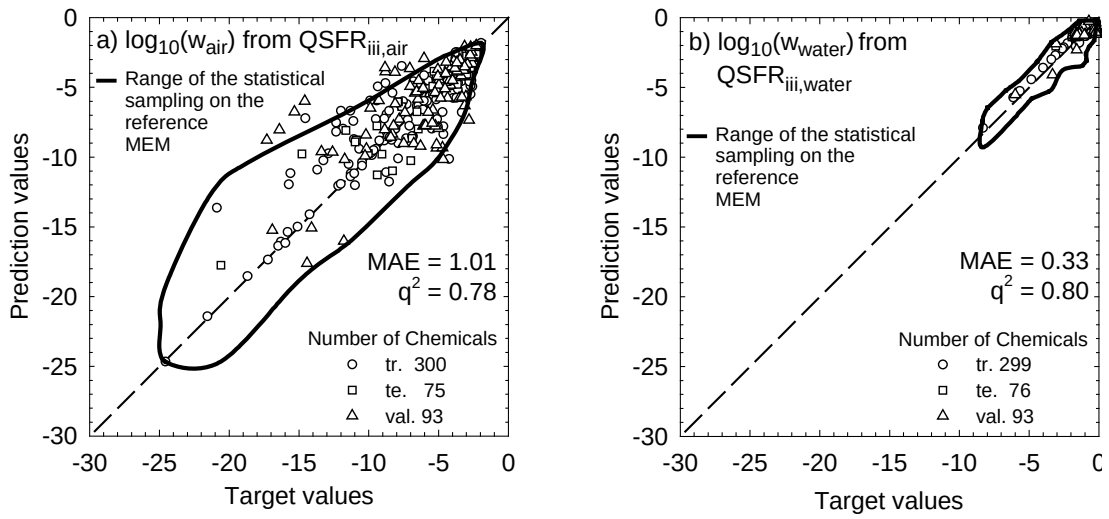


Figure 4-11. Predictions from QSFRs, based on SVRs using optimal molecular information as input (Set iii), for air (a) and water compartments (b), considering emissions in water.
MW and simple counts of molecular constituents provide reasonable information for emulating a MEM by QSFR.

Despite of the fact that QSFR models using simple counts of molecular constituents cannot distinguish between isomers that have identical descriptors (in the reference pollution scenario there are 175 chemicals having such peculiarity: 150 and 25 chemicals out of the 375 work and 93 validation chemicals, respectively), the real fate differences between these chemicals are not extreme and fate predictions from QSFRs are reasonably correct. Chemicals with structures extremely different than those used in the training of the QSFR models may still have reasonable fate predictions from the latter, as long as their molecular constituents are represented in the training set. Molecular constituent counts have a great advantage: they can be easily retrieved or calculated, known the molecular formula or structural code of new chemicals (e.g., SMILES, InChl, OpenSMILES, etc.); this makes them suitable for simple and rapid fate screenings. Since the molecular formula of a chemical is invariable and SVRs yield the same model given the same training data and parameters (unlike ANNs, which adjust internal parameters in search of a local minimum error), QSFRs using these two features can be reproduced easily and exchanged between modelers, analysts and collaborators.

4.5 Fate predictions from class-tailored QSFRs

QSAR models are expected to perform good predictions for chemicals not used in their training; but, that ideal becomes unpractical when innumerable factors have to be adjusted (Johnson, 2008). QSFRs rely on the same methodology used for developing standard QSARs. So, QSAR and QSFR models tend to yield good results when estimating activity or fate for chemicals with appreciable similarity to the chemicals used in the development of the models. This implies that new chemicals to

ISBN:978-84-693-4597-9 /DL:T.1010-2010

assess must be within the domain of applicability of a model (Weaver and Gleeson, 2008), but this is a discussion postponed for the following section (Section 4.6).

In experiments within the NOMIRACLE project, QSFRs were trained to predict the fate of chemicals in 2-class schemes considering: the SOM algorithm clustering both partitioning and degradation data (Martínez et al., 2008c; Annex A.a3); and, the k-means algorithm clustering water degradation data (Martínez et al., 2008b; Annex A.b3). In this section, the development of class-tailored QSFRs is discussed further with basis on new simulations (Martínez et al. 2010; Annex A.1) considering not only classes derived from either water degradation (Section 4.5.1), but also classes from key molecular features (4.5.2).

4.5.1 Chemical families based on key physicochemical properties.

Chemicals with similar physicochemical properties can have very similar environmental fate behavior and, grounded on this idea, experiments within the NOMIRACLE project were performed, here presented in Annexes A.a3 (Martínez et al., 2008c) and A.b3 (Martínez et al., 2008b). In the first experiment, a SOM mapping both partition coefficients and degradation rates was clustered for creating chemical classes for which individual QSFR models were trained and tested (Martínez et al., 2008c; Annex A.a3), it was found that k_{water} was influencing the development of the classes. So, in the second experiment, chemical classes were created automatically by the K-means algorithm with basis on k_{water} (Martínez et al., 2008b; Annex A.b3). In both experiments it was found that test chemicals that were correctly classified by supervised classifier algorithms got their fate predictions improved when using a class-specific QSFR model instead of a general QSFR model. In the same manner, chemicals that were assigned incorrect classes got highly erroneous fate estimations from the use of improper class-specific QSFR models.

In Section 4.1, the similarity of chemicals emitted in water, with respect to all the inputs and outputs of the reference MEM of the Netherlands (Chapter 3, Annex A.Ca1), was studied with a SOM (Figures 4-2, 4-3 and 4-4), noticing that k_{water} strongly influences the mass ratios in every compartment of the system and leads to the clustering of the SOM into two well-defined sections (Figure 4-4). With such information it is possible to generate two QSFR models compartment, one for chemicals with high degradability in water (Class H) and the other for chemicals with low degradability in water (Class L).

Table 4-3 shows the performances on class-tailored QSFRs predicting fate in air and water from MW and 38 constituent counts (set iii), referred to chemicals with high or low k_{water} values (Classes H or L, respectively). Note that the correct classification of all chemicals has been used for obtaining the performances in Table 4-3 (the actual classes of the validation chemicals are identified by evaluating these chemicals on the SOM). With respect to general QSFRs using the set iii of descriptors (Table 4-2), the class-tailored QSFRs in Table 4-3 yielded improved fate predictions (with respect to

ISBN:978-84-693-4597-9 /DL:T.1010-2010

Table 4-3. SVR prototypes of QSFRs dedicated for chemicals with high (Class H) or low (Class L) k_{water} values, for estimating fate in air and water compartments, considering emissions in water.

		Air compartment		Water compartment	
		$QSFR_{iii,air,H}$	$QSFR_{iii,air,L}$	$QSFR_{iii,water,H}$	$QSFR_{iii,water,L}$
Attributes	set	iii	iii	iii	iii
	total	39	39	39	39
	type*	CC	CC	CC	CC
Number of available chemicals	training+	90[a1]	211[a1]	90[a1]	211[a1]
	test+	41[a1]	33[a1]	41[a1]	33[a1]
	validation	50	43	50	43
SVR parameters	C	300	50	1	300
	γ	0	0	0	0
	ε	$1.0x10^{-2}$	$1.0x10^{-2}$	$1.0x10^{-2}$	$1.0x10^{-1}$
	p	$1.0x10^{-3}$	$1.0x10^{-3}$	$1.0x10^{-3}$	$1.0x10^{-5}$
Support vectors	total	90	210	88	196
Prediction performances on normalized data‡	R^2_{tr}	0.96	0.86	0.56	0.92
	R^2_{te}	0.87	0.78	0.20	0.79
	R^2_{val}	0.61	0.37	0.48	0.51
	MAE_{tr}	0.03	0.07	0.07	0.04
	MAE_{te}	0.08	0.07	0.05	0.05
	MAE_{val}	0.13	0.19	0.10	0.07
10-fold CV on normalized data	R^2_{10CV}	0.81	0.73	0.35	0.66
	MAE_{10CV}	0.10	0.11	0.07	0.07
LOO on normalized data	R^2_{LOO}	0.82	0.73	0.31	0.70
	MAE^2_{LOO}	0.09	0.11	0.07	0.07
Prediction performances on denormalized data	q^2_{tr}	0.96	0.85	0.32	0.92
	q^2_{te}	0.87	0.75	0.15	0.73
	q^2_{val}	0.57	0.35	0.30	0.33
	MAE_{tr}	0.32	0.84	0.28	0.16
	MAE_{te}	0.85	0.75	0.20	0.22
	MAE_{val}	1.50	2.14	0.40	0.29

* Type of input variables: CC = MW and simple counts of molecular constituent.
\+ Chemicals selected with a single SOMs, presented in this Annex: [a1] = Annex D.a1.
‡ Prediction performances obtained during the tuning of the SVR algorithm in each case, presented in Annex D.b1.

performances on denormalized data) for mass ratios in air for chemicals with high k_{water} ($QSFR_{iii,air,H}$); and, mass ratios in water for chemicals with low k_{water} ($QSFR_{iii,water,L}$).

The QSFR approach implies solely the use of molecular information in the absence of reliable physicochemical properties (Figure 1-4). This implies that for a practical application of the models of Table 4-3, it is necessary to predict the class of new chemicals for knowing which model to use in every case. A QSBR model for k_{water} is required. EPIsuite (SRC, 2008) includes linear and non-linear QSBR models of MITI-I degradability tests, respectively, BIOWIN 5 and BIOWIN 6 (Tunkel et al., 2000); so, a simple QSBR has been built for correlating the reference $\log_{10}(k_{water})$ values of the 301 training chemicals (selected with the SOM of Figures 4-2 to 4-4) to the probability degradation predictions of both BIOWIN models as Figure 4-12 indicates. The correlation has the form:

ISBN:978-84-693-4597-9 /DL:T.1010-2010

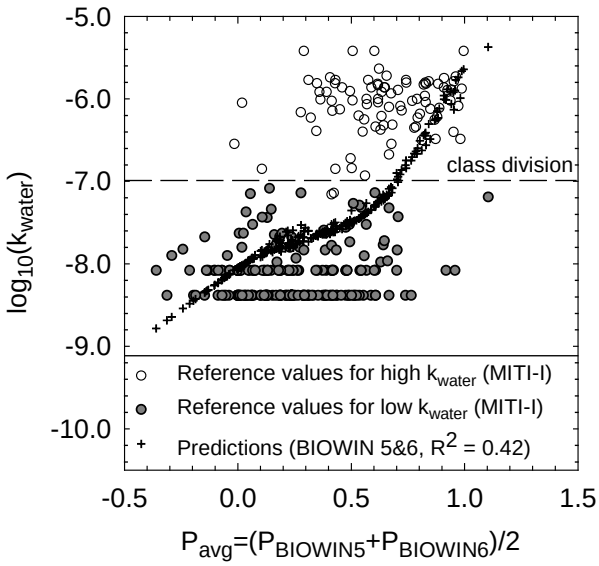


Figure 4-12. Correlation of $\log_{10}(k_{water})$ to degradation probabilities from BIOWIN 5 and BIOWIN 6, for identifying high or low degradability in water.

This figure uses a correlation to separate 301 training chemicals according to their degradability in water, high (H) or low (L).

$$\log_{10}(k_{water}) = 1.02P_{BIOWIN5} + 1.50P^5_{BIOWIN6} - 8.06 \qquad (4\text{-}3)$$

that, despite of its poor correlation coefficient ($R^2 = 0.42$), can be used to identify to which chemical family, or original SOM cluster, a new chemical belongs to. This equation (Equation 4-1) can be used as a simple classification rule: chemicals with $\log_{10}(k_{water}) > -7$ or $\log_{10}(k_{water}) \leq -7$ can be considered to have, respectively, high or low degradability in water as Figure 4-12 shows. Then, for the reference scenario, we can implement ensembles of QSFRs and rules with the form:

$$N_{[-1,1]}(\log_{10}(w_g)) = \begin{cases} f_{QSFR}\left(N_{[-1,1]}(d_1),\ldots,N_{[-1,1]}(d_L)\right)_{g,H} & \text{if} \quad \log_{10}(k_{water}) > -7 \\ f_{QSFR}\left(N_{[-1,1]}(d_1),\ldots,N_{[-1,1]}(d_L)\right)_{g,L} & \text{if} \quad \log_{10}(k_{water}) \leq -7 \end{cases} \qquad (4\text{-}4)$$

for every compartment g, where $\log_{10}(k_{water})$ must be estimated by equation 4-1. Figure 4-13 shows fate prediction of ensembles of this type (Equation 4-4) for the air compartment (Figure 4-13a) and the water compartment (Figure 4-13b), based on the QSFRs developed for chemicals with similar degradability in water of Table 4-3 but predicting the chemical class of a chemicals with Equation 4-1. The improved fate predictions that can be made by training individual QSFRs with chemicals showing similar properties and fate (as Table 4-3) are neutralized by the errors of chemicals wrongly classified, with fate predicted by inappropriate QSFR models as Figure 4-13 shows: most chemicals have predictions very close to their target values, but others have extremely wrong predictions. For the 301 training chemicals, 74 test chemicals and 93 validation chemicals used, the relation of correct-incorrect classified chemicals have been of, respectively, 238 to 63 (79 % to 21 %), 54 to 20 (73 % to 27 %) and 70 to 23 (75 % to 25 %) chemicals.

ISBN:978-84-693-4597-9 /DL:T.1010-2010

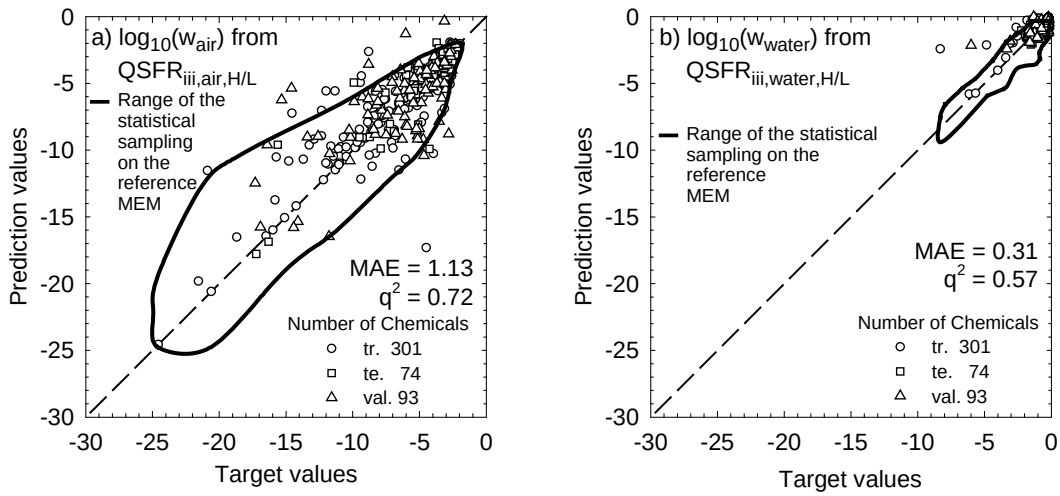


Figure 4-13. Predictions from pairs of specialized QSFRs, based on SVRs using optimal molecular information as input (Set iii) for chemicals with high or low degradability in water, for air (a) and water compartments (b), considering emissions in water.
The fate prediction of a chemical in every compartment can be performed by a QSFR for either high or low degradability in water. However, new chemicals assigned wrong classes can therefore be evaluated with the wrong QSFR model, yielding to highly erroneous predictions.

Figure 4-12 offers an insight of the lack of functionality between very similar degradation values and chemicals with very different degradation probabilities (or, let's say, tracing back relationships to the SMILES notation through BIOWIN 5 & 6, chemicals with very different molecular structures. The functionality between degradability data and molecular information is a problem that remains to be solved (Aronson et al., 2006).

There is the risk of assessing the fate of a new chemical with inappropriate QSFR models if the criterion to define the class of a chemical is simply based on physicochemical properties. Especially with respect to degradation data, chemicals can show similar properties despite of having very different molecular structures; and, thus, the molecular structure of a chemical to assess has a great chance to be out of the domain of applicability of a QSFR. Even if having similar properties, the training chemicals used in the development of a model may contain molecular structures differing greatly from those of new chemicals. This suggests that a better criterion to assess the applicability of available QSFRs should rely on molecular structures rather than on physicochemical properties.

4.5.2 Chemical families based on key molecular features.

The definition of chemical classes depending on exact molecular features, like chemical composition (Martínez et al., 2010; Annex A.1), represents an unambiguous approach for determining the class of a new chemical in the boundaries of available

ISBN:978-84-693-4597-9 /DL:T.1010-2010

classes. Discrete values counting the number of chemical constituents allow the implementation of rules for class predictions with a 100 % of true positives, while continuous values measuring physicochemical properties lead to rules with rates of true positives much lower than 100 % because of the uncertainty of property predictions from QSPRs and QSBRs (Section 4.5.1).

Focusing again on the reference pollution scenario, it was considered the development of QSFRs for chemical classes depending on molecular information. Different criteria could be proposed for creating chemical families with respect to molecular structure, but the performance of any class-tailored QSFR is conditioned by the availability of sufficient training data. In a preliminary screening of the 375 work chemicals of the reference scenario, it was observed that 39 chemicals are composed of solely carbon and hydrogen atoms, while the remaining 336 chemicals have at least one heteroatom (bromine, chlorine, fluorine, nitrogen, oxygen, phosphorus or sulphur atoms). These two groups constitute a starting point for creating two chemical classes, but there is a somewhat unbalanced distribution of chemicals if solely 39 chemicals in the first class are available for creating the training and test data sets of QSFRs. An adjustment can be made to create two chemical families with somewhat similar structure but enough training samples, adding oxygen to the class of chemicals formed with carbon and hydrogen. This way 146 work chemicals are identified to be constituted by carbon, hydrogen or oxygen as the only type of heteroatoms (Class X); while 229 chemicals have a least one heteroatom different than oxygen (Class Y). With this final clustering, a fair class proportion was achieved without sacrificing much with respect to the general properties of the clustered chemicals.

Note that chemicals in class X, having or not oxygen as heteroatom, can be described with a reduced set of descriptors (set iv): MW, 4 atom counts (all atoms, carbon, hydrogen and oxygen), 3 bond counts (all bonds, single bonds and double bonds), 7 functional group counts (aldehyde, carbonyl, carboxyl, ether, hydroxyl, methyl and methylene) and 8 ring counts (all rings, aromatic rings, small rings, 5 membered, aromatic 5 membered, 6 membered, aromatic 6 membered and 7-12 membered). The chemicals in class Y, with any type of heteroatoms, are described with MW and the 38 constituent counts of the set iii of descriptors, Section 4.4). Table 4-4 shows the performances of individual QSFRs predicting fate in air or water from MW and constituent counts, for chemicals of class X and chemicals of class Y.

Based on the same training and test chemicals selected for the models of Figure 4-11 (for the air and water compartment, respectively, $QSFR_{iii,air}$ and $QSFR_{iii,water}$), two QSFRs were developed per compartment, one for class X and the other for class Y. Then, rules for selecting QSFRs, in terms of the presence of heteroatoms (Class X or Y), were implemented for every compartment g (air or water) as follows:

$$N_{[-1,1]}\left(\log_{10}\left(w_g\right)\right) = \begin{cases} f_{QSFR}\left(N_{[-1,1]}(d_1),\ldots,N_{[-1,1]}(d_L)\right)_{g\text{-}X} & \text{if} \quad \text{Class X} \\ f_{QSFR}\left(N_{[-1,1]}(d_1),\ldots,N_{[-1,1]}(d_L)\right)_{g\text{-}Y} & \text{if} \quad \text{Class Y} \end{cases} \tag{4-5}$$

in which a QSFR is selected with an exact criterion based on the amounts of atoms in its molecular formula. Figure 4-14 shows fate predictions from ensembles of QSFRs considering the content or not, of atoms different than carbon, hydrogen or oxygen

ISBN:978-84-693-4597-9 /DL:T.1010-2010

Table 4-4. SVR prototypes of QSFRs dedicated for organic chemicals containing oxygen atoms (Class X) or any type of heteroatoms (Class Y), for estimating fate in air and water compartments, considering emissions in water.

		Air compartment		Water compartment	
		$QSFR_{iv,air,X}$	$QSFR_{iii,air,Y}$	$QSFR_{iv,water,X}$	$QSFR_{iii,water,Y}$
Attributes	set	iv	iii	iv	iii
	total	23	39	23	39
	type*	CC	CC	CC	CC
Number of available chemicals	training+	119[a6]	181[a6]	119[a7]	180[a7]
	test+	27[a6]	48[a6]	27[a7]	49[a7]
	validation	36	57	36	57
SVR parameters	C	300	300	75	0
	γ	0	0	0	1
	ε	$1.0x10^{-3}$	$1.0x10^{-3}$	$1.0x10^{-1}$	$1.0x10^{-3}$
	p	$1.0x10^{-4}$	$1.0x10^{-2}$	$1.0x10^{-3}$	$1.0x10^{-2}$
Support vectors	total	119	165	108	156
Prediction performances on normalized data‡	R^2_{tr}	0.89	0.93	0.89	0.97
	R^2_{te}	0.93	0.91	0.73	0.72
	R^2_{val}	0.68	0.48	0.78	0.23
	MAE_{tr}	0.06	0.05	0.05	0.02
	MAE_{te}	0.06	0.06	0.04	0.07
	MAE_{val}	0.18	0.18	0.09	0.13
10-fold CV on normalized data	R^2_{10CV}	0.74	0.77	0.58	0.47
	MAE_{10CV}	0.12	0.10	0.07	0.12
LOO on normalized data	R^2_{LOO}	0.76	0.78	0.67	0.43
	MAE^2_{LOO}	0.11	0.11	0.07	0.12
Prediction performances on denormalized data	q^2_{tr}	0.89	0.92	0.88	0.97
	q^2_{te}	0.91	0.91	0.71	0.72
	q^2_{val}	0.61	0.44	0.72	0.23
	MAE_{tr}	0.46	0.59	0.20	0.06
	MAE_{te}	0.41	0.70	0.18	0.19
	MAE_{val}	1.30	2.00	0.35	0.39

* Type of input variables: CC = MW and simple counts of molecular constituent.
+ Chemicals selected with specific SOMs, presented in: [a6] = Annex D.a6 and [a7] = Annex D.a7.
‡ Prediction performances obtained during the tuning of the SVR algorithm in each case, presented in Annex D.b1.

(Equation 4-5) for the air compartment (Figure 4-14a) and the water compartment (Figure 4-14b). These models, denominated $QSFR_{iii,air,X/Y}$ and $QSFR_{iii,water,X/Y}$, respectively, yielded better fate predictions than those from simple QSFRs (Figure 4-11), as the application of rules and class-tailored QSFRs (Equation 14) produced higher q^2 and lower MAE values. On average, better fate predictions have been achieved for all chemicals (training, test and validation) in air and in water (Figure 4-14).

Table 4-5 compares q^2 and MAE measurements for fate predictions from molecular information for the air and water compartments resulting from each of the approaches considered in this work: 1000 Monte-Carlo realizations over the reference MEM (MC-MEM) for simulating uncertainty in QSPRs and QSBRs (Figure 4-8), simple QSFRs (Figure 4-11); QSFRs for chemical classes derived from degradability (Figure

ISBN:978-84-693-4597-9 /DL:T.1010-2010

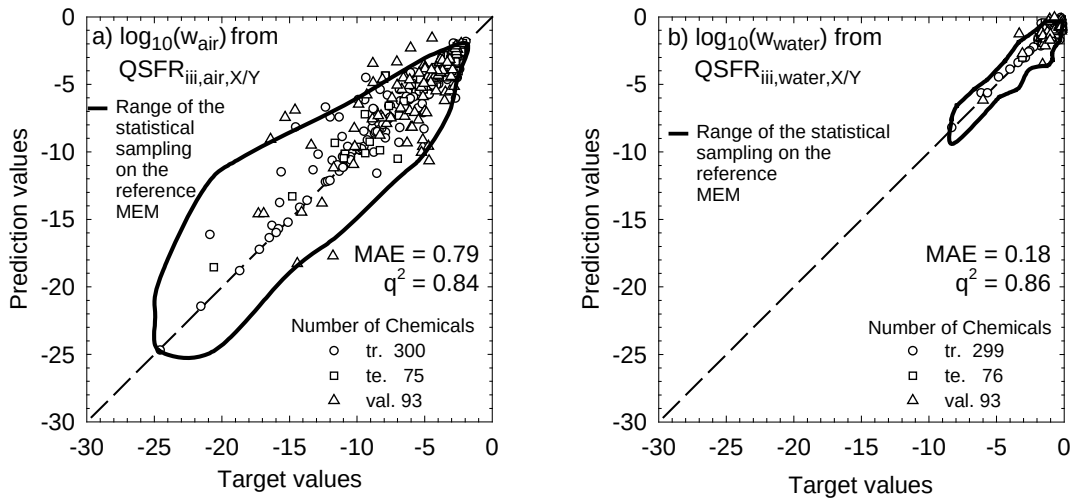


Figure 4-14. Predictions from pairs of specialized QSFRs, based on SVRs using optimal molecular information as input for chemicals of classes X and Y, for air (a) and water compartments (b), considering emissions in water.

Predicting the fate of chemicals with basis on their composition allows accurate class predictions, allowing the use of appropriate class-tailored QSFR models.

Table 4-5. Performance measurements of fate estimation approaches relying on molecular information, for emissions in water.

Compartment	Fate estimation approach	Performance measure	Performances* per data set[0, iii-A, iii-W]			
			Training set (N = 299-301)	Test set (N = 74-76)	Validation set (N = 93)	All sets (N = 468)
Air	MC-MEM	q^2	0.85[a6]	0.87[a6]	0.90	0.87
		MAE	0.85[a6]	0.77[a6]	0.74	0.82
	$QSFR_{iii,air}$	q^2	0.85[a6]	0.86[a6]	0.42	0.78
		MAE	0.81[a6]	0.81[a6]	1.83	1.01
	$QSFR_{iii,air,H/L}$	q^2	0.79[a1]	0.75[a1]	0.44	0.72
		MAE	0.93[a1]	0.98[a1]	1.87	1.13
	$QSFR_{iii,air,X/Y}$	q^2	0.92[a6]	0.91[a6]	0.50	0.84
		MAE	0.54[a6]	0.59[a6]	1.73	0.79
Water	MC-MEM	q^2	0.84[a7]	0.56[a7]	0.87	0.82
		MAE	0.18[a7]	0.19[a7]	0.18	0.18
	$QSFR_{iii,water}$	q^2	0.86[a1]	0.60[a1]	0.63	0.80
		MAE	0.30[a1]	0.34[a1]	0.42	0.33
	$QSFR_{iii,water,H/L}$	q^2	0.63[a7]	0.27[a7]	0.40	0.57
		MAE	0.28[a7]	0.32[a7]	0.41	0.31
	$QSFR_{iii,water,X/Y}$	q^2	0.94[a7]	0.78[a7]	0.60	0.86
		MAE	0.11[a7]	0.19[a7]	0.38	0.18

* q^2 and MAE measurements on logarithmic mass ratios, retrieved from Figure 4-8, Figure 4-11, Figure 4-13 and Figure 4-14. The training and test data sets contain: [a1] 301 training chemicals and 74 test chemicals selected with a SOM based on the set 0 of properties and 5 mass ratios (Annex D.a1). [a6] 300 training chemicals and 75 test chemicals selected with a SOM based on the set iii of descriptors and mass ratios in air (Annex D.a6). [a7] 299 training chemicals and 76 test chemicals selected with a SOM based on the set iii of descriptors and mass ratios in water (Annex D.a7).

ISBN:978-84-693-4597-9 /DL:T.1010-2010

4-13); and QSFRs for chemical classes derived from molecular composition (Figure 4-14). In general, the discrimination and posterior assessment of chemicals with respect to their chemicals composition improved the generalization capability of SVRs linking fate to molecular structure ($QSFR_{iii,air,X/Y}$ and $QSFR_{iii,water,X/Y}$), when compared to simple QSFRs ($QSFR_{iii,air}$ and $QSFR_{iii,water}$) or QSFR for classes depending on properties ($QSFR_{iii,air,H/L}$ and $QSFR_{iii,water,H/L}$) as higher q^2 and low MAE show (Table 4-5). The QSFR models dedicated to specific chemical compositions ($QSFR_{iii,air,X/Y}$ and $QSFR_{iii,water,X/Y}$) performed as well as the reference MEM under uncertainty conditions (MC-MEM). Best overall resulted when assessing the similarity of chemicals in terms of invariable molecular information. The improvements are due to not only the grouping of chemicals with similar composition but also to the 100 % rate of true positives resulting from the class prediction, favoring the evaluation of new chemicals with the most appropriate QSFR model.

4.6 DOA of QSFRs

QSFR models follow the same limitations of QSAR models (Johnson, 2008), for instance, predictions beyond the DOA of the models should be avoided. The DOA of any model is primarily defined by its training chemicals (Weaver and Gleeson, 2008); so, identifying the DOA of an existing QSFR model it is possible to assess, approximately, how appropriate the model is for a new chemical.

Reasonable estimations of the DOA of a model can be performed by measuring distances or probability density distributions of training data vectors to new data vectors (Schroeter et al., 2007), coming either from validation purposes or assessing new chemicals of concern. Since the SOM algorithm is based on the distances between data vectors in a multivariate space (Kohonen et al., 1996), we can use it to define the DOA of QSFR models. As stated in Table 2-3, a work chemical is included in the training data set of a QSFR either when having the lowest or highest mass ratio among all other work chemicals; or, having the lowest or highest quantization error (q_{error}) in the SOM unit. The work chemicals not following such description form the test data set of the QSFR. So, the DOA border of a QSFR model can be defined by the total of training chemicals exhibiting the largest q_{error} with respect to their BMUs, while any chemicals with lower q_{error} values is located within the DOA of the model.

The selection of training chemicals with regard to a SOM is demonstrated in Figure 6, which takes as example three SOM units clustering 3, 4 and 5 work chemicals, respectively: units K15, O2 and U8. This SOM, used for selecting training and test chemicals for $QSFR_{iii,air}$ (Figure 4-11a), is comprised by 40 dimensions (one compartmental mass ratio in air and the set iii of descriptors). Within a single SOM unit, the more similarities between chemicals, in terms of structure and fate, the lower the differences between their q_{error} values; e.g., between dieldrin and endrin in unit K15, or between ethanal, butanal and propanal in unit U8. So, selecting the training chemicals, in each SOM unit, as the ones with the lowest and highest q_{error}

ISBN:978-84-693-4597-9 /DL:T.1010-2010

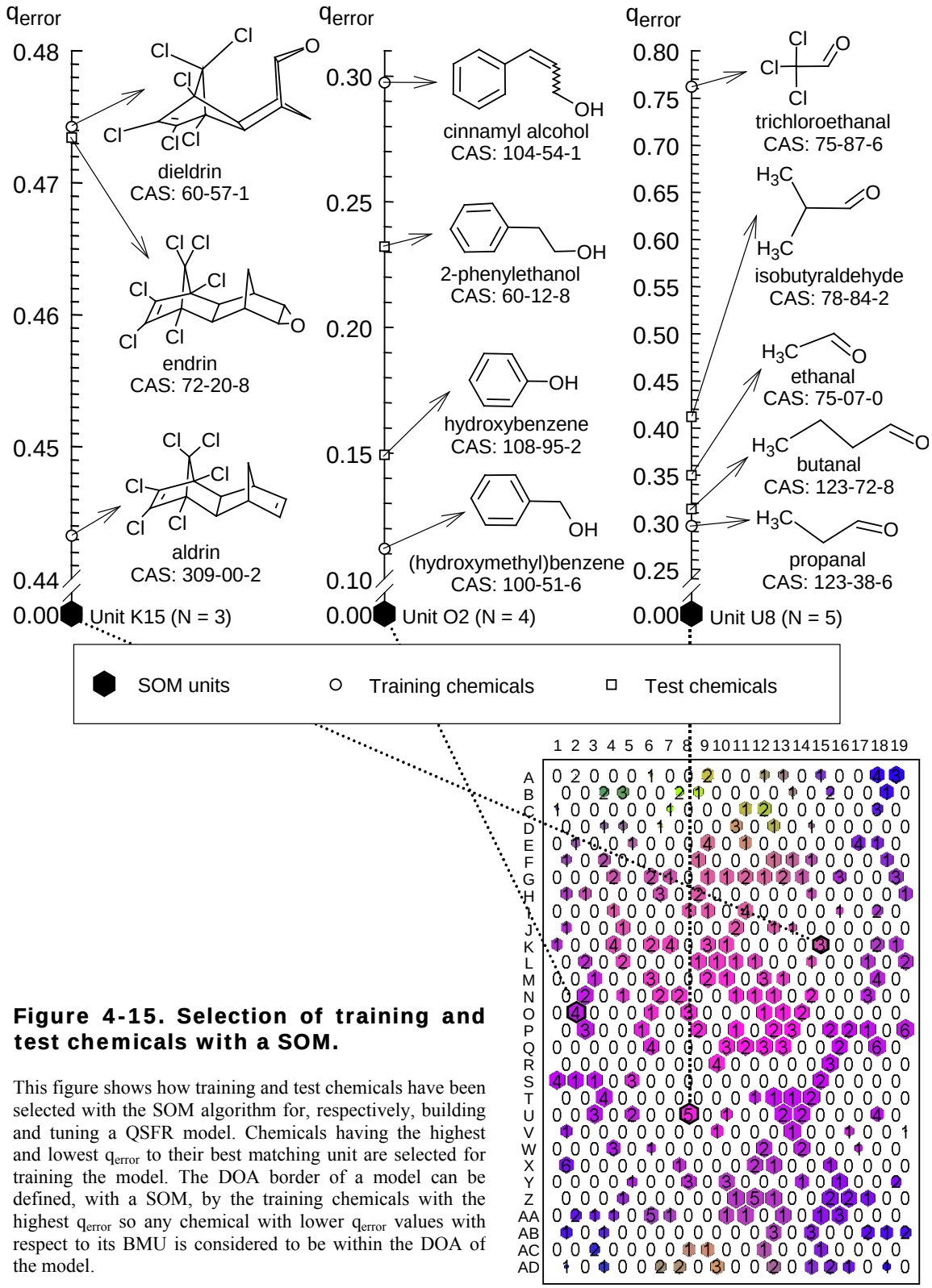


Figure 4-15. Selection of training and test chemicals with a SOM.

This figure shows how training and test chemicals have been selected with the SOM algorithm for, respectively, building and tuning a QSFR model. Chemicals having the highest and lowest q_{error} to their best matching unit are selected for training the model. The DOA border of a model can be defined, with a SOM, by the training chemicals with the highest q_{error} so any chemical with lower q_{error} values with respect to its BMU is considered to be within the DOA of the model.

forces the diversity of the training set and assures the vicinity of the test chemical respect to the training ones. Such vicinity can be practically considered the domain of applicability (DOA) of subsequently trained QSFR models, ranging from every SOM unit to their corresponding farthest clustered training chemicals. Filled SOM units

ISBN:978-84-693-4597-9 /DL:T.1010-2010

Table 4-6. Performance measurements of specialized QSFR models for chemicals in and out the DOAs of the models, in air and water compartments, considering emissions in water.

Model	DOA case	Parameters	Chemicals in DOA			Chemicals out DOA		
			Test	Val	Test & val	Test	Val	Test & val
$QSFR_{air,X/Y}$	I	Chemicals	62	30	92	13	63	76
		q^2	0.93	0.56	0.87	0.70	0.47	0.50
		MAE	0.55	1.11	0.73	0.79	2.03	1.82
	II	Chemicals	36	16	52	39	77	116
		q^2	0.95	0.65	0.89	0.86	0.48	0.59
		MAE	0.45	1.05	0.63	0.73	1.87	1.49
	III	Chemicals	36	12	48	39	81	120
		q^2	0.95	0.78	0.92	0.86	0.47	0.59
		MAE	0.45	0.79	0.54	0.73	1.87	1.50
$QSFR_{water,X/Y}$	I	Chemicals	56	24	80	20	69	89
		q^2	0.84	0.87	0.86	0.57	0.31	0.40
		MAE	0.15	0.31	0.20	0.29	0.40	0.38
	II	Chemicals	44	19	63	32	74	106
		q^2	0.86	0.81	0.83	0.69	0.26	0.44
		MAE	0.13	0.38	0.21	0.26	0.38	0.34
	III	Chemicals	40	12	53	36	81	117
		q^2	0.91	0.94	0.93	0.66	0.28	0.42
		MAE	0.12	0.28	0.16	0.26	0.39	0.35

Table 4-7. Performance measurements of specialized QSFR models for chemicals in and out the DOAs of the models, in air and water compartments, considering emissions in air.

Model	DOA case	Parameters	Chemicals in DOA			Chemicals out DOA		
			Test	Val	Test & val	Test	Val	Test & val
$QSFR_{air,X/Y}$	I	Chemicals	62	30	92	13	63	76
		q^2	0.95	0.41	0.87	0.51	0.41	0.43
		MAE	0.21	0.50	0.31	0.42	0.88	0.80
	II	Chemicals	36	16	52	39	77	116
		q^2	0.97	0.68	0.92	0.84	0.41	0.53
		MAE	0.16	0.40	0.23	0.33	0.83	0.67
	III	Chemicals	36	12	48	39	81	120
		q^2	0.97	0.76	0.94	0.84	0.41	0.53
		MAE	0.16	0.34	0.20	0.33	0.82	0.66
$QSFR_{water,X/Y}$	I	Chemicals	56	24	80	20	69	89
		q^2	0.90	0.86	0.89	0.73	0.42	0.53
		MAE	0.29	0.31	0.29	0.53	0.58	0.57
	II	Chemicals	44	19	63	32	74	106
		q^2	0.92	0.74	0.84	0.81	0.26	0.61
		MAE	0.27	0.53	0.35	0.46	0.51	0.49
	III	Chemicals	40	12	53	36	81	117
		q^2	0.93	0.92	0.92	0.80	0.41	0.61
		MAE	0.25	0.33	0.27	0.46	0.54	0.51

clustering two or more work chemicals contribute with a maximum of two training chemicals; while, SOM units clustering one chemical only make one contribution.

Note that, as explained in Table 2-3, the size of any optimal SOM was set to guarantee a number of training chemicals approximately equal to 80 % of the 375 work chemicals available (per compartment); so, the number of training chemicals for every simple QSFR model in this study is about 300 (Table 4-2).

ISBN:978-84-693-4597-9 /DL:T.1010-2010

The DOA of the best compartmental QSFR models of Table 4-5, $QSFR_{iii,air,X/Y}$ and $QSFR_{iii,water,X/Y}$, were defined according to three different cases:

I) The first approach employs the SOMs used in the selection of training and test data sets. Because the q_{error} of the work chemicals within each SOM unit have been used for selecting the training chemicals, the training chemical with the highest q_{error} defines the DOA border. The original SOM (Figure 5) had 40 dimensions (MW, 38 constituent counts and a mass ratio). When presenting new chemicals to the SOM the mass ratio is assumed unknown, so only 39 out of 40 variables are used for classification purposes (only molecular descriptors), the error of assessing new chemicals with one dimension missing is not significant given the relation 39:1 of available-unavailable dimensions.

II) The second approach employs a new SOM, but applying a principal component analysis (Pearson, 1901) on the 39 molecular descriptors. It was found that five principal components accounted for about 59 % of cumulative variance, so the new SOM was trained with these five principal components and, again, the DOA border was defined with the highest q_{error} of the training chemicals in each SOM unit.

III) The third approach implies the intersection of the first two approaches.

Table 4-6 shows q^2 and MAE performance measurements for models $QSFR_{air,X/Y}$ and $QSFR_{water,X/Y}$, for test and validation chemicals belonging or not to the DOAs defined above. In the first two cases (I and II), test or validation chemicals with quantization errors higher to those of the upper bounding training chemicals are considered to be out the DOA of the models. Since the numbers of chemicals within the DOAs from the first (I) and second (II) cases differ because of the different variables considered and the errors of each SOM, their intersection (III) is preferred because more restrictive conditions are achieved. So, using the third case (III) of Table 4-6, its has been estimated that the fate of about 48 and 53 "new" (test and validation) chemicals can be optimally predicted by, respectively, $QSFR_{air,X,Y}$ (with q^2 = 0.92 and MAE = 0.54) and $QSFR_{water,X,Y}$ (with q^2 = 0.93 and MAE = 0.16). By assessing that new chemicals are within the DOA of a QSFR model, the probability of having acceptable fate predictions is notoriously increased.

All results discussed so far are referred to emissions in the water compartment (Annex D.c2). To check that the QSFR case can be applied to other emission compartments, specific models were tuned for emissions in the air compartment (Annex D.b2) for the same training chemicals of the models in Table 4-4, yielding comparable fate predictions (Annex D.c3). Additionally, Table 4-7 shows q^2 and MAE performances for air-emission models using the same training, test and validation data sets already used for yielding the performances of water-emission models of Table 4-6 with respect to DOAs. Table 4-7 shows similar trends than those in Table 4-6: chemicals within the DOA of every model have more reliable fate predictions than those chemicals out of the DOA.

Concerning the compartment in which emissions take place, another observation can be made. Comparing the performances of environmental fate predictions in air and water, while emitting chemicals in one of these two compartments (Tables 4-6 and 4-7, for emissions in water and air, respectively), it was observed that best predictive

ISBN:978-84-693-4597-9 /DL:T.1010-2010

performances were achieved for a single compartment when emissions occur in itself and not in other compartment. Such trend is confirmed for both the water compartment (considering 53 chemicals within the DOAs in case III: for emissions in water, the performances in water were: q^2 = 0.93 and MAE = 0.16; for emissions in air, the performances in water were: q^2 = 0.92 and MAE = 0.27) and the air compartment (considering 48 chemicals within the DOAs in case III: for emissions in water, the performances in air were: q^2 = 0.92 and MAE = 0.54; for emissions in air, the performances in air were: q^2 = 0.94 and MAE = 0.20) of the scenario considered.

Since these QSFR models are emulators of the MEM used to generate their training data, they inherited its functionality. In Section 4.2, the reference MEM when propagating uncertainty in its input properties showed higher variations in compartments in which emissions were not taking place.

ISBN:978-84-693-4597-9 /DL:T.1010-2010

References

Aronson D, Boethling R, Howard P, Stiteler W. Estimating biodegradation half-lives for use in chemical screening. Chemosphere 2006; 63: 1953.

Basheer IA, Hajmeer M. Artificial neural networks: fundamentals, computing, design, and application. Journal of Microbiological Methods 2000; 43: 3.

Boethling RS, Howard PH, Meylan WM. Finding and estimating chemical property data for environmental assessment. Environmental Toxicology and Chemistry 2004; 23: 2290-2308.

Bredow T, Jug K. Theory and range of modern semiempirical molecular orbital methods. Theoretical Chemistry Accounts: Theory, Computation, and Modeling (Theoretica Chimica Acta) 2005; 113: 1.

Citra MJ. Incorporating Monte Carlo analysis into multimedia environmental fate models. Environmental Toxicology and Chemistry 2004; 23: 1629-1633.

Devillers J. A decade of research in environmental QSAR. SAR and QSAR in Environmental Research 2003; 14: 1 - 6.

Eisenberg JNS, Bennett DH, McKone TE. Chemical Dynamics of Persistent Organic Pollutants: A Sensitivity Analysis Relating Soil Concentration Levels to Atmospheric Emissions. Environ. Sci. Technol. 1998; 32: 115-123.

Gouin T, Mackay D, Webster E, Wania F. Screening Chemicals for Persistence in the Environment. Environ. Sci. Technol. 2000; 34: 881-884.

Hall MA. Correlation-based Feature Selection for Machine Learning. Department of Computer Science. Ph.D. thesis. The University of Waikato, Hamilton, New Zealand, 1999.

Jain AK, Duin RPW, Mao J. Statistical Pattern Recognition: A Review. IEEE Transactions on Pattern Analysis and Machine Intelligence 2000; 22: pp. 4-37.

Johnson SR. The Trouble with QSAR (or How I Learned To Stop Worrying and Embrace Fallacy). Journal of Chemical Information and Modeling 2008; 48: 25-26.

Kawamoto K, MacLeod M, Mackay D. Evaluation and comparison of multimedia mass balance models of chemical fate: application of EUSES and ChemCAN to 68 chemicals in Japan. Chemosphere 2001; 44: 599.

Klöpffer W, Wagner B. Persistence revisited. Environmental Science and Pollution Research 2007; 14: 141.

Kühne R, Breitkopf C, Schüürmann G. Error propagation in fucacity level-III models in the case of uncertain physicochemical properties. Environmental Toxicology and Chemistry 1997; 16: 2067-2069.

Kühne R, Ebert R-U, Schüürmann G. Estimation of Compartmental Half-lives of Organic Compounds - Structural Similarity versus EPI-Suite. QSAR & Combinatorial Science 2007; 26: 542-549.

Limpert E, Stahel WA, Abbt M. Log-normal Distributions across the Sciences: Keys and Clues. BioScience 2001; 51: pp. 341–352.

Mackay D, Shiu W-Y, Ma KC. Illustrated Handbook of Physical-Chemical Properties and Environmental Fate for Organic Chemicals: Lewis Publishers Inc., 1992.

Mager H, Mager PP. Validation of QSARs: Some Reflections. Quantitative Structure-Activity Relationships 1992; 11: 518-521.

Martínez I, Espinosa G, Grifoll J, Cohen Y, Giralt F. Modelling chemical multimedia partitioning with neural networks. SETAC Europe 16th Annual Meeting, The Hague, The Netherlands, 2006a.

Martínez I, Espinosa G, Rallo R, Grifoll J, Cohen Y, Giralt F. A Method for Modeling Chemical Multimedia Partitioning with Neural Networks and Classifiers. AICHE Annual Meeting, San Francisco, United States, 2006b.

Martínez I, Espinosa G, Rallo R, Grifoll J, Cohen Y, Giralt. F. Estimation of environmental multimedia partitioning of pollutants from molecular descriptors using artificial neural networks. SETAC Europe 17th Annual

ISBN:978-84-693-4597-9 /DL:T.1010-2010

Meeting, Oporto, Portugal, 2007a.

Martínez I, Grifoll J, Giralt F, Rallo R, Espinosa G. Report on the feasibility of predicting multimedia chemical partitioning with artificial neural network models by using functional group counts as input information. Universitat Rovira i Virgili, Tarragona, Spain, 2008a. NOMIRACLE Report 2.4.13

Martínez I, Grifoll J, Giralt J, Rallo R and Cohen Y. Multimedia environmental chemical transport and distribution from molecular information. STOTEN. To be submitted in 2010.

Martínez I, Grifoll J, Giralt F, Rallo R, Espinosa G, Cohen Y. Clustering the chemical space to estimate environmental multimedia partitioning of pollutants with Kernel methods and molecular descriptors. SETAC Europe 18th Annual Meeting, Warsawa, Poland, 2008b.

Martínez I, Grifoll J, Rallo R. Cognitive neural network-based intelligent system to identify the most important variables for the differences found in partitioning behaviour, transport pathways and exposure routes between chemicals. Universitat Rovira i Virgili, Tarragona, Spain, 2006c. NOMIRACLE Report 2.4.4

Martínez I, Grifoll J, Rallo R, Giralt F. Report on the most suitable artificial neural network architectures and molecular descriptors to estimate environmental multimedia behavior, including a sensitivity analysis of the effect of compartment sizes on multimedia concentrations. Universitat Rovira i Virgili, Tarragona, Spain, 2007b. NOMIRACLE Report 2.4.9

Martínez I, Grifoll J, Rallo R, Giralt F. Report on the most suitable deterministic and probabilistic algorithms to pre-classify chemicals into families according to their partitioning with the aim of better predicting multimedia concentrations on artificial neural networks for each chemical family. Universitat Rovira i Virgili, Tarragona, Spain, 2008c. NOMIRACLE Report 2.4.12

OECD. Guidance Document on the Validation of (Quantitative) Structure-Activity Relationship [(Q)SAR] Models. OECD Series on Testing and Assessment 69., 2007.

Raymond JW, Rogers TN, Shonnard DR, Kline AA. A review of structure-based biodegradation estimation methods. Journal of Hazardous Materials 2001; 84: 189.

Smola AJ, Schölkopf B. A tutorial on support vector regression. Statistics and Computing 2004; 14: 199.
SRC. EPI Suite v4.00. SRC, 2008.

Taskinen J, Yliruusi J. Prediction of physicochemical properties based on neural network modelling. Advanced Drug Delivery Reviews 2003; 55: 1163.

Toose L, Woodfine DG, MacLeod M, Mackay D, Gouin J. BETR-World: a geographically explicit model of chemical fate: application to transport of [alpha]-HCH to the Arctic. Environmental Pollution 2004; 128: 223.

Tunkel J, Howard PH, Boethling RS, Stiteler W, Loonen H. Predicting ready biodegradability in the Japanese Ministry of International Trade and Industry test. Environmental Toxicology and Chemistry 2000; 19: 2478-2485.

Weaver S, Gleeson MP. The importance of the domain of applicability in QSAR modeling. Journal of Molecular Graphics and Modelling 2008; 26: 1315.

Weininger D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. Journal of Chemical Information and Computer Sciences 1988; 28: 31-36.

Weininger D, Weininger A, Weininger JL. SMILES. 2. Algorithm for generation of unique SMILES notation. Journal of Chemical Information and Computer Sciences 1989; 29: 97-101.

Witten IH, Frank E. Data Mining: Practical machine learning tools and techniques. San Francisco, U.S.: Morgan Kaufmann, 2005.

Wold S. Answer to Mager and Mager. Quantitative Structure-Activity Relationships 1992; 11: 522.

ISBN:978-84-693-4597-9 /DL:T.1010-2010

ISBN:978-84-693-4597-9 /DL:T.1010-2010

Chapter 5
Conclusions

The accuracy of the environmental assessment of chemical pollutants by means of QSFR models is markedly controlled by the training data of the latter. The assessment of new chemicals lying within the domain of applicability of these models is better than the assessment of chemicals not following such rule. Since the availability of training data is critical for the performance of any QSFR with respect to new chemicals, ways for updating the training data of any model should be considered, aiming to enhance the coverage of the known chemical space.

ISBN:978-84-693-4597-9 /DL:T.1010-2010

5.1 Conclusions

It is possible to assess the environmental fate of chemical pollutants from molecular information by two different approaches: first, estimating missing physicochemical properties with QSPR and QSBR models for assessing chemicals with MEMs; and, second, assessing chemicals directly with QSFR models. Whenever the uncertainty in key properties estimated by QSPRs and QSBRs can affect fate predictions from MEM, fate predictions from QSFR models can be a valid alternative as long as the chemicals to assess lie within the DOA of these models.

In this work, it was demonstrated for the reference scenario that:

- It is possible to screen the fate of chemicals under level III conditions by mapping both the inputs and outputs of a MEM in a SOM, a multivariate unsupervised algorithm, for grouping chemicals in terms of their properties and environmental distribution.

- MEM models can perform very uncertain fate predictions when several key properties, like partition coefficients and degradation rates, show large uncertainties.

- QPFR models can perform accurate fate predictions from few physicochemical properties. The shortcoming of these models is that they require as input variables key partition coefficient and degradation rates, which are precisely very difficult to obtain from experiments and literature, making unpractical the QPFR approach.

- QSFR models can perform fate predictions from molecular information, at different levels of accuracy. QSFR models that use as input counts of molecular constituents (atoms, bonds, functional groups and rings) give more accurate fate predictions than QSFR models using theoretical molecular descriptors. Physicochemical properties are solely required for work chemicals, while molecular data are required for both work and new chemicals.

- QSFR models can be tailored to predict the fate of specific chemicals classes, for allowing clearer relationships between chemicals sharing similar behavior. The best way for creating such classes implies the use of invariable molecular information, like chemical composition, instead of physicochemical or molecular properties that can vary due to estimation or measurement procedures. Rules using chemical composition allow class predictions with true positives rates of 100%, allowing the selection of tailored QSFR models when appropriate.

- New chemicals are best predicted by a QSFR model when they lie within the DOA of the model, defined by its training chemicals. For assessing the location of new chemicals in the chemical space with respect to the DOA, procedures involving multivariate Euclidean distances can be employed. The

ISBN:978-84-693-4597-9 /DL:T.1010-2010

SOM algorithm can be used to assess the location of the chemicals while they are characterized by either the same input variables of the QSFR model or a selection of the principal components of such variables.

- Predictions from QSFR models for a given compartment are better when the emissions take place in such compartment, due to a less uncertainty from partitioning and degradation processes in neighboring compartments.

5.2 Applicability of QSFR models

Individual physicochemical properties can be estimated from different QSPR and QSBR approaches for feeding a given MEM. But, when the number of key properties to estimate is elevated, an increase in the uncertainty of the resulting environmental fate predictions must be expected. Every property estimation method propagates certain level of uncertainty into MEMs; therefore, the simultaneous use of several estimated properties in a MEM implies that fate predictions can be affected by high levels of uncertainty (Section 4-2).

For extending the applicability of MEMs to chemicals lacking of several key properties, the establishment of QSFRs constitute a simple, but effective, approach that requires the linkage of fate estimations to molecular information from available training chemicals (Figure 1-3). It must be noticed, that QSFRs should not be considered definitive substitutes of MEMs, as the former must be developed with training data generated, in part, by the latter. The DOA of a QSFR model, like that of any QSAR model, is highly dependent on the training chemicals used in its development. Estimating the fate of a wide range of new chemicals with the QSFR approach requires not only a wide number of training chemicals but also the mapping of wide sections of the chemical space.

Figure 5-1 shows a scheme of how the DOA of a QSAR is located in the existing chemical space. We can say that a region of the chemical space is known when the molecular structures and physicochemical properties of chemicals located in it are practically known. In an analogous matter, the DOA of a QSAR is a subsection of the known chemical space, occupied by the training chemicals of the model. However, since the selection of the training chemicals of a model is a compromise between data availability and modeling criteria, the density of training chemicals within the DOA may vary from point to point affecting the capacity of the model to estimate the activity of chemicals not used in the model training. Wherever the density of training chemicals is high the possibilities of estimating accurately the activity of new, but enclosed, neighboring chemicals is high as well. For this reason, it is crucial to estimate how well the activity of new chemicals can be predicted by checking if they are within the DOA of an available QSAR model.

QSFR models should be viewed as dynamic tools that allow environmental fate estimations from available work data (molecular information and fate estimation examples for a set of work chemicals) that can be updated as more data and better learning algorithms become available. This implies that the applicability of a QSFR

ISBN:978-84-693-4597-9 /DL:T.1010-2010

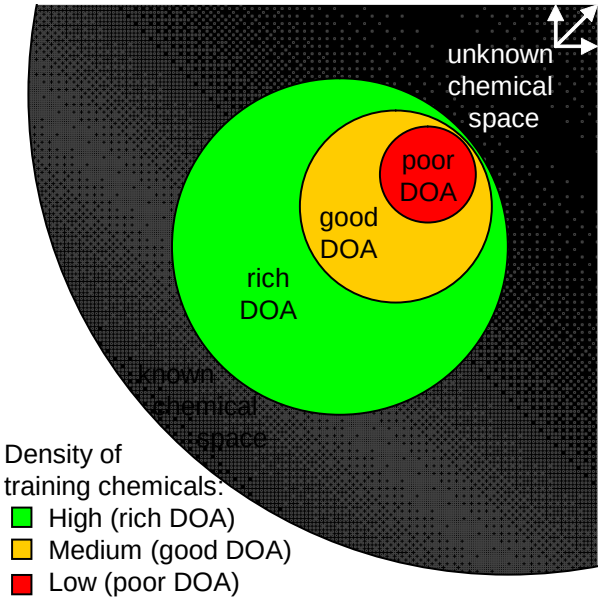


Figure 5-1. The DOA of QSFR models in the chemical space
The DOA of a QSFR model is delimited by its training chemicals, the denser the delimited section (in terms of training data samples) the richer the DOA region and the better the predictions for new chemicals not being part the model and lying in such region.

should be circumscribed to the moment in which a fate assessment is about to be performed and no physicochemical properties are available for a set of chemicals of concern.

5.3 Future work

It has been demonstrated that QSFRs can estimate the final environmental distribution or fate of a chemical pollutant lying within their DOA. The QSFR approach can be further refined as it is used for environmental assessments and better data and algorithms become available. Below, some research lines are proposed for future QSFR developments. They can be carried out sequentially or simultaneously.

Make the QSFR approach available to the average environmental modeler. The next step in the evolution of the QSFR approach is to have this methodology available for standard multimedia environmental assessments. For achieving it, the QSFR approach should be implemented in a way that any user (modelers, decision-makers, regulators, etc.) can exploit its advantages with little training and data manipulation.

Nowadays, open source software packages, with simple graphical user interfaces, are available for molecular modeling (Geldenhuys et al., 2006) and data mining with supervised and unsupervised learning algorithms (Mierswa et al., 2006; Witten and Frank, 2005). Such tools are free to use by anyone that understands how they work, while paid software packages offer extra functions and capacities. With some additional programming, both molecular and data mining software could be linked to standard MEMs for allowing QSFR-based fate predictions in situ by any user. In such case, both a graphical user interface and a standardized routine should be available for guiding inexperienced users to estimate the environmental distribution of chemicals with QSFRs. If few physicochemical properties for a chemical are missing, the

ISBN:978-84-693-4597-9 /DL:T.1010-2010

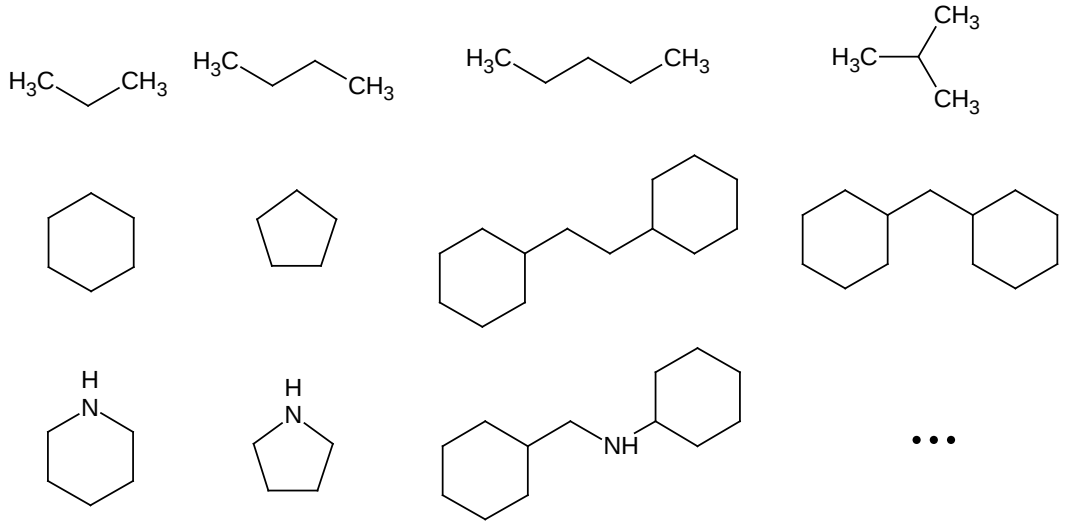


Figure 5-2. Scheme of possible molecular frameworks for creating class-tailored QSFRs.

If the common framework of several chemicals, conformed by chains and rings, is used to create class-tailored QSFR models the capacity of these models can be used to discriminate the environmental fate of new chemicals from small structural differences rather than from large ones.

QSPR/QSBR approach could be used for estimating them and later assessing the chemical with the MEM of preference.

Implementing a QSFR-based fate estimation routine as a plug-in or embedded code in standard MEMs can reduce modeling time, facilitating the assessment of chemicals in several scenarios. This would also facilitate further research for continuing the evaluation and improvement of the QSFR approach as new modeling techniques and data become available.

Enhance the DOA of the QSFR models. The DOA of a model is defined by its training data set. The wider and denser the DOA of a QSFR model the higher the chances of predicting properly the fate of new chemicals, especially when they lie within the DOA. For these reason it is of capital importance to collect reliable physicochemical properties for training chemicals, they will be used later in a MEM for generating examples of environmental fate or distribution, the target variables of QSFRs. A great effort should be carried for setting databases of well known chemicals to contain every physicochemical property determined experimentally under the same conditions for all chemicals.

Another way of enhancing the DOA of QSFR models is restricting them to very specific chemical classes, tied to the availability of physicochemical properties for the training chemicals of a class of concern. It is known that the molecular structures of a great number of chemicals share identifiable framework shapes (Lipkus et al., 2008) and that the synthesis of new chemicals from them is still possible like, for example, the case of heteroaromatic rings (Pitt et al., 2009). So, the QSFR approach could be

ISBN:978-84-693-4597-9 /DL:T.1010-2010

specifically applied to chemical classes defined by well known framework shapes, like shown in Figure 5-2, letting learning models differentiate chemicals through small, rather than huge, structural differences.

Compile physicochemical properties in universal databases for their use in MEMs. Current datasets of physicochemical properties are referred to chemicals under conditions that may differ from those required by current MEMs. For these reason, standard MEMs contain databases with both experimental and estimated properties compiled for limited sets of chemicals, limited to the MEM for which the latter have been compiled for. For assessing new chemicals, there are several methods for estimating partitioning properties (Boethling et al., 2004) but the availability of both experimental and estimated degradation data is still poor (Aronson et al., 2006; Klöpffer and Wagner, 2007; Kühne et al., 2007). The development of a universal and updatable database would significantly improve not only the applicability of MEMs but also the applicability of QSFRs, extending the DOA of these models. At present, such universal database may seem highly idealized, but its applicability would be beyond any doubt.

Perform further research on the input information to use in QSFRs. For every new model it is necessary to design, compute and select molecular descriptors. Molecular descriptors counting the number of constituents (atoms, bonds, functional groups and rings) were found to be a better source of information for QSFR models than semi-empirical molecular descriptors describing average molecular properties (Section 4.4). This represents a clear advantage over models using as input semi-empirical descriptors, as these usually vary depending on the specific MO method (Bredow and Jug, 2005) used to estimate them. Constituent counts can be easily computed when molecular structure is known. However, when using constituent counts as molecular descriptors, the environmental fate of some isomeric chemicals cannot be distinguished as they may happen to have the same descriptors and somewhat different behavior.

There is a recent research trend in the field of QSARs that aims to replace the use of molecular descriptors directly by molecular structures, represented as graphs (Goulon et al., 2007). It proposes the use of graph machines, which implies that for each example in a data set a mathematical function (graph machine) is built, reflecting the structure of the molecule under consideration; it is the combination of identical parameterized functions, like, for example, feed forward neural network.

The sections of a molecule can provide relevant information about its tendency to distribute in the environment. So, it would be interesting to investigate the effect of replacing molecular descriptors by graph machines as inputs to QSFRs. It should be expected an increase in the generalization capacity of the models as more relevant structural information could be available.

ISBN:978-84-693-4597-9 /DL:T.1010-2010

References

Aronson D, Boethling R, Howard P, Stiteler W. Estimating biodegradation half-lives for use in chemical screening. Chemosphere 2006; 63: 1953.

Boethling RS, Howard PH, Meylan WM. Finding and estimating chemical property data for environmental assessment. Environmental Toxicology and Chemistry 2004; 23: 2290-2308.

Bredow T, Jug K. Theory and range of modern semiempirical molecular orbital methods. Theoretical Chemistry Accounts: Theory, Computation, and Modeling (Theoretica Chimica Acta) 2005; 113: 1.

Geldenhuys WJ, Gaasch KE, Watson M, Allen DD, Van der Schyf CJ. Optimizing the use of open-source software applications in drug discovery. Drug Discovery Today 2006; 11: 127.

Goulon A, Picot T, Duprat A, Dreyfus G. Predicting activities without computing descriptors: graph machines for QSAR. SAR and QSAR in Environmental Research 2007; 18: 141.

Klöpffer W, Wagner B. Persistence revisited. Environmental Science and Pollution Research 2007; 14: 141.

Kühne R, Ebert R-U, Schüürmann G. Estimation of Compartmental Half-lives of Organic Compounds - Structural Similarity versus EPI-Suite. QSAR & Combinatorial Science 2007; 26: 542-549.

Lipkus AH, Yuan Q, Lucas KA, Funk SA, Bartelt WF, Schenck RJ, et al. Structural Diversity of Organic Chemistry. A Scaffold Analysis of the CAS Registry. J. Org. Chem. 2008; 73: 4443-4451.

Mierswa I, Wurst M, Klinkenberg R, Scholz M, Euler T. YALE: Rapid Prototyping for Complex Data Mining Tasks. In: Ungar L, Craven M, Gunopulos D, Eliassi-Rad T, editors. Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-06). ACM, Philadelphia, PA, USA, 2006, pp. 935-940.

Pitt WR, Parry DM, Perry BG, Groom CR. Heteroaromatic Rings of the Future. Journal of Medicinal Chemistry 2009; 52: 2952.

Witten IH, Frank E. Data Mining: Practical machine learning tools and techniques. San Francisco, U.S.: Morgan Kaufmann, 2005.

ISBN:978-84-693-4597-9 /DL:T.1010-2010

ISBN:978-84-693-4597-9 /DL:T.1010-2010

Annex A

Research works on QPFRs and QSFRs

ISBN:978-84-693-4597-9 /DL:T.1010-2010

Multimedia environmental chemical transport and distribution from molecular information

Izacar Martínez[1], Jordi Grifoll[1], Francesc Giralt[*,1], Robert Rallo[2] and Yoram Cohen[3].

[1] Departament d'Enginyeria Quimica, Grup de Recerca de Fenomens de Transport, Universitat Rovira i Virgili, Av. Paisos Catalans, 26, 43007 Tarragona, Catalunya, Spain
[2] Departament d'Enginyeria Informatica i Matematiques, Grup de Recerca de Fenomens de Transport, Universitat Rovira i Virgili, Av. Paisos Catalans, 26, 43007 Tarragona, Catalunya, Spain.
[3] Department of Chemical and Biomolecular Engineering, University of California (UCLA), 5531 Boelter Hall, 405 Hilgard Avenue, Los Angeles, CA 90095, United States.
* Corresponding author: Telephone number: +34 977559638; Fax number: +34 977559621; E-mail address: fgiralt@urv.cat (Francesc Giralt).

To be submitted in 2010 to STOTEN

Abstract

Except for common priority chemical pollutants of current concern, environmental key physicochemical properties tend to be unavailable to a wide spectrum of chemicals. This paper analyses the prospect of assessing the environmental distribution of chemicals directly from their molecular information rather than from multimedia models using several physicochemical properties estimated from QSARs. To this end, predictions of chemical partitioning, expressed in dimensionless compartmental mass ratios, have been compared between: a) SimpleBox 3, a Level III Fugacity model, propagating reported uncertainty of key physicochemical properties via statistical sampling; and, b) support vector regressions acting as quantitative structure fate relationships (QSFRs), predicting mass ratios from a set of molecular descriptors comprised by MW and 38 counts of molecule constituents (atoms, bonds, functional groups and rings). These assessments comprised 455 chemicals (including priority chemicals) emitted in a single medium (air or water), in a fixed geographical scenario representing the Netherlands as a set of five compartments (air, water, sediments, soil and vegetation). Out of the 455 chemicals, 375 were used for training and testing QSFR models, while 80 were reserved for the external validation of the models. Training and test chemicals were selected from the set of 375 working chemicals by means of the self-organizing map (SOM) algorithm. Clustering chemicals into classes concerning their molecular composition, the performance of class-tailored QSFRs improved. Additionally, the domain of applicability (DOA) of these models, conformed by their training chemicals, was assessed with SOMs, to demonstrate that mass ratios of new chemicals (test and validation) within the DOAs are well predicted (in air: q^2 = [0.89 , 0.94], MAE = [0.20, 0.69]; in water: q^2 = [0.84, 0.94], MAE = [0.15, 0.35]) compared to those of outlying chemicals (in air: q^2 = [0.68, 0.75], MAE = [1.12, 1.33]; in water: q^2 = [0.38, 0.43], MAE = [0.33, 0.36]).

Keywords: Multimedia environmental model; uncertainty analysis; quantitative structure fate relationships; molecular descriptors; support vector regression; domain of applicability.

ISBN:978-84-693-4597-9 /DL:T.1010-2010

1. Introduction

Multimedia environmental models (MEMs) are routinely used to estimate the environmental distribution of chemical pollutants based on their physicochemical properties, site-specific parameters and emission rates (Cohen, 1986; Mackay, 2001; Cohen and Cooter, 2002a, 2002b). In addition to geographic site-parameters (Cohen, 1986; Cohen and Cooter, 2002a; Mackay, 2001) and emission rates and sources (Breivik et al., 2004; Breivik et al., 2006; Lohmann et al., 2007; Cohen and Cooter, 2002a), MEMs serve to screen chemicals with respect to their persistence in the environment and to provide information needed to estimate the exposures and associated risks to human and ecological receptors.

The reliability of predictions of chemical partitioning from MEMs are affected by model formulation (i.e., system definition, included environmental processes, calculation methods, etc.) and the uncertainties introduced via model parameters (Webster et al., 2004) including estimates of physicochemical parameters (Breivik and Wania, 2003; Cohen and Cooter, 2002a, 2002b). In particular, uncertainty in partitioning and degradation parameters can significantly affect MEM predictions (Citra, 2004; Eisenberg et al., 1998; Kawamoto et al., 2001; Kühne et al., 1997; Toose et al., 2004). Even small structural differences can lead to large differences in chemical activity (Nikolova and Jaworska, 2003). Therefore, it is imperative to develop reliable methods for estimating chemical physicochemical properties with careful considerations of data quality and diversity (Furusjö et al., 2006), and accurate discrimination of chemical descriptors that serve to characterize the chemicals (Cronin and Schultz, 2003, Stouch et al., 2003).

The lack of adequate physicochemical and toxicological information for most commercial chemicals and the risk that they may represent for human health and the environment has motivated the development of new regulatory efforts (Tickner et al., 2005) such as REACH in the European Union and the Inventory Update Rule (US-EPA, 2006) in the United States. These rules aim to collect information about the characteristics, emission rates and existing volumes of commercial chemicals for facilitating their screening and deciding whether to authorize or ban their production. Compiling all mandatory data will be a formidable task given the large number of chemicals that may be of concern. For example, in September 2009, the CAS registry, one of the largest substance registry databases, reported its 50-millionth unique chemical (Toussant, 2009). It is accepted that the regulatory assessments of the multimedia distribution of chemicals for which physicochemical properties are lacking will require the use of estimation methods that rely on quantitative structure activity relationships (QSARs) (Fjodorova et al., 2008; Worth et al., 2007).

QSARs are accepted worldwide in standard environmental assessments and decision-making tasks (Cronin et al., 2003; Walker et al., 2002). QSARs are based on establishing quantitative relations between the target physicochemical (Hugo, 2002), or toxicological properties (Devillers, 2003; Mackay et al., 2003; Mackay and Webster, 2003) of chemicals and their molecular information. However, uncertainties are often associated with the use of QSARs, especially for chemicals that deviate in their molecular structure from those used in the QSAR development (Taskinen and Yliruusi, 2003). In general, partitioning data (Boethling et

ISBN:978-84-693-4597-9 /DL:T.1010-2010

al., 2004; Mackay, 2000) are more readily available (from experiments or estimations) relative to degradation data (Aronson et al., 2006; Howard et al., 1991; Klöpffer and Wagner, 2007; Raymond et al., 2001). Selecting appropriate chemical descriptors is crucial for the development of accurate QSARs as demonstrated, for example, for vapor pressure (Godavarthy et al., 2006; Yaffe and Cohen, 2001), water solubility (Yaffe et al., 2001), Henry's law constant (Modarresi et al., 2007; Yaffe et al., 2003) and octanol-water partition coefficient (Yaffe et al., 2002). QSAR development must consider the selection of model input features (Saeys et al., 2007), often from a large number of descriptors (Bredow and Jug, 2005; Burden et al., 2009; Duca and Hopfinger, 2001; Senese et al., 2004; Todeschini and Consonni, 2000), the selection and tuning of learning algorithms for building relationships (Basheer and Hajmeer, 2000; Xu et al., 2006), the risk of overtraining (Byvatov et al., 2003), the external validation of the models (Golbraikh and Tropsha, 2002; OECD, 2007; Schüürmann et al., 2008) and the definition of applicability domains (Weaver and Gleeson, 2008).

There are essentially two possible approaches to estimate the set of chemical properties required for modeling the environmental multimedia distribution of chemicals. The first is to estimate the properties of each required chemical parameter from independent QSPR models. The second is to consider a single QSPR for the collective chemical properties whereby given a set of chemical descriptors the various environmentally relevant physicochemical properties and reaction rate parameters are predicted by the single QSPR. However, different levels of uncertainty can be present in any of these approaches.

There is the need of assessing the fate of chemicals when physicochemical properties are unavailable or extremely noisy, even when using QSPR-based estimation approaches. For this reason, an alternative approach can be conveniently employed when a given regulatory multimedia model is used for a given emission scenario for specific geographical and meteorological settings. Such approach is usually expected to be linked to the molecular structure of chemicals. Preliminary proposals have considered the implementation of QSPRs in standard MEMs (Breivik and Wania, 2003; Zukowska et al., 2006) or the establishment of structure fate relationships by partial orders (Brüggemann et al., 2006) . Here, we propose the training of machine-learning models (Witten and Frank, 2005) to map directly output of MEMs (in terms of chemical concentrations or media mass distribution) to relevant chemical descriptors. The resulting correlation model, which is referred to herein as a quantitative-structure-fate-relation (QSFR), has the advantage of providing direct information on the environmental distribution of chemicals using a consistent set of chemical descriptors with respect to chemically relevant multimedia model properties.

Note that the term environmental fate is often associated to the processes by which chemicals move and are transformed in the environment, but it has also been associated only to the transformation processes. In this later case, the first meaning is referred as fate and transport. In this paper we tried to avoid the use of this term, but we have included it, in its first meaning, to identify the kind of activity we try to describe. So, the QSARs developed here have been called quantitative structure fate relationships (QSFR).

ISBN:978-84-693-4597-9 /DL:T.1010-2010

The present paper reports on the prospect of assessing the environmental fate of chemicals directly from their molecular information (using QSFRs trained with learned MEM model output) instead of a MEM using all properties estimated by QSPRs and QSBRs (Figure 1). To this end, environmental chemical distributions for a set environmental scenario were compiled (for a set of 455 chemicals) using SimpleBox 3 (SB3) (Brandes et al., 1996; den Hollander and van de Meent, 2004; den Hollander et al., 2004; van de Meent, 1993), considering the range of values of key physicochemical properties (Boethling et al., 2004; Kühne et al., 2007) via statistical sampling. The mass distribution of a set of working chemicals, expressed in mass ratios, along with a selected set of chemical descriptors were then employed to derive QSFR models based on support vector regressions (SVR) (Drucker et al., 1996). Figure 1 depicts these two possible approaches: predict chemical properties from QSPRs and chemical descriptors to feed a multimedia model to obtain final concentrations or predict these concentrations directly from chemical descriptors using QSFRs. Of course, these QSFRs have to be developed by means of data obtained following the first approach for well known chemicals. The QSFR approach was contrasted with predictions from the SB3 model affected by variations in its input physicochemical properties. This study demonstrated that the environmental distribution of chemicals not used to develop the models can be reasonably predicted by QSFRs when these new chemicals to assess lie within the domain of applicability (DOA) of the latter.

2. Scenario for chemical multimedia distributions

2.1. Multimedia model

Multimedia environmental simulations were carried out, using the Level III (steady state with mass transfer limitations) fugacity model SB3 (Brandes et al., 1996; den Hollander and van de Meent, 2004; den Hollander et al., 2004; van de Meent, 1993), to assess the multimedia distribution of 455 chemicals (Martínez, 2010) in the Netherlands as a model environment represented for a reference emission rate of 1 ton/yr in a specific medium. A total of 375 working chemicals were used for training and testing QSFR models, while 80 chemicals were reserved for model validation.

Using site-specific parameters previously reported for the Netherlands (Struijs and Peijnenburg, 2002), this geographic region was described for SB3 (den Hollander and van de Meent, 2004; den Hollander et al., 2004) usage by a set of 5 homogeneous compartments at the regional scale of this MEM: air, water (including fresh and sea water), sediments (including fresh water sediments and sea water sediments), soil (including natural, agricultural and other soil) and vegetation (including natural and agricultural vegetation).

The steady state compartmental chemical mass distributions calculated from the SB3 model are expressed as the dimensionless mass ratio of the chemical mass in the compartment relative to the total amount of the chemical, m_t (g) emitted over a period of one year:

$$w_{n,g} = \frac{C_{n,g} V_g}{m_t} \quad (1)$$

where $C_{n,g}$ (g/m^3) is the steady state concentration of a pollutant n in compartment g of volume V_g (m^3). It is noted that in the present steady-state (Level III) model, the variation of mass partitioning among the different chemicals is governed

ISBN:978-84-693-4597-9 /DL:T.1010-2010

only by their physicochemical, transport and degradation constants since all other parameters are invariant.

2.2 Physicochemical properties

The SB3 model requires a total of 6 physicochemical, 2 transport and 4 degradation parameters for Level III type simulations. The physicochemical parameters included molecular weight (MW, g/mol), melting point (T_m, K), vapor pressure (P_v, Pa), octanol-water partition coefficient (K_{ow}, dimensionless), air-water partition coefficient (K_{aw}, dimensionless), and the solid-water partition coefficient (K_{sw}, dimensionless). The chemical degradation parameters in air (k_{air}, 1/s), water (k_{water}, 1/s), sediment (k_{sed}, 1/s), and soil (k_{soil}, 1/s) media were all for first-order kinetics, and the fundamental transport coefficients were the mass diffusivity of the chemical in air (D_{air}, m^2/s) and water (D_{water}, m^2/s). SB3 uses T_m and P_v to calculate internally the air-aerosol partition coefficients according to Junge (1977).

Some parameters (MW, T_m, P_v, and K_{ow}) have been directly retrieved from the PHYSPROP database (SRC, 2008), while a set of parameters (K_{aw}, K_{sw}, D_{air}, D_{water}, k_{air}) has been estimated from data in such database and another set of parameters (k_{water}, k_{sed}, k_{soil}) has been estimated from MITI-I biodegradability tests (NITE, 2006).

K_{aw} values were estimated from Henry's law constants values divided by the ideal gas constant (8.314 J/(mol·K)) and the reference temperature (298.15 K) (Mackay, 2001). Assuming for solids an average organic carbon content of 2 % and a solid soil density of 2.5 kg/L, K_{sw} values were estimated from K_{ow} values (European Commission, 2003). D_{air} and D_{water} values were estimated considering that diffusivity coefficients vary inversely with the square root of the MW and using as references the diffusion coefficient of water in air and the diffusion coefficient of oxygen in water (Schwarzenbach et al., 2003).

Air degradation was considered a result of reaction of chemicals with hydroxyl radicals at a rate given by:

$$r_{air} = k_{OH\cdot} C_{OH\cdot} C_{n,air} \quad (2)$$

where r_{air} ($g/m^3 \cdot s$) is the degradation rate in air, $k_{OH\cdot}$ ($m^3/g \cdot s$) is the second-order reaction constant (SRC, 2008) and $C_{OH\cdot}$ (g/m^3) is the concentration of hydroxyl radicals in air. Considering a global average concentration of hydroxyl radicals of $C_{OH\cdot} = 2.66 \times 10^{-11}$ g/m^3 (Prinn et al., 2001), pseudo first-order degradation rate constants has been calculated from:

$$k_{air} = k_{OH\cdot} C_{OH\cdot} \quad (3)$$

The degradation rate constant in water, k_{water}, was estimated from results of MITI-I biodegradability tests (NITE, 2006). The MITI-I tests are expressed as the degradation percentage of chemical samples (deg%) over time periods (t) ranging from 2 to 4 weeks, with sample mass determined by direct methods (using total organic carbon, high performance liquid chromatography and gas chromatography) and indirect methods (measuring biological oxygen demand). In the current work, k_{water} values were estimated as follows:

$$k_{water} = \left(\frac{-1}{t}\right) \ln\left(1 - \frac{deg\%}{100}\right)\left(\frac{1}{604800}\right) \quad (4)$$

where t (weeks) is the range period of a test and deg% the degradability percentages determined by the biological oxygen demand (BOD) methodology. Only compounds for which their degradation percentage between the BOD method and the total organic carbon method has been within 10% were included

ISBN:978-84-693-4597-9 /DL:T.1010-2010

into the working and validation chemical sets. Results from chromatographic techniques were not used because they have not been found as reliable as results using BOD (Sedykh and Klopman, 2007). For modeling consistency, when using Equation 4 all deg% values experimentally reported to be higher than 100 % or lower than 0 %, due to error measurements in the MITI-I tests, have been set to be equal to, respectively, 99 % (extremely fast degradability) or 1% (extremely low degradability).

Noticing that Aronson and Howard (1999) indicated that degradation half lives in water are similar to those in soil and that degradation rates in soil tend to be 3 to 4 times faster that degradation rates in flooded soil, in this study k_{soil} values were estimated to be equal to k_{water} values while k_{sed} values were assumed to be 3.5 times slower than k_{soil} (considering the flooded soil as a surrogate of the sediment compartment).

2.3 Molecular information

Molecular information consisting of 39 molecular descriptors was compiled for each of the 455 study chemicals by means of the CACHE molecular simulations package (Fujitsu, 2004). The set of 39 molecular descriptors included molecular weight, 10 atom counts (all atoms, bromine, carbon, chlorine, fluorine, hydrogen, nitrogen, oxygen, phosphorus, and sulfur), 4 bond counts (all bonds, single bonds, double bonds and triple bonds), 16 group counts (aldehyde, amide, amine, sec-amine, carbonyl, carboxyl, cyano, ether, hydroxyl, methyl, methylene, nitro, nitroso, sulfide, sulfone, and thiol), 8 ring counts (all rings, aromatic rings, small rings, 5 membered, aromatic 5 membered, 6 membered, aromatic 6 membered and 7-12 membered).

3. Methods

3.1 Uncertainty assessment of the MEM

For simulating the effect of uncertainties in physicochemical properties, as estimated from QSPRs or QSBRs, on the resulting chemical distribution in the environment, a series of SB3 model simulations were carried out for all 455 chemicals applying 1000 random combinations (Monte Carlo simulations) of the following independent chemical properties: T_m, P_v, H, K_{ow}, k_{air} and k_{water}. Because K_{aw}, K_{sw}, k_{sed}, k_{soil} are estimated properties, they vary as result of the variation of the independent properties. Finally, D_{air} and D_{water} are not subject of variation because they have been estimated from MW.

The uncertainty sources, in terms of statistical distributions, assigned to the varying independent properties are listed in Table 1. For T_m, P_v, H, K_{ow} standard deviations were taken from statistics of widely recommended QSPRs (Boethling et al., 2004), considering results for external validation chemicals where possible. For k_{air} and k_{water} the statistical distributions were taken from QSBRs (Kühne et al., 2007). It has been assumed that the mean value of every distribution coincides with the property value compiled as described in Section 2.2. Finally, it has been assumed that a variable follows a normal distribution if the standard deviation given by Boethling et al. (2004) is in unit variables. When the standard deviation is given in logarithmic units, a lognormal distribution has been considered. Although the standard deviation of P_v is given in terms of mmHg, a lognormal distribution has been used to avoid negative values in chemicals with very low P_v.

ISBN:978-84-693-4597-9 /DL:T.1010-2010

The outputs of the SB3 model from 1000 random combinations for each chemical in terms of dimensionless mass ratios, as schematized in Figure 2 for Endrin, were used to generate a database. This database provided an estimation of the output distribution that one can expect when using recommended QSPRs and QSBRs to estimate the environmental distribution of chemicals. This database was used as a reference for comparing the predictions of the QSFR approach depicted in Figure 1.

3.2 QSFR model development

In this study, QSFRs have been developed to estimate the output of the SB3 for each compartment of the reference pollution scenario. It is expected that these QSFR models will perform better than or at least in a similarly to the SB3 model when fed with properties estimated from several QSPRs and QSBRs. The QSFR relates the chemical mass ratio w_g in a specific environmental compartment, to the chemical's set of molecular descriptors $d_1,...,d_L$.

Fundamentals. Given N chemicals (characterized by K properties) emitted in a geographic region described by G compartments, a reference MEM can be considered to be a multivariate function of the form:

$$\mathbf{C} = f(\mathbf{P},\mathbf{E},\mathbf{S}) \quad (5)$$

where **C** is a matrix of mass ratio predictions of size NxG**, P** is a matrix of physicochemical properties of size NxK, **E** is a matrix of emission rates of size NxG and **S** is a matrix of site-specific parameters. When **E** and **S** remain constant, the chemical distribution in the environment can be solely analyzed in terms of **P**, the collection of physicochemical properties of chemicals to assess.

When key physicochemical properties are unavailable for chemicals of concern (**P** is unknown), and alternative multimedia environmental models can be developed, as explained below, from L molecular descriptors in a matrix **D** (of size NxL) by means of QSFRs of the form:

$$\mathbf{C} \approx f_{\mathrm{QSFR}}(\mathbf{D})$$

In order to develop the QSFR model as expressed by Eq. (6), a set of N_{tr} training chemicals (with $N_{tr} < N$) is required for which all properties and molecular structures are known. The model is then adjusted to emulate the output of the reference MEM (Eq. 5), by tuning its internal parameters with respect to a set of N_{te} test chemicals. Its performance on new chemicals is later evaluated with a set of N_{val} validation chemicals.

Data pre-processing. All input and output variables with values that span more than two orders of magnitude were logarithmically (base 10) scaled and then normalized in the range [-1,1] as follows:

$$N_{[-1,1]}(y_i) = 2\left(\frac{y_i - y_{min}}{y_{max} - y_{min}}\right) - 1 \quad (7)$$

where y_i is a value to be normalized and $N_{[-1,1]}(y_i)$ is its normalized counterpart. y_{min} and y_{max} are, respectively, the minimum and maximum values in the working data set. Since the available molecular information span less than two orders of magnitude, all molecular descriptors have been directly normalized in the range [-1,1] with no prior logarithmic scaling.

ISBN:978-84-693-4597-9 /DL:T.1010-2010

Training, test and validation data sets. To build a QSFR model, the original set of 375 working chemicals was split into a training data set and a test data set. In every case, about 80 % of the working chemicals have been dedicated to training every QSFR model, while the resting 20 % of working chemicals have been reserved for testing its performance while tuning its parameters. The data selection scheme, based on the Self-Organizing Map (SOM) algorithm (Kohonen et al., 1996), has been used to force the diversity of the training data set and to ensure a proper representation of the test data set in the former. The SOM is a procedure for mapping and clustering high-dimensional data by fitting an optimal number of units (also called neurons, cells or nodes) to the data, while minimizing the Euclidean distance between units and data points (i.e. minimizing the mean quantization error), and keeping the vicinity of units in both the map and the data space (i.e. minimizing the mean topological error). The procedure for selecting the training and test data of a single QSFR has been as follows:

First, SOMs of different sizes were trained to fit the 375 working chemicals in the input-target space of the desired QSFR model using the SOM toolbox 5 for Matlab (Vesanto et al., 2000). All SOMs have been set to have toroidal shapes and hexagonal lattices (and not other shapes and lattices) to minimize both the mean quantization error ($\bar{q}_{error}$) and the mean topological error ($\bar{t}_{error}$), while inspecting that such errors are the lowest for each SOM size. The dimensions of the SOM comprise all molecular descriptors selected and the target variable of the QSFR. Note that $\bar{q}_{error}$ and $\bar{t}_{error}$ are estimated, respectively, as (Uriarte and Martín, 2005):

$$\bar{q}_{error} = \frac{1}{N_{wk}} \sum_{n=1}^{N_{wk}} \left\| \mathbf{x}_i - \mathbf{m}_{\mathbf{x}_i} \right\| \quad (8)$$

and

$$\bar{t}_{error} = \frac{1}{N_{wk}} \sum_{n=1}^{N_{wk}} u(\mathbf{x}_i) \quad (9)$$

where: N_{wk} is the number of work data vectors; $\mathbf{m}_{\mathbf{x}_i}$ is the best matching unit (BMU) the corresponding data vector $\mathbf{x}_i$; and, $u(\mathbf{x}_i)$ is a function that yields 1 if the BMU and the next BMU of $\mathbf{x}_i$ are adjacent and, 0 otherwise.

Second, for each trained SOM, chemicals have been included into training data sets when showing the lowest or highest quantization error with respect to their corresponding BMUs. Also, chemicals having extreme values (the lowest or highest values in the whole working data set) in target variables have been included in the training data sets. All remaining working chemicals not following such characteristics have been moved to the corresponding test data set.

Finally, only one pair of training and test data sets is considered for the development of a single QSFR, when the number of training chemicals is about 80 % (±5 %) the number of working chemicals. The bigger the SOM, the higher the number of training chemicals and the lower the number of test chemicals proposed by the algorithm. By setting about 20% of work chemicals for a test data set, it is possible to tune the parameters of the supervised learning algorithm conforming the QSFR to generalize well for chemicals represented in the training data set, but not used in the model.

The 80 validation chemicals have not been used in any stage of the development of QSFRs. However, it has been

ISBN:978-84-693-4597-9 /DL:T.1010-2010

assured that each of the physicochemical properties and molecular descriptors of these 80 chemicals are within the value ranges of the attributes that characterize the 375 working chemicals.

Supervised learning algorithms. Support vector regressions (SVR) (Drucker et al., 1996) using RBF kernel functions have been used to build QSFRs, per compartment g, with basis on the fixed training data set. The QSFR models have the form:

$$N_{[-1,1]}(\log_{10}(w_g)) = f_{QSFR}(N_{[-1,1]}(d_1),\ldots,N_{[-1,1]}(d_L))$$

where the function f_{QSFR} represents a SVR that links normalized molecular descriptors to normalized logarithmic mass ratios. The ε-SVR implementation in the software package RapidMiner 4.4 (Mierswa et al., 2006) was used.

For every compartment and sets of input features considered, an iterative evaluation of 4000 models has been implemented for tuning the parameters of an optimal SVR model (Mierswa et al., 2006). For every combination of parameters, a SVR is developed with the training data set and evaluated on the test and validation data sets. An optimal SVR model is selected when having the lowest mean absolute error (MAE) on the test data set among the SVRs with the 10 highest squared correlation (R^2) values on the test data set. This criteria aims to select a model with optimal generalization capabilities based on chemicals not included in the training set, but somehow represented in it. The MAE and R^2 values measure the performance of the SVR models comparing the target and prediction values of the N chemicals of a data set (tr = training, te = test or val = validation) as follows:

$$MAE_{set} = \frac{\sum_{n=1}^{Nset} |t_n - p_n|}{N_{set}} \text{ ; set = tr, te or val.} \quad (11)$$

and

$$R^2_{set} = \frac{\left(\sum_{n=1}^{Nset}(t_n - \bar{t})(p_n - \bar{p})\right)^2}{\left(\sum_{n=1}^{Nset}(t_n - \bar{t})^2\right)\left(\sum_{n=1}^{Nset}(p_n - \bar{p})^2\right)} \text{ ; set = tr, te or val.} \quad (12)$$

where t_n and p_n are, respectively, the target (MEM output) and predicted (SVR output) values of normalized logarithmic mass ratios of a chemical n in a given compartment. The overbar indicates averages running over all the N_{set} chemicals of a given data set (set = tr, te or val).

Having selected a SVR model for an optimal set of parameters, its accuracy is estimated by means of both a 10-fold cross validation (CV) and a leave one out (LOO) validation procedure running over all the 375 working chemicals. Note that the evaluation of the SVRs is based on normalized logarithms of mass ratios.

Data post-processing and model performance. After evaluating QSFR models (Equation 10) with simple data sets (training, test and validation) and with 10-fold CV and LOO validation procedures, the final normalized predictions for all chemicals have been denormalized (Equation 7), yielding logarithmic mass ratios.

For measuring the performance of a compartmental QSFR model with respect to a single data set, its predictions

ISBN:978-84-693-4597-9 /DL:T.1010-2010

have been compared with respect to the target values, i.e., the logarithmic mass ratios originally generated by SB3.

The differences between targets and predictions are estimated, in average, calculating a mean absolute error measure as follows:

$$MAE_{set} = \frac{1}{N}\sum_{n=1}^{Nset}\left|\log_{10}\left(w_n^{target}\right) - \log_{10}\left(w_n^{predicted}\right)\right|; \text{ set = tr, te or val.} \quad (13)$$

the lower the MAE of a data set, the lower the differences between the targets and predictions of all chemicals in the set.

The predictive performance of a model is assessed in terms of the predictive squared coefficient (q^2), as suggested by Schüürmann et al. (2008) as follows:

$$q_{set}^2 = 1 - \frac{\sum_{n=1}^{Nset}\left(\log_{10}\left(w_{n,g}^{predicted}\right) - \log_{10}\left(w_{n,g}^{target}\right)\right)^2}{\sum_{n=1}^{Nset}\left(\log_{10}\left(w_{n,g}^{target}\right) - \frac{1}{N}\sum_{n=1}^{Nset}\log_{10}\left(w_{n,g}^{target}\right)\right)^2}; \text{ set = tr, te or val.} \quad (14)$$

with the q^2 coefficient varying in the range $(-\infty,1]$. Models with q^2 values closer to 1 have a high predictive performance, while models having q^2 values equal or lower than zero have predictions worst than the simply average of all targets.

4. Results and discussions

Results and discussions for QSFR models emulating the reference scenario are presented below, considering emissions in the water compartment. QSFR models developed and tested for predicting mass ratios in the air and water compartment of the scenario are presented in Section 4.1. Finally, in section 4.2, the clustering of chemicals is used to discuss about how QSFRs can be improved and in which conditions should be used, respectively, by training class-tailored models and being sure that new chemicals fall within the DOA of the models. Also, the performance of air-emission models is briefly presented.

4.1 Chemical distribution assessment

Feature selection. In this study, two types of molecular descriptors were tested as input for the QSFR models: molecular properties calculated from a semi-empirical molecular orbital method and simple counts of molecular constituents.

A wide variety of molecular descriptors (topological, electronic, geometric, etc.), derived from semi-empirical approximations of the molecular orbital (MO) theory (Bredow and Jug, 2005), have been widely used as input in a wide variety of property estimation methods (Devillers, 2003; Raymond et al., 2001; Taskinen and Yliruusi, 2003). So, following the methodology of Section 3.2, preliminary QSFR models were developed using as input combinations of MW and 22 semi-empirical descriptors estimated with CACHE (Fujitsu, 2004) applying the Parameterized Model 3 (PM3) of the MO theory (James, 1989). The descriptors used were selected by means of the CFS filtering algorithm (Hall, 1999) from an initial set of variables comprising MW and 22 semi-empirical descriptors: heat of formation (ΔH_f, kcal/mol), molar refractivity (MR, m^3/mol), polarizability (PO, $Å^3$), total hybridization dipole moment (μ_{hyb}, debye), total point charge dipole moment (μ_{pc}, debye), total sum dipole moment (μ, debye), area (Area, $Å^2$), volume (Vol, $Å^3$), number of filled

ISBN:978-84-693-4597-9 /DL:T.1010-2010

levels (NFL), highest occupied molecular orbital energy (HOMO, eV), lowest occupied molecular orbital energy (LUMO, eV), ionization potential (IP, eV), electron affinity (EA, eV), connectivity indexes ($^0\chi$, $^1\chi$, $^2\chi$), valence connectivity indexes ($^0\chi^v$, $^1\chi^v$, $^2\chi^v$) and kappa alpha shape indexes ($^1\kappa$, $^2\kappa$, $^3\kappa$).

As a second type of input variables, counts of molecular constituents were tested on QSFR models. It is known that fragment contributions have proven to be of great help in the development of QSPRs (Boethling et al., 2004) and QSBRs (Raymond et al., 2001) for a wide range of chemicals. Such is the case of the models traditionally included in EPI suite™ (SRC, 2008). So, it seems plausible predicting the environmental distribution of chemicals directly from molecular information via QSFRs (Equation 10) as schematized in Figure 1, grounded on counts of molecule constituents (atoms, bonds, functional groups and rings). Supported on such idea, the QSFR models of this work were developed (as explained in Section 3.2) to use as input MW and counts of molecular constituents. Table 2 lists the molecular descriptors considered and their minimum and maximum values in the working and validation data sets.

With respect to the 80 validation chemicals, preliminary QSFR models of the air and water compartments yielded better performances using MW and counts of molecular constituents (in air: $q^2 = 0.64$ and MAE = 1.34, in water: $q^2 = 0.68$ and MAE = 0.39) instead of combinations of MW and molecular properties from semi-empirical MO estimations (in air: $q^2 = [-0.09, 0.15]$ and MAE = [2.30, 2.57], in water: $q^2 = [0.27, 0.49]$ and MAE = [0.46, 0.47]). So, final QSFR models, presented below, were built using MW and counts of molecular constituents.

Selection of training and test chemicals. A critical step in the development of optimal QSFRs has been the selection of their training and test chemicals, which affect the generalization capability of the resulting models to new chemicals. Their selection was possible by screening the work chemicals (acting as example vectors of multimedia environmental modeling for the scenario, for which all inputs and targets are known) with the SOM algorithm. Figure 3 shows two SOMs clustering the 375 work chemicals of the reference scenario in terms of the mass ratios of single compartments and their molecular descriptors (MW and 38 non-zero counts of molecule constituents). The first SOM (Figure 3a) clusters chemicals for the air compartment, while the other (Figure 3b) does the clustering for the water compartment. Every SOM fits as close as possible the work chemicals in their corresponding multivariable space (comprised by 40 dimensions: one compartmental mass ratio and 39 descriptors) by clustering neighboring chemicals in their BMUs.

A work chemical is included in the training data set when it has the lowest or highest quantization error (q_{error}) with respect to its BMU. Also, a work chemical is included in the training data set, when it has the lowest or highest mass ratio among all other work chemicals. The work chemicals not following any of these two cases form the test data set of the corresponding QSFR. The selection of training chemicals with regard to a SOM is demonstrated in Figure 4, which takes as example three SOM units from Figure 3a clustering 3, 4 and 5 work chemicals, respectively: units K15, O2 and U8. Within a single SOM unit, the more similarities between chemicals, in terms of structure and environmental distribution, the lower the

ISBN:978-84-693-4597-9 /DL:T.1010-2010

differences between their q_{error} values; e.g., between dieldrin and endrin in unit K15, or between ethanal, butanal and propanal in unit U8. So, selecting the training chemicals, in each SOM unit, as the ones with the lowest and highest q_{error} forces the diversity of the training set and assures the vicinity of the test chemical respect to the training ones.

The domain of applicability (DOA) of a QSFR is defined here as the set of q_{error} ranges covered within each non-empty BMU by the BMU itself till the farthest training chemical clustered in the BMU (the one with the largest q_{error}).

Filled SOM units that cluster two or more work chemicals contribute with a maximum of two training chemicals; while, SOM units clustering one chemical only make one contribution. Note that, as explained in Section 3.2, the size of any optimal SOMs was set to guarantee a number of training chemicals approximately equal to 80 % of the 375 work chemicals available (per compartment); so, the number of training chemicals proposed by the SOMs of Figure 3 for developing QSFRs for the air and water compartments of the reference scenario, resulted to be in total 300 (80.0 %) and 299 (79.7 %), respectively.

Prediction of environmental distributions. QSFRs modeling the air and water compartment of the reference scenario were developed by building SVRs that relate molecular descriptors and mass ratios for the training chemicals (Equation 10).

For assessing the generalization capacity of QSFRs in more realistic conditions, mass ratio predictions must be evaluated for chemicals not used at all in the development of the models. To this end, the 80 validation chemicals were used (Section 2).

Figure 5 compares target values (reference mass ratios) generated by the MEM of the scenario to predictions resulting from two approaches. First, scatter plots from the use of optimized QSFRs (from MW and 38 non-zero count of molecular constituents); and, second, the ranges of the output obtained from the Monte Carlo simulation described in Section 4.1. While Figure 5a is referred to a specific QSFR for air ($QSFR_{air}$), Figure 5b is referred to a QSFR for water ($QSFR_{water}$).

At first sight, it can be noticed in Figure 5 that mass ratios resulting from the QSFRs tend to cover prediction ranges somewhat similar to those from the reference MEM running under Monte Carlo realizations (MC-MEM), presented in Section 4.1. The most deviated predictions from QSFRs tend to be close to the limits delimited by the variation ranges of the MEM, especially for the air compartment in which mass ratios tend to be very small and sensitive to input uncertainties in both estimation approaches.

Depending on the "real" mean reference property values of a chemical, the random property values generated by statistical distributions of standard property estimation methods (Table 1) produced variations in the outputs of MC-MEM of up to 12 logarithmic units. In the same manner it can be inferred that, when estimating input variables from available QSPRs and QSBRs, the output of standard MEMs should undergo a similar variability.

Table 3 shows for MC-MEM, average q^2 and MAE measurements computed considering 1000 realizations for all the 455 chemicals, giving for the air compartment, q^2_{mean} = 0.88 and MAE_{mean} = 0.80; while, for the water compartment, q^2_{mean} = 0.86 and MAE_{mean} = 0.17. Table 3 also shows the

ISBN:978-84-693-4597-9 /DL:T.1010-2010

predictive performances of $QSFR_{air}$ and $QSFR_{water}$ per data set, in terms of MAE and q^2 as defined in equations 13 and 14. The predictive capacity of the QSFR models of Figure 5 tends to be high for sets of chemicals located within the boundaries of the DOA of a QSFR, which is the case of the training and test data sets (with chemicals previously selected with the SOMs of Figure 3). Lower predictive abilities in these QSFR models can be expected for a set of new chemicals (validation set), when some chemicals fall out of the DOA. The QSFR for water (Figure 5b) generalizes much better that the QSFR for air (Figure 5a), evidenced by the minimal dispersion in the mass ratios from the former model. This is markedly supported by the overall performances of these models, including all the 455 chemicals of the scenario (comprising the training, test and validation sets simultaneously): for air, $q^2 = 0.82$ and MAE = 0.91; while, for water, $q^2 = 0.81$ and MAE = 0.32.

QSFR models using simple counts of molecular constituents, as the ones we propose here, cannot distinguish between isomers that have in common the exact number of bonds, functional groups and ring structures (with these characteristics, there are 81 working and 10 validation isomeric chemicals out of the 375 working and 80 validation chemicals, respectively). That characteristic is not a serious drawback because transport and degradation properties for these isomers are not extremely different, at least in our working and validation data sets. On the other hand, molecular constituent counts have a great advantage, they can be easily retrieved or calculated known the molecular formula or structural code of new chemicals (e.g., SMILES, InChl, OpenSMILES, etc.); this makes them suitable for simple and rapid screenings. Since the constituents (atoms, bonds, groups and rings) of a chemical are counted without errors and SVRs yield the same model if given the same training data and parameters (unlike ANNs, which adjust internal parameters in search of a local minimum error), QSFRs using these two features can be reproduced easily. This represents a clear advantage over models using as input semi-empirical descriptors, as these usually vary depending on the specific MO method (Bredow and Jug, 2005) used to estimate them.

4.2 Assessment of the Chemical Domain

QSFRs models for classes of chemicals. For improving the prediction performance of the QSFRs described in section 4.2, it was considered the development of QSFRs specialized in very specific classes of chemicals. This implies, first, the definition of chemical classes (families) and, second, the development and use of specialized QSFR models (one per chemical class). In a practical distribution assessment of new chemicals, it would be necessary to identify to which chemical class they belong to for later using the appropriate QSFR model.

Different criteria could be proposed for creating chemical families with respect to molecular structure, but the performance of any class-tailored QSFR is hampered by the availability of sufficient training data. In a preliminary screening of the 375 work chemicals of the reference scenario, it was observed that 39 chemicals are composed of solely carbon and hydrogen atoms, while the remaining 336 chemicals have at least one heteroatom (bromine, chlorine, fluorine, nitrogen, oxygen, and phosphorus or sulfur atoms). These two groups constitute a starting point for creating two

ISBN:978-84-693-4597-9 /DL:T.1010-2010

chemical classes, but there is a somewhat unbalanced distribution of chemicals if solely 39 chemicals in the first class are available for creating the training and test data sets of QSFRs. An adjustment can be made to create two chemical families with somewhat similar structure but enough training samples, adding oxygen to the preliminary class of chemicals formed solely by carbon and hydrogen. This way, 146 work chemicals were identified to be constituted by carbon and hydrogen with no heteroatoms or only oxygen (Class X); while 229 chemicals have a least one heteroatom different than oxygen (Class Y). With this final clustering, a fair class proportion was achieved without sacrificing much with respect to the general properties of the clustered chemicals. Note that chemicals in class X can be described with solely MW, 4 atom counts (all atoms, carbon, hydrogen and oxygen), 3 bond counts (all bonds, single bonds and double bonds), 7 functional group counts (aldehyde, carbonyl, carboxyl, ether, hydroxyl, methyl and methylene) and 8 ring counts (all rings, aromatic rings, small rings, 5 membered, aromatic 5 membered, 6 membered, aromatic 6 membered and 7-12 membered). While the chemicals in class Y are described with MW and the 38 constituent counts listed in Table 2 (like in the QSFR models of Section 4.2).

For optimal results, specific training and test data sets should be used every time a new SVR is trained. But, for comparison purposes, the same training and test chemicals previously selected for the models $QSFR_{air}$ and $QSFR_{water}$ (using the SOMs of Figure 3) were maintained when developing class-tailored QSFRs for classes X and Y. Then, four class-tailored model were developed: $QSFR_{air,X}$, $QSFR_{air,Y}$, $QSFR_{water,X}$ and $QSFR_{water,Y}$.

Logarithmic mass ratios were predicted for each chemical, according to its chemical class (X or Y), using the appropriate model per compartment. Below, the results for both classes (X and Y) are presented together for each compartment using the acronyms $QSFR_{air,X/Y}$ (i.e. using the models $QSFR_{air,X}$ or $QSFR_{air,Y}$) and $QSFR_{water,X/Y}$ (i.e. using the models $QSFR_{water,X}$ or $QSFR_{water,Y}$).

Figure 6 shows predictions of logarithmic mass ratios for the air compartment (Figure 6a) and the water compartment (Figure 6b), using the models $QSFR_{air,X/Y}$ and $QSFR_{water,X/Y}$, respectively. A general improvement has been achieved with respect to the predictions of simple QSFRs ($QSFR_{air}$ and $QSFR_{water}$ in Figure 5), as the application of class-tailored QSFRs ($QSFR_{air,X/Y}$ and $QSFR_{water,X/Y}$ in Figure 6) yielded higher q^2 and lower MAE values, as shown in Table 3 for air (from $q^2 = 0.82$ and MAE = 0.91 in $QSFR_{air}$ to $q^2 = 0.88$ and MAE = 0.68 in $QSFR_{air,X/Y}$) and water (from $q^2 = 0.81$ and MAE = 0.32 in $QSFR_{water}$ to $q^2 = 0.87$ and MAE = 0.17 in $QSFR_{water,X/Y}$).

Considering all data sets simultaneously, on average, the results of $QSFR_{air,X/Y}$ and $QSFR_{water,X/Y}$ are very close to those from MC-MEM (Table 3). The discrimination and posterior assessment of chemicals with respect to their chemicals composition (using classes X and Y) improved the generalization capability of SVRs linking chemical distribution and molecular structure, when compared to the processing of all available chemicals with a simple SVR (Equation 10) as Table 3 shows. Also, Figure 6 displays the majority of scatter points closer to the diagonals of each subplot than those in Figure 5.

ISBN:978-84-693-4597-9 /DL:T.1010-2010

With respect to the 80 validation chemicals, q^2 and MAE measurements improved slightly in air and got somewhat deteriorated in water when making predictions with class-tailored models ($QSFR_{air,X/Y}$ and $QSFR_{water,X/Y}$). Please note that this is the result of reducing the number of training chemicals per SVR when implementing chemicals classes (about half the training chemicals selected for $QSFR_{air}$ and $QSFR_{water}$), incrementing the chances of having some test and validation chemicals out the DOA of the class-tailored models. This implies that such outlying chemicals have singularities that are better covered by the totality of training chemicals available at the time of the assessments. Remember that as stated above, the training and test data sets selected from SOMs (Figure 3) for general QSFR models (Figure 5) were kept unchanged for training specialized QSFRs (Figure 6). This allows a direct comparison of the performance indexes on each data set in Table 3. An additional improvement on these indexes should be expected if the selection of the training and test sets for the class-tailored QSFR models were performed after clustering the chemicals in individual SOMs (one per class and compartment), but this would make impossible the comparison of the two approaches under the same conditions.

Domain of Aplicability. QSFR models follow the same limitations of QSAR models (Johnson, 2008). For instance, predictions beyond the DOA of the models should be avoided. The DOA of any model is primarily defined by its training chemicals (Weaver and Gleeson, 2008); so, identifying the DOA of an existing QSFR model it is possible to assess, approximately, how appropriate it is for a new chemical (Kühne et al., 2009)).

Reasonable estimations of the DOA of a model can be performed by measuring distances or probability density distributions of training data vectors to new data vectors (Schroeter et al., 2007), coming either from validation purposes or assessing new chemicals of concern. Since the SOM algorithm is based on the distances between data vectors in a multivariate space (Kohonen et al., 1996), we can use it to define the DOA of the QSFR models. Three different SOM-based approaches have been used to define the DOA:

(i) Using the SOMs used in the selection of training and test data sets. Because the q_{error} of the work chemicals within each SOM unit have been used for selecting the training chemicals, the training chemical with the highest q_{error} defines the DOA border. The original SOM (Figure 3) had 40 dimensions (MW, 38 constituent counts and a mass ratio). When presenting new chemicals to the SOM, the mass ratio is unknown so only 39 out of 40 variables are used for classification purposes, the error of assessing new chemicals with one dimension missing is not significant given the relation 39:1 of available-unavailable dimensions.

(ii) Applying a principal component analysis (Pearson, 1901) on the 39 input variables, it was found that five principal components accounted for about 59 % of cumulative variance. We trained a SOM with these five principal components and, again, defined a DOA with the highest q_{error} of the training chemicals in each SOM unit.

(iii) Intersecting the first two approaches.

ISBN:978-84-693-4597-9 /DL:T.1010-2010

Table 4 shows q^2 and MAE performance measurements for models $QSFR_{air,X/Y}$ and $QSFR_{water,X/Y}$, for test and validation chemicals emitted in water belonging or not to the DOAs defined above.

In the first two approaches (1 and 2), test or validation chemicals with quantization errors higher to those of the upper bounding training chemicals are considered to be out the DOA of the models. Since the numbers of chemicals within the DOAs from the first (1) and second (2) approaches differ because of the different variables considered and the errors of each SOM, their intersection (3) is preferred because more restrictive conditions are achieved. So, as shown in Table 4, using the third approach (3), it has been estimated that the mass ratios of about 48 and 50 "new" (test and validation) chemicals can be optimally predicted by, respectively, $QSFR_{air,X,Y}$ (with $q^2 = 0.92$ and MAE = 0.54) and $QSFR_{water,X,Y}$ (with $q^2 = 0.94$ and MAE = 0.15). By assessing that new chemicals are within the DOA of a QSFR model, the probability of having acceptable predictions is notoriously increased.

All results discussed so far resulted from emissions in the water compartment. To check that the present QSFR approach can be applied to other emission compartments, specific models were developed for emissions in the air compartment. Table 4 also shows q^2 and MAE performances for air-emission models using the same training, test and validation data sets already used in the water-emission models. Only indexes for chemicals within the different DOAs considered are shown, demonstrating that similar results are obtained irrespective of the emission compartment. The distribution of new chemicals can be reasonably predicted, as long as they lie within the DOA of a QSFR model.

5. Conclusions

Assessing the environmental concentrations of chemical pollutants from molecular information can be performed by two different approaches (Figure 1): The first approach implies estimating missing physicochemical properties from available QSPR and QSBR models for assessing chemicals of concern with standard MEMs. The second approach, proposed here, implies developing QSFR models that link concentrations to molecular information for assessing chemicals of concern known their molecular structure. When the uncertainty of key properties estimated by QSPRs and QSBRs can affect the outputs of standard MEMs, QSFR models can be an alternative for the latter if enough representative training chemicals are available for developing these models. Since QSFRs rely on the same methodology employed in the development of QSARs, the concentrations of chemicals of concern can be predicted with appreciable accuracy if they are within the DOA of available QSFR models.

Acknowledgements

This research was financially supported by the European Union (NOMIRACLE, European Commission, FP6 contract Nº 003956) and the Generalitat de Catalunya (2009SGR1529).

References

Aronson D, Boethling R, Howard P, Stiteler W. Estimating biodegradation half-lives for use in chemical screening. Chemosphere 2006; 63: 1953.

ISBN:978-84-693-4597-9 /DL:T.1010-2010

Aronson D, Howard PH. Evaluating Potential POP/PBT Compounds for Environmental Persistence. SRC, North Syracuse, United states, 1999.

Basheer IA, Hajmeer M. Artificial neural networks: fundamentals, computing, design, and application. Journal of Microbiological Methods 2000; 43: 3.

Boethling RS, Howard PH, Meylan WM. Finding and estimating chemical property data for environmental assessment. Environmental Toxicology and Chemistry 2004; 23: 2290-2308.

Brandes LJ, den Hollander H, van de Meent D. SimpleBox 2.0: a nested multimedia fate model for evaluating the environmental fate of chemicals. RIVM, Bilthoven, The Netherlands, 1996, pp. 156.

Bredow T, Jug K. Theory and range of modern semiempirical molecular orbital methods. Theoretical Chemistry Accounts: Theory, Computation, and Modeling (Theoretica Chimica Acta) 2005; 113: 1.

Breivik K, Alcock R, Li Y-F, Bailey RE, Fiedler H, Pacyna JM. Primary sources of selected POPs: regional and global scale emission inventories. Environmental Pollution 2004; 128: 3.

Breivik K, Vestreng V, Rozovskaya O, Pacyna JM. Atmospheric emissions of some POPs in Europe: a discussion of existing inventories and data needs. Environmental Science & Policy 2006; 9: 663.

Breivik K, Wania F. Expanding the Applicability of Multimedia Fate Models to Polar Organic Chemicals. Environmental Science and Technology 2003; 37: 4934.

Brüggemann R, Restrepo G, Voigt K. Structure-Fate Relationships of Organic Chemicals Derived from the Software Packages E4CHEM and WHASSE. Journal of Chemical Information and Modeling 2006; 46: 894.

Burden FR, Polley MJ, Winkler DA. Toward Novel Universal Descriptors: Charge Fingerprints. Journal of Chemical Information and Modeling 2009; 49: 710-715.

Byvatov E, Fechner U, Sadowski J, Schneider G. Comparison of Support Vector Machine and Artificial Neural Network Systems for Drug/Nondrug Classification. J. Chem. Inf. Comput. Sci. 2003; 43: 1882-1889.

Citra MJ. Incorporating Monte Carlo analysis into multimedia environmental fate models. Environmental Toxicology and Chemistry 2004; 23: 1629-1633.

Cohen Y. Pollutants in a multimedia environment. In: Cohen Y, editor. Workshop on pollutant transport and accumulation in a multimedia environment. Plenum Press, New York, Santa Mónica, California, 1986.

Cohen Y, Cooter EJ. Multimedia Environmental Distribution of Toxics (Mend-Tox). I: Hybrid Compartmental-Spatial Modeling Framework. Practice Periodical of Hazardous, Toxic, and Radioactive Waste Management 2002a; 6: 70-86.

Cohen Y, Cooter EJ. Multimedia Environmental Distribution of Toxics (Mend-Tox). II: Software Implementation and Case Studies. Practice Periodical of Hazardous, Toxic, and Radioactive Waste Management 2002b; 6: 87-101.

Cronin MT, Walker JD, Jaworska JS, Comber MH, Watts CD, Worth AP. Use of QSARs in international decision-making frameworks to predict health effects of chemical substances. Environmental Health Perspectives 2003; 111: 1376-90.

ISBN:978-84-693-4597-9 /DL:T.1010-2010

Cronin MTD, Schultz TW. Pitfalls in QSAR. Journal of Molecular Structure: THEOCHEM 2003; 622: 39.

den Hollander HA, van de Meent D. Appendix to SimpleBox 3.0: A multimedia mass balance model for evaluating the environmental fate of chemicals. RIVM, 2004.

den Hollander HA, van Eijkeren JCH, van de Meent D. SimpleBox 3.0. RIVM, Bilthoven, The Netherlands, 2004.

Devillers J. A decade of research in environmental QSAR. SAR and QSAR in Environmental Research 2003; 14: 1 - 6.

Drucker H, Burges CJC, Kaufman L, Smola A, Vapnik V. Support Vector Regression Machines. Advances in Neural Information Processing Systems 1996: 155-161.

Duca JS, Hopfinger AJ. Estimation of Molecular Similarity Based on 4D-QSAR Analysis: Formalism and Validation. Journal of Chemical Information and Computer Sciences 2001; 41: 1367-1387.

Eisenberg JNS, Bennett DH, McKone TE. Chemical Dynamics of Persistent Organic Pollutants: A Sensitivity Analysis Relating Soil Concentration Levels to Atmospheric Emissions. Environ. Sci. Technol. 1998; 32: 115-123.

European Commission. Technical Guidance Document on Risk Assessment, Part III. Institute for Health and Consumer Protection, European Chemicals Bureau, 2003.

Fjodorova N, Novich M, Vrachko M, Smirnov V, Kharchevnikova N, Zholdakova Z, et al. Directions in QSAR Modeling for Regulatory Uses in OECD Member Countries, EU and in Russia. Journal of Environmental Science and Health, Part C 2008; 26: 201 - 236.

Fujitsu BGo. CAChe Software. BioSciences Group, Fujitsu Computer Systems, Beaverton, 2004.

Furusjö E, Svenson A, Rahmberg M, Andersson M. The importance of outlier detection and training set selection for reliable environmental QSAR predictions. Chemosphere 2006; 63: 99.

Godavarthy SS, Robinson JRL, Gasem KAM. SVRC-QSPR model for predicting saturated vapor pressures of pure fluids. Fluid Phase Equilibria 2006; 246: 39.

Golbraikh A, Tropsha A. Beware of q2! Journal of Molecular Graphics and Modelling 2002; 20: 269.

Hall MA. Correlation-based Feature Selection for Machine Learning. Department of Computer Science. Ph.D. thesis. The University of Waikato, Hamilton, New Zealand, 1999.

Howard PS, Boethling RS, Jarvis WF, Meylan WM, Michalenko EM. Handbook of environmental degradation rates. Chelsea, MI: Lewis Publications, 1991.

Hugo K. From Narcosis to Hyperspace: The History of QSAR. Quantitative Structure-Activity Relationships 2002; 21: 348-356.

Jain AK, Duin RPW, Mao J. Statistical Pattern Recognition: A Review. IEEE Transactions on Pattern Analysis and Machine Intelligence 2000; 22: pp. 4-37.

James JPS. Optimization of parameters for semiempirical methods II. Applications. Journal of Computational Chemistry 1989; 10: 221-264.

Johnson SR. The Trouble with QSAR (or How I Learned To Stop Worrying and Embrace Fallacy). Journal of Chemical Information and Modeling 2008; 48: 25-26.

ISBN:978-84-693-4597-9 /DL:T.1010-2010

Junge CE. Fate of pollutants in the air and water environment: Wiley - Interscience, 1977.

Kawamoto K, MacLeod M, Mackay D. Evaluation and comparison of multimedia mass balance models of chemical fate: application of EUSES and ChemCAN to 68 chemicals in Japan. Chemosphere 2001; 44: 599.

Klöpffer W, Wagner B. Persistence revisited. Environmental Science and Pollution Research 2007; 14: 141.

Kohonen T, Oja E, Simula O, Visa A, Kangas J. Engineering applications of the self-organizing map. Proceedings of the IEEE 1996; 84: 1358.

Kühne R, Breitkopf C, Schüürmann G. Error propagation in fucacity level-III models in the case of uncertain physicochemical properties. Environmental Toxicology and Chemistry 1997; 16: 2067-2069.

Kühne R, Ebert R-U, Schüürmann G. Estimation of Compartmental Half-lives of Organic Compounds - Structural Similarity versus EPI-Suite. QSAR & Combinatorial Science 2007; 26: 542-549.

Kühne R; Ebert RU; Schüürmann G. Chemical Domain of QSAR Models form Atom-Centered Fragments. J. Chem. Inf. Model. 2009; 49: 2660-2669.

Lijzen JPA, Rikken MGJ. European Union System for the Evaluation of Substances 2.0 (EUSES 2.0); background report. RIVM, Bilthoven, the Netherlands., 2004, pp. 454.

Lohmann R, Breivik K, Dachs J, Muir D. Global fate of POPs: Current and future research directions. Environmental Pollution 2007; 150: 150.

Mackay D. Multimedia Environmental Models - The Fugacity Approach. Boca Ratón: Lewis Publishers, 2001.

Mackay D, Hubbarde J, Webster E. The role of QSARs and fate models in chemical hazard and risk assessment. QSAR & Combinatorial Science 2003; 22: 106-112.

Mackay D, Shiu W-Y, Ma KC. Illustrated Handbook of Physical-Chemical Properties and Environmental Fate for Organic Chemicals: Lewis Publishers Inc., 1992.

Mackay D, Wan-Yiu Shiu. Physical-Chemical Properties and Environmental Fate Handbook on CD-ROM. Chapman & Hall / CDCnetBASE, 2000.

Mackay D, Webster E. A perspective on environmental models and QSARs. SAR and QSAR in Environmental Research 2003; 14: 7.

Martínez I. Quantitative structure fate relationships for multimedia environmental analysis. Ph.D. thesis. Universitat Rovira i Virgili, Tarragona, Spain., 2010.

Mierswa I, Wurst M, Klinkenberg R, Scholz M, Euler T. YALE: Rapid Prototyping for Complex Data Mining Tasks. In: Ungar L, Craven M, Gunopulos D, Eliassi-Rad T, editors. Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-06). ACM, Philadelphia, PA, USA, 2006, pp. 935-940.

Modarresi H, Modarress H, Dearden JC. QSPR model of Henry's law constant for a diverse set of organic chemicals based on genetic algorithm-radial basis function network approach. Chemosphere 2007; 66: 2067.

Nikolova N, Jaworska J. Approaches to Measure Chemical Similarity - a Review. QSAR & Combinatorial Science 2003; 22: 1006-1026.

ISBN:978-84-693-4597-9 /DL:T.1010-2010

NITE. Chemical Risk Information Platform (CHRIP). National Institute of Technology and Evaluation.

OECD. Guidance Document on the Validation of (Quantitative) Structure-Activity Relationship [(Q)SAR] Models. OECD Series on Testing and Assessment 69., 2007.

Pearson K. On lines and planes of closest fit to systems of points in space. The London, Edinburgh and Dublin Philosophical Magazine and Journal of Science, sixth series 1901; 2: 559-572.

Prinn RG, Huang J, Weiss RF, Cunnold DM, Fraser PJ, Simmonds PG, et al. Evidence for substantial variations of atmospheric hydroxyl radicals in the past two decades. Science 2001; 292: 1882.

Raymond JW, Rogers TN, Shonnard DR, Kline AA. A review of structure-based biodegradation estimation methods. Journal of Hazardous Materials 2001; 84: 189.

Saeys Y, Inza I, Larranaga P. A review of feature selection techniques in bioinformatics. Bioinformatics 2007; 23: 2507-2517.

Schroeter T, Schwaighofer A, Mika S, Ter Laak A, Suelzle D, Ganzer U, et al. Estimating the domain of applicability for machine learning QSAR models: a study on aqueous solubility of drug discovery molecules. Journal of Computer-Aided Molecular Design 2007; 21: 651.

Schüürmann G, Ebert R-U, Chen J, Wang B, Ku□hne R. External Validation and Prediction Employing the Predictive Squared Correlation Coefficient - Test Set Activity Mean vs Training Set Activity Mean. Journal of Chemical Information and Modeling 2008; 48: 2140-2145.

Schwarzenbach RP, Gschwend PM, Imboden DM. Environmental organic chemistry: John Wiley & Sons, 2003.

Sedykh A, Klopman G. Data analysis and alternative modelling of MITI-I aerobic biodegradation. SAR and QSAR in Environmental Research 2007; 18: 693 - 709.

Senese CL, Duca J, Pan D, Hopfinger AJ, Tseng YJ. 4D-Fingerprints, Universal QSAR and QSPR Descriptors. Journal of Chemical Information and Computer Sciences 2004; 44: 1526-1539.

Smola AJ, Schölkopf B. A tutorial on support vector regression. Statistics and Computing 2004; 14: 199.

SRC. Interactive PhysProp Database Demo. Syracuse Research Corporation.

SRC. EPI Suite v4.00. SRC, 2008.

Stouch TR, Kenyon JR, Johnson SR, Chen X-Q, Doweyko A, Li Y. In silico ADME/Tox: why models fail. Journal of Computer-Aided Molecular Design 2003; 17: 83.

Struijs J, Peijnenburg WJGM. Predictions by the multimedia environmental fate model SimpleBox compared to field data: Intermedia concentration ratios of two phthalate esters. RIVM, Bilthoven, 2002, pp. 62.

Taskinen J, Yliruusi J. Prediction of physicochemical properties based on neural network modelling. Advanced Drug Delivery Reviews 2003; 55: 1163.

Tickner J, Geiser K, Coffin M. The U.S. Experience in Promoting Sustainable Chemistry (9 pp). Environmental Science and Pollution Research 2005; 12: 115.

Todeschini R, Consonni V. Handbook of Molecular Descriptors: Wiley-VCH, 2000.

ISBN:978-84-693-4597-9 /DL:T.1010-2010

Toose L, Woodfine DG, MacLeod M, Mackay D, Gouin J. BETR-World: a geographically explicit model of chemical fate: application to transport of [alpha]-HCH to the Arctic. Environmental Pollution 2004; 128: 223.

Toussant M. A scientific milestone. Chemical & Engineering News 2009; 87: 3.

Uriarte EA, Martín FD. Topology Preservation in SOM. International Journal of Applied Mathematics and Computer Sciences 2005; 1: 19.

US-EPA. Inventory Update Rule. Office of Pollution Prevention and Toxics, Environmental Protection Agency http://www.epa.gov/oppt/iur/, Washington, 2006.

van de Meent D. SIMPLEBOX: a generic multimedia fate evaluation model. RIVM, Bilthoven, The Netherlands, 1993.

Vermeire T, Rikken M, Attias L, Boccardi P, Boeije G, Brooke D, et al. European union system for the evaluation of substances: the second version. Chemosphere 2005; 59: 473.

Vermeire TG, Jager DT, Bussian B, Devillers J, den Haan K, Hansen B, et al. European Union System for the Evaluation of Substances (EUSES). Principles and structure. Chemosphere 1997; 34: 1823.

Vesanto J, Himberg J, Alhoniemi E, Parhankangas J. SOM Toolbox for Matlab 5, 2000.

Walker JD, Carlsen L, Hulzebos E, Simon-Hettich B. Global Government applications of analogues, SARs and QSARs to predict aquatic toxicity, chemical or physical properties, environmental fate parameters and health effects of organic chemicals. SAR and QSAR in Environmental Research 2002; 13: 607.

Weaver S, Gleeson MP. The importance of the domain of applicability in QSAR modeling. Journal of Molecular Graphics and Modelling 2008; 26: 1315.

Webster E, Mackay D, Di Guardo A, Kane D, Woodfine D. Regional differences in chemical fate model outcome. Chemosphere 2004; 55: 1361.

Willighagen EL, Wehrens R, Buydens LMC. Molecular Chemometrics. Critical Reviews in Analytical Chemistry 2006; 36: 189 - 198.

Witten IH, Frank E. Data Mining: Practical machine learning tools and techniques. San Francisco, U.S.: Morgan Kaufmann, 2005.

Worth AP, Bassan A, De Bruijn J, Saliner AG, Netzeva T, Patlewicz G, et al. The role of the European Chemicals Bureau in promoting the regulatory use of (Q)SAR methods. SAR and QSAR in Environmental Research 2007; 18: 111.

Xu Y, Zomer S, Brereton RG. Support Vector Machines: A Recent Method for Classification in Chemometrics. Critical Reviews in Analytical Chemistry 2006; 36: 177 - 188.

Yaffe D, Cohen Y. Neural network based temperature-dependent quantitative structure property relations (QSPRs) for predicting vapor pressure of hydrocarbons. Journal of Chemical Information and Computer Sciences 2001; 41: 463.

Yaffe D, Cohen Y, Espinosa G, Arenas A, Giralt F. A Fuzzy ARTMAP Based on Quantitative Structure - Property Relationships (QSPRs) for Predicting Aqueous

ISBN:978-84-693-4597-9 /DL:T.1010-2010

Solubility of Organic Compounds. Journal of Chemical Information and Computer Sciences 2001; 41: 1177.

Yaffe D, Cohen Y, Espinosa G, Arenas A, Giralt F. Fuzzy ARTMAP and Back-Propagation Neural Networks Based Quantitative Structure-Property Relationships (QSPRs) for Octanol-Water Partition Coefficient of Organic Compounds. J. Chem. Inf. Model. 2002; 42: 162-183.

Yaffe D, Cohen Y, Espinosa G, Giralt F, Arenas A. A fuzzy ARTMAP-based quantitative structure-property relationship (QSPR) for the Henry's Law constant of organic compounds. Journal of Chemical Information and Computer Sciences 2003; 43: 85.

Zukowska B, Breivik K, Wania F. Evaluating the environmental fate of pharmaceuticals using a level III model based on poly-parameter linear free energy relationships. Science of The Total Environment 2006; 359: 177.

ISBN:978-84-693-4597-9 /DL:T.1010-2010

Table 1. Statistical distributions assigned to independent properties affecting the reference pollution scenario.

input	**Assumed distribution for simulations**	Typical uncertainty distribution of parameters predicted by QSPRs and QSBRs			
		Data set	**Statistic parameters*,+**	**Units**	**Source**
T_m	Normal	validation	SD = 58.00	K	(Boethling et al., 2004)
P_v	Log-normal	validation	SD = 0.717	mmHg	(Boethling et al., 2004)
H	Log-normal	training	SD = 0.440	$\log_{10}(atm \cdot m^3/mol)$	(Boethling et al., 2004)
K_{ow}	Log-normal	validation	SD = 0.427	$\log_{10}(-)$	(Boethling et al., 2004)
k_{air}	Discrete	training	P(0) = 0.48, P(±1) = 0.37, P(±2) = 0.13, P(±>2) = 0.02	-	(Kühne et al., 2007)
k_{water}	Discrete	training	P(0) = 0.52, P(±1) = 0.35, P(±2) = 0.08, P(±>2) = 0.05	-	(Kühne et al., 2007)

* For QSPRs, the parameters have been reported in standard deviations, SD, in logarithmic values when noted.

+ For QSBRs, the reported parameters are probabilities, P(C), that indicate if a chemical has been classified as member of a degradation class C (0 = correct class, ±1 = neighbor category predicted, ±2 = two categories differing and ±>2 = more than two categories differing) in the 9-class scale proposed by Mackay et al. (1992).

ISBN:978-84-693-4597-9 /DL:T.1010-2010

Table 2. Molecular descriptors used in the QSFRs of this study.

Count	Symbol	Working data set		Validation data set	
		min	max	min	max
Molecular weight (g/mol)	MW	44.05	959.17	85.11	402.49
Count of all atoms	AC_{all}	5	89	10	81
Count of bromine atoms	$AC_{bromine}$	0	10	0	3
Count of carbon atoms	AC_{carbon}	1	32	3	26
Count of chlorine atoms	$AC_{chlorine}$	0	8	0	3
Count of fluorine atoms	$AC_{fluorine}$	0	27	0	3
Count of hydrogen atoms	$AC_{hydrogen}$	0	60	3	54
Count of nitrogen atoms	$AC_{nitrogen}$	0	6	0	3
Count of oxygen atoms	AC_{oxygen}	0	8	0	8
Count of phosphorus atoms	$AC_{phosphorus}$	0	1	0	1
Count of suplhur atoms	$AC_{sulphur}$	0	4	0	2
Count of all bonds	BC_{all}	4	88	10	80
Count of single bonds	BC_{single}	4	88	9	80
Count of double bonds	BC_{double}	0	18	0	8
Count of triple bonds	BC_{triple}	0	2	0	2
Count of aldehyde groups	$GC_{aldehyde}$	0	1	0	1
Count of amide groups	GC_{amide}	0	2	0	2
Count of amine groups	GC_{amine}	0	2	0	2
Count of sec-amine groups	$GC_{sec\text{-}amine}$	0	2	0	2
Count of carbonyl groups	$GC_{carbonyl}$	0	2	0	2
Count of carboxyl groups	$GC_{carboxyl}$	0	2	0	2
Count of cyano groups	GC_{cyano}	0	2	0	2
Count of ether groups	GC_{ether}	0	4	0	3
Count of hydroxyl groups	$GC_{hydroxyl}$	0	4	0	2
Count of methyl groups	GC_{methyl}	0	9	0	7
Count of methylene groups	$GC_{methylene}$	0	3	0	0
Count of nitro groups	GC_{nitro}	0	3	0	1
Count of nitroso groups	$GC_{nitroso}$	0	1	0	0
Count of sulfide groups	$GC_{sulfide}$	0	4	0	2
Count of sulfone groups	$GC_{sulfone}$	0	1	0	1
Count of thiol groups	GC_{thiol}	0	1	0	1
Count of all rings	RC_{all}	0	12	0	2
Count of aromatic rings	$RC_{aromatic}$	0	4	0	2
Count of small rings	RC_{small}	0	7	0	0
Count of 5-membered rings	$RC_{5\text{-}m}$	0	4	0	1
Count of aromatic 5-membered rings	$RC_{a\text{-}5\text{-}m}$	0	2	0	0
Count of 6-membered rings	$RC_{6\text{-}m}$	0	4	0	2
Count of aromatic 6-membered rings	$RC_{a\text{-}6\text{-}m}$	0	4	0	2
Count of (7-12)-membered rings	$RC_{7\text{-}12\text{-}m}$	0	2	0	1

ISBN:978-84-693-4597-9 /DL:T.1010-2010

Table 3. Performance measurements of different concentration estimation approaches for air and water when the chemical is emitted in water.

Compartment	Estimation approach	Performance measure	Performances per data set*			
			Training set	Test set	Validation set	All sets
Air	MC-MEM	q^2	0.88	0.87	0.86	0.88
		MAE	0.80	0.79	0.82	0.80
Air	$QSFR_{air}$	q^2	0.85	0.86	0.64	0.82
		MAE	0.81	0.81	1.34	0.91
Air	$QSFR_{air,X/Y}$	q^2	0.92	0.91	0.68	0.88
		MAE	0.54	0.59	1.30	0.68
Water	MC-MEM	q^2	0.89	0.79	0.78	0.86
		MAE	0.16	0.15	0.22	0.17
Water	$QSFR_{water}$	q^2	0.86	0.60	0.68	0.81
		MAE	0.30	0.34	0.39	0.32
Water	$QSFR_{water,X/Y}$	q^2	0.94	0.78	0.62	0.87
		MAE	0.11	0.19	0.36	0.17

* The number of chemicals per data set varies per compartment. For the air compartment there are 300 training, 75 test and 80 validation chemicals; and, for the water compartment there are 299 training, 76 test and 80 validation chemicals.

ISBN:978-84-693-4597-9 /DL:T.1010-2010

Table 4. Performance measurements of specialized QSFRs for the air and water compartments (removing 13 outlying validation chemicals)

Model	DOA	Parameters	Emissions in water						Emissions in air		
			Chemicals within DOA			Chemicals out DOA			Chemicals within DOA		
			te	val	te, val	te	val	te, val	te	val	te, val
$QSFR_{air,X/Y}$	(i)	Chemicals	62	29	91	13	51	64	62	29	91
		q^2	0.93	0.69	0.89	0.70	0.67	0.68	0.95	0.54	0.89
		MAE	0.55	0.99	0.69	0.79	1.47	1.33	0.21	0.46	0.29
	(ii)	Chemicals	36	15	51	39	65	104	36	15	51
		q^2	0.95	0.64	0.89	0.86	0.68	0.75	0.97	0.65	0.92
		MAE	0.45	1.03	0.62	0.73	1.36	1.12	0.16	0.41	0.23
	(iii)	Chemicals	36	12	48	39	68	107	36	12	48
		q^2	0.95	0.78	0.92	0.86	0.66	0.73	0.97	0.76	0.94
		MAE	0.45	0.79	0.54	0.73	1.39	1.15	0.16	0.34	0.20
$QSFR_{water,X/Y}$	(i)	Chemicals	56	21	77	20	59	79	56	21	77
		q^2	0.84	0.88	0.87	0.57	0.28	0.38	0.90	0.88	0.90
		MAE	0.15	0.30	0.19	0.29	0.39	0.36	0.29	0.30	0.29
	(ii)	Chemicals	44	16	60	32	64	96	44	16	60
		q^2	0.86	0.81	0.84	0.69	0.21	0.43	0.92	0.73	0.84
		MAE	0.13	0.38	0.20	0.26	0.36	0.33	0.27	0.58	0.35
	(iii)	Chemicals	40	10	50	36	70	106	40	10	50
		q^2	0.91	0.95	0.94	0.66	0.25	0.40	0.93	0.93	0.93
		MAE	0.12	0.25	0.15	0.26	0.38	0.34	0.25	0.32	0.27

ISBN:978-84-693-4597-9 /DL:T.1010-2010

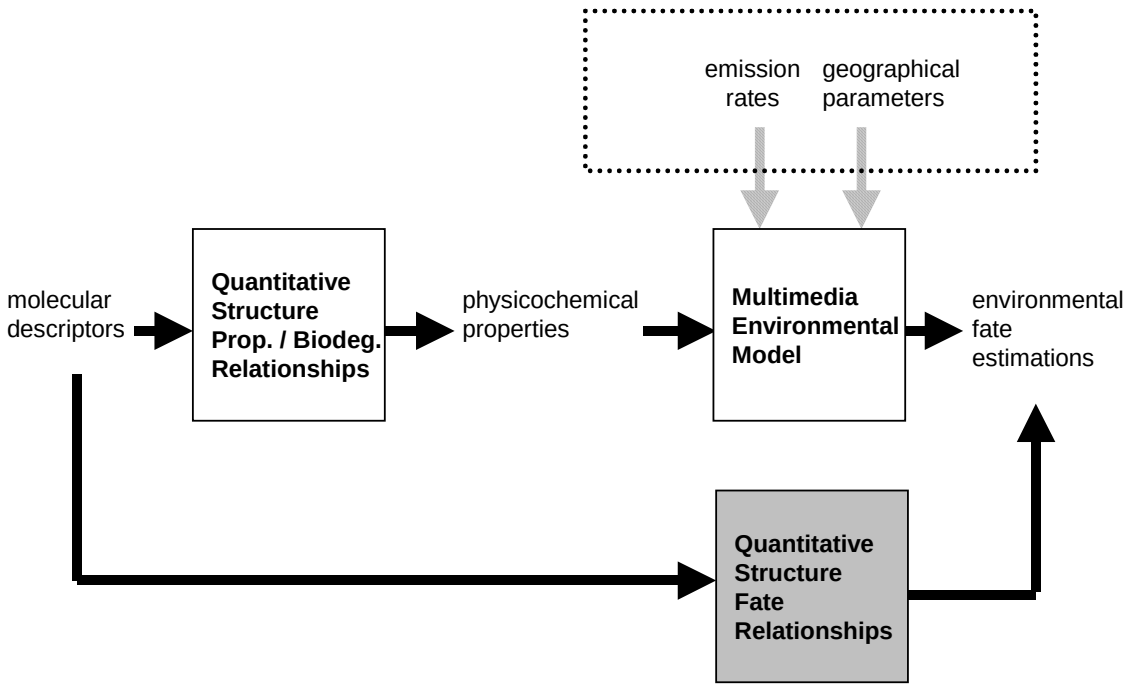


Figure 1. Two approaches for assessing environmental chemical partitioning from molecular information.

ISBN:978-84-693-4597-9 /DL:T.1010-2010

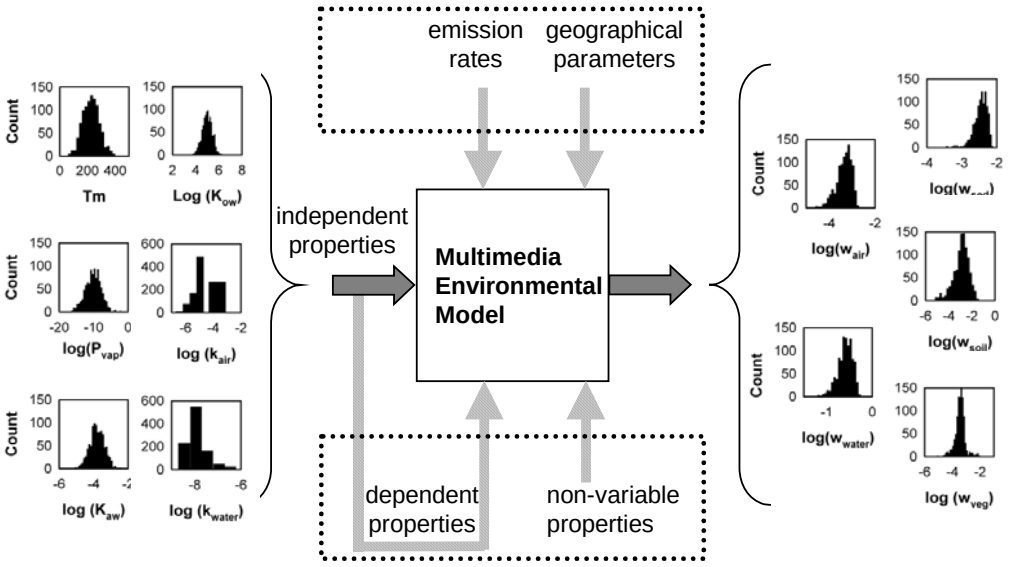


Figure 2. Random realizations of the Monte Carlo approach on SB3 for endrin.

ISBN:978-84-693-4597-9 /DL:T.1010-2010

Annex A.1 Paper on QSFRs to be submitted to Science of the Total Environment (STOTEN) in 2010 (continued)

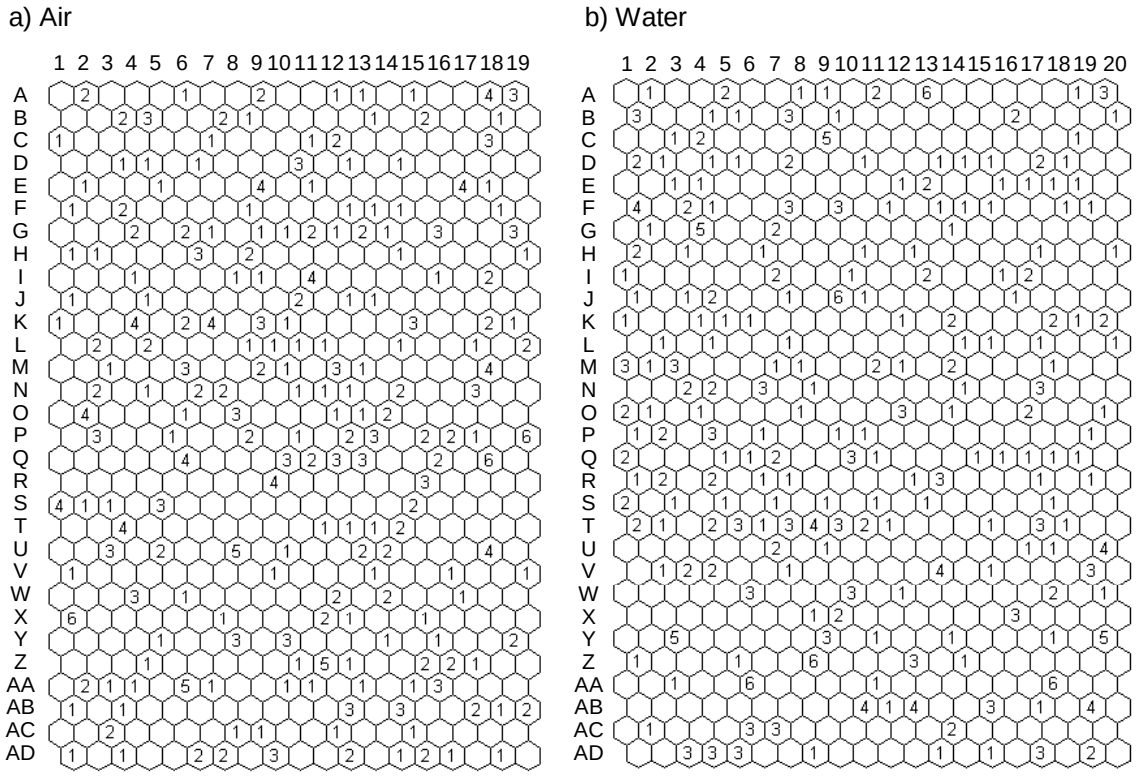


Figure 3. Clustering of the 375 work chemicals of the reference scenario in two SOMs.

ISBN:978-84-693-4597-9 /DL:T.1010-2010

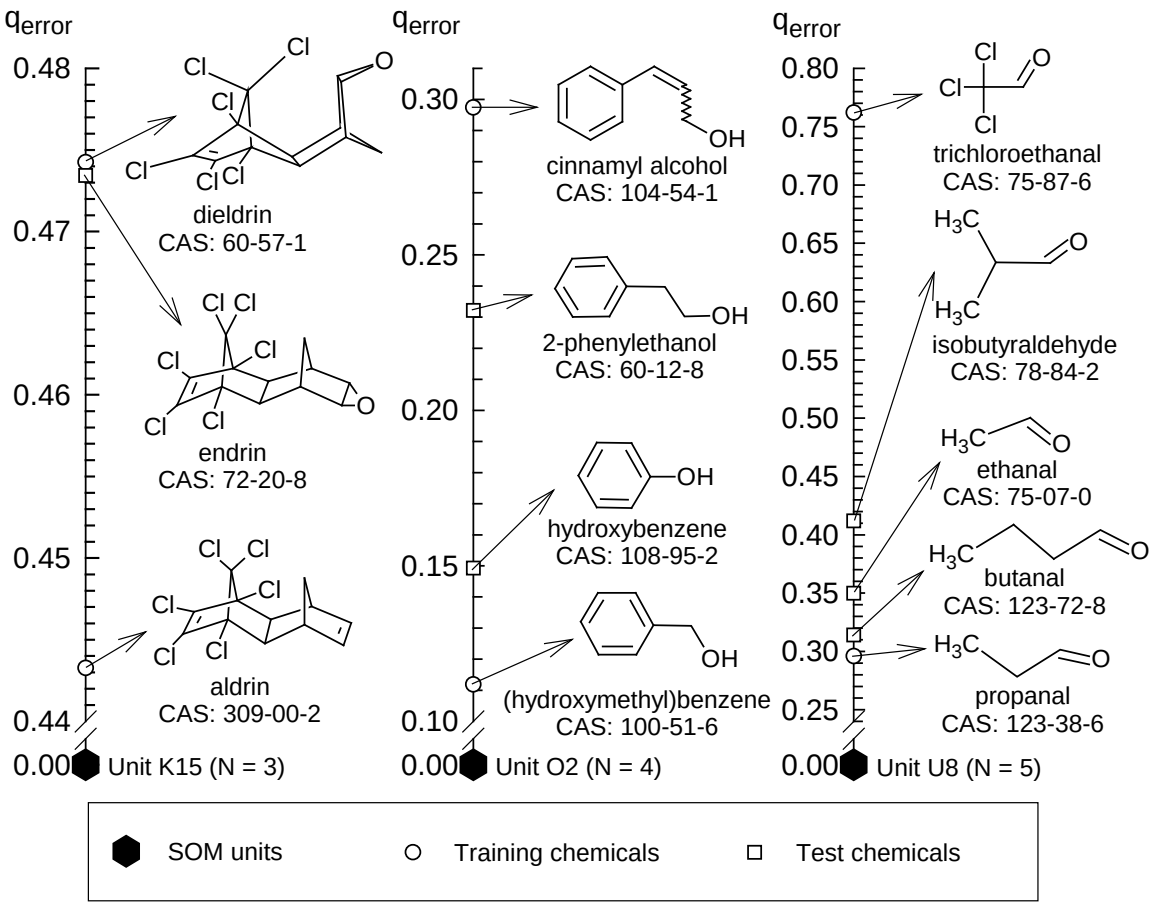


Figure 4. Distances of work chemicals to units of the SOM for the air compartment, expressed in quantization errors (q_{error}).

ISBN:978-84-693-4597-9 /DL:T.1010-2010

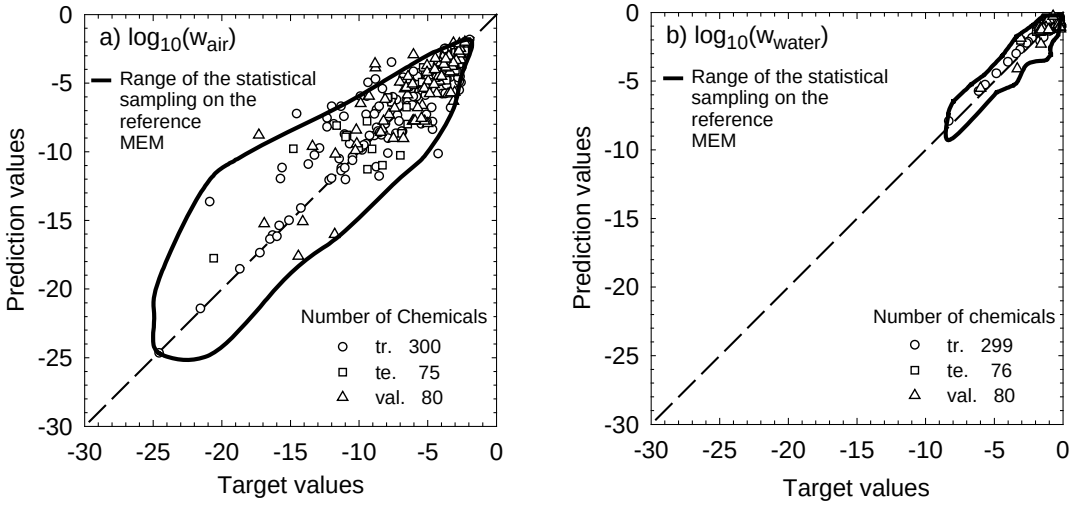


Figure 5. Predicted logarithmic mass ratios in: a) air, by means of $QSFR_{air}$ (with overall performances of $q^2 = 0.82$ and MAE = 0.91); and, b) water, by means of $QSFR_{water}$ (with overall performances of $q^2 = 0.81$ and MAE = 0.32).

ISBN:978-84-693-4597-9 /DL:T.1010-2010

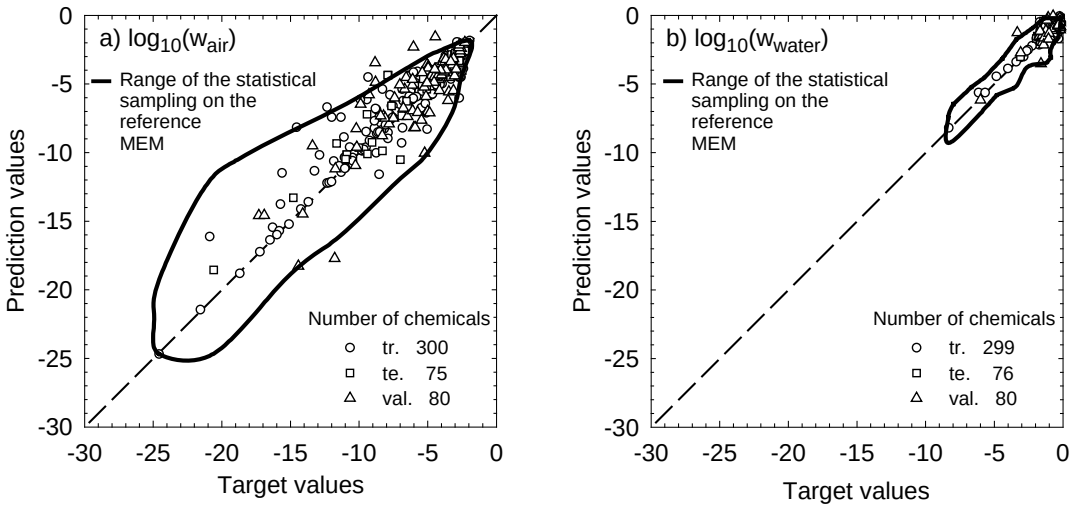


Figure 6. Predicted logarithmic mass ratios in: a) air, by means of $QSFR_{air,X/Y}$ (with overall performances of $q^2 = 0.88$ and MAE = 0.68); and, b) water, by means of $QSFR_{water,X/Y}$ (with overall performances of $q^2 = 0.87$ and MAE = 0.17).

ISBN:978-84-693-4597-9 /DL:T.1010-2010

ISBN:978-84-693-4597-9 /DL:T.1010-2010

Annex B

Program scripts used in this study

ISBN:978-84-693-4597-9 /DL:T.1010-2010

Annex B.1 Matlab script for training SOMs of different size with input and target variables of QPFRs or QSFRs.

```
function [ REPORT ] = Train_SOMs_for_MEM_screening(sD, selX, selY,...
   rem, maxC, perc);
% ----------------------Train_SOMs_for_MEM_screening-----------------------
% MATLAB script that trains SOMs of different size for the same work data,
% referred to the inputs and outputs of a Multimedia Environmental Model
% (MEM).
%
%
% Requirements
%
%      To have installed the "SOM Toolbox for Matlab Version 2"
%      http://www.cis.hut.fi/projects/somtoolbox/
%
%      To be, along with the script "SOM_analysis", in a folder with
%      sufficient hard disk space.
%
% Syntax
%
%      [ REPORT ] = SOMs_for_MEM_screening(sD, selX, selY...
%      rem, maxC, perc);
%
%
%      where:
%
%   sD is a data struct containing N chemicals (rows) characterized by
%      both X attributes and Y labels (columns). Given X+Y columns , the
%      attributes must be located in the first columns (1 to X),
%      while the label(s) must be located in the last columns (X+1 to
%      X+Y). The format of sD is that same as that used in the SOM
%      toolbox.
%
%  selX is a row vector listing all selected attributes (first columns)
%
%  selY is a row vector listing all selected targets (last columns)
%
%   rem is a row vector listing all the chemicals (rows) to be ignored
%
%  maxC is a scalar indicating then number of partitions to evaluate on
%      every SOM
%
%  perc is a row vector indicating the SOM sizes to be considered in terms
%      of percentages. In this script, the number of SOM units is
%      approximated to a percentage of the work chemicals available (e.g.
%      entering 100 percent would force a SOM to have approximately as
%      many units as chemicals available).
%
%
% Description
%
%      "Train_SOMs_for_MEM_screening" performs the following actions:
%
%      a) takes for a set of work chemicals, both chemical attributes
%      (e.g. properties, molecular descriptors) and multimedia
%      environmental fate estimations (e.g. concentrations, mass
%      fractions,fugacities, etc.).
%
%      b) applies a linear normalization to the data in the range [-1,1]
%
%      c) applies a loop in which SOMs of different size are trained and
%      save in .zip files
%
%      Every SOM generates a selection of training and test chemicals for
```

ISBN:978-84-693-4597-9 /DL:T.1010-2010

Annex B.1 Matlab script for training SOMs of different size with input and target variables of QPFRs or QSFRs (continued).

```
%       supervised learning algorithms to be used as QPFRs or QSFRs. A
%       training chemical is selected for training QSPRs or QSFRs when
%       having extreme values (lowest or highest) in its associated
%       quantization error (compared to chemicals in the same SOM unit)
%       or in its fate prediction (compared to all chemicals in the work
%       data set). The chemical space can be visualized by means of the component
%       planes of every SOM.
%
%       The bigger the SOM, the higher the number of training chemicals
%       selected. Several SOM sizes can be tested to inspect the number of
%       training chemicals selected, along with the quantization or
%       topological erors of each SOM (which indicate approximately the
%       goodness of the work data fitting).
%
% Warning
%
%       This is an script used for numerical experiments, it comes with no
%       warranty. The user is advised to check the code before using it and
%       be sure that there is enough hard disk space for saving the .zip
%       files in the work directory.
%
% Example:
%
%       %Given a set of 375 work chemicals and 93 validation chemicals,
%       %from the example case studied obtained the
%       %former can be classified as training and test chemicals for QSFRs
%       %for the air compartment of a given pollution scenario as follows:
%
%       % Loading the example data
%       load example_QSFR_data.mat
%
%       % Selecting MW and non-zero counts of molecular consituents
%       selX = [14 37:42 44:46 48:58 60:68 70:78]
%
%       % Selecting mass fractions in air as target variable
%       selY = [82]
%
%       % Discarding validation chemicals from the SOM analysis
%       rem = [376:468]
%
%       % Clustering of each SOM into 2 clusters maximum
%       maxC = 2
%
%       % Percentages of SOM units, related to the amount of work chemicals:
%       % 25%, 50%, 75%, 100%.
%       perc = [25 50 75 100]
%
%       % Running the script for training SOMs of varied size
%       [ REPORT ] = Train_SOMs_for_MEM_screening(sD, selX, selY,...
%                   rem, maxC, perc);
%
if nargin < 6
  error('Wrong number of arguments.');
end
if nargin > 6
  error('Wrong number of arguments.');
end

% Settings

    input_features = [selX selY]
    unused_samples = rem
```

ISBN:978-84-693-4597-9 /DL:T.1010-2010

Annex B.1 Matlab script for training SOMs of different size with input and target variables of QPFRs or QSFRs (continued).

```
    max_clusters = maxC
    percentages = perc

% Removal of unused attributes
    [dummy var] = size(sD.data);
    used_features = input_features;
    unused_features = setdiff([1:var],used_features);
    sD = som_modify_dataset(sD,'removecomp',unused_features);

% Removal of unused samples
    sD = som_modify_dataset(sD,'removesamp',unused_samples);

% Data normalization
    sDn = sD;
    sDn.data = sDn.data*0;
    p = sD.data';
    [pn,minp,maxp] = premnmx(p);
    sDn.data = pn';
    sD = sDn;
    clear sDn;

% Training different SOMs

    % Setting different SOM sizes to evaluate
    [samples dummy] = size(sD.data);
    percentages = percentages*(1/100);
    units = round(percentages*samples);
    clear side percentages;

    % Creating matrices for storing SOM outputs
    [dummy no] = size(units);
    qerrors = zeros(no,1);
    terrors = zeros(no,1);
    tr_percentage = zeros(no,1);
    te_percentage = zeros(no,1);
    tr_No = zeros(no,1);
    te_No = zeros(no,1);

for j=1:no;
    unit=units(1,j)
    sM = som_make(sD,'randinit','batch', 'munits',unit,'lattice','hexa',...
        'shape','toroid','neigh','gaussian');
    [qe,te] = som_quality(sM,sD);
    qerrors(j,1)=qe;
    terrors(j,1)=te;
    save map_and_data.mat sD sM input_features;

    %%%%%%%%% A) VISUALIZATION %%%%%%%%%%

    % Figure 01: SOM
    figure;
    som_show(sM);
    print('-dtiff',['Figure01_SOM.tiff'])
    close all;

    % Figure 02: U-Matrix
    figure;
    som_show(sM,'umat','all');
    print('-dtiff',['Figure02_Umat.tiff'])
    close all;

    % Figure 03: SOM components
```

ISBN:978-84-693-4597-9 /DL:T.1010-2010

Annex B.1 Matlab script for training SOMs of different size with input and target variables of QPFRs or QSFRs (continued).

```
figure;
som_show(sM,'comp','all','norm','d','bar','horiz');
print('-dtiff',['Figure03_components.tiff'])
close all;

% Figure 04: SOM geometry
figure;
som_show(sM,'empty','Labels','norm','d');
som_show_add('label',sM.labels,'textsize',8,'textcolor','r');
print('-dtiff',['Figure04_empty.tiff'])
close all;

% Figure 05: Color code in the SOM
f1=figure;
[Pd,V,me,l] = pcaproj(sD,2); Pm = pcaproj(sM,V,me); % PC-projection
Code = som_colorcode(Pm); % color coding
hits = som_hits(sM,sD);  % hits
U = som_umat(sM); % U-matrix
Dm = U(1:2:size(U,1),1:2:size(U,2)); % distance matrix
Dm = 1-Dm(:)/max(Dm(:)); Dm(find(hits==0)) = 0; % clustering info

som_cplane(sM,Code,Dm);
hold on
som_grid(sM,'Label',cellstr(int2str(hits)),...
        'Line','none','Marker','none','Labelcolor','k');
hold off
title('Color code')
print('-dtiff',['Figure05_colorcode.tiff'])
close all;

% Figure 06: PCA projection
figure;
som_grid(sM,'Coord',Pm,'MarkerColor',Code,'Linecolor','k');
hold on, plot(Pd(:,1),Pd(:,2),'k+'), hold off, axis tight, axis equal
title('PC projection')
print('-dtiff',['Figure06_PC_projection.tiff'])
close all;

% Figure 07: Labels in the SOM
figure;
som_cplane(sM,'none')
hold on
som_grid(sM,'Label',sM.labels,'Labelsize',8,...
        'Line','none','Marker','none','Labelcolor','r');
hold off;
title('Labels');
print('-dtiff',['Figure07_labels.tiff'])
close all;

save color_code.mat Code sM sD Dm input_features selX selY;

%%%%%%%%% B) CLUSTERING %%%%%%%%%%

% Figure 08: Davies-Boulding index vs number of clusters
figure;
[c,p,err,ind] = kmeans_clusters(sM, max_clusters);
plot(1:length(ind),ind,'x-')
[dummy,i] = min(ind)
cl = p{i};
hold on;
xlabel('number of clusters');
ylabel('Davies-Boulding index');
print('-dtiff',['Figure08_DB_indexes.tiff'])
```

ISBN:978-84-693-4597-9 /DL:T.1010-2010

Annex B.1 Matlab script for training SOMs of different size with input and target variables of QPFRs or QSFRs (continued).

```
close all;

% Figure 09: Colorcode in the SOM
figure;
som_cplane(sM,Code,Dm);
title('color code');
print('-dtiff',['Figure09_Colorcode.tiff'])
close all;

% Figure 10: Partitions in the SOM
figure;
som_cplane(sM,cl);
title('partitions');
print('-dtiff',['Figure10_Partitions.tiff'])
close all;

%%%%%%%%% C) REPORT %%%%%%%%%%

[ Data_clusters, Data_for_QSFRs ] = SOM_analysis( sM, sD, cl, selX, selY);

tr_percentage(j,1) = Data_for_QSFRs.est_training_percentage;
te_percentage(j,1) = Data_for_QSFRs.est_testing_percentage;
tr_No(j,1) = Data_for_QSFRs.no_training_chemicals
te_No(j,1) = Data_for_QSFRs.no_testing_chemicals

if j<100, ceros='0', end;
if j<10, ceros='00', end;
%zip(['map_',num2str(ceros),num2str(j)],'map_and_data.mat');
%delete map_and_data.mat;

zip(['map_',num2str(ceros),num2str(j)],...
    {...
      '01_BMUs.csv',...
      '02_Qerrs.csv',...
      '03_Training.csv',...
      '04_minQerrs.csv',...
      '05_maxQerrs.csv',...
      '06_above_meanQerr.csv',...
      '07_Extreme.csv',...
      '08_Clusters.csv',...
      'Figure01_SOM.tiff',...
      'Figure02_Umat.tiff',...
      'Figure03_components.tiff',...
      'Figure04_empty.tiff',...
      'Figure05_colorcode.tiff',...
      'Figure06_PC_projection.tiff',...
      'Figure07_labels.tiff',...
      'Figure08_DB_indexes.tiff',...
      'Figure09_Colorcode.tiff',...
      'Figure10_Partitions.tiff',...
      'example_QSFR_data.mat',...
      'map_and_data.mat',...
      'color_code.mat'...
    });

% Deleting all files produced
delete(...
      '01_BMUs.csv',
      '02_Qerrs.csv',
      '03_Training.csv',
      '04_minQerrs.csv',
      '05_maxQerrs.csv',
      '06_above_meanQerr.csv',
```

ISBN:978-84-693-4597-9 /DL:T.1010-2010

Annex B.1 Matlab script for training SOMs of different size with input and target variables of QPFRs or QSFRs (continued).

```
        '07_Extreme.csv',...
        '08_Clusters.csv',...
        'Figure01_SOM.tiff',...
        'Figure02_Umat.tiff',...
        'Figure03_components.tiff',...
        'Figure04_empty.tiff',...
        'Figure05_colorcode.tiff',...
        'Figure06_PC_projection.tiff',...
        'Figure07_labels.tiff',...
        'Figure08_DB_indexes.tiff',...
        'Figure09_Colorcode.tiff',...
        'Figure10_Partitions.tiff',...
        'map_and_data.mat',...
        'color_code.mat'...
        );
end;
close all;

% Plotting quantization errors for all SOMs tested
figure;
plot([1:no]',qerrors);
title('Qerrors vs # of Map');
xlabel('# of Map');
ylabel('Qerror');
print('-djpeg',['Qerrors_graph.jpeg'])
close all;

% Plotting topological errors for all SOMs tested
figure;
plot([1:no]',terrors);
title('Terrors vs # of Map');
xlabel('# of Map');
ylabel('Terror');
print('-djpeg',['Terrors_graph.jpeg'])
close all;

% Plotting the percentage of training chemicals derived from the work data
% for all chemicals
figure;
plot([1:no]',tr_percentage);
title('TR percentage vs # of Map');
xlabel('# of Map');
ylabel('Terror');
print('-djpeg',['TR_percentage_graph.jpeg'])
close all;

% Saving a MATLAB file with the characteristics of every SOM
REPORT.percentages = perc';
REPORT.SOMs = [1:no]';
REPORT.qerrors = qerrors;
REPORT.terrors = terrors;
REPORT.training_chemicals = tr_No;
REPORT.tr_percentage = tr_percentage;
REPORT.test_chemicals = te_No;
REPORT.te_percentage = te_percentage;
save REPORT.mat REPORT;
```

ISBN:978-84-693-4597-9 /DL:T.1010-2010

Annex B.2 Matlab script for evaluating iteratively different SOM clusterings.

```
% -------------------------Iterate_SOM_clustering-------------------------
% MATLAB script that performs the clustering of a selected SOM in an
% iterative manner, using the Davies-Bouldin algorithm.
%
%
% Requirements
%
%       To have installed the "SOM Toolbox for Matlab Version 2.0beta"
%       http://www.cis.hut.fi/projects/somtoolbox/
%
%       To be in a folder, along with a MATLAB data file called
%       "color_code.mat" (that can be obtained from the .zip file of a SOM
%       trained with the script "Train_SOMs_for_MEM_screening"),
%       with sufficient hard disk space.
%
%
% Description
%
%       "Iterate_SOM_clustering" performs a repeated clustering of a SOM
%       with basis on the Davies-Bouldin (DB) algorithm and compares the
%       number of clusters and the associated DB indexes associated to each
%       resulting SOM clustering. The SOM clustering with the lowest DB
%       index is the optimal one.
%
%
% Warning
%
%       This is an script used for numerical experiments, it comes with no
%       warranty. The user is advised to check the code before using it and
%       be sure that there is enough hard disk space for saving the .zip
%       files in the work directory.
%
% Example:
%
%       %First make use of the function  "Train_SOMs_for_MEM_screening" for
%       %generating various SOMs, select one of them and take the
%       %"color_code.mat" file associated to it and place it in a
%       %folder along with this script.
%
%       % Especify the maximum number of clusters, for example: 2
%       maxC = 2
%
%       % Especify the maximum number of iterations, for example: 10
%       maxC = 10
%
%       % Running the script for iterating the clustering of the SOM
%       run('Iterate_SOM_clustering')

load color_code.mat;
max_clusters = maxC;
iterations = iter;

report_clusters = zeros(iterations,1);
report_DB_indexes = zeros(iterations,1);

for QQ=1:iterations;

% Figure: Davies-Boulding index vs number of clusters
figure;
[c,p,err,ind] = kmeans_clusters(sM, max_clusters); % find at most 7 clusters
plot(1:length(ind),ind,'x-')
[dummy,i] = min(ind)
cl = p{i};
```

ISBN:978-84-693-4597-9 /DL:T.1010-2010

Annex B.2 Matlab script for evaluating iteratively different SOM clusterings (continued).

```
hold on;
xlabel('number of clusters');
ylabel('Davies-Boulding index');
print('-dtiff',['Figure01_DB_indexes.tiff'])
close all;

figure;
som_cplane(sM,Code,Dm);
title('color code');
print('-dtiff',['Figure02_Colorcode.tiff'])
close all;

figure;
som_cplane(sM,cl);
title('partitions');
print('-dtiff',['Figure03_Partitions.tiff'])
close all;

   report_clusters(QQ,1) = i;
   report_DB_indexes(QQ,1) = min(ind);

save davis_boulding.mat;

% Zipping results
   if j<100, ceros='0', end;
   if j<10, ceros='00', end;

   zip(['iteration',num2str(ceros),num2str(QQ)],...
       {...
         'davis_boulding.mat',...
         'Figure01_DB_indexes.tiff',...
         'Figure02_Colorcode.tiff',...
         'Figure03_Partitions.tiff'...
       });

       % Deleting all files produced
   delete(...
         'davis_boulding.mat',...
         'Figure01_DB_indexes.tiff',...
         'Figure02_Colorcode.tiff',...
         'Figure03_Partitions.tiff'...
         );

end;

close all;

figure;
plot([1:iterations]',report_clusters);
title('Number of clusters vs Iterations');
xlabel('Iterations');
ylabel('Number of clusters');
print('-dtiff',['Clusters_graph.tiff'])
close all;

figure;
plot([1:iterations]',report_DB_indexes);
title('Davis-Boulding indexes vs iterations (rect. lattice)');
xlabel('Iterations');
ylabel('Davis-Boulding indexes');
print('-dtiff',['Davis-Boulding_graph.tiff'])
close all;

save comparison.mat;
return;
```

ISBN:978-84-693-4597-9 /DL:T.1010-2010

Annex B.3 Matlab script for evaluating new chemicals in a trained SOM.

```
% -------------------------Evaluate_new_chemicals--------------------------
%
% MATLAB script that performs the evaluation of both old and new data in
% a SOM already trained by the function "Train_SOMs_for_MEM_screening"
%
%
% Requirements
%
%       To have installed the "SOM Toolbox for Matlab Version 2.0beta"
%       http://www.cis.hut.fi/projects/somtoolbox/
%
%       To be in a folder, along with a MATLAB data file called
%       "davis_boulding.mat" (that can be obtained from the .zip file of
%       any of the clustering iterations performed with the script
%       "Iterate_SOM_clustering") and another containing the original
%       data on evaluation.
%
%
% Description
%
%       "Evaluate_new_chemicals" enters the attributes and labels in a
%       trained SOM for any set of chemicals, indicating their corresponding
%       clustering.
%
%
% Warning
%
%       This is an script used for numerical experiments, it comes with no
%       warranty. The user is advised to check the code before using it and
%       be sure that there is enough hard disk space.
%
%
% Example:
%
%       % First make use of the function "Train_SOMs_for_MEM_screening" for
%       % generating various SOMs, select one of them and generate a SOM
%       % clustering with the script "Iterate_SOM_clustering". Select the
%       % clustering iteration for which the Davies-Bouldin index is a minimum
%       % and take the file "davis_boulding.mat", it contains variables
%       % required here.
%
%
%       % Load the clustering data
%       load davis_boulding.mat;
%
%       % Load all data (this file contains both work and validation
%       % chemicals)
%       load example_QSFR_data.mat
%
%       % Running the script for knowing the clustering of new chemicals
%       run('Evaluate_new_chemicals')


% Loading the clustering
load davis_boulding.mat;

% Loading all data
load example_QSFR_data.mat

% Variables of the SOM
   input_features = [selX selY]
```

ISBN:978-84-693-4597-9 /DL:T.1010-2010

Annex B.3 Matlab script for evaluating new chemicals in a trained SOM (continued).

```
% Removal of unused features
[samples var] = size(sD.data);
used_features = input_features; % union(input_features,target_features;)
unused_features = setdiff([1:var],used_features)
sD = som_modify_dataset(sD,'removecomp',unused_features)

[ Data_clusters, Data_for_QSFRs ] = SOM_analysis( sM, sD, cl, selX, selY);
save classification_val.mat Data_clusters Data_for_QSFRs;
```

ISBN:978-84-693-4597-9 /DL:T.1010-2010

Annex B.4 RapidMiner script for simple validation of SVRs with different parameter combinations.

```
<operator name="Root" class="Process" expanded="yes">
   <description text="#ylt#p#ygt# Often the different operators have many parameters and it is not clear which
parameter values are best for the learning task at hand. The parameter optimization operator helps to find an optimal
parameter set for the used operators. #ylt#/p#ygt#  #ylt#p#ygt# The inner crossvalidation estimates the performance
for each parameter set. In this experiment two parameters of the SVM are tuned. The result can be plotted in 3D
(using gnuplot) or in color mode. #ylt#/p#ygt#  #ylt#p#ygt# Try the following: #ylt#ul#ygt# #ylt#li#ygt#Start the
experiment. The result is the best parameter set and the performance which was achieved with this parameter
set.#ylt#/li#ygt# #ylt#li#ygt#Edit the parameter list of the ParameterOptimization operator to find another parameter
set.#ylt#/li#ygt# #ylt#/ul#ygt# #ylt#/p#ygt# "/>
   <operator name="Work chemicals" class="ExcelExampleSource">
      <parameter key="excel_file"       value="C:\SimpleVal\test_data.xls"/>
      <parameter key="first_row_as_names"       value="true"/>
      <parameter key="create_label"     value="true"/>
      <parameter key="label_column"  value="41"/>
      <parameter key="create_id"        value="true"/>
   </operator>
   <operator name="ParameterOptimization" class="GridParameterOptimization" expanded="yes">
      <list key="parameters">
        <parameter key="LibSVMLearner.C"       value="0,1,5,10,25,50,75,100,150,300"/>
        <parameter key="LibSVMLearner.epsilon"
            value="0.000001,0.00001,0.0001,0.001,0.01,0.1,0.25,0.50,0.75,0.90"/>
        <parameter key="LibSVMLearner.p"
            value="0.000001,0.00001,0.0001,0.001,0.01,0.1,0.25,0.50,0.75,0.90"/>
        <parameter key="LibSVMLearner.gamma"             value="0,1,5,10"/>
      </list>
      <operator name="SimpleValidation" class="SimpleValidation" expanded="yes">
         <parameter key="keep_example_set"      value="true"/>
         <parameter key="create_complete_model"            value="true"/>
         <parameter key="split_ratio"    value="0.5"/>
         <parameter key="sampling_type"           value="linear sampling"/>
         <operator name="LibSVMLearner" class="LibSVMLearner">
            <parameter key="keep_example_set"  value="true"/>
            <parameter key="svm_type"value="epsilon-SVR"/>
            <parameter key="degree"     value="1"/>
            <parameter key="gamma"    value="0"/>
            <parameter key="coef0"      value="0"/>
            <parameter key="C"            value="300"/>
            <parameter key="nu"           value="0.0"/>
            <parameter key="epsilon"    value="0.000001"/>
            <parameter key="p"            value="0.000001"/>
            <list key="class_weights">
            </list>
         </operator>
         <operator name="ApplierChain" class="OperatorChain" expanded="yes">
            <operator name="IOMultiplier" class="IOMultiplier">
               <parameter key="number_of_copies"            value="2"/>
               <parameter key="io_object"           value="Model"/>
            </operator>
            <operator name="SupportVectorCounter" class="SupportVectorCounter">
            </operator>
            <operator name="Test" class="ModelApplier">
               <list key="application_parameters">
               </list>
            </operator>
            <operator name="Evaluation" class="RegressionPerformance">
               <parameter key="main_criterion"    value="absolute_error"/>
               <parameter key="root_mean_squared_error"  value="true"/>
               <parameter key="absolute_error"     value="true"/>
               <parameter key="relative_error"      value="true"/>
               <parameter key="relative_error_lenient"        value="true"/>
               <parameter key="relative_error_strict"           value="true"/>
```

Gray-shaded code: code to be modified by the user.

ISBN:978-84-693-4597-9 /DL:T.1010-2010

Annex B.4 RapidMiner script for simple validation of SVRs with different parameter combinations (continued).

```
            <parameter key="normalized_absolute_error" value="true"/>
            <parameter key="root_relative_squared_error"  value="true"/>
            <parameter key="squared_error"  value="true"/>
            <parameter key="correlation"  value="true"/>
            <parameter key="squared_correlation"  value="true"/>
            <parameter key="prediction_average"  value="true"/>
            <parameter key="spearman_rho"  value="true"/>
            <parameter key="kendall_tau"  value="true"/>
          </operator>
        </operator>
      </operator>
      <operator name="Log" class="ProcessLog">
        <parameter key="filename"  value="C:\SimpleVal\Comparison_table.log"/>
        <list key="log">
          <parameter key="VALIDATION_applycount"  value="operator.SimpleValidation.value.applycount"/>
          <parameter key="gamma"  value="operator.LibSVMLearner.parameter.gamma"/>
          <parameter key="C"  value="operator.LibSVMLearner.parameter.C"/>
          <parameter key="epsilon"  value="operator.LibSVMLearner.parameter.epsilon"/>
          <parameter key="p"value="operator.LibSVMLearner.parameter.p"/>
          <parameter key="EVAL_absolute_error"  value="operator.Evaluation.value.absolute_error"/>
          <parameter key="EVAL_correlation"  value="operator.Evaluation.value.correlation"/>
          <parameter key="EVAL_squared_correlation"  value="operator.Evaluation.value.squared_correlation"/>
          <parameter key="EVAL_prediction_average"  value="operator.Evaluation.value.prediction_average"/>
          <parameter key="EVAL_Suppor_Vectors"
           value="operator.SupportVectorCounter.value.support_vectors"/>
          <parameter key="VALIDATION_performance"  value="operator.SimpleValidation.value.performance"/>
          <parameter key="VALIDATION_performance1" value="operator.SimpleValidation.value.performance1"/>
          <parameter key="VALIDATION_performance2" value="operator.SimpleValidation.value.performance2"/>
          <parameter key="VALIDATION_performance3" value="operator.SimpleValidation.value.performance3"/>
          <parameter key="VALIDATION_deviation"  value="operator.SimpleValidation.value.deviation"/>
          <parameter key="VALIDATION_variance"  value="operator.SimpleValidation.value.variance"/>
        </list>
        <parameter key="persistent"  value="true"/>
      </operator>
    </operator>
</operator>
```

Gray-shaded code: code to be modified by the user.

ISBN:978-84-693-4597-9 /DL:T.1010-2010

Annex B.5 RapidMiner script for training a SVR with optimized parameters and later perfroming fate predictions for all chemicals simultaneously.

```
<operator name="Root" class="Process" expanded="yes">
  <operator name="1 Loading training data" class="ExcelExampleSource">
    <parameter key="excel_file"       value="C:\SVRs\training_data.xls"/>
    <parameter key="sheet_number"  value="4"/>
    <parameter key="first_row_as_names"       value="true"/>
    <parameter key="create_label"     value="true"/>
    <parameter key="label_column"  value="25"/>
    <parameter key="create_id"        value="true"/>
    <parameter key="decimal_point_character"value=""/>
  </operator>
  <operator name="2 Training a SVR" class="LibSVMLearner">
    <parameter key="svm_type"       value="epsilon-SVR"/>
    <parameter key="degree"           value="1"/>
    <parameter key="gamma"           value="1.0"/>
    <parameter key="C"     value="10.0"/>
    <parameter key="nu"    value="0.0"/>
    <parameter key="epsilon"           value="0.25"/>
    <parameter key="p"      value="1.0E-5"/>
    <list key="class_weights">
    </list>
  </operator>
  <operator name="3 Saving the trained SVR model" class="ModelWriter">
    <parameter key="model_file"      value="C:\SVRs\SVR.mod"/>
    <parameter key="output_type"     value="XML"/>
  </operator>
  <operator name="4 Cleaning the memory" class="MemoryCleanUp">
  </operator>
  <operator name="5 Loading all data (training, test and validation data)" class="ExcelExampleSource">
    <parameter key="excel_file"       value="C:\SVRs\training_test_and_validation_data.xls"/>
    <parameter key="sheet_number"  value="5"/>
    <parameter key="first_row_as_names"       value="true"/>
    <parameter key="create_label"     value="true"/>
    <parameter key="label_column"  value="25"/>
    <parameter key="create_id"        value="true"/>
    <parameter key="decimal_point_character"value=""/>
  </operator>
  <operator name="6 Loading the SVR model" class="ModelLoader">
    <parameter key="model_file"      value="C:\SVR\SVR.mod"/>
  </operator>
  <operator name="7 Predicting fate for all chemicals" class="ModelApplier" breakpoints="after">
    <list key="application_parameters">
    </list>
  </operator>
  <operator name="8 Saving predictions to Excel file" class="ExcelExampleSetWriter">
    <parameter key="excel_file"       value="C:\SVR\fate_predictions.xls"/>
  </operator>
  <operator name="9 Measuring the overall performance" class="RegressionPerformance">
    <parameter key="main_criterion"value="absolute_error"/>
    <parameter key="root_mean_squared_error"           value="true"/>
    <parameter key="absolute_error" value="true"/>
    <parameter key="relative_error"  value="true"/>
    <parameter key="relative_error_lenient"     value="true"/>
    <parameter key="relative_error_strict"       value="true"/>
    <parameter key="normalized_absolute_error"           value="true"/>
    <parameter key="root_relative_squared_error"          value="true"/>
    <parameter key="squared_error"  value="true"/>
    <parameter key="correlation"      value="true"/>
    <parameter key="squared_correlation"       value="true"/>
    <parameter key="prediction_average"        value="true"/>
    <parameter key="spearman_rho"  value="true"/>
    <parameter key="kendall_tau"     value="true"/>
  </operator>
```

Gray-shaded code: code to be modified by the user.

ISBN:978-84-693-4597-9 /DL:T.1010-2010

Annex B.5 RapidMiner script for training a SVR with optimized parameters and later perfroming fate predictions for all chemicals simultaneously (continued).

```
   <operator name="10 Saving overall performance results" class="PerformanceWriter">
      <parameter key="performance_file"          value="C:\SVR\performance.per"/>
   </operator>
</operator>
```

Gray-shaded code: code to be modified by the user.

ISBN:978-84-693-4597-9 /DL:T.1010-2010

Annex B.6 RapidMiner script for performing a 10-fold cross validation on a SVR with optimized parameters.

```
<operator name="Root" class="Process" expanded="yes">
   <description text="#ylt#p#ygt# In many cases not the learned model is of interest but the accuracy of the model. One possible solution to estimate the predictiveness of the learned model is to apply it to labeled test data and calculate the number of prediction errors (or other performance criteria). Since labeled data is rare, other approaches to estimate the performance of a learning scheme are often used. This process demonstrates #yquot#cross validation#yquot# in RapidMiner.#ylt#/p#ygt#  #ylt#table#ygt# #ylt#tr#ygt# #ylt#td#ygt##ylt#icon#ygt#groups/24/validation#ylt#/icon#ygt##ylt#/td#ygt# #ylt#td#ygt##ylt#p#ygt#Cross validation divides the labelled data in training and test sets. Models are learned on training data and applied on test data. The prediction errors are calculated and averaged for all subsets. This building block can be used as inner operator for several wrappers like feature generation / selection operators. #ylt#/p#ygt##ylt#/td#ygt# #ylt#/tr#ygt# #ylt#/table#ygt#    #ylt#p#ygt# This is the first example of a more complex process. The operators build a tree structure. For now it is enough to accept that the cross validation operator demands an example set as input and delivers a vector of performance values as output. Additionally it manages the division into training and test examples. The training examples are used as input for the training learner which delivers a model. This model and the test examples form the input of the applier chain which delivers the performance for this test set. The results for all possible test sets are collected by the cross validation operator. Finally the average is built and delivered as result. #ylt#/p#ygt#     #ylt#p#ygt#One of the hardest things for the RapidMiner beginner is often to get an idea of the #ylt#b#ygt#data flow#ylt#/b#ygt#. The solution is surprisingly simple: the data flow resembles a depth-first-search through the tree structure. For example, after processing the training set with the first child of the cross validation the learned model, is delivered to the second child (the applier chain). This basic data flow idea is always the same for all processes and thinking in this flow will become very convenient for the experienced user.#ylt#/p#ygt# #ylt#p#ygt#Try the following:#ylt#/p#ygt# #ylt#ul#ygt##ylt#li#ygt#Start the process. The result is a performance estimation of the learning scheme on the input data.#ylt#/li#ygt#  #ylt#li#ygt#Select the Evaluation operator and select other performance criteria. The main criterion is used for performance comparisons, for example in a wrapper.#ylt#/li#ygt#  #ylt#li#ygt#Replace the cross validation #yquot#XVal#yquot# by other evaluation schemes and run the process with them. Alternatively you can check how other learners perform on this data and replace the Training operator.#ylt#/li#ygt##ylt#/ul#ygt#"/>
   <operator name="ExcelExampleSource" class="ExcelExampleSource">
      <parameter key="excel_file"       value="C:\10fold-CV\work_data.xls"/>
      <parameter key="sheet_number"  value="2"/>
      <parameter key="first_row_as_names"       value="true"/>
      <parameter key="create_label"     value="true"/>
      <parameter key="label_column"  value="41"/>
      <parameter key="create_id"        value="true"/>
   </operator>
   <operator name="XVal" class="XValidation" expanded="yes">
      <parameter key="sampling_type" value="shuffled sampling"/>
      <operator name="Training" class="LibSVMLearner">
         <parameter key="svm_type"     value="epsilon-SVR"/>
         <parameter key="degree"        value="1"/>
         <parameter key="C"  value="300.0"/>
         <parameter key="epsilon"       value="0.1"/>
         <parameter key="p"  value="1.0E-5"/>
         <list key="class_weights">
         </list>
      </operator>
      <operator name="ApplierChain" class="OperatorChain" expanded="yes">
         <operator name="Test" class="ModelApplier">
            <list key="application_parameters">
            </list>
         </operator>
         <operator name="Evaluation" class="RegressionPerformance">
            <parameter key="keep_example_set"  value="true"/>
            <parameter key="main_criterion"       value="absolute_error"/>
            <parameter key="root_mean_squared_error"      value="true"/>
            <parameter key="absolute_error"       value="true"/>
            <parameter key="relative_error"       value="true"/>
            <parameter key="relative_error_lenient"           value="true"/>
            <parameter key="relative_error_strict"value="true"/>
            <parameter key="normalized_absolute_error"    value="true"/>
            <parameter key="root_relative_squared_error"   value="true"/>
            <parameter key="squared_error"        value="true"/>
```

Gray-shaded code: code to be modified by the user.

ISBN:978-84-693-4597-9 /DL:T.1010-2010

Annex B.6 RapidMiner script for performing a 10-fold cross validation on a SVR with optimized parameters.

```
            <parameter key="correlation"            value="true"/>
            <parameter key="squared_correlation"value="true"/>
            <parameter key="prediction_average" value="true"/>
            <parameter key="spearman_rho"        value="true"/>
            <parameter key="kendall_tau"            value="true"/>
        </operator>
        <operator name="PerformanceWriter" class="PerformanceWriter">
            <parameter key="performance_file"    value="C:\10fold-CV\10fold-CV_performance.per"/>
        </operator>
    </operator>
  </operator>
</operator>
```

Gray-shaded code: code to be modified by the user.

ISBN:978-84-693-4597-9 /DL:T.1010-2010

Annex B.7 RapidMiner script for performing a LOO validation on a SVR with optimized parameters.

```
<operator name="Root" class="Process" expanded="yes">
   <description text="#ylt#p#ygt# In many cases not the learned model is of interest but the accuracy of the model. One possible solution to estimate the predictiveness of the learned model is to apply it to labeled test data and calculate the number of prediction errors (or other performance criteria). Since labeled data is rare, other approaches to estimate the performance of a learning scheme are often used. This process demonstrates #yquot#cross validation#yquot# in RapidMiner.#ylt#/p#ygt#  #ylt#table#ygt# #ylt#tr#ygt# #ylt#td#ygt##ylt#icon#ygt#groups/24/validation#ylt#/icon#ygt##ylt#/td#ygt# #ylt#td#ygt##ylt#p#ygt#Cross validation divides the labelled data in training and test sets. Models are learned on training data and applied on test data. The prediction errors are calculated and averaged for all subsets. This building block can be used as inner operator for several wrappers like feature generation / selection operators. #ylt#/p#ygt##ylt#/td#ygt# #ylt#/tr#ygt# #ylt#/table#ygt#    #ylt#p#ygt# This is the first example of a more complex process. The operators build a tree structure. For now it is enough to accept that the cross validation operator demands an example set as input and delivers a vector of performance values as output. Additionally it manages the division into training and test examples. The training examples are used as input for the training learner which delivers a model. This model and the test examples form the input of the applier chain which delivers the performance for this test set. The results for all possible test sets are collected by the cross validation operator. Finally the average is built and delivered as result. #ylt#/p#ygt#     #ylt#p#ygt#One of the hardest things for the RapidMiner beginner is often to get an idea of the #ylt#b#ygt#data flow#ylt#/b#ygt#. The solution is surprisingly simple: the data flow resembles a depth-first-search through the tree structure. For example, after processing the training set with the first child of the cross validation the learned model, is delivered to the second child (the applier chain). This basic data flow idea is always the same for all processes and thinking in this flow will become very convenient for the experienced user.#ylt#/p#ygt# #ylt#p#ygt#Try the following:#ylt#/p#ygt# #ylt#ul#ygt##ylt#li#ygt#Start the process. The result is a performance estimation of the learning scheme on the input data.#ylt#/li#ygt#  #ylt#li#ygt#Select the Evaluation operator and select other performance criteria. The main criterion is used for performance comparisons, for example in a wrapper.#ylt#/li#ygt#  #ylt#li#ygt#Replace the cross validation #yquot#XVal#yquot# by other evaluation schemes and run the process with them. Alternatively you can check how other learners perform on this data and replace the Training operator.#ylt#/li#ygt##ylt#/ul#ygt#"/>
   <operator name="ExcelExampleSource" class="ExcelExampleSource">
      <parameter key="excel_file"       value="C:\LOO\work_data.xls"/>
      <parameter key="sheet_number"  value="2"/>
      <parameter key="first_row_as_names"       value="true"/>
      <parameter key="create_label"     value="true"/>
      <parameter key="label_column"  value="41"/>
      <parameter key="create_id"        value="true"/>
   </operator>
   <operator name="XVal" class="XValidation" expanded="yes">
      <parameter key="leave_one_out" value="true"/>
      <parameter key="sampling_type" value="shuffled sampling"/>
      <operator name="Training" class="LibSVMLearner">
         <parameter key="svm_type"     value="epsilon-SVR"/>
         <parameter key="degree"        value="1"/>
         <parameter key="C"  value="300.0"/>
         <parameter key="epsilon"       value="0.1"/>
         <parameter key="p"  value="1.0E-5"/>
         <list key="class_weights">
         </list>
      </operator>
      <operator name="ApplierChain" class="OperatorChain" expanded="yes">
         <operator name="Test" class="ModelApplier">
            <list key="application_parameters">
            </list>
         </operator>
         <operator name="Evaluation" class="RegressionPerformance">
            <parameter key="main_criterion"       value="absolute_error"/>
            <parameter key="root_mean_squared_error"      value="true"/>
            <parameter key="absolute_error"       value="true"/>
            <parameter key="relative_error"       value="true"/>
            <parameter key="relative_error_lenient"           value="true"/>
            <parameter key="relative_error_strict"value="true"/>
            <parameter key="squared_error"        value="true"/>
            <parameter key="correlation"          value="true"/>
```

Gray-shaded code: code to be modified by the user.

ISBN:978-84-693-4597-9 /DL:T.1010-2010

Annex B.7 RapidMiner script for performing a LOO validation on a SVR with optimized parameters (continued).

```
            <parameter key="squared_correlation"value="true"/>
            <parameter key="prediction_average" value="true"/>
            <parameter key="spearman_rho"       value="true"/>
            <parameter key="kendall_tau"        value="true"/>
        </operator>
        <operator name="PerformanceWriter" class="PerformanceWriter">
            <parameter key="performance_file"   value="C:\LOO\ LOO_performance.per"/>
        </operator>
    </operator>
  </operator>
</operator>
```

Gray-shaded code: code to be modified by the user.

ISBN:978-84-693-4597-9 /DL:T.1010-2010

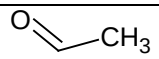
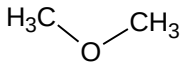
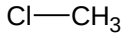
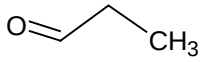
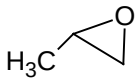
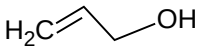
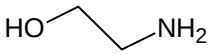
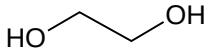
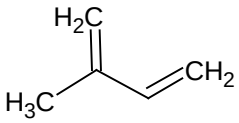
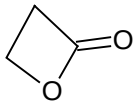
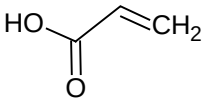
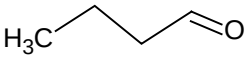
ISBN:978-84-693-4597-9 /DL:T.1010-2010

Annex C

Work and validation chemicals used in this work

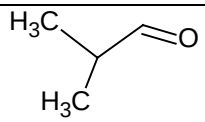
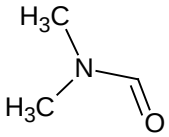
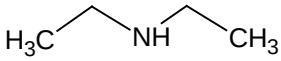
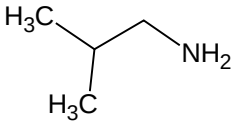
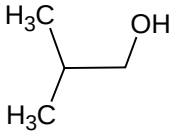
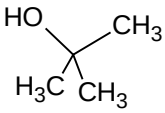
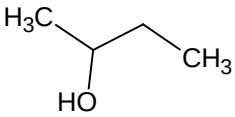
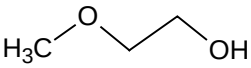
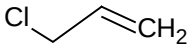
ISBN:978-84-693-4597-9 /DL:T.1010-2010

Annex C.1. List of 375 work chemicals used in this study.

ID	CAS	SMILES code	MW (g/mol)	3D molecular representation and name
w001	75-07-0	O=CC	44.05	 ethanal
w002	115-10-6	O(C)C	46.07	 dimethyl ether
w003	74-87-3	ClC	50.49	 chloromethane
w004	123-38-6	O=CCC	58.08	 propanal
w005	75-56-9	O1C(C)C1	58.08	 1,2-epoxypropane
w006	107-18-6	OC\C=C	58.08	 2-propen-1-ol
w007	141-43-5	OCCN	61.08	 ethanolamine
w008	107-21-1	OCCO	62.07	 1,2-ethanediol
w009	110-00-9	o1cccc1	68.08	 oxacyclopentadiene (furan)
w010	78-79-5	C=C(\C=C)C	68.12	 methylbutadiene
w011	57-57-8	O=C1OCC1	72.06	 beta-propiolactone
w012	79-10-7	O=C(O)\C=C	72.06	 propenoic acid
w013	123-72-8	O=CCCC	72.11	 butanal

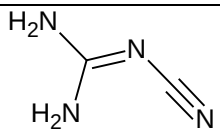
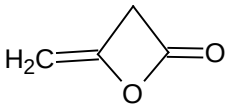
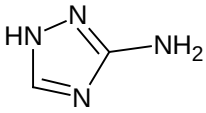
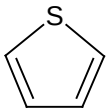
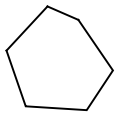
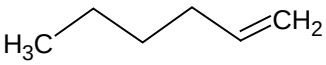
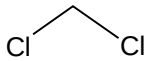
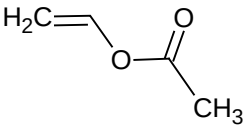
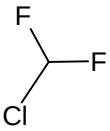
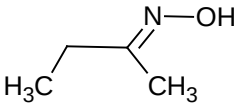
ISBN:978-84-693-4597-9 /DL:T.1010-2010

Annex C.1. List of 375 work chemicals used in this study (continued).

ID	CAS	SMILES code	MW (g/mol)	3D molecular representation and name
w014	78-84-2	O=CC(C)C	72.11	 isobutyraldehyde
w015	68-12-2	O=CN(C)C	73.10	 n,n'-dimethylformamide
w016	109-73-9	NCCCC	73.14	 butylamine
w017	109-89-7	N(CC)CC	73.14	 diethylamine
w018	78-81-9	NCC(C)C	73.14	 isobutylamine
w019	78-83-1	OCC(C)C	74.12	 2-methyl-1-propanol
w020	75-65-0	OC(C)(C)C	74.12	 2-methyl-2-propanol
w021	78-92-2	OC(C)CC	74.12	 2-butanol
w022	109-86-4	OCCOC	76.10	 2-methoxyethanol
w023	107-05-1	ClC\C=C	76.53	 3-chloro-1-propene
w024	71-43-2	c1ccccc1	78.12	 benzene

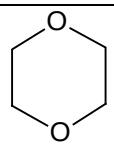
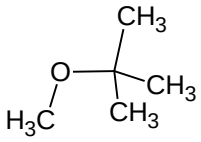
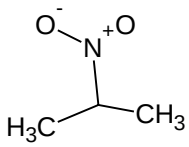
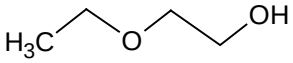
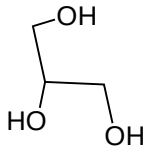
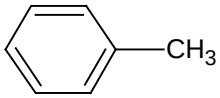
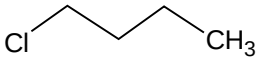
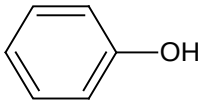
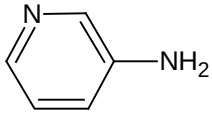
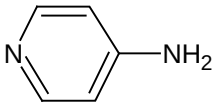
ISBN:978-84-693-4597-9 /DL:T.1010-2010

Annex C.1. List of 375 work chemicals used in this study (continued).

ID	CAS	SMILES code	MW (g/mol)	3D molecular representation and name
w025	461-58-5	N#C\N=C(/N)N	84.08	 2-cyanoguanidine
w026	674-82-8	O=C1O\C(=C)C1	84.08	 diketene
w027	61-82-5	n1cnnc1N	84.08	 amitrole
w028	110-02-1	s1cccc1	84.14	 thiophene
w029	110-82-7	C1CCCCC1	84.16	 cyclohexane
w030	592-41-6	C=C\CCCC	84.16	 1-hexene
w031	75-09-2	ClCCl	84.93	 dichloromethane
w032	108-05-4	O=C(O\C=C)C	86.09	 ethenyl ethanoate
w033	75-45-6	ClC(F)F	86.47	 chlorodifluoromethane
w034	96-29-7	N(/O)=C(/C)CC	87.12	 2-butanone oxime

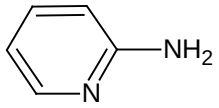
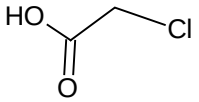
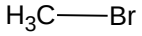
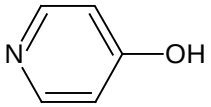
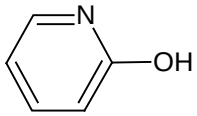
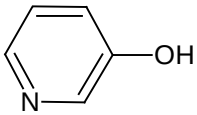
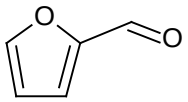
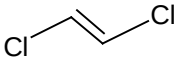
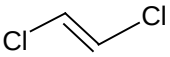
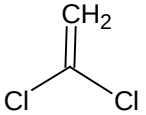
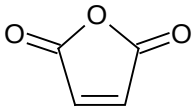
ISBN:978-84-693-4597-9 /DL:T.1010-2010

Annex C.1. List of 375 work chemicals used in this study (continued).

ID	CAS	SMILES code	MW (g/mol)	3D molecular representation and name
w035	123-91-1	O1CCOCC1	88.11	 1,4-dioxane
w036	1634-04-4	O(C(C)(C)C)C	88.15	 methyl t-butyl ether
w037	79-46-9	[O-][N+](=O)C(C)C	89.10	 2-nitropropane
w038	110-80-5	OCCOCC	90.12	 2-ethoxyethanol
w039	56-81-5	OCC(O)CO	92.10	 1,2,3-propanetriol
w040	108-88-3	c1ccccc1C	92.14	 methylbenzene
w041	109-69-3	ClCCCC	92.57	 1-chlorobutane
w042	108-95-2	Oc1ccccc1	94.11	 hydroxybenzene
w043	462-08-8	n1cccc(N)c1	94.12	 3-aminopyridine
w044	504-24-5	n1ccc(N)cc1	94.12	 4-aminopyridine

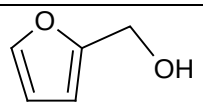
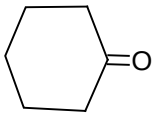
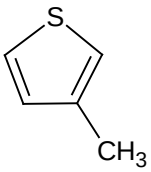
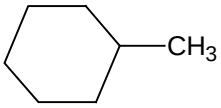
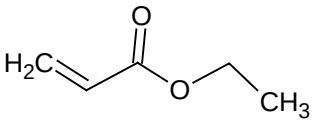
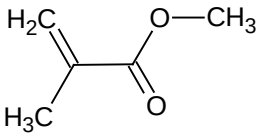
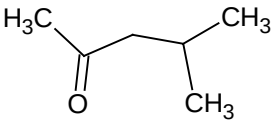
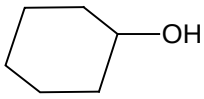
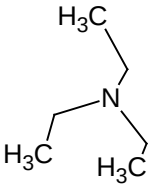
ISBN:978-84-693-4597-9 /DL:T.1010-2010

Annex C.1. List of 375 work chemicals used in this study (continued).

ID	CAS	SMILES code	MW (g/mol)	3D molecular representation and name
w045	504-29-0	n1ccccc1N	94.12	 2-aminopyridine
w046	79-11-8	ClCC(=O)O	94.50	 chloroethanoic acid
w047	74-83-9	BrC	94.94	 bromomethane
w048	626-64-2	N1=CC=C(C=C1)O	95.10	 4-hydroxypyridine
w049	72762-00-6	Oc1ccccn1	95.10	 2-hydroxypyridine
w050	109-00-2	Oc1cccnc1	95.10	 3-hydroxypyridine
w051	98-01-1	O=Cc1occc1	96.09	 furfural
w052	156-59-2	Cl[C@H]=CCl	96.94	 (z)-1,2-dichloroethene
w053	156-60-5	Cl[C@H]=CCl	96.94	 (e)-1,2-dichloroethene
w054	75-35-4	Cl/C(Cl)=C	96.94	 1,1-dichloroethene
w055	108-31-6	O=C1OC(=O)C=C1	98.06	 maleic anhydride

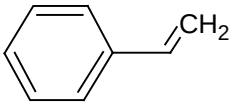
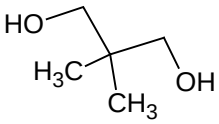
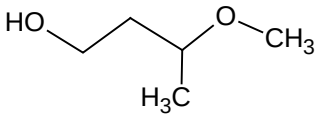
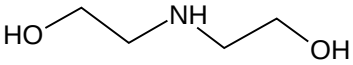
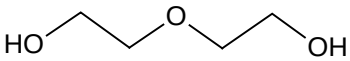
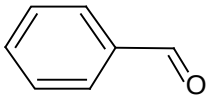
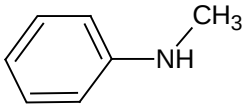
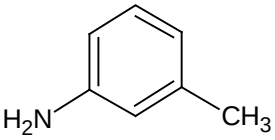
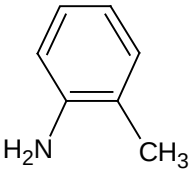
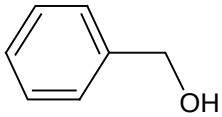
ISBN:978-84-693-4597-9 /DL:T.1010-2010

Annex C.1. List of 375 work chemicals used in this study (continued).

ID	CAS	SMILES code	MW (g/mol)	3D molecular representation and name
w056	98-00-0	OCc1occc1	98.10	 2-hydroxymethylfuran
w057	108-94-1	O=C1CCCCC1	98.15	 cyclohexanone
w058	616-44-4	s1ccc(c1)C	98.17	 3-methylthiophene
w059	108-87-2	CC1CCCCC1	98.19	 methylcyclohexane
w060	140-88-5	O=C(OCC)\C=C	100.12	 ethyl acrylate
w061	80-62-6	O=C(OC)\C(=C)C	100.12	 methyl methacrylate
w062	108-10-1	O=C(C)CC(C)C	100.16	 4-methyl-2-pentanone
w063	108-93-0	OC1CCCCC1	100.16	 cyclohexanol
w064	121-44-8	N(CC)(CC)CC	101.19	 triethylamine

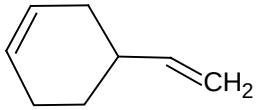
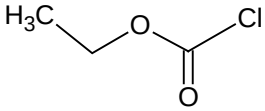
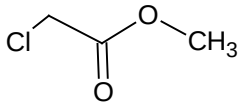
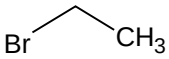
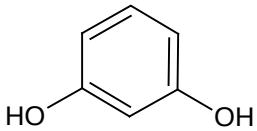
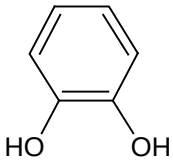
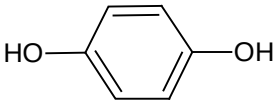
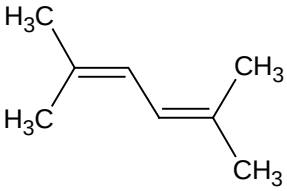
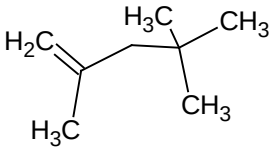
ISBN:978-84-693-4597-9 /DL:T.1010-2010

Annex C.1. List of 375 work chemicals used in this study (continued).

ID	CAS	SMILES code	MW (g/mol)	3D molecular representation and name
w065	100-42-5	C=C\c1ccccc1	104.15	 ethenylbenzene
w066	126-30-7	OCC(C)(C)CO	104.15	 1,3-propanediol, 2,2-dimethyl-
w067	2517-43-3	OCCC(OC)C	104.15	 3-methoxy-1-butanol
w068	111-42-2	OCCNCCO	105.14	 diethanolamine
w069	111-46-6	OCCOCCO	106.12	 3-oxapentane-1,5-diol
w070	100-52-7	O=Cc1ccccc1	106.13	 benzaldehyde
w071	100-61-8	N(c1ccccc1)C	107.16	 n-methylaniline
w072	108-44-1	Nc1cc(ccc1)C	107.16	 m-toluidine
w073	95-53-4	Nc1ccccc1C	107.16	 o-toluidine
w074	100-51-6	OCc1ccccc1	108.14	 (hydroxymethyl)benzene

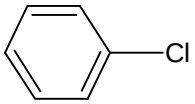
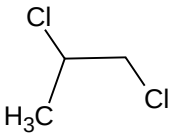
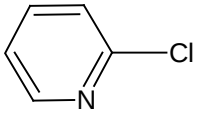
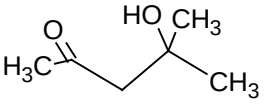
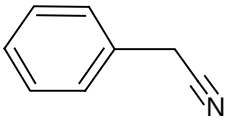
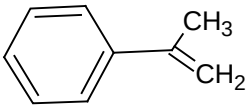
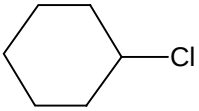
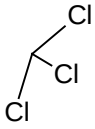
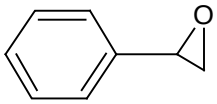
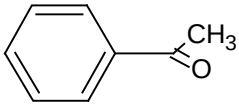
ISBN:978-84-693-4597-9 /DL:T.1010-2010

Annex C.1. List of 375 work chemicals used in this study (continued).

ID	CAS	SMILES code	MW (g/mol)	3D molecular representation and name
w075	100-40-3	C=C\C1CC=CCC1	108.18	 4-vinylcyclohexene
w076	541-41-3	ClC(=O)OCC	108.53	 ethyl chlorocarbonate
w077	96-34-4	ClCC(=O)OC	108.53	 methyl chloroacetate
w078	74-96-4	BrCC	108.97	 bromoethane
w079	108-46-3	Oc1cccc(O)c1	110.11	 1,3-dihydroxybenzene
w080	120-80-9	Oc1ccccc1O	110.11	 1,2-dihydroxybenzene
w081	123-31-9	Oc1ccc(O)cc1	110.11	 1,4-dihydroxybenzene
w082	764-13-6	C(=C\C=C(/C)C)(\C)C	110.20	 2,5-dimethyl-2,4-hexadiene
w083	107-39-1	C=C(/C)CC(C)(C)C	112.22	 2,4,4-trimethyl-1-pentene

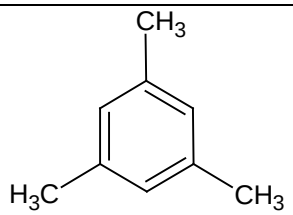
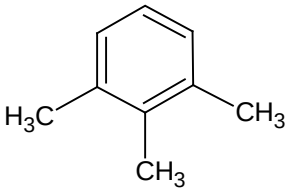
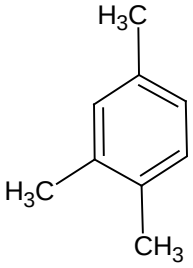
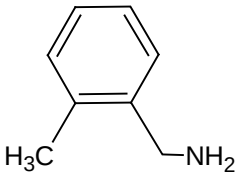
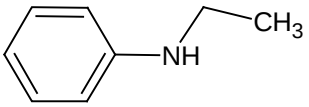
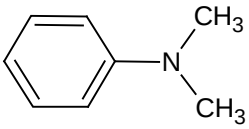
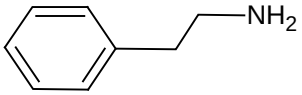
ISBN:978-84-693-4597-9 /DL:T.1010-2010

Annex C.1. List of 375 work chemicals used in this study (continued).

ID	CAS	SMILES code	MW (g/mol)	3D molecular representation and name
w084	108-90-7	Clc1ccccc1	112.56	 chlorobenzene
w085	78-87-5	ClCC(Cl)C	112.99	 1,2-dichloropropane
w086	109-09-1	Clc1ncccc1	113.55	 2-chloropyridine
w087	123-42-2	O=C(C)CC(O)(C)C	116.16	 4-hydroxy-4-methyl-2-pentanone
w088	140-29-4	N#CCc1ccccc1	117.15	 phenylacetonitrile
w089	98-83-9	C=C(\c1ccccc1)C	118.18	 alpha-methylstyrene
w090	542-18-7	ClC1CCCCC1	118.61	 chlorocyclohexane
w091	67-66-3	ClC(Cl)Cl	119.38	 trichloromethane
w092	96-09-3	O2C(c1ccccc1)C2	120.15	 styrene oxide
w093	98-86-2	O=C(c1ccccc1)C	120.15	 1-phenylethanone

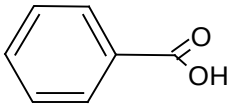
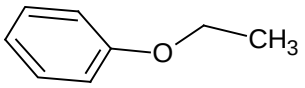
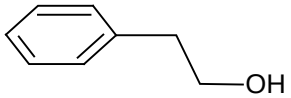
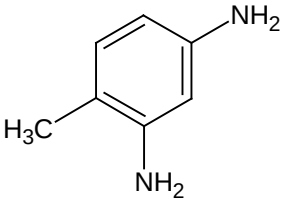
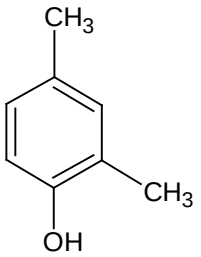
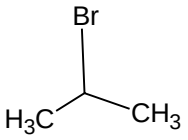
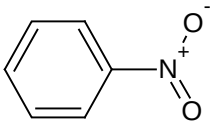
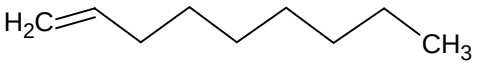
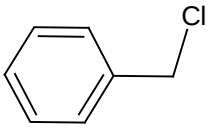
ISBN:978-84-693-4597-9 /DL:T.1010-2010

Annex C.1. List of 375 work chemicals used in this study (continued).

ID	CAS	SMILES code	MW (g/mol)	3D molecular representation and name
w094	108-67-8	c1c(cc(cc1C)C)C	120.20	 1,3,5-trimethylbenzene
w095	526-73-8	c1(cccc(c1C)C)C	120.20	 1,2,3-trimethylbenzene
w096	95-63-6	c1c(ccc(c1C)C)C	120.20	 1,2,4-trimethylbenzene
w097	89-93-0	NCc1ccccc1C	121.18	 benzenemethanamine, 2-methyl-
w098	103-69-5	N(c1ccccc1)CC	121.18	 n-ethylaniline
w099	121-69-7	N(c1ccccc1)(C)C	121.18	 n,n-dimethylaniline
w100	64-04-0	NCCc1ccccc1	121.18	 2-phenylethylamine

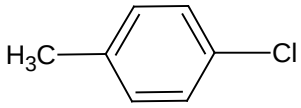
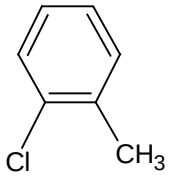
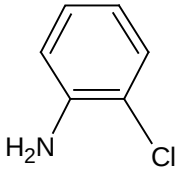
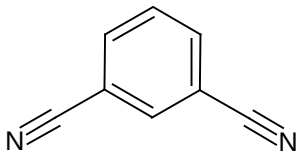
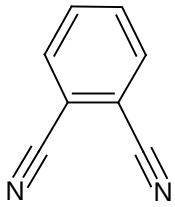
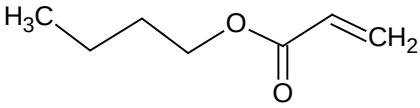
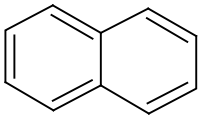
ISBN:978-84-693-4597-9 /DL:T.1010-2010

Annex C.1. List of 375 work chemicals used in this study (continued).

ID	CAS	SMILES code	MW (g/mol)	3D molecular representation and name
w101	65-85-0	O=C(O)c1ccccc1	122.12	 benzoic acid
w102	103-73-1	O(c1ccccc1)CC	122.17	 ethoxybenzene
w103	60-12-8	OCCc1ccccc1	122.17	 2-phenylethanol
w104	95-80-7	Nc1cc(N)c(cc1)C	122.17	 2,4-toluenediamine
w105	105-67-9	Oc1ccc(cc1C)C	122.17	 2,4-dimethylphenol
w106	75-26-3	BrC(C)C	122.99	 2-bromopropane
w107	98-95-3	[O-][N+](=O)c1ccccc1	123.11	 nitrobenzene
w108	124-11-8	C=C\CCCCCCC	126.24	 1-nonene
w109	100-44-7	ClCc1ccccc1	126.59	 a-chlorotoluene

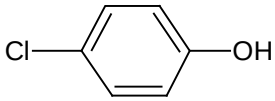
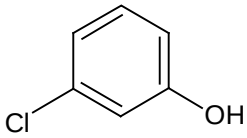
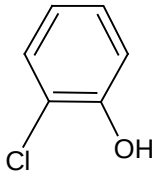
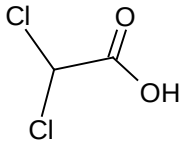
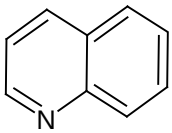
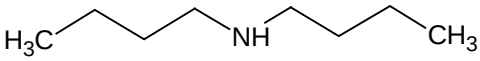
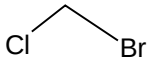
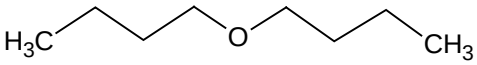
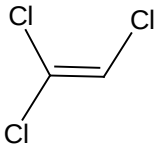
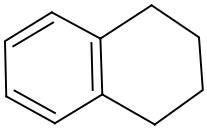
ISBN:978-84-693-4597-9 /DL:T.1010-2010

Annex C.1. List of 375 work chemicals used in this study (continued).

ID	CAS	SMILES code	MW (g/mol)	3D molecular representation and name
w110	106-43-4	Clc1ccc(cc1)C	126.59	 p-chlorotoluene
w111	95-49-8	Clc1ccccc1C	126.59	 1-chloro-2-methylbenzene
w112	95-51-2	Clc1ccccc1N	127.57	 2-chloroaniline
w113	623-26-7	N#Cc1ccc(C#N)cc1	128.13	 1,4-benzenedicarbonitrile
w114	626-17-5	N#Cc1cccc(C#N)c1	128.13	 1,3-dicyanobenzene
w115	91-15-6	N#Cc1ccccc1C#N	128.13	 1,2-dicyanobenzene
w116	141-32-2	O=C(OCCCC)\C=C	128.17	 butyl acrylate
w117	91-20-3	c12ccccc1cccc2	128.18	 naphthalene

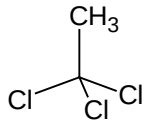
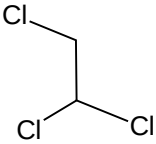
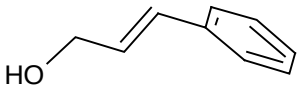
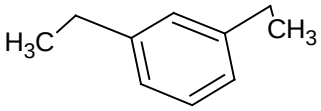
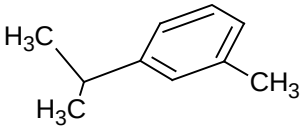
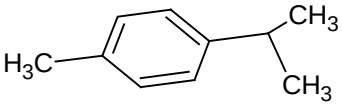
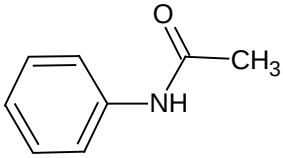
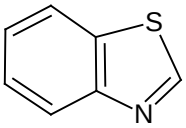
ISBN:978-84-693-4597-9 /DL:T.1010-2010

Annex C.1. List of 375 work chemicals used in this study (continued).

ID	CAS	SMILES code	MW (g/mol)	3D molecular representation and name
w118	106-48-9	Clc1ccc(O)cc1	128.56	 4-chlorophenol
w119	108-43-0	Clc1cc(O)ccc1	128.56	 3-chlorophenol
w120	95-57-8	Clc1ccccc1O	128.56	 2-hydroxychlorobenzene
w121	79-43-6	ClC(Cl)C(=O)O	128.94	 dichloroethanoic acid
w122	91-22-5	n1cccc2ccccc12	129.16	 benzo[b]pyridine
w123	111-92-2	N(CCCC)CCCC	129.25	 dibutylamine
w124	74-97-5	BrCCl	129.38	 chlorobromomethane
w125	142-96-1	O(CCCC)CCCC	130.23	 dibutyl ether
w126	79-01-6	Cl[C@H]=C(Cl)Cl	131.39	 trichloroethene
w127	119-64-2	c1ccc2c(c1)CCCC2	132.21	 1,2,3,4-tetrahydronaphthalene

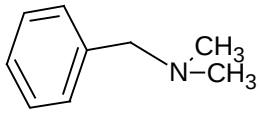
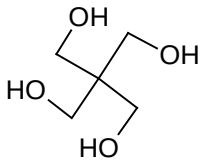
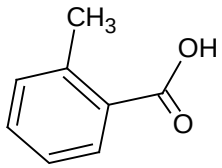
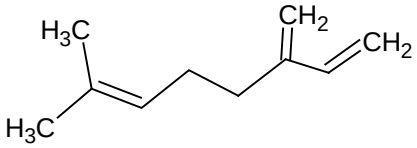
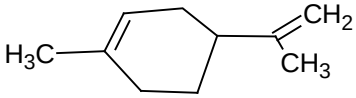
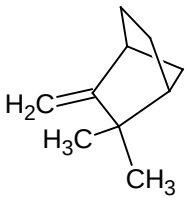
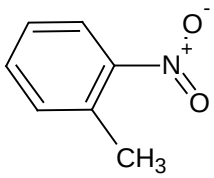
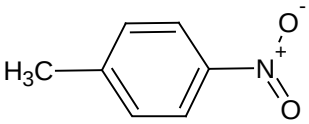
ISBN:978-84-693-4597-9 /DL:T.1010-2010

Annex C.1. List of 375 work chemicals used in this study (continued).

ID	CAS	SMILES code	MW (g/mol)	3D molecular representation and name
w128	77-73-6	C2=CC3C1C=CC(C1)C3C2	132.21	 dicyclopentadiene
w129	71-55-6	ClC(Cl)(Cl)C	133.41	 1,1,1-trichloroethane
w130	79-00-5	ClCC(Cl)Cl	133.41	 1,1,2-trichloroethane
w131	104-54-1	OC/C=C/c1ccccc1	134.18	 cinnamyl alcohol
w132	141-93-5	c1ccc(cc1CC)CC	134.22	 m-diethylbenzene
w133	535-77-3	c1ccc(cc1C(C)C)C	134.22	 m-cymene
w134	99-87-6	c1cc(ccc1C(C)C)C	134.22	 1-isopropyl-4-methylbenzene
w135	103-84-4	O=C(Nc1ccccc1)C	135.17	 acetanilide
w136	95-16-9	n1c2ccccc2sc1	135.20	 benzothiazole

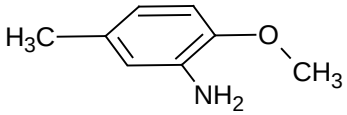
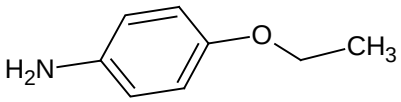
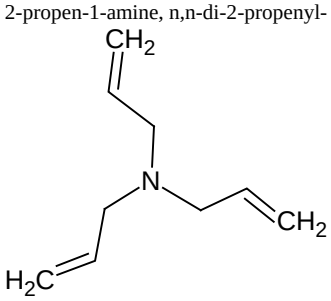
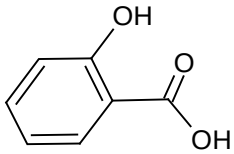
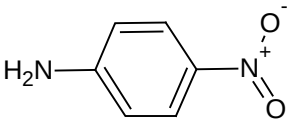
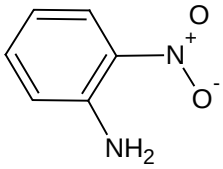
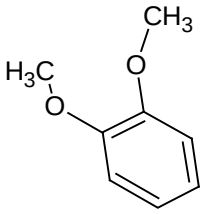
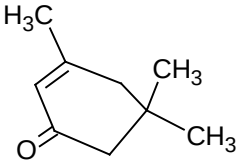
ISBN:978-84-693-4597-9 /DL:T.1010-2010

Annex C.1. List of 375 work chemicals used in this study (continued).

ID	CAS	SMILES code	MW (g/mol)	3D molecular representation and name
w137	103-83-3	N(C)(Cc1ccccc1)C	135.21	 n,n-dimethylbenzylamine
w138	115-77-5	OCC(CO)(CO)CO	136.15	 pentaerythritol
w139	118-90-1	O=C(O)c1ccccc1C	136.15	 o-toluic acid
w140	123-35-3	C=C\C(=C)CC\C=C(/C)C	136.24	 3-methylene-7-methyl-1,6-octadiene (myrcene)
w141	138-86-3	C(=C)(\C)C1CC=C(C)CC1	136.24	 limonene
w142	79-92-5	C2(=C)\C1CCC(C1)C2(C)C	136.24	 camphene
w143	88-72-2	O=[N+]([O-])c1ccccc1C	137.14	 2-nitrotoluene
w144	99-99-0	O=[N+]([O-])c1ccc(cc1)C	137.14	 4-nitrotoluene

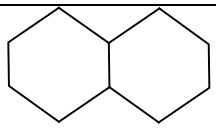
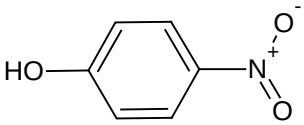
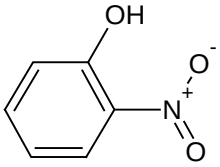
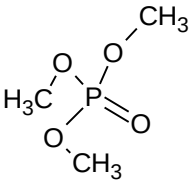
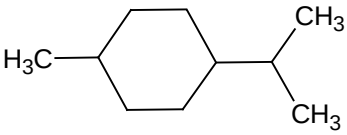
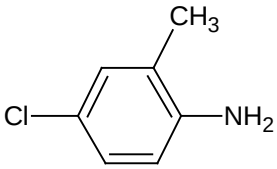
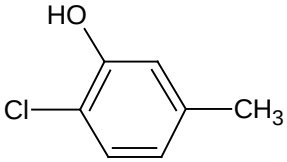
ISBN:978-84-693-4597-9 /DL:T.1010-2010

Annex C.1. List of 375 work chemicals used in this study (continued).

ID	CAS	SMILES code	MW (g/mol)	3D molecular representation and name
w145	120-71-8	O(c1ccc(cc1N)C)C	137.18	 2-methoxy-5-methylbenzenamine
w146	156-43-4	O(c1ccc(cc1)N)CC	137.18	 2-propen-1-amine, n,n-di-2-propenyl-
w147	102-70-5	C(=C)\CN(C\C=C)C\C=C	137.23	 2-propen-1-amine, n,n-di-2-propenyl-
w148	69-72-7	O=C(O)c1ccccc1O	138.12	 salicylic acid
w149	100-01-6	O=[N+]([O-])c1ccc(N)cc1	138.13	 4-nitroaniline
w150	88-74-4	O=[N+]([O-])c1ccccc1N	138.13	 2-nitroaniline
w151	91-16-7	O(c1ccccc1OC)C	138.17	 1,2-dimethoxybenzene
w152	78-59-1	O=C1C=C(CC(C)(C)C1)C	138.21	 3,5,5-trimethyl-2-cyclohexen-1-one

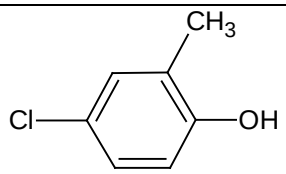
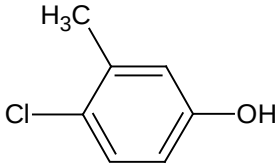
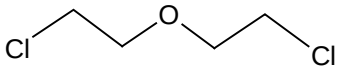
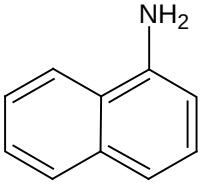
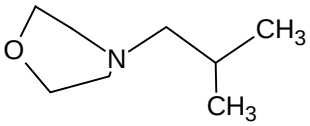
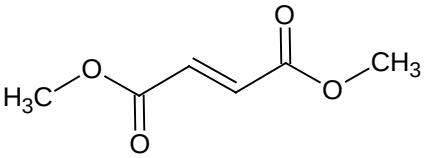
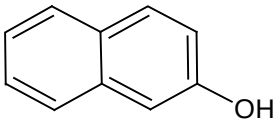
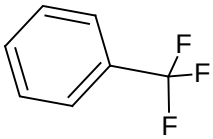
ISBN:978-84-693-4597-9 /DL:T.1010-2010

Annex C.1. List of 375 work chemicals used in this study (continued).

ID	CAS	SMILES code	MW (g/mol)	3D molecular representation and name
w153	91-17-8	C1CCC2CCCCC2C1	138.25	 decahydronaphthalene
w154	100-02-7	O=[N+]([O-])c1ccc(O)cc1	139.11	 4-nitrophenol
w155	88-75-5	O=[N+]([O-])c1ccccc1O	139.11	 2-nitrophenol
w156	2243-27-8	N#CCCCCCCCC	139.24	 nonanonitrile
w157	512-56-1	O=P(OC)(OC)OC	140.08	 trimethyl phosphate
w158	99-82-1	CC1CCC(C(C)C)CC1	140.27	 p-menthane
w159	95-69-2	Clc1cc(c(N)cc1)C	141.60	 2-methyl-4-chloroaniline
w160	615-74-7	Clc1ccc(cc1O)C	142.59	 2-chloro-5-methylphenol

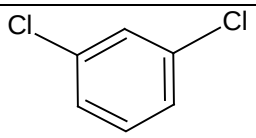
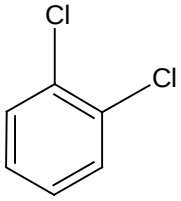
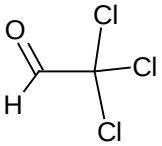
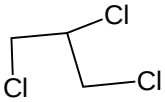
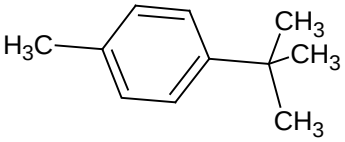
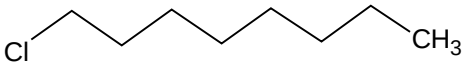
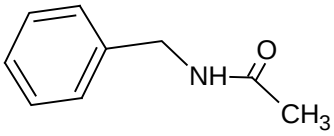
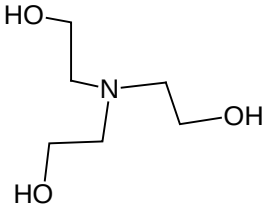
ISBN:978-84-693-4597-9 /DL:T.1010-2010

Annex C.1. List of 375 work chemicals used in this study (continued).

ID	CAS	SMILES code	MW (g/mol)	3D molecular representation and name
w161	1570-64-5	Clc1cc(c(O)cc1)C	142.59	 2-methyl-4-chlorophenol
w162	59-50-7	Clc1ccc(O)cc1C	142.59	 3-methyl-4-chlorophenol
w163	111-44-4	ClCCOCCCl	143.01	 1,5-dichloro-3-oxapentane
w164	134-32-7	c1cccc2cccc(N)c12	143.19	 1-naphthylamine
w165	10315-98-7	O1CCN(CC(C)C)CC1	143.23	 n-isobutylmorpholine
w166	624-48-6	O=C(OC)\C=C\C(=O)OC	144.13	 methyl maleate
w167	135-19-3	Oc2ccc1c(cccc1)c2	144.17	 2-naphthol
w168	98-08-8	FC(F)(F)c1ccccc1	146.11	 (trifluoromethyl)benzene

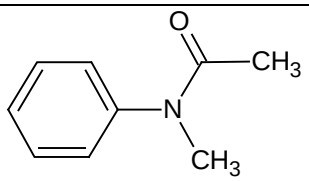
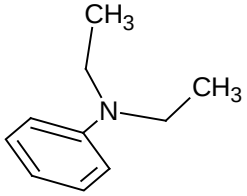
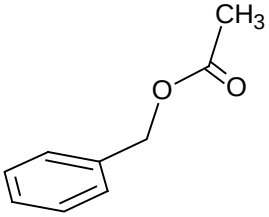
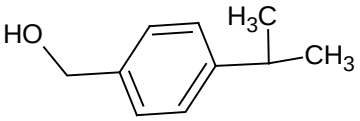
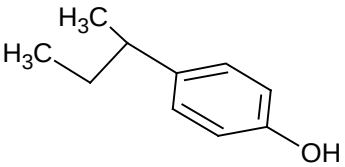
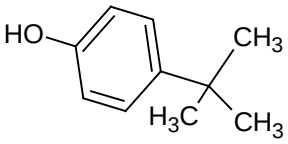
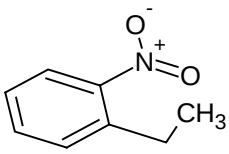
ISBN:978-84-693-4597-9 /DL:T.1010-2010

Annex C.1. List of 375 work chemicals used in this study (continued).

ID	CAS	SMILES code	MW (g/mol)	3D molecular representation and name
w169	541-73-1	Clc1cccc(Cl)c1	147.00	 1,3-dichlorobenzene
w170	95-50-1	Clc1ccccc1Cl	147.00	 1,2-dichlorobenzene
w171	75-87-6	ClC(Cl)(Cl)C=O	147.39	 trichloroethanal
w172	96-18-4	ClCC(Cl)CCl	147.43	 1,2,3-trichloropropane
w173	98-51-1	c1cc(ccc1C(C)(C)C)C	148.25	 p-(t-butyl)toluene
w174	111-85-3	ClCCCCCCCC	148.68	 1-chlorooctane
w175	588-46-5	O=C(NCc1ccccc1)C	149.19	 n-benzylacetamide
w176	102-71-6	OCCN(CCO)CCO	149.19	 triethanolamine

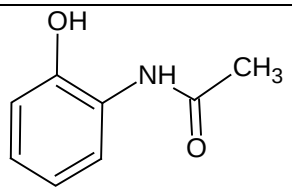
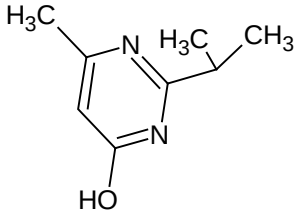
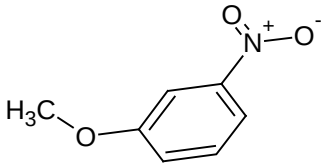
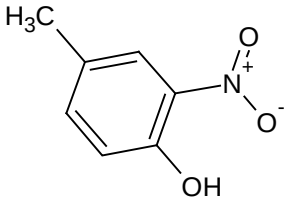
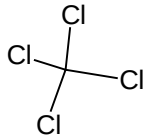
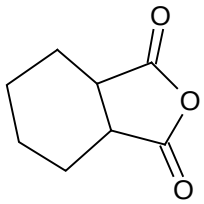
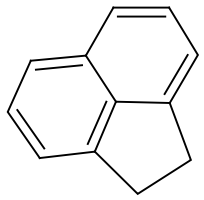
ISBN:978-84-693-4597-9 /DL:T.1010-2010

Annex C.1. List of 375 work chemicals used in this study (continued).

ID	CAS	SMILES code	MW (g/mol)	3D molecular representation and name
w177	579-10-2	O=C(N(c1ccccc1)C)C	149.19	 n-methylacetanilide
w178	91-66-7	N(c1ccccc1)(CC)CC	149.24	 n,n-diethylaniline
w179	140-11-4	O=C(OCc1ccccc1)C	150.18	 benzyl acetate
w180	536-60-7	OCc1ccc(cc1)C(C)C	150.22	 benzenemethanol, 4-(1-methylethyl)-
w181	99-71-8	Oc1ccc(cc1)C(C)CC	150.22	 p-(sec-butyl)phenol
w182	98-54-4	Oc1ccc(cc1)C(C)(C)C	150.22	 4-tert-butylphenol
w183	612-22-6	O=[N+]([O-])c1ccccc1CC	151.17	 2-ethylnitrobenzene

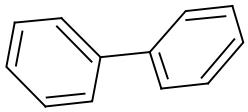
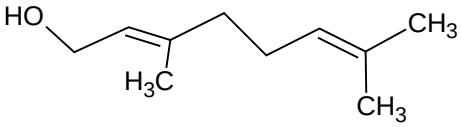
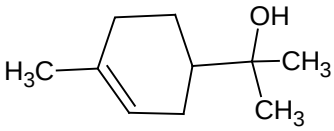
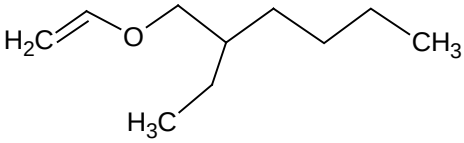
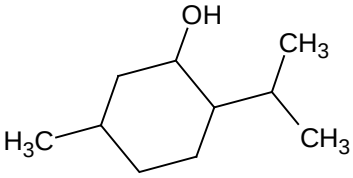
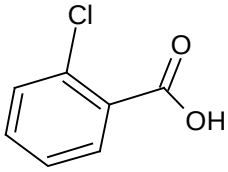
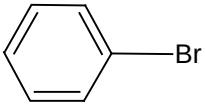
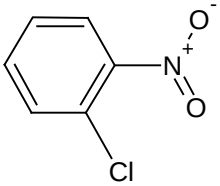
ISBN:978-84-693-4597-9 /DL:T.1010-2010

Annex C.1. List of 375 work chemicals used in this study (continued).

ID	CAS	SMILES code	MW (g/mol)	3D molecular representation and name
w184	614-80-2	O=C(Nc1ccccc1O)C	151.17	 o-hydroxyacetanilide
w185	2814-20-2	n1c(cc(O)nc1C(C)C)C	152.20	 6-methyl-2-(propan-2-yl)pyrimidin-4-ol
w186	555-03-3	[O-][N+](=O)c1cccc(OC)c1	153.14	 m-nitroanisole
w187	119-33-5	O=[N+]([O-])c1cc(ccc1O)C	153.14	 4-methyl-2-nitrophenol
w188	56-23-5	ClC(Cl)(Cl)Cl	153.82	 tetrachloromethane
w189	85-42-7	O=C1OC(=O)C2CCCCC12	154.17	 1,3-isobenzofurandione, hexahydro-
w190	83-32-9	c2cc1cccc3c1c(c2)CC3	154.21	 acenaphthene

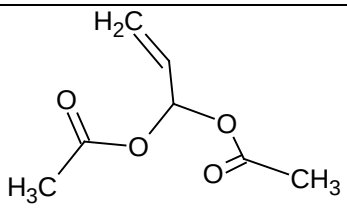
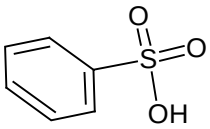
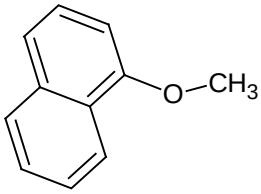
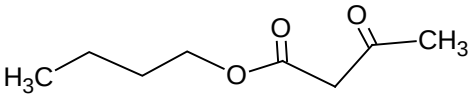
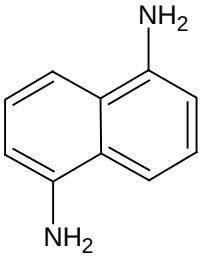
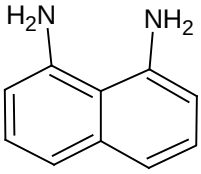
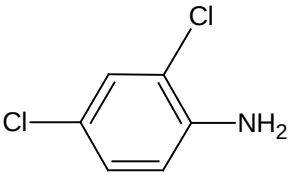
ISBN:978-84-693-4597-9 /DL:T.1010-2010

Annex C.1. List of 375 work chemicals used in this study (continued).

ID	CAS	SMILES code	MW (g/mol)	3D molecular representation and name
w191	92-52-4	c1cc(ccc1)c2ccccc2	154.21	 biphenyl
w192	106-24-1	OC/C=C(/CC/C=C(/C)C)C	154.25	 geraniol
w193	98-55-5	OC(C)(C)C1CC=C(C)CC1	154.25	 alpha-terpineol
w194	103-44-6	O(\C=C)CC(CCCC)CC	156.27	 vinyl 2-ethylhexyl ether
w195	2216-51-5	OC1CC(CCC1C(C)C)C	156.27	 menthol (l)
w196	118-91-2	Clc1ccccc1C(=O)O	156.57	 2-chlorobenzoic acid
w197	108-86-1	Brc1ccccc1	157.01	 bromobenzene
w198	88-73-3	O=[N+]([O-])c1ccccc1Cl	157.56	 2-chloro-1-nitrobenzene

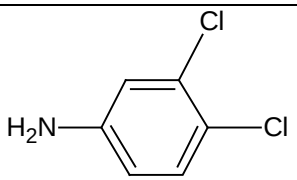
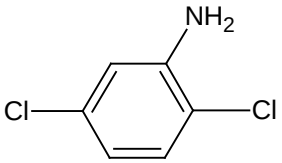
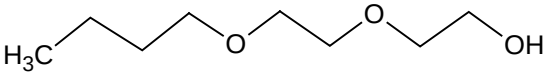
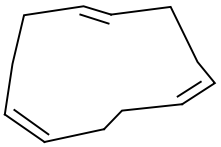
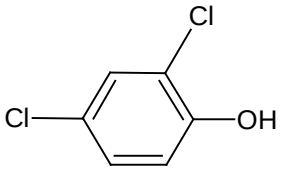
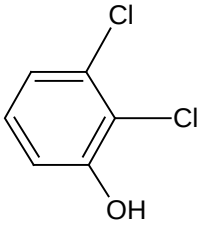
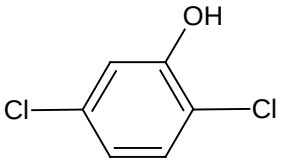
ISBN:978-84-693-4597-9 /DL:T.1010-2010

Annex C.1. List of 375 work chemicals used in this study (continued).

ID	CAS	SMILES code	MW (g/mol)	3D molecular representation and name
w199	869-29-4	O=C(OC(OC(=O)C)/C=C)C	158.16	 allylidenediacetate
w200	98-11-3	O=S(=O)(O)c1ccccc1	158.18	 benzenesulfonic acid
w201	2216-69-5	O(c2cccc1ccccc12)C	158.20	 naphthalene, 1-methoxy-
w202	591-60-6	O=C(OCCCC)CC(=O)C	158.20	 butanoic acid, 3-oxo-, butyl ester
w203	2243-62-1	c1ccc(c2cccc(N)c12)N	158.20	 1,5-diaminonaphthalene
w204	479-27-6	Nc1cccc2cccc(N)c12	158.20	 1,8-naphthalenediamine
w205	554-00-7	Clc1cc(Cl)c(N)cc1	162.02	 2,4-dichloroaniline

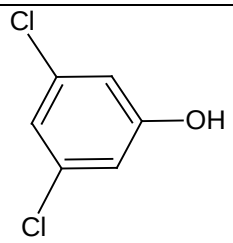
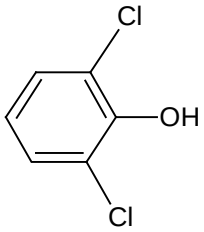
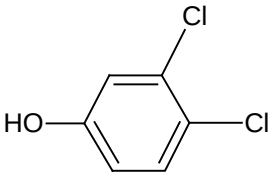
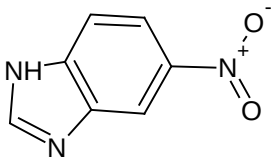
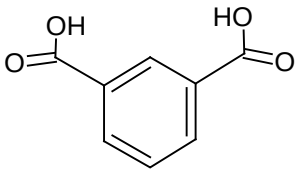
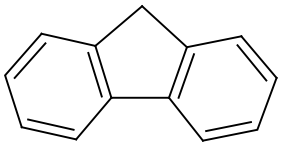
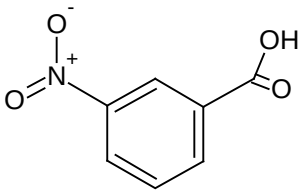
ISBN:978-84-693-4597-9 /DL:T.1010-2010

Annex C.1. List of 375 work chemicals used in this study (continued).

ID	CAS	SMILES code	MW (g/mol)	3D molecular representation and name
w206	95-76-1	Clc1ccc(N)cc1Cl	162.02	 3,4-dichloroaniline
w207	95-82-9	Clc1ccc(Cl)c(N)c1	162.02	 2,5-dichloroaniline
w208	112-34-5	O(CCCC)CCOCCO	162.23	 diethylene glycol mono-n-butyl ether
w209	4904-61-4	C1=CCCC=CCCC=CCC1	162.28	 1,5,9-cyclododecatriene
w210	120-83-2	Clc1cc(Cl)c(O)cc1	163.00	 2.4-dichlorophenol
w211	576-24-9	Clc1c(O)cccc1Cl	163.00	 2,3-dichlorophenol
w212	583-78-8	Clc1ccc(Cl)c(O)c1	163.00	 2,5-dichlorophenol

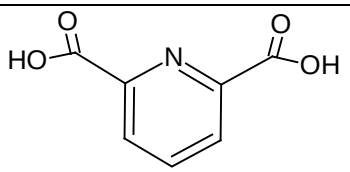
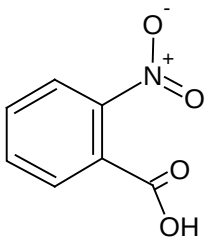
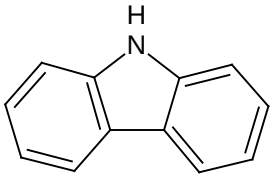
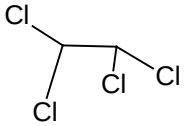
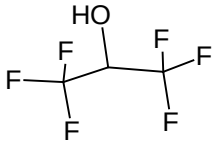
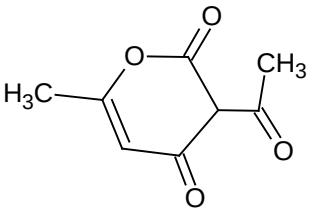
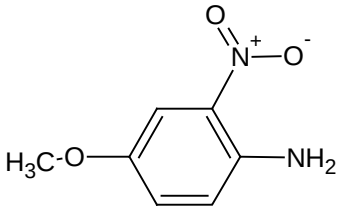
ISBN:978-84-693-4597-9 /DL:T.1010-2010

Annex C.1. List of 375 work chemicals used in this study (continued).

ID	CAS	SMILES code	MW (g/mol)	3D molecular representation and name
w213	591-35-5	Clc1cc(O)cc(Cl)c1	163.00	 3,5-dichlorophenol
w214	87-65-0	Clc1cccc(Cl)c1O	163.00	 2,6-dichlorophenol
w215	95-77-2	Clc1ccc(O)cc1Cl	163.00	 3,4-dichlorophenol
w216	94-52-0	[O-][N+](=O)c1cc2ncnc2cc1	163.14	 1h-benzimidazole, 5-nitro-
w217	121-91-5	O=C(O)c1cccc(C(=O)O)c1	166.13	 isophthalic acid
w218	86-73-7	c1cccc3c1c2c(cccc2)C3	166.22	 2,3-benzindene
w219	121-92-6	O=[N+]([O-])c1cc(ccc1)C(=O)O	167.12	 m-nitrobenzoic acid

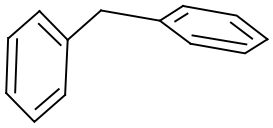
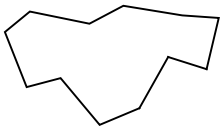
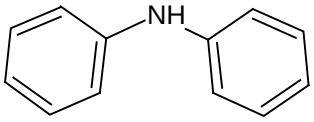
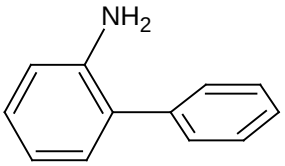
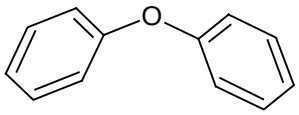
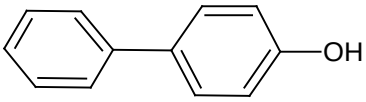
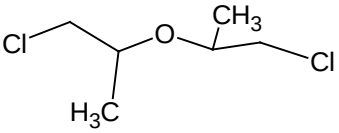
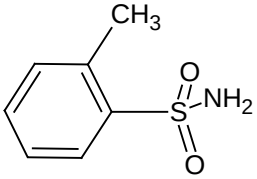
ISBN:978-84-693-4597-9 /DL:T.1010-2010

Annex C.1. List of 375 work chemicals used in this study (continued).

ID	CAS	SMILES code	MW (g/mol)	3D molecular representation and name
w220	499-83-2	O=C(O)c1nc(C(=O)O)ccc1	167.12	 pyridine-2,6-dicarboxylic acid
w221	552-16-9	O=[N+]([O-])c1ccccc1C(=O)O	167.12	 2-nitrobenzoic acid
w222	86-74-8	c1cccc2nc3ccccc3c12	167.22	 carbazole
w223	79-34-5	ClC(Cl)C(Cl)Cl	167.85	 1,1,2,2-tetrachloroethane
w224	920-66-1	FC(F)(F)C(O)C(F)(F)F	168.04	 1,1,1,3,3,3-hexafluoro-2-propanol
w225	520-45-6	O=C(C)C1C(=O)C=C(C)OC1=O	168.15	 dehydroacetic acid
w226	96-96-8	COc1ccc(N)c(c1)[N+]([O-])=O	168.15	 benzenamine, 4-methoxy-2-nitro-

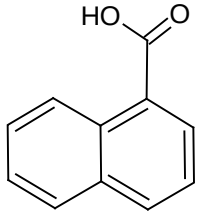
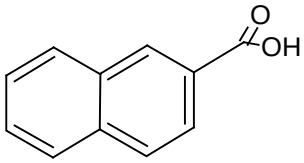
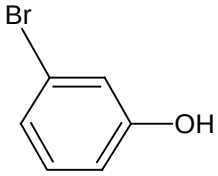
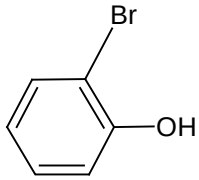
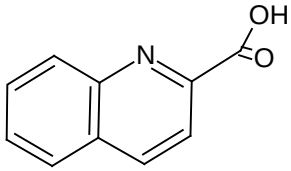
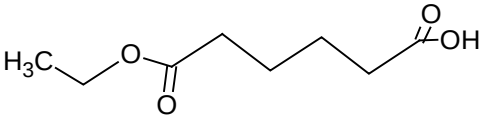
ISBN:978-84-693-4597-9 /DL:T.1010-2010

Annex C.1. List of 375 work chemicals used in this study (continued).

ID	CAS	SMILES code	MW (g/mol)	3D molecular representation and name
w227	101-81-5	c1c(cccc1)Cc2ccccc2	168.24	 benzene, 1,1'-methylenebis-
w228	294-62-2	C1CCCCCCCCCCC1	168.33	 cyclododecane
w229	122-39-4	c1ccccc1Nc2ccccc2	169.23	 diphenylamine
w230	90-41-5	c2c(c1ccccc1N)cccc2	169.23	 2-aminobiphenyl
w231	101-84-8	O(c1ccccc1)c2ccccc2	170.21	 diphenyl ether
w232	92-69-3	Oc2ccc(c1ccccc1)cc2	170.21	 p-phenylphenol
w233	108-60-1	CC(CCl)OC(C)CCl	171.07	 bis(2-chloroisopropyl)ether
w234	88-19-7	O=S(=O)(N)c1ccccc1C	171.22	 o-methylbenzenesulfonamide

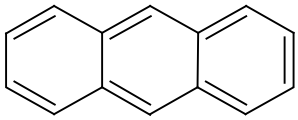
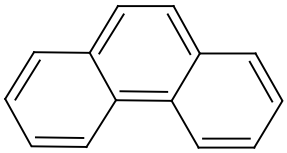
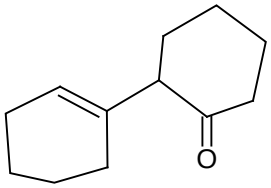
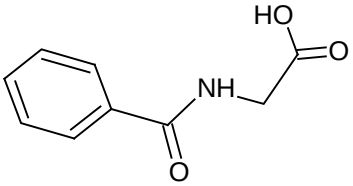
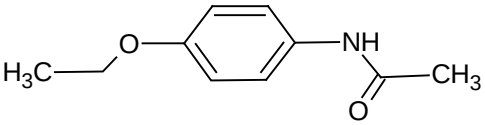
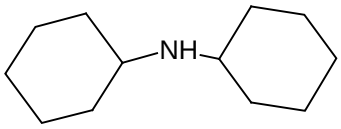
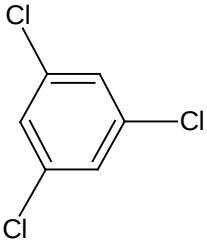
ISBN:978-84-693-4597-9 /DL:T.1010-2010

Annex C.1. List of 375 work chemicals used in this study (continued).

ID	CAS	SMILES code	MW (g/mol)	3D molecular representation and name
w235	86-55-5	O=C(O)c2cccc1ccccc12	172.19	 1-napthoic acid
w236	93-09-4	O=C(O)c2ccc1c(cccc1)c2	172.19	 2-naphthoic acid
w237	106-41-2	Brc1ccc(O)cc1	173.01	 4-bromophenol
w238	591-20-8	Brc1cc(O)ccc1	173.01	 m-bromophenol
w239	95-56-7	Brc1ccccc1O	173.01	 o-bromophenol
w240	93-10-7	O=C(O)c1nc2ccccc2cc1	173.17	 2-quinolinecarboxylic acid
w241	626-86-8	O=C(OCC)CCCCC(=O)O	174.20	 hexanedioic acid, monoethyl ester

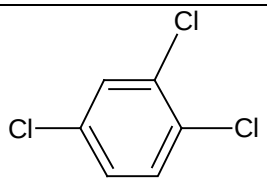
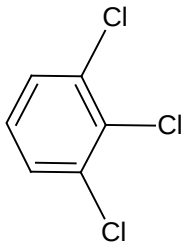
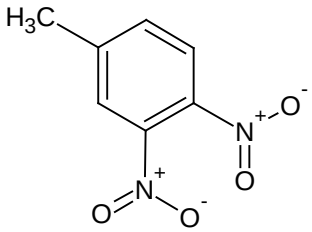
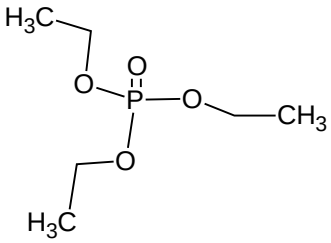
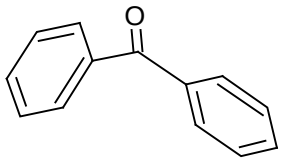
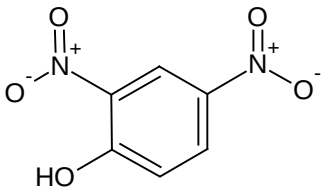
ISBN:978-84-693-4597-9 /DL:T.1010-2010

Annex C.1. List of 375 work chemicals used in this study (continued).

ID	CAS	SMILES code	MW (g/mol)	3D molecular representation and name
w242	120-12-7	c3ccc2cc1ccccc1cc2c3	178.24	 anthracene
w243	85-01-8	c3cc2ccc1ccccc1c2cc3	178.24	 phenanthrene
w244	1502-22-3	O=C2CCCCC2C1=CCCCC1	178.28	 cyclohexanone, 2-(1-cyclohexen-1-yl)-
w245	495-69-2	O=C(NCC(=O)O)c1ccccc1	179.18	 hippuric acid
w246	62-44-2	O=C(Nc1ccc(OCC)cc1)C	179.22	 p-phenacetin
w247	101-83-7	N(C1CCCCC1)C2CCCCC2	181.32	 dicyclohexylamine
w248	108-70-3	Clc1cc(Cl)cc(Cl)c1	181.45	 1,3,5-trichlorobenzene

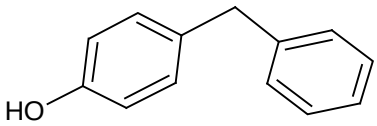
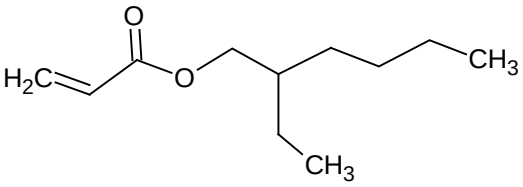
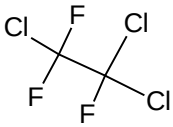
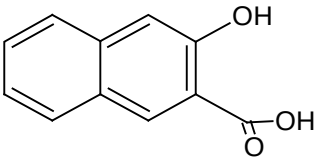
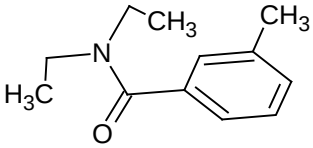
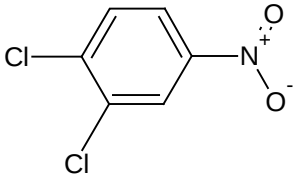
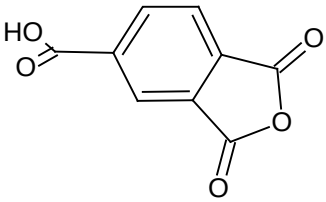
ISBN:978-84-693-4597-9 /DL:T.1010-2010

Annex C.1. List of 375 work chemicals used in this study (continued).

ID	CAS	SMILES code	MW (g/mol)	3D molecular representation and name
w249	120-82-1	Clc1cc(Cl)c(Cl)cc1	181.45	 1,2,4-trichlorobenzene
w250	87-61-6	Clc1cccc(Cl)c1Cl	181.45	 1,2,3-trichlorobenzene
w251	610-39-9	O=[N+]([O-])c1cc(ccc1[N+]([O-])=O)C	182.14	 1,2-dino2 4-methyl benzene
w252	78-40-0	O=P(OCC)(OCC)OCC	182.16	 triethyl phosphate
w253	119-61-9	O=C(c1ccccc1)c2ccccc2	182.22	 benzophenone
w254	51-28-5	O=[N+]([O-])c1cc(ccc1O)[N+]([O-])=O	184.11	 2,4-dinitrophenol

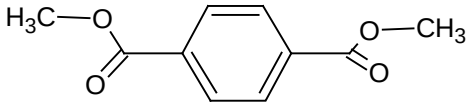
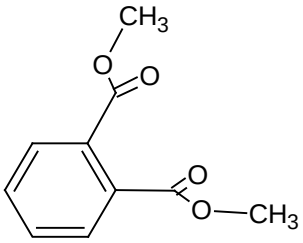
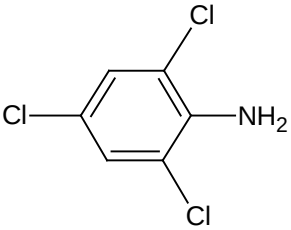
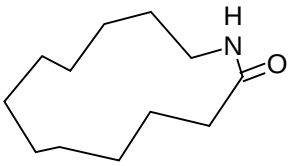
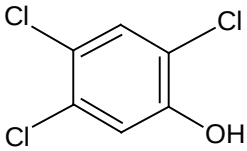
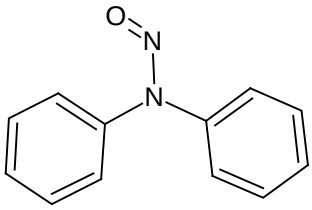
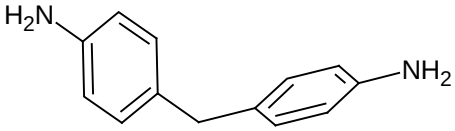
ISBN:978-84-693-4597-9 /DL:T.1010-2010

Annex C.1. List of 375 work chemicals used in this study (continued).

ID	CAS	SMILES code	MW (g/mol)	3D molecular representation and name
w255	101-53-1	Oc1ccc(cc1)Cc2ccccc2	184.24	 4-hydroxydiphenylmethane
w256	103-11-7	O=C(OCC(CCCC)CC)\C=C	184.28	 2-ethylhexyl acrylate
w257	76-13-1	ClC(F)(F)C(Cl)(Cl)F	187.38	 1,1,2-trichlorotrifluoroethane
w258	106-93-4	BrCCBr	187.86	 1,2-dibromoethane
w259	92-70-6	O=C(O)c2cc1c(cccc1)cc2O	188.18	 2-naphthalenecarboxylic acid, 3-hydroxy-
w260	134-62-3	O=C(N(CC)CC)c1cccc(c1)C	191.28	 deet [n,n,-diet-3-me benzamide]
w261	99-54-7	Clc1ccc([N+]([O-])=O)cc1Cl	192.00	 3,4-dichloronitrobenzene
w262	552-30-7	O=C(O)c1ccc2C(=O)OC(=O)c2c1	192.13	 trimellitic anhydride

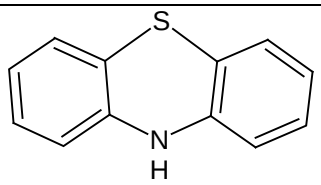
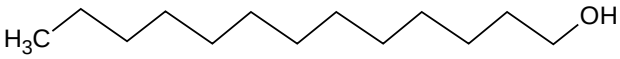
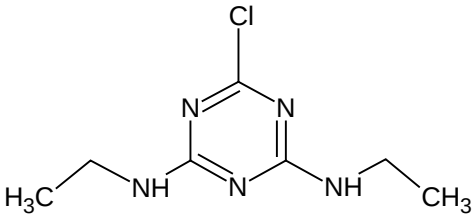
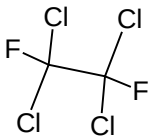
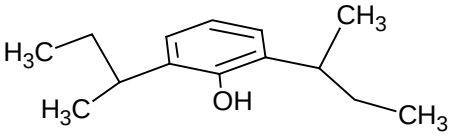
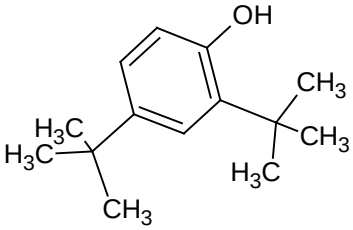
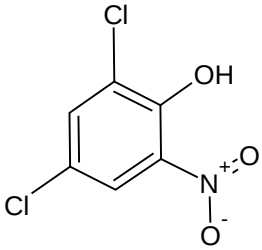
ISBN:978-84-693-4597-9 /DL:T.1010-2010

Annex C.1. List of 375 work chemicals used in this study (continued).

ID	CAS	SMILES code	MW (g/mol)	3D molecular representation and name
w263	120-61-6	O=C(OC)c1ccc(C(=O)OC)cc1	194.19	 dimethylterephthalate
w264	131-11-3	O=C(OC)c1ccccc1C(=O)OC	194.19	 dimethyl phthalate
w265	634-93-5	Clc1cc(Cl)cc(Cl)c1N	196.46	 2,4,6-trichloroaniline
w266	947-04-6	O=C1NCCCCCCCCCCC1	197.32	 azacyclotridecan-2-one
w267	95-95-4	Clc1cc(O)c(Cl)cc1Cl	197.45	 2,4,5-trichlorophenol
w268	86-30-6	O=NN(c1ccccc1)c2ccccc2	198.23	 diphenylnitrosamine
w269	101-77-9	c1(ccc(N)cc1)Cc2ccc(N)cc2	198.27	 di-(p-aminophenyl)methane

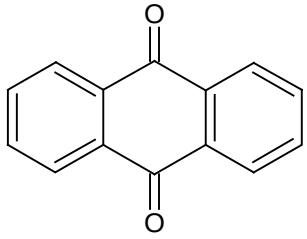
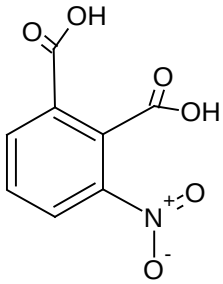
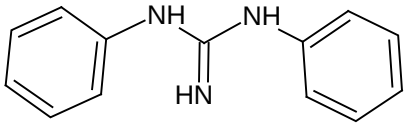
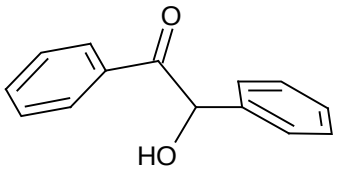
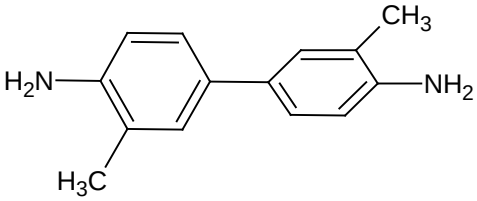
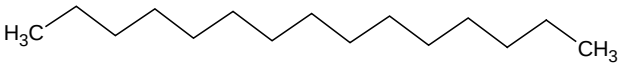
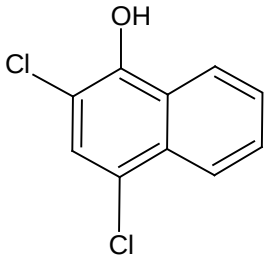
ISBN:978-84-693-4597-9 /DL:T.1010-2010

Annex C.1. List of 375 work chemicals used in this study (continued).

ID	CAS	SMILES code	MW (g/mol)	3D molecular representation and name
w270	92-84-2	S2c1ccccc1Nc3c2cccc3	199.27	 phenothiazine
w271	112-70-9	OCCCCCCCCCCCCC	200.37	 1-tridecanol
w272	122-34-9	Clc1nc(nc(n1)NCC)NCC	201.66	 simazine
w273	76-12-0	ClC(Cl)(F)C(Cl)(Cl)F	203.83	 1,1,2,2-tetrachlorodifluoroethane
w274	5510-99-6	Oc1c(cccc1C(CC)C)C(C)CC	206.33	 2,6-di-sec-butylphenol
w275	96-76-4	Oc1ccc(cc1C(C)(C)C)C(C)(C)C	206.33	 2,4-di-t-butylphenol
w276	609-89-2	Clc1cc(Cl)cc([N+]([O-])=O)c1O	208.00	 2,4-dichloro-6-nitrophenol

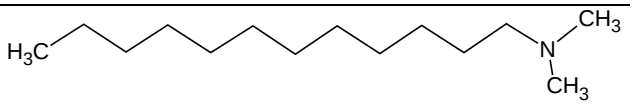
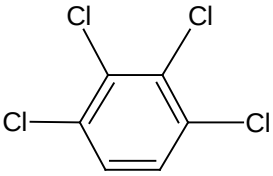
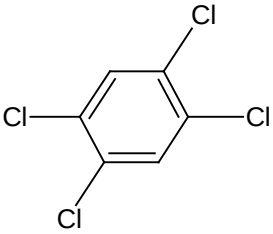
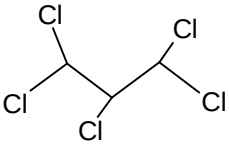
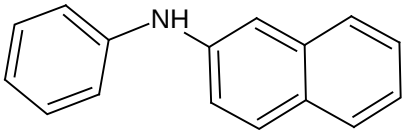
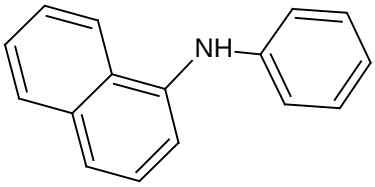
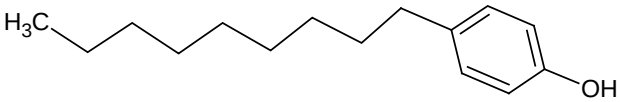
ISBN:978-84-693-4597-9 /DL:T.1010-2010

Annex C.1. List of 375 work chemicals used in this study (continued).

ID	CAS	SMILES code	MW (g/mol)	3D molecular representation and name
w277	84-65-1	O=C2c1c(cccc1)C(=O)c3c2cccc3	208.22	 anthraquinone
w278	603-11-2	O=[N+]([O-])c1cccc(c1C(=O)O)C(=O)O	211.13	 3-nitrophthalic acid
w279	102-06-7	[N@H]=C(Nc1ccccc1)Nc2ccccc2	211.27	 n,n'-diphenylguanidine
w280	119-53-9	O=C(c1ccccc1)C(O)c2ccccc2	212.25	 benzoin
w281	119-93-7	c2(c1ccc(N)c(c1)C)ccc(N)c(c2)C	212.30	 bianisidine
w282	629-62-9	C(CCCCCC)CCCCCCCC	212.42	 pentadecane
w283	2050-76-2	Clc2c1ccccc1c(O)c(Cl)c2	213.06	 2,4-dichloro-1-naphthol

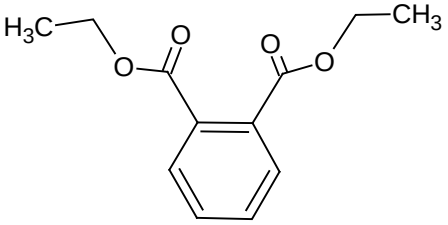
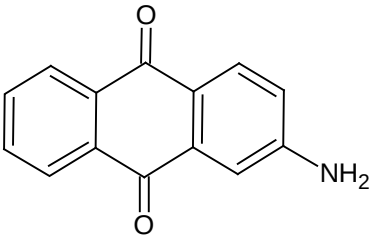
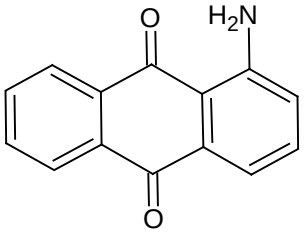
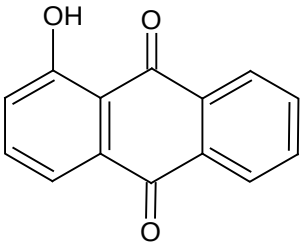
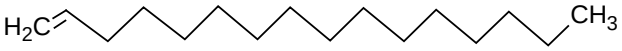
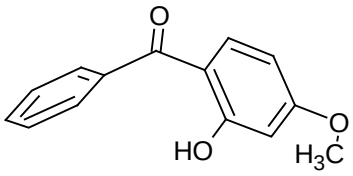
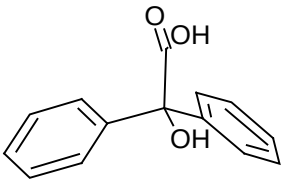
ISBN:978-84-693-4597-9 /DL:T.1010-2010

Annex C.1. List of 375 work chemicals used in this study (continued).

ID	CAS	SMILES code	MW (g/mol)	3D molecular representation and name
w284	112-18-5	N(CCCCCCCCCCCC)(C)C	213.41	 n,n-dimethyldodecylamine
w285	634-66-2	Clc1ccc(Cl)c(Cl)c1Cl	215.89	 1,2,3,4-tetrachlorobenzene
w286	95-94-3	Clc1c(Cl)cc(Cl)c(Cl)c1	215.89	 1,2,4,5-tetrachlorobenzene
w287	15104-61-7	ClC(Cl)C(Cl)C(Cl)Cl	216.32	 1,1,2,3,3-pentachloropropane
w288	135-88-6	c3c(Nc1ccccc1)cc2ccccc2c3	219.29	 n-phenyl-2-naphthylamine
w289	90-30-2	c3c(Nc1ccccc1)c2ccccc2cc3	219.29	 1-naphthalenamine, n-phenyl-
w290	25154-52-3	Oc1ccc(cc1)CCCCCCCCC	220.36	 nonylphenol (isomer mixture)

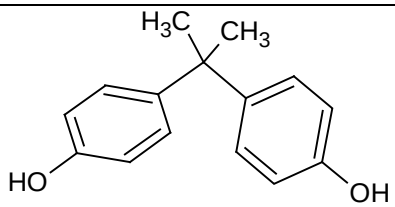
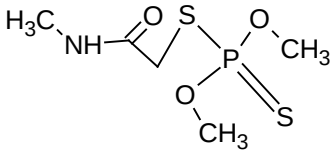
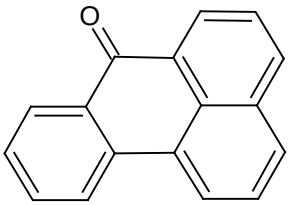
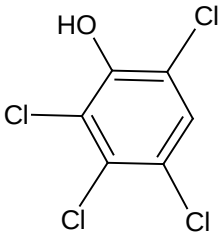
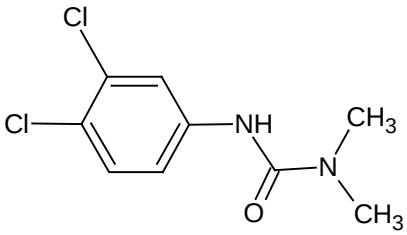
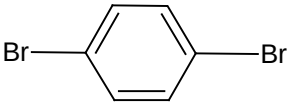
ISBN:978-84-693-4597-9 /DL:T.1010-2010

Annex C.1. List of 375 work chemicals used in this study (continued).

ID	CAS	SMILES code	MW (g/mol)	3D molecular representation and name
w291	84-66-2	O=C(OCC)c1ccccc1C(=O)OCC	222.24	 diethyl phthalate
w292	117-79-3	O=C2c1c(cccc1)C(=O)c3c2ccc(N)c3	223.23	 2-aminoanthraquinone
w293	82-45-1	O=C3c1ccccc1C(=O)c2c3cccc2N	223.23	
w294	129-43-1	O=C2c1ccccc1C(=O)c3c2cccc3O	224.22	 1-hydroxyanthraquinone
w295	629-73-2	C=C\CCCCCCCCCCCCCC	224.43	 1-hexadecene
w296	131-57-7	O=C(c1ccc(cc1O)OC)c2ccccc2	228.25	 oxybenzone
w297	76-93-7	OC(C(=O)O)(c1ccccc1)c2ccccc2	228.25	 benzilic acid

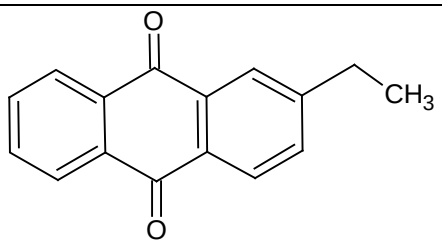
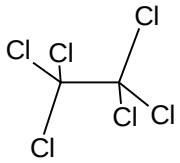
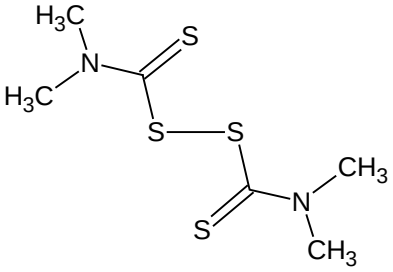
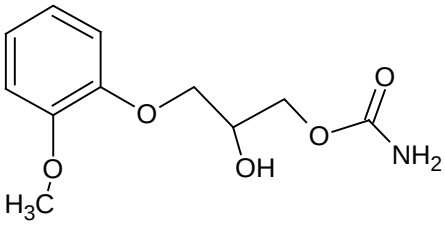
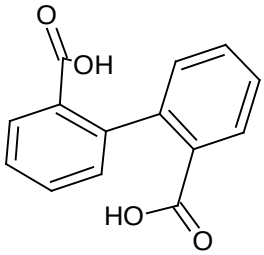
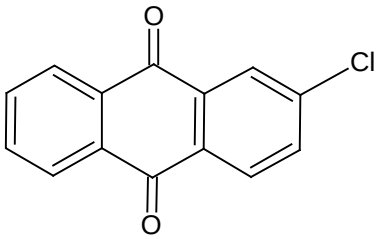
ISBN:978-84-693-4597-9 /DL:T.1010-2010

Annex C.1. List of 375 work chemicals used in this study (continued).

ID	CAS	SMILES code	MW (g/mol)	3D molecular representation and name
w298	80-05-7	Oc1ccc(cc1)C(c2ccc(O)cc2)(C)C	228.29	 diphenylolpropane
w299	60-51-5	O=C(NC)CSP(=S)(OC)OC	229.26	 dimethoate
w300	82-05-3	O=C3c4ccccc4c2cccc1cccc3c12	230.27	 benzanthrone
w301	58-90-2	Clc1c(O)c(Cl)c(Cl)c(Cl)c1	231.89	 2,3,4,6-tetrachlorophenol
w302	330-54-1	Clc1ccc(NC(=O)N(C)C)cc1Cl	233.10	 diuron
w303	106-37-6	Brc1ccc(Br)cc1	235.91	 1,4-dibromobenzene

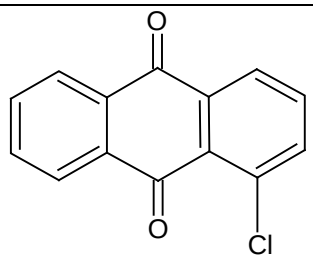
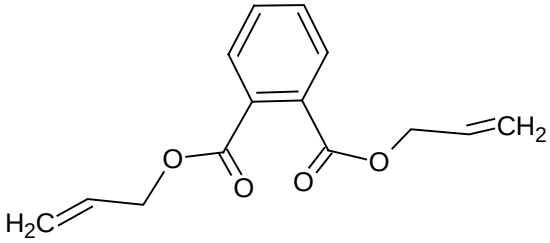
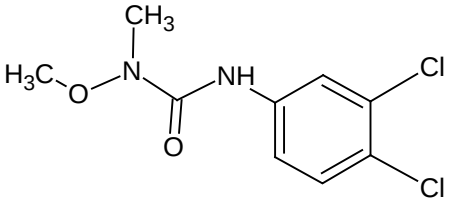
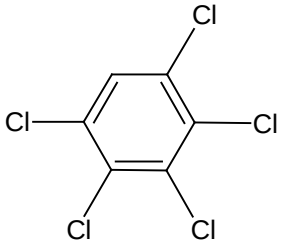
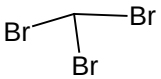
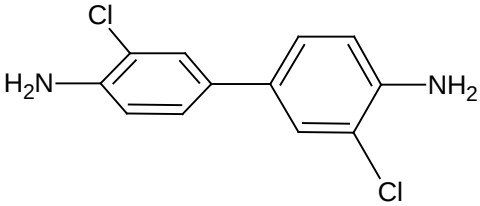
ISBN:978-84-693-4597-9 /DL:T.1010-2010

Annex C.1. List of 375 work chemicals used in this study (continued).

ID	CAS	SMILES code	MW (g/mol)	3D molecular representation and name
w304	84-51-5	O=C2c1c(cccc1)C(=O)c3c2ccc(c3)CC	236.27	 2-ethylanthraquinone
w305	67-72-1	ClC(Cl)(Cl)C(Cl)(Cl)Cl	236.74	 hexachloroethane
w306	137-26-8	CN(C)C(=S)SSC(=S)N(C)C	240.43	 thiram
w307	532-03-6	O=C(OCC(O)COc1cccccc1OC)N	241.25	 1,2-propanediol, 3-(2-methoxyphenoxy)-, 1-carbam
w308	482-05-3	O=C(O)c2c(c1ccccc1C(=O)O)cccc2	242.23	 1,1'-biphenyl -2,2'-dicarboxylic acid
w309	131-09-9	Clc3ccc2C(=O)c1c(cccc1)C(=O)c2c3	242.66	 9,10-anthracenedione, 2-chloro-

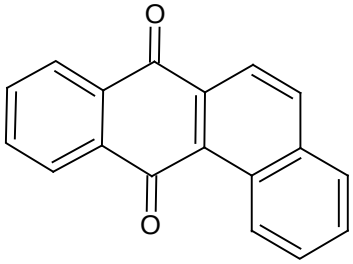
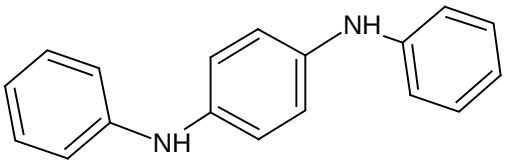
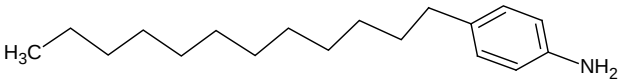
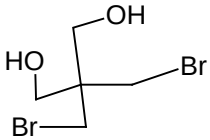
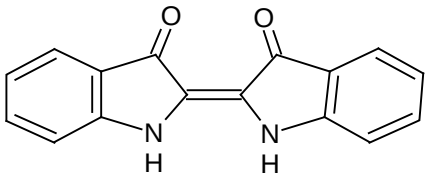
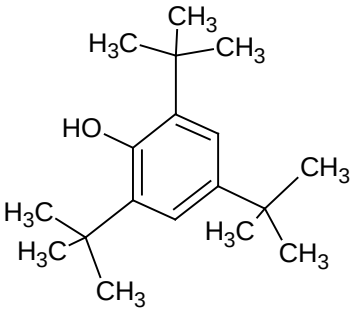
ISBN:978-84-693-4597-9 /DL:T.1010-2010

Annex C.1. List of 375 work chemicals used in this study (continued).

ID	CAS	SMILES code	MW (g/mol)	3D molecular representation and name
w310	82-44-0	O=C2c1ccccc1C(=O)c3c2cccc3Cl	242.66	 1-chloroanthraquinone
w311	131-17-9	O=C(OC\C=C)c1ccccc1C(=O)OC\C=C	246.27	 diallylphthalate
w312	330-55-2	Clc1ccc(NC(=O)N(OC)C)cc1Cl	249.10	 linuron
w313	608-93-5	Clc1cc(Cl)c(Cl)c(Cl)c1Cl	250.34	 pentachlorobenzene
w314	75-25-2	BrC(Br)Br	252.73	 tribromomethane
w315	91-94-1	Clc2cc(c1ccc(N)c(Cl)c1)ccc2N	253.13	 3,3'-dichlorobenzidine

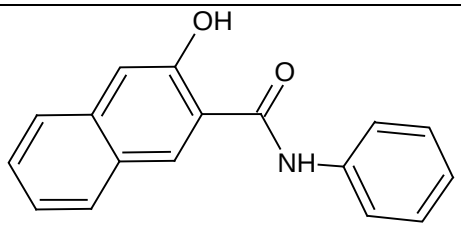
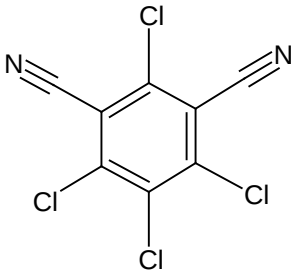
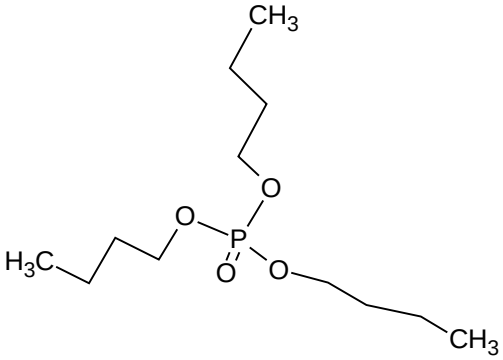
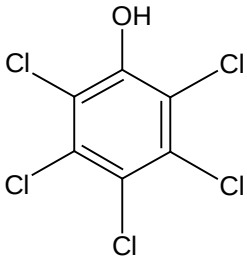
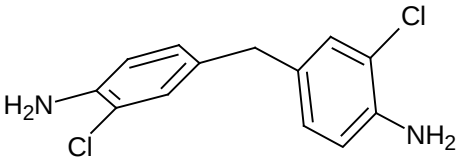
ISBN:978-84-693-4597-9 /DL:T.1010-2010

Annex C.1. List of 375 work chemicals used in this study (continued).

ID	CAS	SMILES code	MW (g/mol)	3D molecular representation and name
w316	2498-66-0	O=C3c1c(ccc2c1cccc2)C(=O)c4c3cccc4	258.28	 benz a anthracene-7,12-dione
w317	74-31-7	c3c(Nc1ccc(cc1)Nc2ccccc2)cccc3	260.36	 n,n'-diphenyl-p-benzenediamine
w318	104-42-7	Nc1ccc(cc1)CCCCCCCCCCCC	261.45	 p-dodecylaniline
w319	3296-90-0	BrCC(CO)(CBr)CO	261.94	 1,3-propanediol, 2,2-bis(brme)-
w320	482-89-3	O=C\4c1ccccc1NC/4=C3\C(=O)c2ccccc2N3	262.27	 2-(1,3-dihydro-3-oxo-2h-indol-2-ylidene)-1,2-di*
w321	732-26-3	Oc1c(cc(cc1C(C)(C)C)C(C)(C)C)C(C)(C)C	262.44	 2,4,6-tri(tert-butyl)phenol

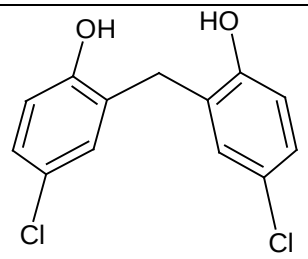
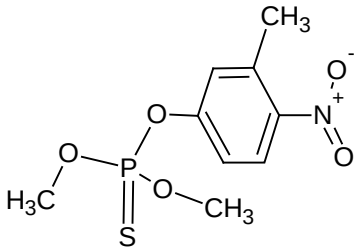
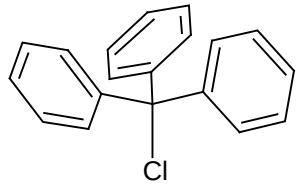
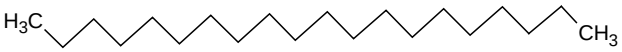
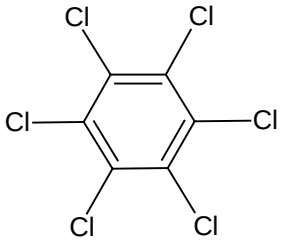
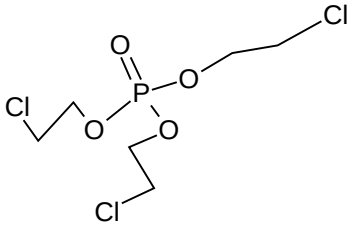
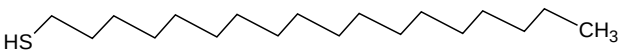
ISBN:978-84-693-4597-9 /DL:T.1010-2010

Annex C.1. List of 375 work chemicals used in this study (continued).

ID	CAS	SMILES code	MW (g/mol)	3D molecular representation and name
w322	92-77-3	O=C(c2cc1c(cccc1)cc2O)Nc3ccccc3	263.30	 naphthol as
w323	1897-45-6	Clc1c(C#N)c(Cl)c(C#N)c(Cl)c1Cl	265.91	 chlorothanonil
w324	126-73-8	O=P(OCCCC)(OCCCC)OCCCC	266.32	 tributylphosphate
w325	87-86-5	Clc1c(O)c(Cl)c(Cl)c(Cl)c1Cl	266.34	 hydroxypentachlorobenzene
w326	101-14-4	Clc1cc(ccc1N)Cc2ccc(N)c(Cl)c2	267.16	 4,4'-methylenebis(2-chloroaniline)

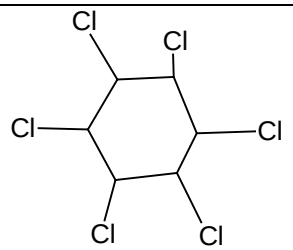
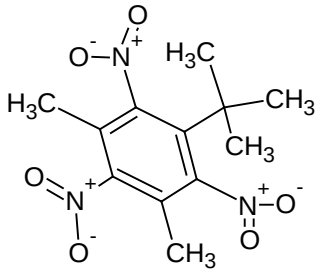
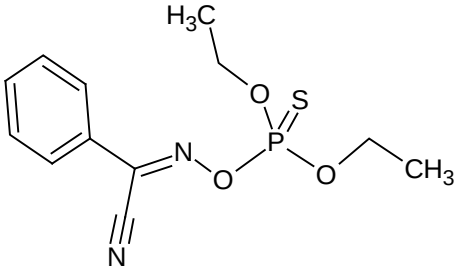
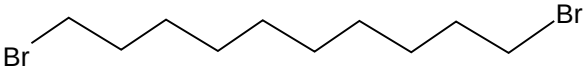
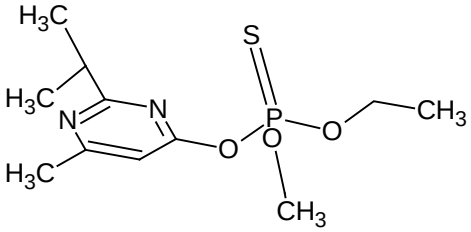
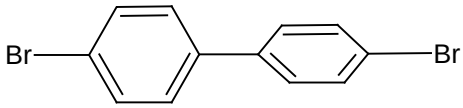
ISBN:978-84-693-4597-9 /DL:T.1010-2010

Annex C.1. List of 375 work chemicals used in this study (continued).

ID	CAS	SMILES code	MW (g/mol)	3D molecular representation and name
w327	97-23-4	Clc1cc(c(O)cc1)Cc2cc(Cl)ccc2O	269.13	 phenol,2,2'-methylenebis 4-chloro-
w328	122-14-5	S=P(Oc1cc(c(cc1)[N+](\[O-\])=O)C)(OC)OC	277.24	 fenitrothion
w329	76-83-5	ClC(c1ccccc1)(c2ccccc2)c3ccccc3	278.78	 benzene, 1,1',1"-(chloromethylidyne)tris-
w330	112-95-8	C(CCCCCCCCCCCCCCCCCC)CC	282.56	 eicosane
w331	118-74-1	Clc1c(Cl)c(Cl)c(Cl)c(Cl)c1Cl	284.78	 hexachlorobenzene
w332	115-96-8	ClCCOP(=O)(OCCCl)OCCCl	285.49	 tri-2-chloroethyl phosphate
w333	2885-00-9	SCCCCCCCCCCCCCCCCCC	286.57	 1-octadecanethiol

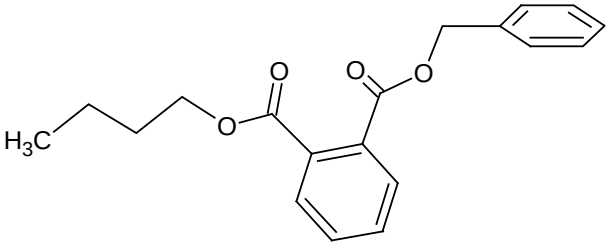
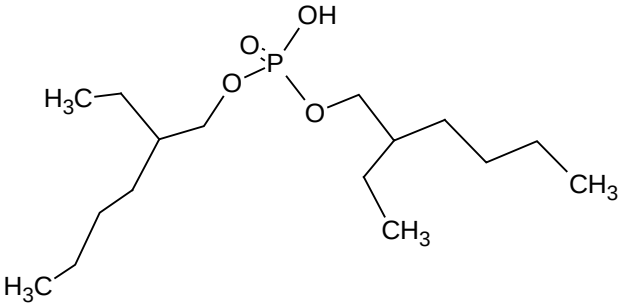
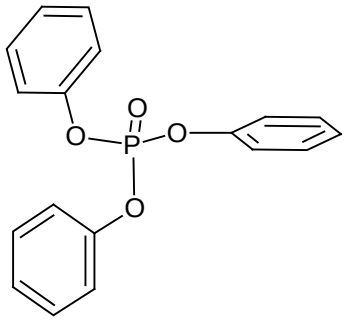
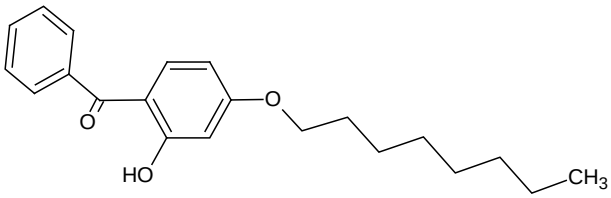
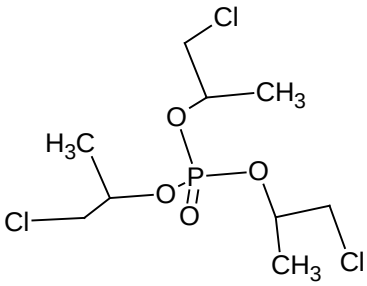
ISBN:978-84-693-4597-9 /DL:T.1010-2010

Annex C.1. List of 375 work chemicals used in this study (continued).

ID	CAS	SMILES code	MW (g/mol)	3D molecular representation and name
w334	608-73-1	ClC1C(Cl)C(Cl)C(Cl)C(Cl)C1Cl	290.83	 1,2,3,4,5,6-hexachlorocyclohexane
w335	81-15-2	O=[N+]([O-])c1c(c(c(c(c1C(C)(C)C)[N+]([O-])=O)C)[N+]([O-])=O)C	297.27	 benzene, 1-(1,1-dimethylethyl)-3,5-dimethyl-2,4,
w336	14816-18-3	N#C/C(=N\OP(=S)(OCC)OCC)c1ccccc1	298.30	 phoxim
w337	4101-68-2	BrCCCCCCCCCCBr	300.08	 1,2-dibromodecane
w338	333-41-5	S=P(OCC)(OCC)Oc1nc(nc(c1)C)C(C)C	304.35	 diazinon
w339	92-86-4	Brc2ccc(c1ccc(Br)cc1)cc2	312.01	 4,4'-dibromobiphenyl

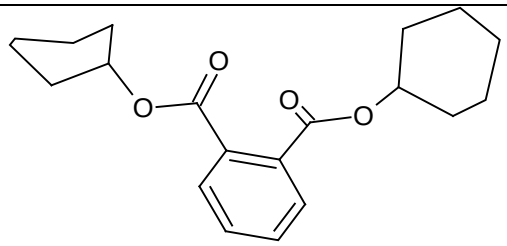
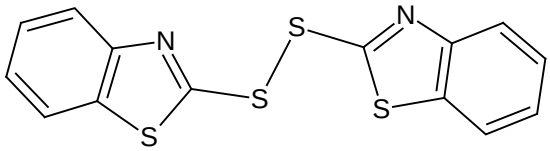
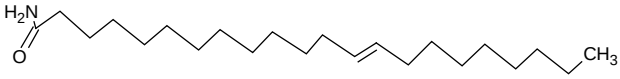
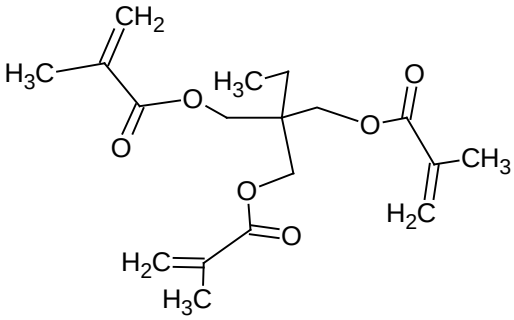
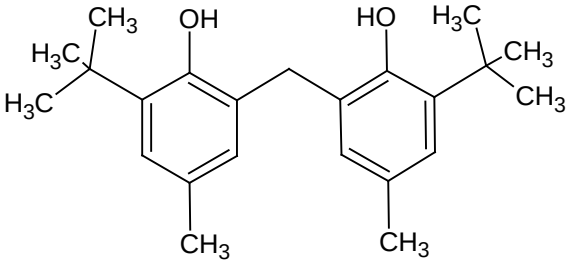
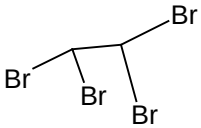
ISBN:978-84-693-4597-9 /DL:T.1010-2010

Annex C.1. List of 375 work chemicals used in this study (continued).

ID	CAS	SMILES code	MW (g/mol)	3D molecular representation and name
w340	85-68-7	O=C(OCc1ccccc1)c2ccccc2C(=O)OCCCC	312.37	 butyl benzyl phthalate
w341	298-07-7	O=P(OCC(CCCC)CC)(O)OCC(CC)CCCC	322.43	 bis(2-ethylhexyl)phosphate
w342	115-86-6	O=P(Oc1ccccc1)(Oc2ccccc2)Oc3ccccc3	326.29	 triphenylphosphate
w343	1843-05-6	O=C(c1ccc(OCCCCCCCC)cc1O)c2ccccc2	326.47	 methanone, 2-hydroxy-4-(octyloxy)phenyl phenyl-
w344	13674-84-5	ClCC(OP(=O)(OC(CCl)C)OC(C)CCl)C	327.57	 2-propanol, 1-chloro-, phosphate (3:1)

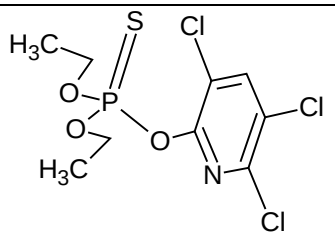
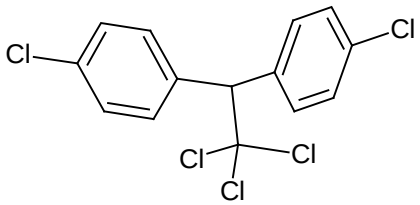
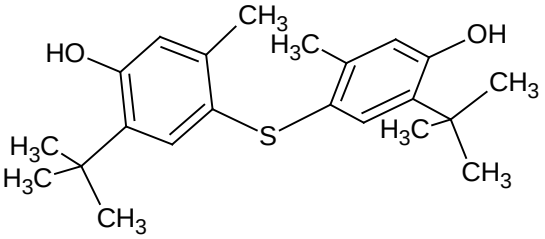
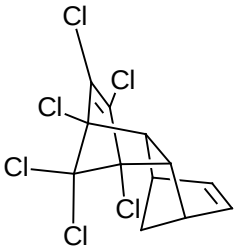
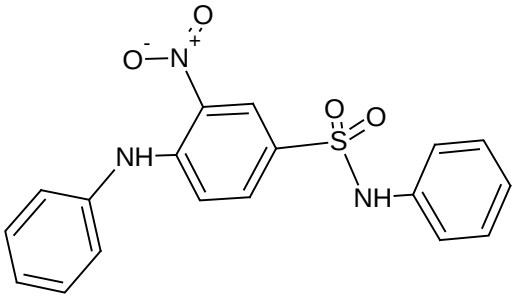
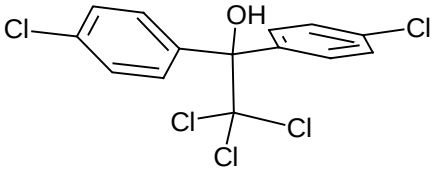
ISBN:978-84-693-4597-9 /DL:T.1010-2010

Annex C.1. List of 375 work chemicals used in this study (continued).

ID	CAS	SMILES code	MW (g/mol)	3D molecular representation and name
w345	84-61-7	O=C(OC1CCCCC1)c3ccccc3C(=O)OC2CCCCC2	330.43	 dicyclohexyl phthalate
w346	120-78-5	n1c4ccccc4sc1SSc2nc3ccccc3s2	332.49	 2,2'-dithiobisbenzothiazole
w347	112-84-5	O=C(N)CCCCCCCCCCC/C=C/CCCCCCCC	337.59	 13-decosenamide (cis)
w348	3290-92-4	O=C(OCC(COC(=O)\C(=C)C)(CC)COC(=O)\C(=C)C)\C(=C)C	338.00	 trimethylolpropane trimethacrylate
w349	119-47-1	Oc1c(cc(cc1C(C)(C)C)C)Cc2cc(cc(c2O)C(C)(C)C)C	340.51	 bis (2-hydroxy-3tert-butyl-5-methylphenyl) metha
w350	79-27-6	BrC(Br)C(Br)Br	345.65	 1,1,2,2-tetrabromoethane

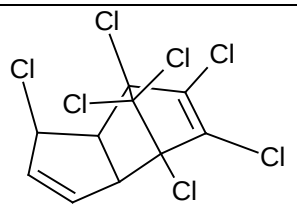
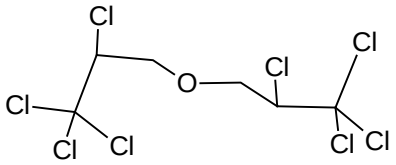
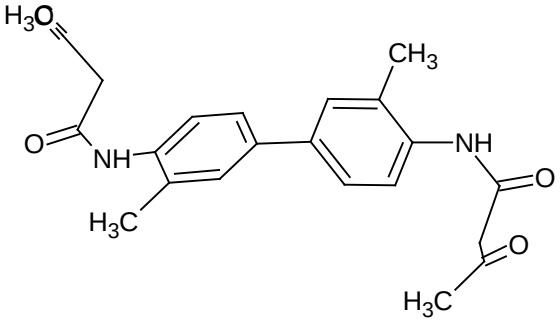
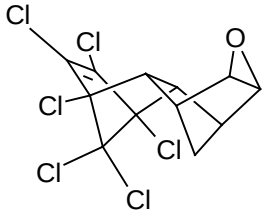
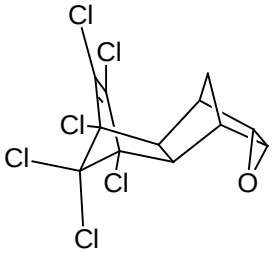
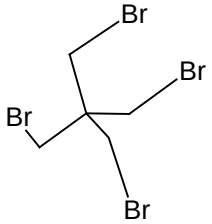
ISBN:978-84-693-4597-9 /DL:T.1010-2010

Annex C.1. List of 375 work chemicals used in this study (continued).

ID	CAS	SMILES code	MW (g/mol)	3D molecular representation and name
w351	2921-88-2	Clc1c(OP(=S)(OCC)OCC)nc(Cl)c(Cl)c1	350.59	 chlorpyrifos
w352	50-29-3	Clc1ccc(cc1)C(c2ccc(Cl)cc2)C(Cl)(Cl)Cl	354.50	 1,1,1-trichloro-2,2-bis-(4-chlorophenyl)ethane
w353	96-69-5	S(c1c(cc(O)c(c1)C(C)(C)C)c2cc(c(O)cc2C)C(C)(C)C	358.55	 4,4'-thiobis(6-tert-butyl-3-cresol)
w354	309-00-2	ClC3=C(Cl)C4(Cl)C2C(C1C=CC2C1)C3(Cl)C4(Cl)Cl	364.92	 aldrin
w355	5124-25-4	O=S(=O)(Nc1ccccc1)c3cc(c(Nc2ccccc2)cc3)[N+]([O-])=O	369.40	 c.i. disperse yellow 42
w356	115-32-2	Clc1ccc(cc1)C(O)(c2ccc(Cl)cc2)C(Cl)(Cl)Cl	370.49	 dicofol

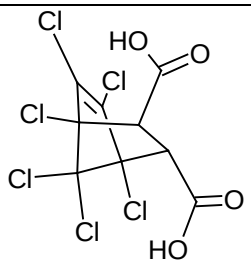
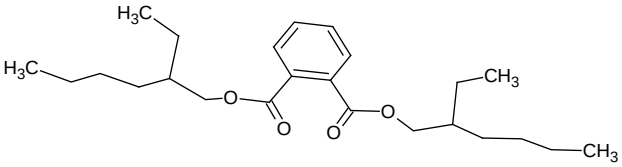
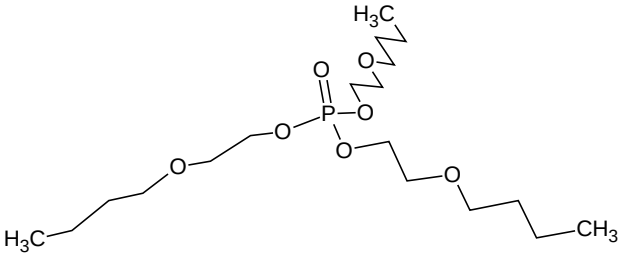
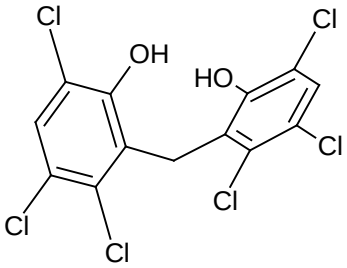
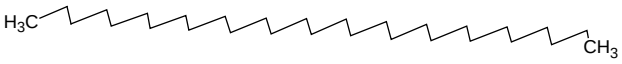
ISBN:978-84-693-4597-9 /DL:T.1010-2010

Annex C.1. List of 375 work chemicals used in this study (continued).

ID	CAS	SMILES code	MW (g/mol)	3D molecular representation and name
w357	76-44-8	ClC2=C(Cl)C3(Cl)C1C=CC(Cl)C1C2(Cl)C3(Cl)Cl	373.32	 heptachlor
w358	127-90-2	ClC(Cl)(Cl)C(Cl)COCC(Cl)C(Cl)(Cl)Cl	377.74	 1,1'-oxybis[2,3,3,3-tetrachloropropane]
w359	91-96-3	O=C(Nc1ccc(cc1C)c2ccc(NC(=O)CC(C)=O)c(C)c2)CC(C)=O	380.45	 c.i. azoic coupling component 5
w360	60-57-1	ClC1=C(Cl)C2(Cl)C(Cl)(Cl)C1(Cl)C4C2C5C3OC3C4C5	380.91	 dieldrin
w361	72-20-8	ClC1=C(Cl)C2(Cl)C(Cl)(Cl)C1(Cl)C4C2C5C3OC3C4C5	380.91	 endrin
w362	3229-00-3	BrCC(CBr)(CBr)CBr	387.74	 propane, 1,3-dibromo-2,2-bis(bromomethyl)-

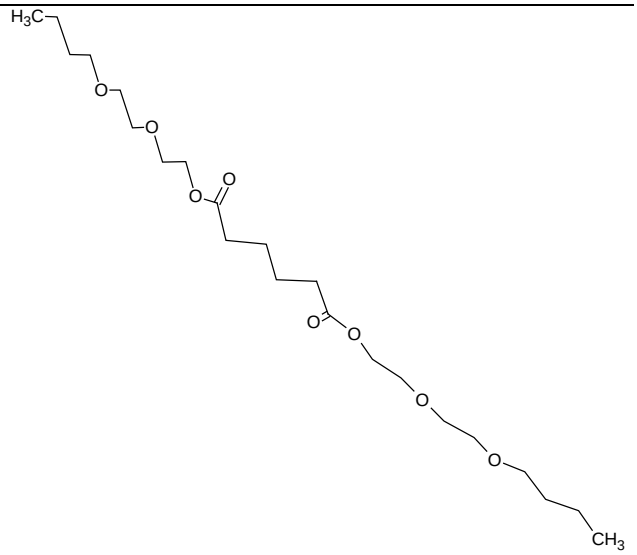
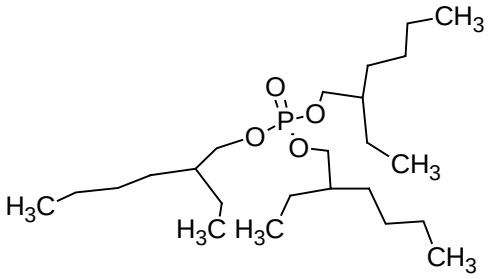
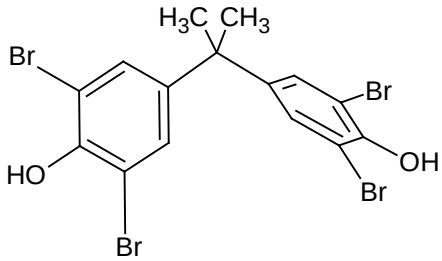
ISBN:978-84-693-4597-9 /DL:T.1010-2010

Annex C.1. List of 375 work chemicals used in this study (continued).

ID	CAS	SMILES code	MW (g/mol)	3D molecular representation and name
w363	115-28-6	ClC2(Cl)C1(Cl)C(Cl)=C(Cl)C2(Cl)C(C(=O)O)C1C(=O)O	388.85	 chlorendic acid
w364	117-81-7	O=C(OCC(CC)CCCC)c1ccccc1C(=O)OCC(CC)CCCC	390.57	 bis(2-ethylhexyl)phthalate
w365	78-51-3	O=P(OCCOCCCC)(OCCOCCCC)OCCOCCCC	398.54	 tri-2-butoxyethyl phosphate
w366	70-30-4	Clc1c(c(O)c(Cl)cc1Cl)Cc2c(O)c(Cl)cc(Cl)c2Cl	406.91	 hexachlorophene
w367	630-03-5	C(CCCCCCCCCCCCCCCCCCCCCCC)CCCCCC	408.80	 nonacosane

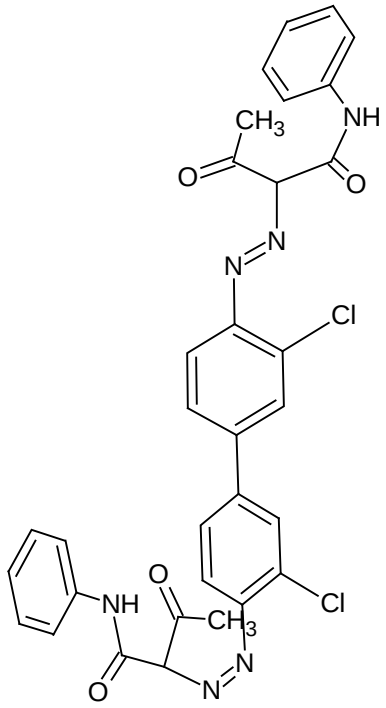
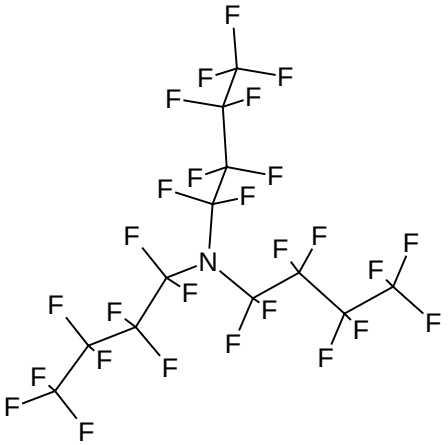
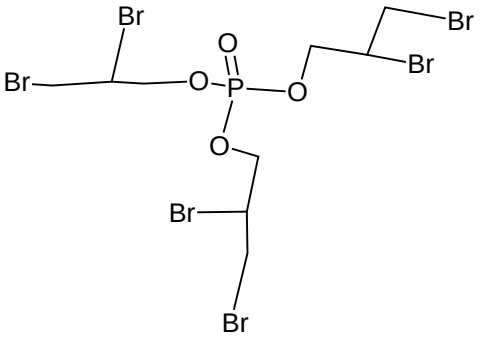
ISBN:978-84-693-4597-9 /DL:T.1010-2010

Annex C.1. List of 375 work chemicals used in this study (continued).

ID	CAS	SMILES code	MW (g/mol)	3D molecular representation and name
w368	141-17-3	O=C(OCCOCCOCCC C)CCCCC(=O)OCCO CCOCCCC	434.58	 bis(2-(2-butoxyethoxy)ethyl) adipate
w369	78-42-2	CCCCC(COP(=O)(O CC(CC)CCCC)OCC(CC)CCCC)CC	434.65	 tris(2-ethylhexyl) phosphate
w370	79-94-7	Brc1cc(cc(Br)c1O)C(c 2cc(Br)c(O)c(Br)c2)(C)C	543.88	 2,2-bis(4-hydroxy-3,5-dibromophenyl)propane

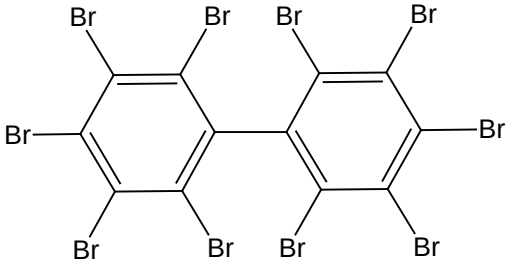
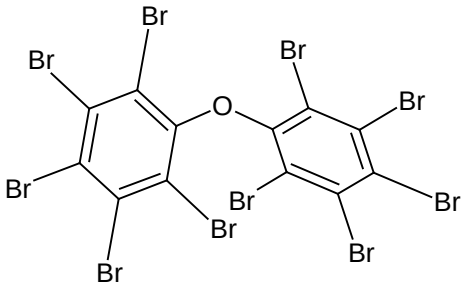
ISBN:978-84-693-4597-9 /DL:T.1010-2010

Annex C.1. List of 375 work chemicals used in this study (continued).

ID	CAS	SMILES code	MW (g/mol)	3D molecular representation and name
w371	6358-85-6	Clc2cc(ccc2/N=N/C(C(=O)C)C(=O)Nc1ccccc1)c4ccc(\N=N/C(C(=O)C)C(=O)Nc3ccccc3)c(Cl)c4	629.51	 c.i. pigment yellow 12
w372	311-89-7	FC(F)(N(C(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)(F)F	671.10	 heptacosafluorotributylamine
w373	126-72-7	BrCC(Br)COP(=O)(OCC(Br)CBr)OCC(Br)CBr	697.62	 tris(2,3-dibromopropyl) phosphate

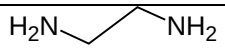
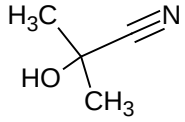
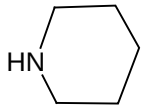
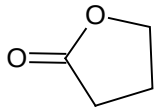
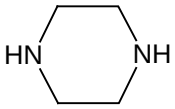
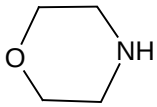
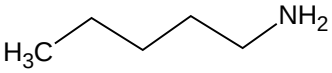
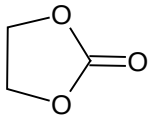
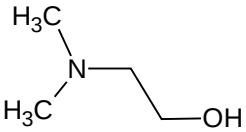
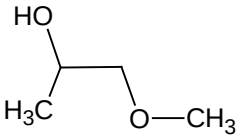
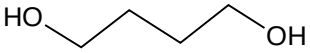
ISBN:978-84-693-4597-9 /DL:T.1010-2010

Annex C.1. List of 375 work chemicals used in this study (continued).

ID	CAS	SMILES code	MW (g/mol)	3D molecular representation and name
w374	13654-09-6	Brc1c(c(Br)c(Br)c(Br)c1Br)c2c(Br)c(Br)c(Br)c(Br)c2Br	943.17	 decabromobiphenyl
w375	1163-19-5	Brc2c(Oc1c(Br)c(Br)c(Br)c(Br)c1Br)c(Br)c(Br)c(Br)c2Br	959.17	 decabromodiphenyl ether

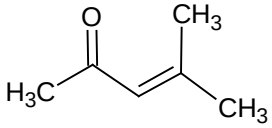
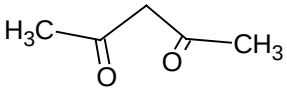
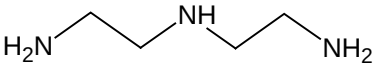
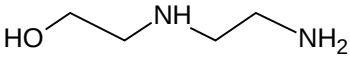
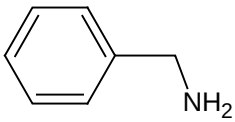
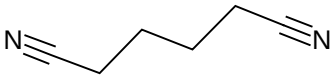
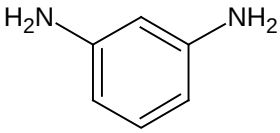
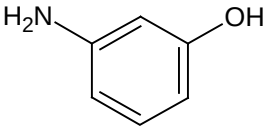
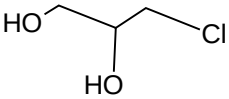
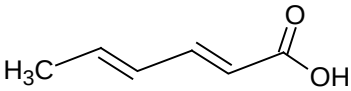
ISBN:978-84-693-4597-9 /DL:T.1010-2010

Annex C.2. List of 93 validation chemicals used in this study.

ID	CAS	SMILES code	MW (g/mol)	3d molecular representation and name
v001	107-15-3	NCCN	60.10	 ethylenediamine
v002	75-86-5	N#CC(O)(C)C	85.11	 acetone cyanohydrin
v003	110-89-4	N1CCCCC1	85.15	 piperidine
v004	96-48-0	O=C1OCCC1	86.09	 gamma-butyrolactone
v005	110-85-0	N1CCNCC1	86.14	 piperazine
v006	110-91-8	O1CCNCC1	87.12	
v007	110-58-7	NCCCCC	87.17	 pentylamine
v008	96-49-1	O=C1OCCO1	88.06	 1,3-dioxolan-2-one
v009	108-01-0	OCCN(C)C	89.14	 2-dimethylaminoethanol
v010	107-98-2	OC(C)COC	90.12	 1-methoxy-2-propanol
v011	110-63-4	OCCCCO	90.12	 1,4-butanediol

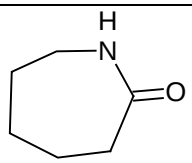
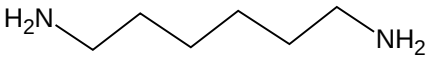
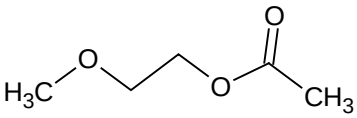
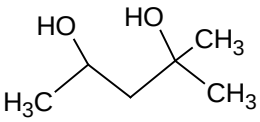
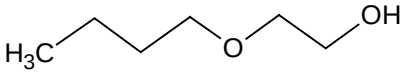
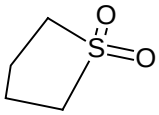
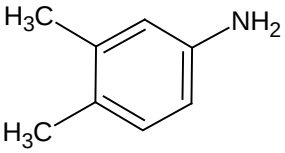
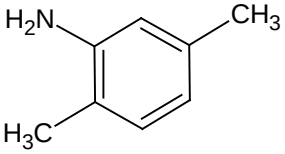
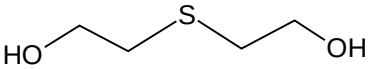
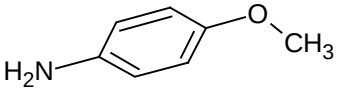
ISBN:978-84-693-4597-9 /DL:T.1010-2010

Annex C.2. List of 93 validation chemicals used in this study (continued).

ID	CAS	SMILES code	MW (g/mol)	3d molecular representation and name
v012	141-79-7	O=C(\C=C(/C)C)C	98.15	 mesityl oxide
v013	123-54-6	O=C(C)CC(=O)C	100.12	 2,4-pentanedione
v014	111-40-0	NCCNCCN	103.17	 diethylenetriamine
v015	111-41-1	OCCNCCN	104.15	 2-(2-aminoethylamino)ethanol
v016	100-46-9	NCc1ccccc1	107.16	 benzylamine
v017	111-69-3	N#CCCCCC#N	108.14	 adiponitrile
v018	108-45-2	Nc1cccc(N)c1	108.14	 1,3-benzenediamine
v019	591-27-5	Oc1cccc(N)c1	109.13	 phenol, 3-amino-
v020	96-24-2	ClCC(O)CO	110.55	 3-chloro-1,2-propanediol
v021	110-44-1	O=C(O)\C=C\C=C\C	112.13	 sorbic acid

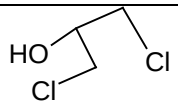
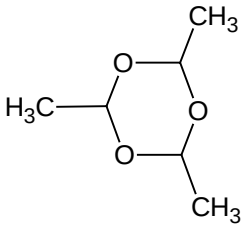
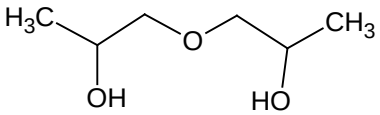
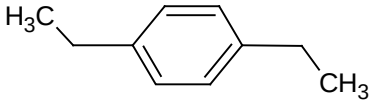
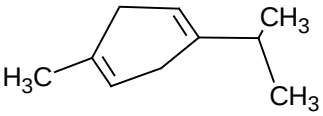
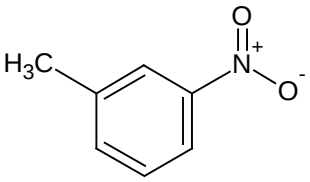
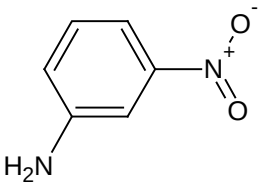
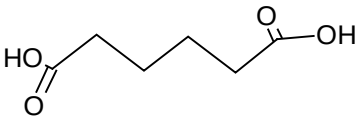
ISBN:978-84-693-4597-9 /DL:T.1010-2010

Annex C.2. List of 93 validation chemicals used in this study (continued).

ID	CAS	SMILES code	MW (g/mol)	3d molecular representation and name
v022	105-60-2	O=C1NCCCCC1	113.16	 caprolactam
v023	124-09-4	NCCCCCCN	116.21	 hexamethylene diamine
v024	110-49-6	O=C(OCCOC)C	118.13	 2-methoxyethyl acetate
v025	107-41-5	OC(C)CC(O)(C)C	118.18	 2-methyl-2,4-pentanediol
v026	111-76-2	OCCOCCCC	118.18	 2-butoxyethanol
v027	126-33-0	O=S1(=O)CCCC1	120.17	 tetrahydrothiophene-1,1-dioxide
v028	95-64-7	Nc1cc(c(cc1)C)C	121.18	 3,4-xylidine
v029	95-78-3	Nc1cc(ccc1C)C	121.18	 2,5-dimethylaniline
v030	111-48-8	OCCSCCO	122.19	 2,2'-thiobisethanol
v031	104-94-9	O(c1ccc(N)cc1)C	123.16	 4-methoxyaniline

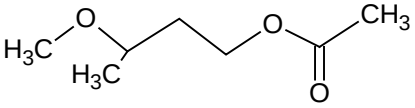
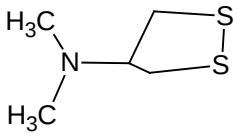
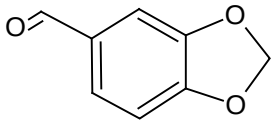
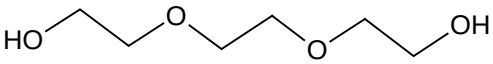
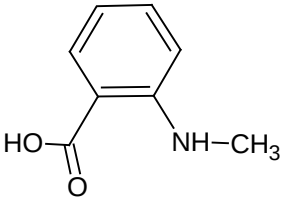
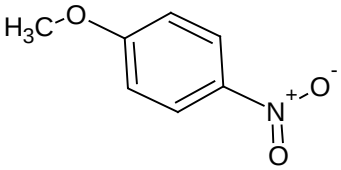
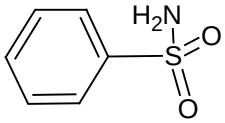
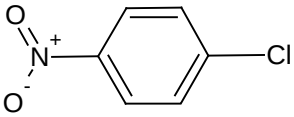
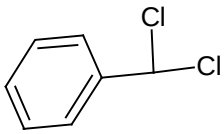
ISBN:978-84-693-4597-9 /DL:T.1010-2010

Annex C.2. List of 93 validation chemicals used in this study (continued).

ID	CAS	SMILES code	MW (g/mol)	3d molecular representation and name
v032	96-23-1	ClCC(O)CCl	128.99	 1,3-dichloro-2-propanol
v033	123-63-7	O1C(OC(OC1C)C)C	132.16	 paraldehyde
v034	110-98-5	OC(C)COCC(O)C	134.18	 1,1'-oxydi-2-propanol
v035	105-05-5	c1cc(ccc1CC)CC	134.22	 p-diethylbenzene
v036	99-85-4	C1=C(C)CC=C(C(C)C)C1	136.24	 gamma-terpinene
v037	99-08-1	Cc1cc(ccc1)[N+](=O)[O-]	137.14	 3-nitrotoluene
v038	99-09-2	O=[N+]([O-])c1cccc(N)c1	138.13	 3-nitroaniline
v039	124-04-9	O=C(O)CCCCC(=O)O	146.14	 hexanedioic acid

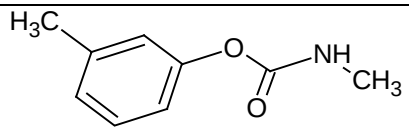
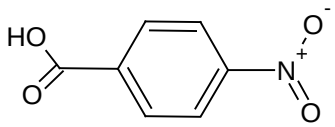
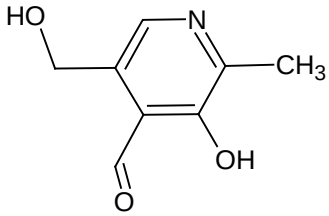
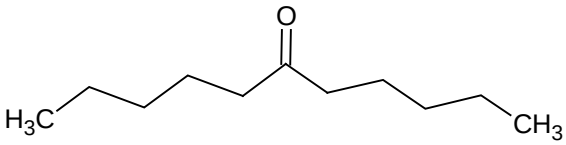
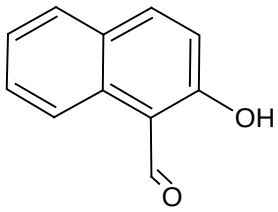
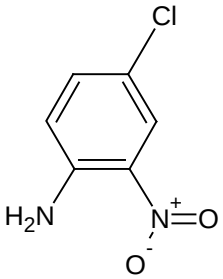
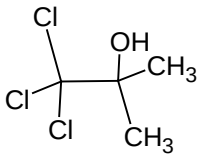
ISBN:978-84-693-4597-9 /DL:T.1010-2010

Annex C.2. List of 93 validation chemicals used in this study (continued).

ID	CAS	SMILES code	MW (g/mol)	3d molecular representation and name
v040	4435-53-4	O=C(OCCC(OC)C)C	146.19	 3-methoxybutyl acetate
v041	1631-58-9	S1SCC(N(C)C)C1	149.28	 nereistoxin
v042	120-57-0	O=Cc1ccc2OCOc2c1	150.14	 piperonal
v043	112-27-6	OCCOCCOCCO	150.18	 3,6-dioxaoctane-1,8-diol
v044	119-68-6	O=C(O)c1ccccc1NC	151.17	 benzoic acid, 2-(methylamino)-
v045	100-17-4	[O-][N+](=O)c1ccc(OC)cc1	153.14	 p-nitroanisole
v046	98-10-2	O=S(=O)(N)c1ccccc1	157.19	 benzenesulfonamide
v047	100-00-5	O=[N+]([O-])c1ccc(Cl)cc1	157.56	 p-chloronitrobenzene
v048	98-87-3	ClC(Cl)c1ccccc1	161.03	 (dichloromethyl)benzene

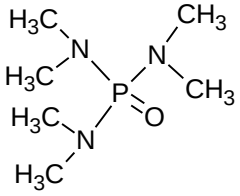
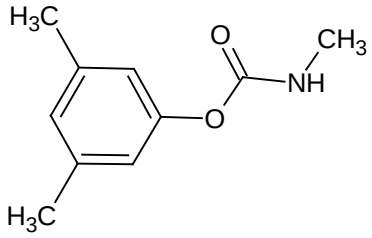
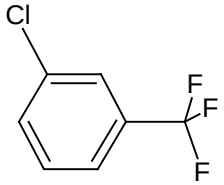
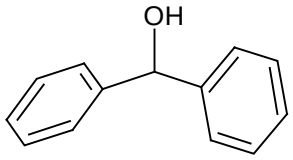
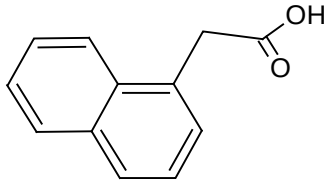
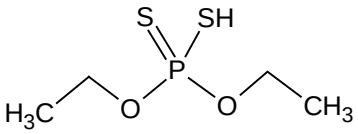
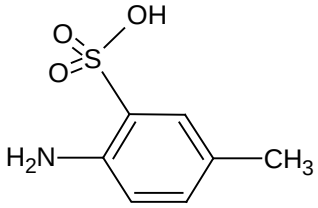
ISBN:978-84-693-4597-9 /DL:T.1010-2010

Annex C.2. List of 93 validation chemicals used in this study (continued).

ID	CAS	SMILES code	MW (g/mol)	3d molecular representation and name
v049	1129-41-5	O=C(Oc1cc(ccc1)C)NC	165.19	 n-methyl-m-tolylcarbamate
v050	62-23-7	O=[N+]([O-])c1ccc(C(=O)O)cc1	167.12	 p-nitrobenzoic acid
v051	66-72-8	O=Cc1c(cnc(c1O)C)CO	167.17	 pyridoxal
v052	927-49-1	O=C(CCCCC)CCCCC	170.30	 6-undecanone
v053	708-06-5	O=Cc1c2c(ccc1O)cccc2	172.19	 1-naphthalenecarboxaldehyde, 2-hydroxy-
v054	89-63-4	Clc1cc([N+]([O-])=O)c(N)cc1	172.57	 4-chloro-2-nitroaniline
v055	57-15-8	ClC(Cl)(Cl)C(O)(C)C	177.46	 b,b,b-trichloro-t-butanol

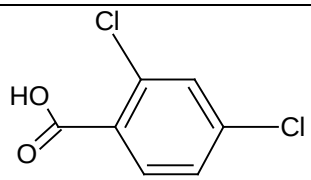
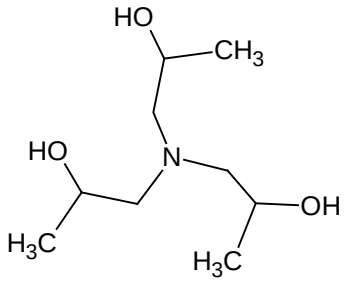
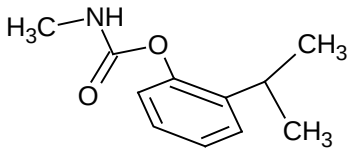
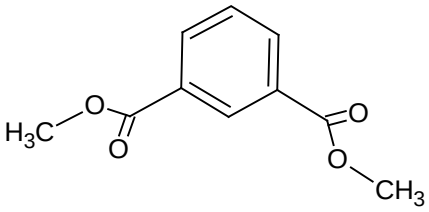
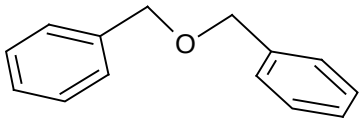
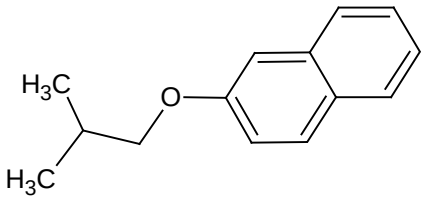
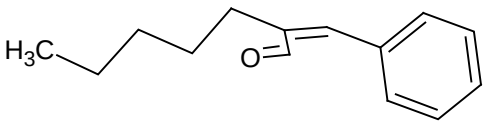
ISBN:978-84-693-4597-9 /DL:T.1010-2010

Annex C.2. List of 93 validation chemicals used in this study (continued).

ID	CAS	SMILES code	MW (g/mol)	3d molecular representation and name
v056	680-31-9	O=P(N(C)C)(N(C)C)N(C)C	179.20	 hexamethylphosphoramide
v057	2655-14-3	O=C(Oc1cc(cc(c1)C)C)NC	179.22	 n-me-3,5-dimethylphenyl carbamate
v058	98-15-7	FC(F)(F)c1cc(Cl)ccc1	180.56	 benzene, 1-chloro-3-(trifluoromethyl)-
v059	91-01-0	OC(c1ccccc1)c2ccccc2	184.24	 benzhydrol
v060	86-87-3	O=C(O)Cc2cccc1ccccc12	186.21	 naphthaleneacetic acid
v061	298-06-6	S=P(OCC)(OCC)S	186.23	 o,o-diethyl dithiophosphate
v062	88-44-8	O=S(=O)(O)c1cc(ccc1N)C	187.22	 2-amino-5-methylbenzenesulfonic acid

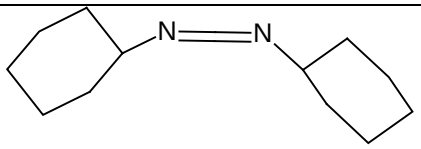
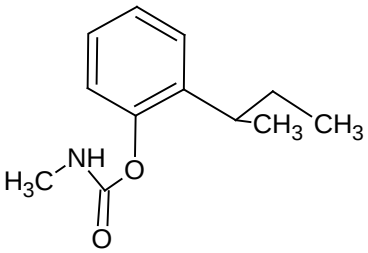
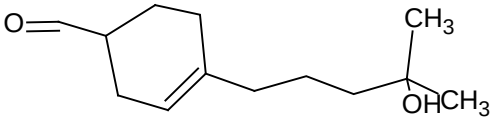
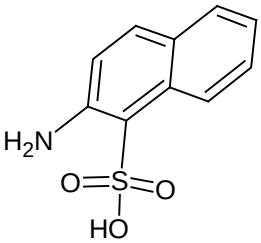
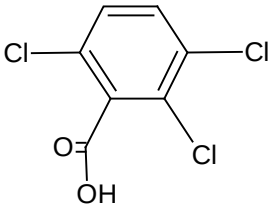
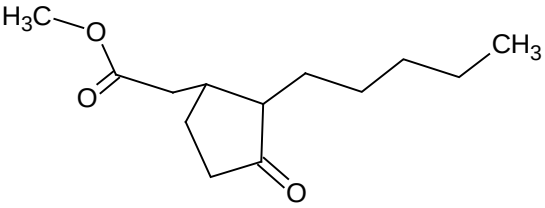
ISBN:978-84-693-4597-9 /DL:T.1010-2010

Annex C.2. List of 93 validation chemicals used in this study (continued).

ID	CAS	SMILES code	MW (g/mol)	3d molecular representation and name
v063	50-84-0	Clc1cc(Cl)ccc1C(=O)O	191.01	 2,4-dichlorobenzoic acid
v064	122-20-3	OC(CN(CC(O)C)CC(O)C)C	191.27	 2-propanol, 1,1',1"-nitrilotris-
v065	2631-40-5	O=C(Oc1ccccc1C(C)C)NC	193.25	 isoprocarb
v066	1459-93-4	O=C(OC)c1cccc(C(=O)OC)c1	194.19	 dimethyl isophthalate
v067	103-50-4	O(Cc1ccccc1)Cc2ccccc2	198.27	 dibenzyl ether
v068	2173-57-1	O(c2ccc1c(cccc1)c2)CC(C)C	200.28	 naphthalene, 2-(2-methylpropoxy)-
v069	122-40-7	O=C/C(=C\c1ccccc1)CCCCC	202.30	 heptanal, 2-(phenylmethylene)-

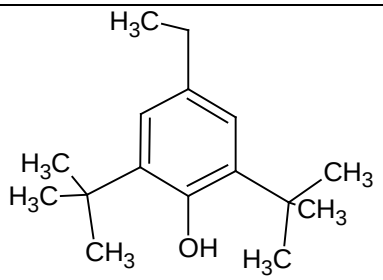
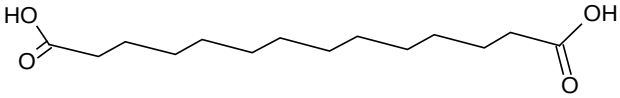
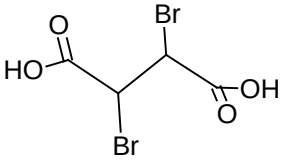
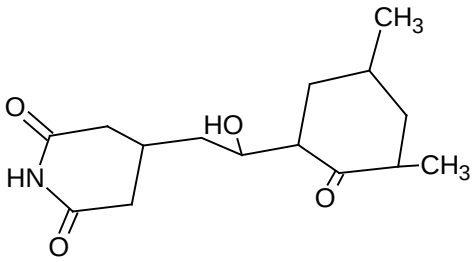
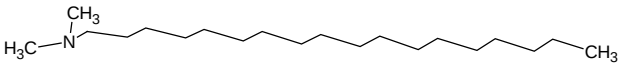
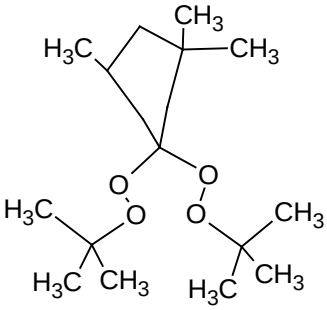
ISBN:978-84-693-4597-9 /DL:T.1010-2010

Annex C.2. List of 93 validation chemicals used in this study (continued).

ID	CAS	SMILES code	MW (g/mol)	3d molecular representation and name
v070	538-75-0	N(=C=N/C1CCCCC1)\C2CCCCC2	206.33	 cyclohexanamine, n,n'-methanetetraylbis-
v071	3766-81-2	O=C(Oc1ccccc1C(C)CC)NC	207.27	 n-methyl o-sec-butyl phenyl carbamate
v072	31906-04-4	O=CC1CC=C(CCCC(O)(C)C)CC1	210.32	 lyral
v073	81-16-3	O=S(=O)(O)c2c(ccc1ccccc12)N	223.25	 2-amino-1-naphthalenesulfonic acid
v074	50-31-7	Clc1c(C(=O)O)c(Cl)ccc1Cl	225.46	 2,3,6-trichlorobenzoic acid
v075	24851-98-7	O=C(OC)CC1CCC(=O)C1CCCCC	226.32	 cyclopentaneacetic acid, 3-oxo-2-pentyl-, methyl

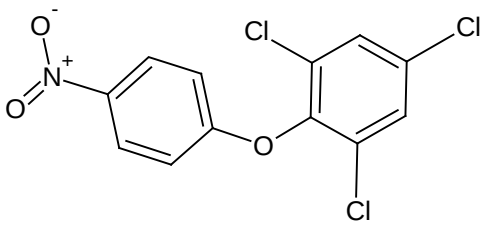
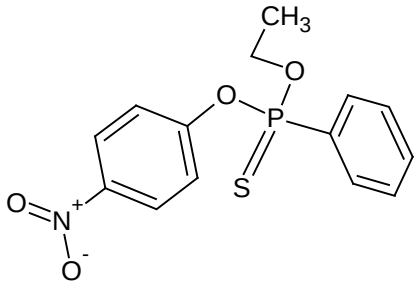
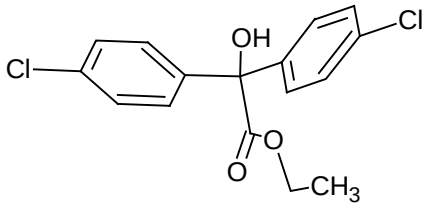
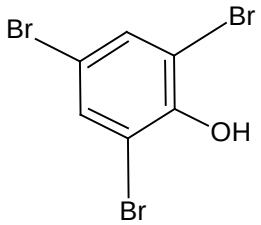
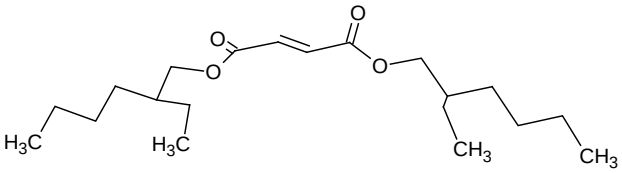
ISBN:978-84-693-4597-9 /DL:T.1010-2010

Annex C.2. List of 93 validation chemicals used in this study (continued).

ID	CAS	SMILES code	MW (g/mol)	3d molecular representation and name
v076	4130-42-1	Oc1c(cc(cc1C(C)(C)C)CC)C(C)(C)C	234.39	 phenol, 2,6-bis(1,1-dimethylethyl)-4-ethyl-
v077	821-38-5	O=C(O)CCCCCCCCCCCCC(=O)O	258.36	 1,12-dodecanedicarboxylic acid
v078	526-78-3	BrC(C(=O)O)C(Br)C(=O)O	275.88	 2,3-dibromosuccinic acid
v079	66-81-9	O=C2NC(=O)CC(CC(O)C1C(=O)C(C)CC(C)C1)C2	281.35	 cycloheximide
v080	124-28-7	N(CCCCCCCCCCCCCCCCCC)(C)C	297.57	 dymanthine
v081	6731-36-8	O(OC1(OOC(C)(C)C)CC(CC(C1)C)(C)C)C(C)(C)C	302.46	 di-tert-butylperoxy-3,3,5-trimethylcyclohexane p

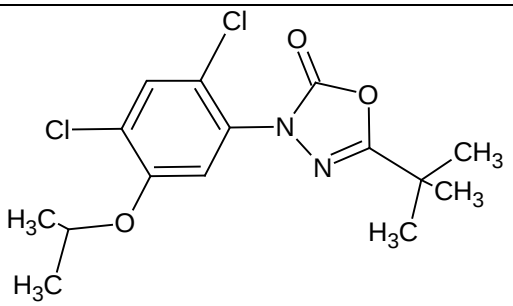
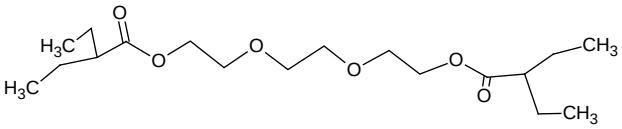
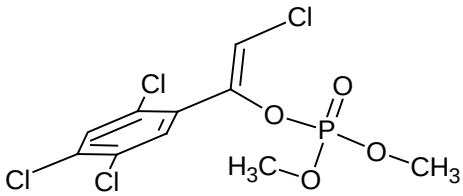
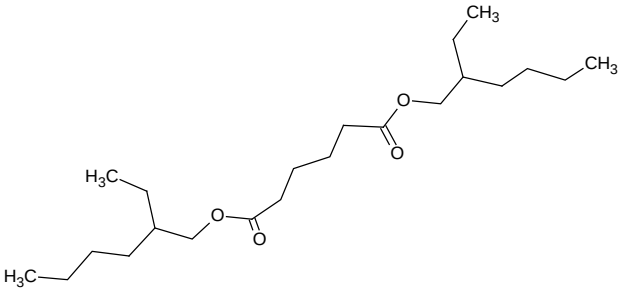
ISBN:978-84-693-4597-9 /DL:T.1010-2010

Annex C.2. List of 93 validation chemicals used in this study (continued).

ID	CAS	SMILES code	MW (g/mol)	3d molecular representation and name
v082	1836-77-7	Clc2cc(Cl)cc(Cl)c2Oc1ccc([N+]([O-])=O)cc1	318.55	 chlornitrofen
v083	2104-64-5	S=P(OCC)(Oc1ccc([N+]([O-])=O)cc1)c2ccccc2	323.31	 epn
v084	510-15-6	Clc1ccc(cc1)C(O)(c2ccc(Cl)cc2)C(=O)OCC	325.19	 chlorobenzilate
v085	118-79-6	Brc1cc(Br)cc(Br)c1O	330.80	 2,4,6-tribromophenol
v086	141-02-6	O=C(OCC(CCCC)CC)\C=C\C(=O)OCC(CC)CCCC	340.51	 2-ethylhexyl fumarate

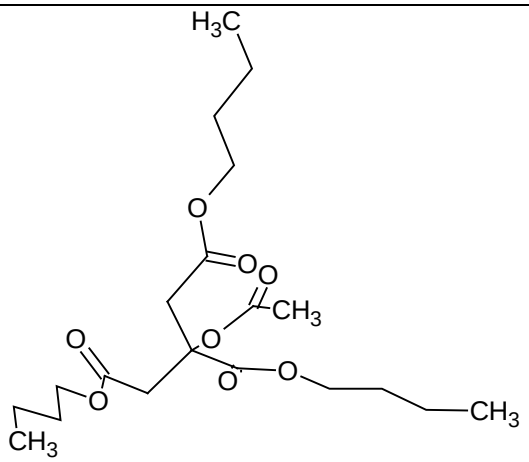
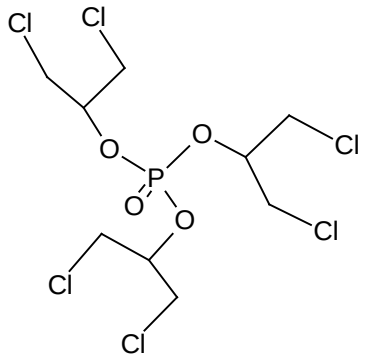
ISBN:978-84-693-4597-9 /DL:T.1010-2010

Annex C.2. List of 93 validation chemicals used in this study (continued).

ID	CAS	SMILES code	MW (g/mol)	3d molecular representation and name
v087	19666-30-9	O=C2OC(=NN2c1c(Cl)cc(Cl)c(OC(C)C)c1)C(C)(C)C	345.23	 oxadiazon
v088	95-08-9	O=C(OCCOCCOCCOC(=O)C(CC)CC)C(CC)CC	346.47	 triethylene glycol bis(2-ethylbutyrate)
v089	961-11-5	Clc1cc(C(OP(=O)(OC)OC)=[C@H]Cl)c(Cl)cc1Cl	365.97	 stirofos
v090	103-23-1	O=C(OCC(CCCC)CC)CCCCC(=O)OCC(CC)CCCC	370.58	 di-2-ethylhexyl adipate
v091	506-52-5	OCCCCCCCCCCCCCCCCCCCCCCCCCC	382.72	 1-hexacosanol

ISBN:978-84-693-4597-9 /DL:T.1010-2010

Annex C.2. List of 93 validation chemicals used in this study (continued).

ID	CAS	SMILES code	MW (g/mol)	3d molecular representation and name
v092	77-90-7	O=C(OCCCC)C(OC(=O)C)(CC(=O)OCCCC)CC(=O)OCCCC	402.49	 acetyl tributyl citrate
v093	13674-87-8	ClCC(OP(=O)(OC(CCl)CCl)OC(CCl)CCl)CCl	430.88	 tris(1,3-dichloroisopropyl) phosphate

ISBN:978-84-693-4597-9 /DL:T.1010-2010