Substance-tailored testing strategies in toxicology : an in silico methodology based on QSAR modeling of toxicological thresholds and Monte Carlo simulations of toxicological testing

Alexandre Pery, Sophie Desmots, Enrico Mombelli

Substance-tailored Testing Strategies in Toxicology: an *in silico* Methodology based on QSAR modeling of toxicological thresholds and Monte Carlo simulations of toxicological testing

Alexandre R.R Péry[a], S. Desmots[b] and E. Mombelli[a*]

INERIS BP2, Parc Technologique Alata 60550 Verneuil-en-Halatte, France

[a]Unité Modèles pour l'Ecotoxicologie et la Toxicologie (METO),

[b]Unité de Toxicologie Expérimentale (TOXI)

[*] Corresponding author. Address : INERIS, BP2, Parc Alata 60550 Verneuil-en-Halatte, France. Tel. +33-3-44-61 81 44 Fax: +33-3-44-55-68-00. E-mail address : enrico.mombelli@ineris.fr

Abstract

The design of toxicological testing strategies aimed at identifying the toxic effects of chemicals without (or with a minimal) recourse to animal experimentation is an important issue for toxicological regulations and for industrial decision-making. This article describes an original approach which enables the design of substance-tailored testing strategies with a specified performance in terms of false-positive and false-negative rates. The outcome of toxicological testing is simulated in a different way than previously published articles on the topic. Indeed, toxicological outcomes are simulated not only as a function of the performance of toxicological tests but also as a function of the physico-chemical properties of chemicals. The required inputs are QSAR predictions for the LOAELs of the toxicological effect of interest and statistical distributions describing the relationship existing between *in vivo* LOAEL values and results from *in vitro* tests.

Our methodology is able to correctly predict the performance of testing strategies designed to analyze the teratogenic effects of two chemicals: Di(2-ethylhexyl)phthalate and Indomethacin. The proposed decision-support methodology can be adapted to any toxicological context as long as a statistical comparison between *in vitro* and *in vivo* results is possible and QSAR models for the toxicological effect of interest can be developed.

Keywords: Decision Analysis, QSAR, teratogenicity, Integrated Testing Strategies, Test batteries, Uncertainty

1. Introduction

In the framework of the European REACH (Registration, Evaluation, and Authorization of Chemicals) regulation it is expected that predictive tools will offer support in the screening and prioritization of chemicals for further toxicological testing (Benfenati, 2007). It is also expected that information obtained via alternatives to conventional animal testing should be considered during the different stages of decision-making provided that its reliability and pertinence can be proved (Grindon et al., 2006).

In this context it is clear that computational models designed to simulate results generated by *in vitro* and *in vivo* toxicological tests could effectively support these expectations thanks to the insight they could provide before any experimental testing is carried out.

The suitability of this computational approach is not limited to a regulatory framework but it can also be a valuable decision-support tool during the selection of candidate chemicals for the design of new products. Usually, several candidates show interesting properties but only a minority among them will maximize the desired property while minimizing unwanted toxicological side effect. It is therefore crucial to weigh the pros and cons relative to options such as the cessation of a research project (because of unmanageable toxicological concerns) or the beginning of a thorough and costly toxicological investigation on few selected chemicals.

Quantitative approaches for decision-making proved to be useful when applied to complex decisions in the toxicological field (Burman and Senn, 2003) but these methods in drug or chemical development are still in their infancy.

Currently, the problem is dealt with by having recourse to testing strategies that integrate different sources of toxicological information, such as data from *in vivo*, *in vitro* and *in silico*, approaches within a sound decision-making framework (Grindon et al. 2006).

This overall approach can permit the toxicological characterization of chemicals with a minimal recourse to animal testing while ensuring a better protection of human health and the environment thanks to an efficient prioritization of chemicals for further toxicological testing. However, it is important noting that, during these stepwise procedures, information obtained at a given point of the strategy is not used to optimize the subsequent evaluation steps. In our opinion, an optimal tiered system would be a system in which the decision to be made after the completion of a step would not only be a choice between “stop testing” or “continue testing” but also a choice aimed at the rational selection of the test to be adopted for the subsequent evaluation.

A theoretical basis for tiered tests has already been proposed based on a classical decision-theoretical framework (Hansson and Ruden, 2007). It consists in utility maximization, this utility being a function of the values of true positives and true negatives and the disvalues of false negatives and false positives. The probability of true positives, true negatives, false positives and false negatives is assumed to be test dependent.

In contrast, the methodology we propose in this paper computes these probabilities as a function of both the performance of toxicological tests and the physico-chemical properties of the chemicals to be analyzed. The rational for this approach is that the statistical simulation of toxicological testing can be carried out with increased certainty if a QSAR prediction can be obtained for a chemical of interest as we will describe in detail in section ****.,

Therefore, according to our methodology, an effective simulation of toxicological test results requires the quantification of the ability for a given test to detect toxicity as a function of the *a priori* probability for a chemical to be toxic.

An essential part of our approach relies, on the simulation of *in vitro* and *in vivo* test results by means of Monte Carlo sampling of two statistical distributions: the distribution that describes how the results yielded by alternative tests compare with results from a toxicological "gold standard" (*in vivo* results obtained via OECD guidelines for this article) and the distribution (obtained by means of QSAR modeling) describing the distribution of "gold standard" Lowest Observed Adverse Effect Levels (LOAELs) for the chemical under investigation. Thanks to this procedure the expected rate of false positives and false negatives is therefore derived for the chemical of interest. During the computational simulations, the uncertainties relative to the results of QSAR modeling and the simulation of toxicological testing are accounted for. At the end of the simulations, the testing strategies whose false positives and false negatives rates are below a specified threshold are selected.

In order to illustrate our methodology, we devised testing strategies with respect to teratogenesis for two molecules: Di(2-ethylhexyl) phthalate (DEHP, used as a plasticizer in polyvinyl chloride plastics) and indomethacin (a non-steroidal anti-inflammatory drug).

The evaluation of offspring for structural abnormalities comprising external, visceral and skeletal examinations is an essential branch of reproductive toxicology (OECD, 2000) and, as of now, three *in vitr*o tests for developmental toxicity testing have been validated by the European Center for the Validation of Alternative Methods (ECVAM) for embryotoxicity screenings. The predictive performance of the tests was deemed to be generally satisfactory. These tests are the embryonic stem-cell test (EST, Genschow et al., 2004), the post-implantation rat whole-embryo culture test (WEC, Piersma et al., 2004) and the micromass test (MMT, Spielmann et al., 2004). Although they are not capable of completely replacing the *in vivo* developmental toxicity tests, their use as screening tests (where positive results would negate the need for further testing) within a testing strategy could reduce the number of animals required (Grindon et al., 2008) and the cost of the assessment of the developmental toxicity of substances requires by the REACH regulation.

2. Materials and Methods

2.1 Selection of chemicals for QSAR modeling

Chemicals were selected from the database Registry of Toxic effects of Chemical Substances (RTCES). It is a database of toxicological information compiled, maintained, and updated by the National institute for Occupational Safety and Health (NIOSH). From RTECS we construct a database in the following way: first select the chemicals having an effect on embryo or fetus development (specific developmental abnormalities following a gavage administration of female rats). Second, we converted lowest published toxic doses (TDLo, the total dose amount administered to the pregnant female) in LOAELs (mg/kg bw/day).

We only selected compounds which have been tested during the pregnancy of animals and the smallest LOAEL was retained if for the same compound, multiple LOAELs were available. LOAELs corresponding to single day exposures were discarded if data on multiple-day exposures were available.

We based our work on LOAEL instead of NOAEL (No observed Adverse Effect Level) mainly because the availability of NOAEL values is rather limited. Secondly, a NOAEL that is not defined with respect to a LOAEL is not helpful. To be useful, NOAEL values have to be characterized as the highest level of exposure at which no adverse effects are detected and it is difficult to ascertain if this threshold has been reached without having defined the level of exposure at which the adverse effects begin to appear (WHO, 2000). Opinions on this subject differ, but the working consensus is that the level of exposure of concern in terms of human health is more adequately described by LOAEL values and this parameter was therefore adopted for our methodology.

2.2 QSAR modeling

The 38 chemicals used to derive a QSAR model are reported in Table 1. LOAEL values were converted into decimal logarithmic values. Chemical structures were built and energy-minimized thanks to the software “Molecular Modeling Pro Plus” licensed by ChemSw inc. Fairfield, CA (USA) and its quantum or molecular-mechanics modules.

Molecules were processed with hydrogens and their stereochemistry coincides with the chirality information displayed by the website ChemIDplus lite (http://chem.sis.nlm.nih.gov/chemidplus/chemidlite.jsp).

The simulations were carried out in a vacuum (i.e. without any water molecule around the organic molecules) and conformational analysis was carried before energy minimization by using the conformational energy routine of the software Molecualr Modeling ProPlus iteratively before the energy minimization with the MM2force field.

After energy minimization, the MOPAC/AM1 semiempirical quantum mechanical calculations were used to generate atom partial charges and further optimize molecular conformations.

2.2.1 Molecular descriptors. The majority of the selected chemicals are likely to induce teratogenic effects trough binding to diverse nuclear receptors but, as previous work showed (Liu et Gramatica, 2007; Giorgi et al., 2009; Wang et al. , 2008), binding to receptors can be adequately described without molecular docking calculations. Therefore, we adopted a receptor-independent approach and adopted the software DRAGON version 5.3 licensed by Talete srl, Milan (Italy) to compute molecular descriptors. Descriptors with a standard deviation less than 5% of the mean were deleted and the remaining 1352 descriptors were used for the subsequent analysis.

2.2.2 Selection of Training and Validation sets. Chemicals were hierarchically clustered according to descriptor similarity by means of the open source software "Cluster 3.0" downloaded from the webpage of the "Laboratory of DNA information analysis" at the Tokyo University (http://bonsai.ims.u-tokyo.ac.jp/~mdehoon/software/cluster/software.htm). Descriptors were autoscaled before analysis. Centered correlation was adopted as similarity metrics when evaluating the clusters. Distances among items were computed by means of an average linkage. Eleven chemicals were chosen from the final branches of the hierarchical tree in order to cover both the descriptor space and the response range (i.e. the dependent variable).

2.2.3 Genetic Algorithms. In order to select the descriptors that better correlates with the LOAEL to be modeled, we used Partial Least Squares regression and Genetic Algorithms (PLS-GA) that are implemented within the "PLS-toolbox" licensed by Eigenvector Research, Inc., WA (USA). The basic parameters used during the selection of the variables were as follows: 120 individuals, mutation rate =0.006 and double cross-over. Maximal number of generations and convergence were kept at a medium level (100% and 50% respectively) to avoid overfitting. The percentage of terms included in the initial variable subsets was set at 10% with a penalty slope for the fitness equal at 0.05 in order to limit the number of variables included in the final model.

The fitness function was implemented as the root-mean-square error of cross validation computed after a PLS regression that was allowed to include a maximum number of 3 latent variables. During the GA the PLS models were validated by means of leave several out cross-validation computed by randomly splitting the dataset into 7 groups. The number of iterations for each cross-validation cycle was set at 10. Leave-several-out (7-groups) cross-validated R^2 (Q^2) values and Response permutation analysis (y-scrambling, 100 permutations) were obtained using the SIMCA-P software version 11.0 licensed by Umetrics, Umea (Sweden).

2.2.4 Probabilistic QSAR modeling. The orthogonal latent variables retained after the GA selection were used to define a Euclidean space that allowed us to derive a QSAR model that yielded predictions in terms of probability for a given chemical to be toxic together with its intrinsic certainty (ranging from 0 to 1) as described in Pery et al. (2009).

In short, training set chemicals were divided into two categories (safe and toxic chemicals) according to a given threshold for LOAEL values. After this categorization, predictive certainty was described by Gaussian functions centered on each training set chemical so that certainty quickly decreased to zero for query chemicals that were far away from the training set. Each training set chemical contributed to predict the probability for the query chemical to be toxic and its contribution to predictive certainty was weighted with respect to the Euclidean distances (measured in the space described by the latent variables) separating the training set chemicals from the query chemical.

The remaining uncertainty (equals to 1 minus the sum of the contributions to certainty from the training set) contributed to the prediction of toxicological probabilities by multiplying the fraction of toxic molecules in the training set. According to this procedure if uncertainty is equal at 1 (i.e. a non-informative QSAR) the toxicological probability coincides with the fraction of toxic molecules within the training set. The estimation of the width σ of the Gaussians representing predictive certainty is obtained by maximizing the toxicological likelihood of the training set. The magnitude of σ is related to the spatial organization of the space described by the latent variables. According to this methodology, when little can be inferred on the basis of chemical similarity, σ value and predictive certainty are very low and the predicted probability for a chemical to be toxic approaches the percentage of toxic chemicals in the training set.

It is important to observe that QSAR modeling in our methodology is also used to construct a statistical distribution that describes how predictive certainty relates to the LOAEL values of training set chemicals. This distribution has a central role during the simulation of testing strategies as described later in the article.

2.3 Toxicological Tests for Teratogenesis

We used data from teratogenic substances evaluated with OECD Guideline 414 and with the three validated *in vitro* assays (EST, WEC and MMT).

The EST is based on the determination of inhibition of differentiation and growth. The embryotoxic potential of chemicals is determined by the evaluation of the inhibition of cardiac muscle differentiation of embryonic stem (ES) cells and the inhibition of growth of ES and 3T3 cells. The EST is performed with permanent cell lines from the mouse.

In the WEC, post implantation rat embryos at early stages of organogenesis are cultured. At day 10 of gestation, pregnant rats are killed and embryos are isolated. Embryos are cultured for 48 hours in culture vessels and subsequently scored. This allows the identification of chemicals that induce embryotoxicity and malformations.

The MM assay is a simple cell culture system, in which development and differentiation of embryonic limb buds cells are studied. Single cell suspensions are prepared from limb buds isolated from 13-days-old embryos. Undifferentiated mesenchyme cells of limb buds will form differentiation of foci of chondrocytes in micromass culture. Teratogenic compounds inhibit the formation of foci and can therefore be detected by a reduced number of foci, or a reduced number of cells within foci. Data which permit to compare *in vivo* and *in vitro* results are presented in Table 2.

2.4 Simulation of testing strategies

In tiered strategies, it is crucial to select test methods for lower tier test that minimize the probability of false negatives, while allowing for some false positives. False negatives can be corrected at higher tiers, whereas false positives will not be corrected since they do not reach higher tiers (Hansson and Ruden, 2007).

The aim of the decision-support methodology we propose is to simulate the final results of testing strategies, by integrating probabilistic QSAR model to the statistical simulation of toxicological testing in order to select only the testing strategies for which the levels of false positives and false negatives are below defined thresholds.

A testing strategy is composed of successive tests to be executed in chronological order. At each step, if the result of the test is positive (i.e. the chemical is flagged as being toxic), then the strategy ends. If the result of the test is negative, then the following step is performed. Once that all the possible tests have been carried out the chemical is considered as not being toxic. In the framework of our methodology, the distinction between toxic and safe chemicals is determined by a comparison between the LOAEL (from *in vivo* rat experiments) values characterizing the chemicals and a given toxicological threshold, which could be, for instance an expected maximum exposure concentration. Chemicals for which LOAEL value is below this threshold are considered as toxic.

The key question that our approach helps to answer is the following: "what battery of tests can assess the toxicity of chemicals of interest while ensuring a desired performance in terms of false positive and negative rates?"

. *The assumptions of the methodology*

The methodology we propose is based on five assumptions

1) LOAEL values referring to the same sex, species and route of administration (rat, female and gavage in our case) adequately describe the toxicity of a molecule.
2) Chemicals can be categorized as "safe or "toxic" according to a given LOAEL threshold.

3) A probability distribution for the LOAEL of a chemical can be derived by QSAR modeling

4) The relative performance of an alternative test with respect to the *in vivo* “gold-standard” (OECD guidelines in our case) can be adequately characterized by statistical distributions that describe the ratio between “gold-standard” LOAEL and results from an *in vitro* tests for a given set of molecules that have been tested by both methods.

5) A testing strategy is composed of successive tests to be executed in chronological order. When the result of a test is positive (i.e. the chemical is flagged as being toxic), the strategy ends. If the result of the test is negative, a following step is simulated.

**.* The underlying logic of the methodology*

A statistical distribution is derived for a given set of molecules that has been tested both *in vivo* and *in vitro*. This distribution describes the ratios of in vivo vs. in vitro results for each molecule composing the aforementioned set and they are an essential input of our methodology. Its role is to enable a comparison of the results yielded by *in vitro* tests and *in vivo* tests that are expressed in different units.

In order to compute false positive/negative rates, the *in vitro* result obtained by means of MC sampling of this distribution (hereafter referred to as *sim-LOAEL*) has to be compared with the LOAEL that an *in vivo* experiments would yield if carried out for the query chemical (hereafter referred to as *ref-LOAEL*). The latter LOAEL value is obtained thanks to the sampling of a probability distribution for the *in vivo* LOAEL of the chemical under investigation that is determined by QSAR modeling.

The simultaneous Monte Carlo sampling of these two distributions enables the simulation of the LOAEL values that an *in vitro* test and the gold-standard test would yield if carried out on the chemical under investigation (more details on the protocol are given in section *** and

***). In conclusion, the *in vitro* result is recognized as being a true/false positive or a true/false negative thanks to a comparison with the *in vivo* LOAEL and a user-defined LOAEL threshold that discriminates between safe and toxic chemicals.

2.5 Simulation of "gold-standard" in vivo testing. Ref-LOAEL values for the query chemicals were simulated on the basis of the predictive certainties that characterize each chemical of the QSAR training set (calibrated with respect to "gold-standard" LOAELs) and that measures the predictive influence of each of them on the final prediction for the query chemical as described at paragraph 2.2.4.

Thanks to this distribution of predictive certainties, the derivation of ref-LOAEL values was then performed in two steps. First, a random number between 0 and 1 is generated. If this number is below the percentage of predictive uncertainty estimated for the query chemical then a LOAEL (referred to as ref-LOAEL$_{\text{first}}$) is equiprobabilistically drawn from the LOAELs characterizing the chemicals that form the QSAR training set. If the number exceeds the percentage of uncertainty, the LOAEL is randomly selected among the LOAELs of the database with a frequency that will be proportional to the predictive certainties (C_t) estimated by QSAR modeling.

In order to better approximate empirical testing we also introduced an extra source of uncertainty. Indeed, performing two times the same test (even a "gold standard") would very likely result in slightly different outcomes. For this reason, the final ref-LOAEL value (ref-LOAEL$_{\text{final}}$) is then sampled from a statistical distribution describing the ratio between two "gold-standard" LOAELs determined for the same chemical during two different runs of the same experimental protocol. To represent this experimental variability for the "gold-standard" test, we selected a loguniform distribution (*i.e.* the logarithm is uniformly distributed) centered on ref-LOAEL$_{\text{first}}$ and with bounds equal at this value multiplied by three (upper bound) and divided by three (lower bound). This choice is related to a gold standard test with selected exposure concentration in geometrical progression with a factor of three.

2.6 Simulation of in vitro testing

We simulated results of the gold standard *in vivo* test and of the three *in vitro* tests described at paragraph *.*. Before carrying out these simulations the distribution of the ratio between results of *in vitro* tests and the “gold-standard” test had to be described for a given set of reference chemicals. The sets of chemicals that were adopted for this pairwise comparison are reported in table 2.

Two different statistical distributions were evaluated for the fitting of the ratios: loguniform and lognormal. The goodness of fit was determined by Chi-square tests and by a method specifically designed to assess uniformity of distribution even in the case where the data seem to be asymmetrically distributed (Afonso and Duarte, 1992). After the distribution was fitted, a ratio was randomly sampled. The sampled ratio was then multiplied by the median of the distributions to obtain “sim-LOAEL since results for *in vitro* tests are not expressed in the same units as “gold-standard” LOAELs”.

When all this information was available, it was then possible to assess if the predicted LOAEL is a true positive, a true negative, a false positive or a false negative, by means of a comparison between ref-LOAEL$_{\mathrm{final}}$, sim-LOAEL and the adopted toxicological threshold. For instance, if we have a LOAEL threshold equal at 50 mg/kg/day, a ref-LOAEL$_{\mathrm{final}}$ equal at 40 mg/kg/day and a sim-LOAEL (e.g. from an *in vitro* test) equal at 80 mg/ml/day, the prediction will be regarded as being a false negative. A flowchart summarizing our methodology is reported in Figure 1.

[FIGURE 1 HERE]

For a given toxicological threshold, we performed 10,000 simulations of LOAELs test results to estimate the rates of false positives and negatives for each test. The false positives and false negatives rates of the battery of tests were calculated as follows. The false negatives rate was equal at the product of the false negative rates of all tests in the battery, whereas the false positives rate of the testing battery was equal at one minus the product of one minus the false positive rate of each test. For instance, for a battery of tests composed by two tests characterized by a false positive and false negative rate equal at 0.1, the testing strategy would have a false negative rate of 0.01 and a false positive rate of 1-(1-0.1)*(1-0.1))= 0.19.

2.7 Case study: DEHP

DEHP, CAS RN 117-81-7) is a high production volume chemical used as a plasticizer in polyvinyl chloride plastics. It is found in a wide variety of consumer products, such as building products, car products, clothing, food packaging, children’s product, and in some medical devices made of polyvinyl chloride (Kavlock et al., 2006).

We carried out our simulation of testing strategies as if no other information than QSAR predictions and statistical distributions describing the performance of toxicological tests were known. The results we obtained were finally compared with available *in vivo* and *in vitro* (WEC test, (Rhee et al., 2002)) data.

We tested four exposure scenarios based on NTP-CERHR Expert panel report on the DEHP (2000). The first one corresponds to the range of general population exposure (3-30 10^{-3} mg/kg/day), the second one (0.6 mg/kg/day) to long term exposure for adult hemodialysis, the third one to the highest possible exposure for medical use (3 mg/kg/day) and a fourth one corresponding to this latter use with a safety factor of 10.

2.8 Case study 2: Indomethacin

Indomethacin (CAS RN 56-86-1) is a non-steroid anti-inflammatory and antipyretic agent (Hart and Boardman, 1963). It works by inhibiting the synthesis of prostaglandins in various tissues. It is known that it can cause constriction of the *ductus arteriosus* with pulmonary hypertension and right ventricular dysfunction in some fetuses and consequently the use for children is considered as dangerous (Lione and Scialli, 1995).

Similarly to what was previously described for DEHP, we simulated testing strategies by integrating information yielded by QSAR modeling and then determined what testing strategy would keep the percentage of false positives and negatives at a reasonable level. This was finally compared with available *in vivo* and *in vitro* (WEC test) data.

We tested four scenarios of exposure. The first one corresponds to the recommended use for an adult (about 25 mg per day, which we translated into 0.5 mg/kg/day). The second one corresponds to the maximum recommended use (200 mg per day, which we translated into 4 mg/kg/day). The third and the fourth one correspond to maximum recommended use with safety factors of 10 and 100.

3. Results

3.1 Statistical distributions for the in vitro tests outputs

Distributions parameters were estimated through mean and standard deviation estimates for the logarithm of the ratio values. For an uniform distribution [-A; A], the standard deviation equals A divided by the square root of 3. Statistical tests did not permit to reject neither lognormal or loguniform distributions but, when comparing the grouping of substances by groups of 5 and the statistical predictions, the loguniform distribution provided a better fit to the data. We obtained, respectively for WEC, MM and EST tests, the following intervals for uniform distribution of the logarithms of the ratio values, [-1.78; 2.58], [-2.42; 2.66] and [-

2.55; 2.73] with result of the *in vivo* test expressed in mg/kg/day, and the result of *in vitro* tests in µg/ml.

3.2 QSAR Modeling

3.2.1 Partial Least Squares analysis

We derived a PLS QSAR model as described in the materials and methods section. The model was characterized by two latent variables and its internal cross-validation (by leave-many-out validation) yielded a cross validated coefficient Q^2_{lso} equal at 0.68 indicating an internally predictive model. The response permutation plot displayed an R^2Y intercept equal at 0.24 and a Q^2Y intercept equal at -0.28 confirming the statistical significance of the model (Eriksson et al. 2003). An inspection of the plot representing the orthogonal latent variables (Fig. 2) revealed that Tetrachlorodibenzodioxin (TCDD) was an outlier whose leverage on the model is very high. Indeed, if this chemical was removed the inclusion of a second latent variable was still useful but Q^2_{lso} fell at 0.40. The probabilistic analysis of the outlier chemical that limited its influence in a rational way (i.e. by defining an applicability domain) will be described in the following paragraph together with the predictions for the external test set of chemicals.

[FIG. 2 HERE]

3.2.2 Probabilistic QSAR modeling

The chemical space defined by the two latent variables of the PLS model allowed the application of the methodology described in Pery et al. (2009). A prerequisite for this methodology is that a toxicological threshold has to be defined in order to categorize molecules as being safe or toxic. For DEHP and Indomethacin the thresholds described in the Materials and Methods section were adopted and such a binarization allowed the estimation of different values for the σ parameter as reported in Table 3 and the prediction of toxicological probabilities and intrinsic certainty for the validation set as reported in Table 4. The values of the sigma parameter showed that the chemical space was structured and its smallest value is 0.24 times the average distance among all the molecules indicating a reasonable predictive quality of the dataset. Moreover, four phthalates populate the immediate neighborhood of DEHP (a phtalate belonging to the validation set) confirming the rational organization of the chemical space.

Only predictions characterized by a certainty equal at 1 were regarded as belonging to the applicability domain of the model. This definition limited the validity of predictions only for query chemicals located in close proximity of the training set molecules. More importantly, it reduced the predictive influence of TCDD only at a region in the space immediately surrounding it (Fig. 2).

Virtual query chemicals located in empty regions of the chemical space (Fig. 2, Table 5) were, as expected, characterized by low certainty values.

Nine out of eleven chemicals initially selected for the validation set had a certainty equal at 1 (including DEHP and Indomethacin) for all the values of the sigma parameter and could therefore be regarded as belonging to the applicability domain of the model. Predictions for DEHP and Indomethacin were always correct and the number of toxic chemicals characterized by a probability lower than 0.5 (i.e. false negatives) and safe chemicals characterized by a probability larger than 0.5 (i.e. false positives) were in the worst cases (Table 4) equal at two. Therefore we considered that the external predictivity of the model was satisfactory for the purpose of this article.

3.2.3 Interpretation of molecular descriptors

The model was characterized by the descriptors reported in table 6 together with their regression coefficients (referring to scaled and centered descriptors and scaled response values). There is not a coefficient that clearly dominates the model. The descriptors try to capture different molecular characteristics (2-D autocorrelations, distribution of electronegativity, topology, shape, distribution of atomic mass) that statistically correlate with LOAEL values without indicating any clear mechanistic rationale for the biological activity. For instance, 2-D autocorrelations (Todeschini and Consonni, 2000) indicate that a positive autocorrelation between atomic masses at a path length of 4 (MATS4m, largest coefficient) and a negative autocorrelation between van der Waals volumes at a path length of 5 (GATS5v) increases the LOAEL. The association of these fragments to a given toxicological mechanism is very difficult because of the structural heterogeneity of the dataset. Similarly, the E3e descriptor that is related to the atom distribution along the third axis for the electronic Sanderson electronegativities (Todeschini and Consonni, 2000) indicates that the three dimensional arrangement of halogens plays a role in modulating the LOAEL. Despite these observations, the three-dimensional arrangement of electronegative atoms for the training set chemicals cannot be directly associated with a unique class of chemical reactivity because the different functional groups that characterize the molecules.

The descriptor RDF050m, that posses the largest negative coefficient, does not provide any clear mechanistic insight either. It simply takes into account the occurrence of some linear dependence between the LOAEL of the chemicals and the molecular distribution of atomic mass calculated at a radius of 0.5 Å, from the geometrical center of each molecule (Todeschini and Consonni, 2000).

WHIM descriptors G2u and G2m (Todeschini and Consonni, 2000) encode information about the symmetry of the molecule and their negative coefficients indicates that the higher the molecular symmetry the smaller the LOAEL values without identifying any mechanistic basis for the model.

The QSAR model developed for this article is therefore purely statistical mainly because of the structural heterogeneity of the training set. Indeed, the training set chemicals could act by means of large number of possible toxicological mechanisms whose thorough description by means of a QSAR model would require a much larger training set. In addition, the fact that LOAEL values cover a broad range of teratogenic effects also renders difficult any possible mechanistic interpretation. Indeed, complex endpoints such as teratogenesis are predictable through QSAR models only to a limited extent mainly because of the high number of potential toxicological mechanisms characterizing the available sets of chemicals (Grindon et al., 2008 ; Enoch et al. 2009). As a consequence, the possibilities of taking decisions on the only basis of QSAR information are markedly reduced, but QSAR models can still be useful in planning further testing, using the method we propose in this paper.

3.3 Simulation of testing strategies

For the two selected molecules, we evaluated all possible strategies whose number of tests ranged from one to four (chosen among the three *in vitro* tests and the reference *in vivo* test). Table 7 presents the results we obtained for each test. The performances of the *in vitro* tests, in terms of false positive and false negative rates were relatively similar.

For DEHP and threshold concentrations from 3 10^{-3} to 0.6 mg/kg/day (-1.52 and -0.22 on a logarithmic scale), the simulation of testing strategies indicated that performing one test (*in vitro* or *in vivo*) was enough to confirm the absence of toxicity, with false negative rate of 0 and false positive rate below 5%, which can be considered as acceptable. In contrast, for exposure concentrations 3 and 30 mg/kg/day, the simulation of testing strategies indicated that only the execution of the *in vivo* test could confirm the absence of toxicity with false negative rates and false positive rates below 10%. Using *in vitro* testing in a test strategy would generate a false positive rate over 20%, which may be unacceptable.

The actual *in vivo* LOAEL for DEHP was measured at 100 mg/kg/day. Moreover, we dispose of results from a WEC test (Rhee et al., 2002), according to which the predicted LOAEL (obtained by applying to the WEC test result the median ratio of the distribution of ratios between WEC test outputs and actual LOAELs) would be 2.5 mg/kg/day.

Basing decision on WEC test output would result in a true negative for exposure concentrations from 3 10^{-3} to 0.6 mg/kg/day, but in a false positive for the two other scenarios. This is coherent with the estimation of statistical performance resulting from our strategies based on QSAR modeling.

For indomethacin and exposure scenarios 0.5 and 4 mg/kg/day (-0.3 and 0.6 on a logarithmic scale), simulation of testing strategies indicated that performing only the *in vivo* test alone can permit to have false positive and negative rates below 0.05. Simulations for a single *in vitro* test indicated a false negative rate of about 0.1 and false positive rates beyond 0.2. For exposure concentration of 40 mg/kg/day (1.60 on a logarithmic scale), performing the *in vivo* test alone can permit to have false positive and negative rates below 0.06. Performing two *in vitro* tests would result in a false positive rate of about 0.16 and false negative rate of about 0.065, which could be reasonable. Performing only one *in vitro* test would increase the false negative rate beyond 0.25 which could be problematic. Performing the three *in vitro* tests would be counterproductive, as the rate of false positive would be over 0.25. For exposure concentration of 400 mg/kg/day, performing two *in vitro* tests would permit to have false positive rate at 0 and false negative rates below 0.03. Performing reference test would lead to false positive rate of 0 and false negative rate around 0.01. Performing only one *in vitro* test would result in a false negative rate above 0.15.

The experimental *in vivo* LOAEL for indomethacin was measured at 1 mg/kg/day and we could therefore compare our results with experimental findings from a WEC test, according to which the estimated LOAEL would be 2.5 mg/kg. Basing decision on WEC test output would consequently yield only true negatives and true positives for the threshold we selected. This is coherent with the conclusions of our simulation study.

As a final remark it is important to point out the added value of simulating testing strategies on the basis of QSAR predictions. Indeed, if toxicological probabilities are simply equal to the percentage of toxic molecules in the dataset (i.e. if nothing could be inferred on the basis of structural similarity) the simulation of testing strategies is less accurate. For instance, for a threshold equal at 0.6 mg/kg/day the simulation of testing strategies for DEHP in the absence of QSAR modeling would indicate that the WEC test is the only valuable *in vitro* test to be performed. Such a result is not accurate. Indeed, the actual LOAEL value for DEHP is 167 times higher than the threshold. This means that all *in vitro* tests would have a rate of false positive below 0.1, as we showed when basing testing strategies simulations on QSAR information.

As far as indomethacin is concerned, the simulation of testing strategies without the support of QSAR results for a threshold of 0.5 mg/kg/day would indicate that performing only a WEC test is enough in order to have a false positive/negative rate below 10%. Interestingly, simulations of test results integrating the information coming from QSAR modeling, indicated that an *in vivo* test is necessary. The latter result is more consistent with experimental results since the adopted threshold is twice as lower than the gold-standard LOAEL (1 mg/kg/day) and carrying out only the WEC test ($LOAEL_{WEC}$ = 2.5 mg/kg/day) would lead to a much higher false positive rate (43%). Moreover, for a threshold equal at 40 mg/kg/day the simulation of testing strategies for Indomethacin without QSAR information would suggest that only an *in vivo* test could yield a false positive/negative rate below 10%. On the other hand, the integration of QSAR information shows that performing two *in vitro* tests could be an acceptable solution. This result is confirmed by the fact that performing a small test battery composed by the WEC test and any other *in vitro* test would prove that Indomethacin is toxic at 40 mg/kg.

4. Discussion

Our methodology enables to integrate QSAR and simulation of testing strategies within a structured workflow. Decision-making could therefore be driven by tests outputs simulated as a function of the physico-chemical properties of the molecule and permitted the selection of the most convenient testing strategy. Our methodology can contribute to the reduction of animal testing requirements by proposing alternatives each time that it is statistically acceptable in terms of toxicological performances.

In their paper about tiered testing strategies, Grindon et al. (2008), propose a toxicological evaluation with the three embryotoxicity assays we use in our study. According to their proposition, if any of these tests gives a positive result, then a decision on Classification and Labeling and/or Risk Assessment should take place and during their discussion they consider that the probability to give unacceptably high levels of false negative data is low. As complementary information to their work, the present study shows that the rate of false positive of the approach by Grindon et al. (2008) can become unacceptable, in particular when all the three *in vitro* tests are performed as we showed, for instance, for indomethacin (false positive rate greater than 25% for an exposure concentration equal at 40 mg/kg/day). Overtesting should be avoided and only one or two *in vitro* tests should be carried out when *in vivo* tests can be replaced.

Since the statistical performances of the *in vitro* tests are relatively similar, when the simulation of testing strategies indicates that one *in vitro* test would be sufficient, the selection of the test to be adopted cannot be based on statistical performances. Therefore, another criterion should be used (i.e. monetary and/or ethical cost). For instance, Grindon et al. (2008) showed that the EST test should be preferred to the two other tests because it is the only one that can be adapted to high-throughput studies without involving the killing of large numbers of pregnant animals.

The QSAR model we derived organized the chemical space in a way that enabled the derivation of reasonable predictions for toxicological probabilities. However, the adopted model was very local and it was therefore crucial to define its applicability domain thanks to the methodology described in Pery et al. (2009). In particular, this methodology, allowed to define the applicability domain of the model for a chemical space that yields reliable predictions by strongly limiting the number of prediction outliers thanks to Gaussians distribution for predictive certainty centered on each training set chemicals

Thanks to this approach it was therefore possible to derive correct probabilistic predictions for the two analyzed molecules while being sure that query chemicals located far from the training set chemicals were characterized by a high predictive uncertainty. As can be seen in Figure 2, the predictive influence of TCDD was limited to the region immediately surrounding the chemical and virtual query chemicals probing empty spaces lying between TCDD and the rest of the training set are characterized by a predictive certainty lower than 1 and therefore outside the applicability domain of the model. (Table 4, Fig. 1).

Our methodology can be adapted to any other toxicological endpoint, provided that a relevant QSAR model can be derived together with available *in vitro* information for several chemicals (about 20) with both reference test (i.e. “gold standard” *in vivo* test) and *in vitro* results. The methodology we presented can also be adapted to toxicological “gold standard” other than tests on animals, such as thresholds derived from clinical trials.

In conclusion, our paper presents an original method coupling QSAR modeling and simulation of toxicologial testing. The selection of a relevant testing strategy is driven not only by test performance, but also by the particular physico-chemical properties of the molecules under investigation. The selection of the optimal testing strategy is then a trade-off between minimization of false statistical results, regulatory considerations, and ethic and monetary costs.

Acknowledgements

This work was supported by the National French Agency for scientific research (ANR) in the framework of the CP2D program (Chimie et Procèdes pour le Développement Durable).

References

Afonso L.A., Duarte P. 1992. Un nouveau test pour la distribution uniforme. Rev Statistiques Appliquée 40, 77-79.

Ahlers J., Stock F., Werschkun B. 2008. Integrated testing and intelligent assessment-new challenges under REACH. Environ Sci Pollut Res 15, 565-572.

Benfenati E. 2007. Predicting toxicity through computers: a changing world. Chem Cent J 1, 32.

Bowden H.C., Tesh J.M., Ross F.W. 1993. Effects of female sex hormones in whole embryo culture. Toxicol. In Vitro 7, 799-802.

Burman C.F., Senn S. 2003. Examples of option values in drug development. Pharm Stat 2, 113-125.

Cicurel L., Schmid B.P. 1988. Postimplantation embryo culture for the assessment of the teratogenic potential and potency of compounds. Experientia 44, 833-840.

Cosenza M.E., Bidanset J. 1995. Effects of chlorpyrifos on neuronal development in rat embryo midbrain micromass cultures. Vet. Hum. Toxicol. 37(2), 118-121.

Cumberland P.F.T., Richold M., Parsons J.F., Pratten M.K. 1994. Further evaluation of a teratogenicity screen unsing an intravitelline injection technique. Toxicol. In Vitro 8(2), 153-166.

Enoch, S.J., Cronin, M.T.D., Madden, J.C., Hewitt, M. 2009. Formation of Structural Categories to allow for Read-Across for Teratogenicity. QSAR Comb. Sci. *in press*.

Genschow E., Spielman H., Scholz G., Pohl I., Seiler A., Clemann N. Bremer S., and Backer K. 2004. Validation of the embryonic stem cell test in the international ECVAM validation study on three *in vitro* embryotoxicity tests. ATLA 32, 209-244.

Golbraikh,A., Tropsha, A. 2002. Beware of q2! J Mol Graph Model. 20, 269-276.

Gramatica, P. 2007. Principles of QSAR models validation: internal and external. QSAR Comb. Sci., 26, 694 – 701.

Gramatica, P. 2004. Evaluation of different statistical approaches to the validation of Quantitative Structure - Activity Relationships. Available online at: http://ecb.jrc.it/DOCUMENTS/QSAR/Report_on_QSAR_validation_methods.pdf,"

Grindon C., Combes R., Cronin M.T.D., Roberts D., Garrod J.F. 2006. Integrated testing strategies for use in the EU REACH system. ATLA 34, 407-427.

Grindon C., Combes R., Cronin M.T.D., Roberts D., Garrod J.F. 2008. Integrated decision-tree testing strategies for developmental and reproductive toxicity with respect to the requirements of the EU REACH legislation. ATLA 36, 65-80.

Hansen D.K., Grafton T.F. 1994. Comparison of dexamethasone-induced embryotoxicity *in vitro* in mouse and rat embryos. Teratog. Carcinog. Mutag. 14, 281-289.

Eriksson, L. Jaworska, J., Worth, A.P., Cronin, M.T., McDowell, R.M. and Gramatica, P. 2003. Methods for reliability and uncertainty assessment and for applicability evaluations of classification and regression-based QSARs. Environ Health Perspect. 111, 1361-1375

Flint O.P., Orton T.C. 1984. An *in vitro* assay for teratogens with cultures of rat embryo midbrain and limb bud cells. Toxicol. Appl. Pharmacol. 76, 383-395.

Hansson S.O., Ruden C. Towards a theory of tiered testing. Regul Toxicol Pharmacol 48, 35-44.

Kavlock R., Barr D., Boekelheide K., Breslin W., Breysse P., Chapin R., Gaido K., Hodgson E., Marcus M., Shea K. and Williams P. 2006. NTP-CERHR Expert Panel update on the reproductive and developmental toxicity of di(2-ethylhexyl) phthalate. Reprod Toxicol 22, 291-299.

Lione A., Scialli A.R. 1995. The developmental toxicity of indomethacin and sulindac. Reprod Toxicol 9, 7-20.

Menegola E., Broccia M.L., Di Renzo F., Giavini E. 2001. Acetaldehyde *in vitro* exposure and apoptosis: a possible mechanism of teratogenesis. Alcohol 23, 35-39.

Morris G.M., Steele C.E. 1976. Comparison of the effects of retinol and retinoic acid on postimplantation rat embryos. Teratology 15, 109-120.

Newall D.R. and Beedles K.E. 1996. The stem-cell test: an *in vitro* assay for teratogenic potential. Results of a blind trial with 25 compounds. Toxicol In Vitro 10, 229-240.

NTP-CERHR Expert Panel report on Di(2-ethylhexyl) phthalate. October, 2000. NTP-CERHR-DEHP-00. Available online at: http://cerhr.niehs.nih.gov/chemicals/dehp/DEHP-final.pdf

OECD. 2000. Guideline for testing of Chemicals. 414 Updated. Prenatal developmental toxicity study.

Pery, A.R.R., Henegar, A., Mombelli E. 2009. Maximum-Likelihood Estimation of Predictive Uncertainty in Probabilistic QSAR modeling. QSAR Comb. Sci. 28, 338-344.

Piersma A.H., Grenschow E., Verhoef A., Spanjersberg M.Q.I., Brown N.A., Brady M., Burns A., Clemann N., Seiler A. and Spielmann H. 2004. Validation of the post-implantation rat whole-embryo culture test in the international ECVAM validation study on three *in vitro* embryotoxicity tests. ATLA 32, 275-307.

Renault J-Y., Meicion C., Cordier A. 1989. Limb bud cell culture for *in vitro* teratogen screening: Validation of an improved assessment method using 51 compounds. Teratog. Carcinog. Mutag. 9, 83-96.

Rhee G.S., Kim S.H., Kim S.S., Sohn K.H., Kwack S.J., Kim B.H., Park K.L. 2002. Comparison of embryotoxicity of ESBO and phthalate esters using an *in vitro* battery system. Toxicol *in vitro* 16, 443-448.

Ritchie H.E., Webster W.S., Eckhoff C., Oakes D.J. 1998. Model predicting the teratogenic potential of retinyl palmitate, using a combined *in vivo*/*in vitro* approach. Teratology 58, 113-123.

Schuurmann, G., Ebert, R. U., Chen, J., Wang, B., Kuhne, R.2008 External validation and prediction employing the predictive squared correlation coefficient test set activity mean vs training set activity mean. J Chem Inf Model. 48, 2140-2145.

Spielmann H., Pohl I., Döring B., Liebsch M., Moldenhauer F. 1997. The embryonic stem cell test, an *in vitro* embryotoxicity test using two permanent cell lines: 3T3 fibroblasts and embryonic stem cells. In Vitro Toxicol. 10(1), 119-127.

Spielmann H., Genschow E., Brown N.A., Piersma A.H., Verhoef A., Spanjersberg M.Q.I., Huuskonen H., Paillard F. and Seiler A. 2004. Validation of the rat limb bud micromass test in the international ECVAM validation study on three *in vitro* embryotoxicity tests. ATLA 32, 245-274.

Todeschini, R., Consonni, V. 2000. Handbook of Molecular Descriptors. Methods and Principles in Medicinal Chemistry (Volume 11). Edited by Mannhold, R. Kubinyi, H. and Timmerman, H. Wiley-VCH Verlag (Germany)

Tropsha, A., Gramatica, P., Gombar, V. K. 2003. The Importance of Being Earnest: Validation is the Absolute Essential for Successful Application and Interpretation of QSPR Models. QSAR Comb. Sci. 22, 69-77.

WHO Regional Publications, European Series, Second Edition. 2000. Air quality guideline for Europe. No 91, 273 pages.

Zur Nieden, N.I., Ruf, L.J., Kempka, G., Hildebrand, H., Ahr, H.J. 2001. Molecular markers in embryonic stem cells. Toxicol *in vitro* 15, 455-461.

Table 1 Chemicals used to develop and test the QSAR model adopted for our investigation. Activities are given as decimal Log of the LOAEL expressed in mg/kg/day. Training set molecules belongs to set 1 and validation set molecules to set 2.

CAS number	Name	LOG LOAEL	Set
119-61-9	Benzophenone	2.00	1
131-70-4	Monobutyl phthalate	2.40	1
138261-41-3	Imidacloprid	1.93	1
1746-01-6	Tetrachlorodibenzodioxin	-3.30	1
17804-35-2	Benomyl	1.75	1
1836-75-5	Nitrofen	0.80	1
22071-15-4	Ketoprofen	0.00	1
24602-86-6	Tridemorph	0.00	1
26171-23-3	Tolmetin	1.31	1
26761-40-0	Diisodecyl phthalate	3.00	1
28553-12-0	Diisononyl phthalate	2.88	1
35065-27-1	2,4,5,2',4',5'-Hexachlorobiphenyl	1.30	1
50-06-6	Phenobarbital	1.60	1
50471-44-8	Vinclozolin	2.00	1
5104-49-4	Flurbiprofen	1.00	1
52315-07-8	Cypermethrin	1.70	1
53164-05-9	Acemetacin	-0.30	1
58-08-2	Caffeine	0.78	1
58-89-9	Lindane	1.30	1
68-22-4	Norethindrone	1.40	1
83-05-6	2,2-Bis(4-chlorophenyl)acetic acid (DDA)	1.70	1
83-79-4	Rotenone	0.70	1
84-69-5	Diisobutyl phthalate	2.70	1
84-74-2	Dibutyl phthalate	2.70	1
87237-48-7	Haloxyfop-etotyl	1.00	1
87-51-4	Indoleacetic acid	1.70	1
94361-06-5	Cyproconazole	2.00	1
107534-96-3	Tebuconazole	1.78	2
117-81-7	Diethylhexyl phthalate	2.00	2
13710-19-5	Tolfenamic acid	0.00	2
32774-16-6	3,4,5,3',4',5'-Hexachlorobiphenyl	-1.52	2
53-86-1	Indomethacin	0.00	2
54-31-9	Furosemide	1.70	2
58138-08-2	Tridiphane	2.00	2
69806-50-4	Fluazifop-butyl	2.30	2
76-44-8	Heptachlor	0.48	2
84-66-2	Diethyl Phthalate	3.39	2
85-68-7	Butyl benzyl phthalate	2.43	2

Table 2. Toxicological data that allowed the derivation of statistical distributions describing the ratio between *in vitro* and *in vivo* results. Such distributions are necessary in order to make possible a comparison between results yielded by different experimental systems which are expressed in different units.

CAS Number	Name	LOAEL	*In vitro* effect level	*In vitro* test	Reference
79-81-2	retinol palmitate	206 mg/kg/day	25 µg/ml	WEC	Ritchie et al., 1998.
4759-48-2	13-cis-retinoic acid	50 mg/kg/day	0,5 µg/ml	WEC	Ritchie et al., 1998.
68-26-8	vitamin A	37,5 mg/kg/day	375 ng/ml	WEC	Ritchie et al., 1998.
302-79-4	all-trans retinoic acid	5 mg/kg/day	0,5 µg/ml	WEC	Morris et al., 1976.
50-78-2	aspirin	250 mg/kg/day	10 µg/ml	WEC	Cicurel et al., 1988.
50-02-2	dexamethasone	0,08 mg/kg/day	5 µg/ml	WEC	Hansen et al., 1994
53-86-1	methazine	1 mg/kg/day	1 µg/ml	WEC	Cuberland et al., 1994.
147-24-0	diphenhydramine	100 mg/kg/day	1 µg/ml	WEC	Cuberland et al., 1994.
58-08-2	caffeine	6 mg/kg/day	100 µg/ml	WEC	Cicurel et al., 1988
305-03-3	chlorambucil	6 mg/kg/day	10 µg/ml	WEC	Cicurel et al., 1988
66-81-9	cycloheximide	1 mg/kg/day	0,03 µg/ml	WEC	Cicurel et al., 1988
50-02-2	dexamethasone	0,08 mg/kg/day	270 µg/ml	WEC	Cicurel et al., 1988
439-14-5	diazepam	500 mg/kg/day	100 µg/ml	WEC	Cicurel et al., 1988
57-41-0	diphenylhydantoin	100 mg/kg/day	100 µg/ml	WEC	Cicurel et al., 1988
96-45-7	N,N'-ethylene thiourea	30 mg/kg/day	100 µg/ml	WEC	Cicurel et al., 1988
51-21-8	fluorouracil	5 mg/kg/day	0,6 µg/ml	WEC	Cicurel et al., 1988
50-35-1	thalidomide	100 mg/kg/day	1000 µg/ml	WEC	Cicurel et al., 1988
58-55-9	theophylline	258,6 mg/kg/day	100 µg/ml	WEC	Cicurel et al., 1988
75-07-0	acetaldehyde	240 mg/kg/day	45 µg/ml	WEC	Menegola et al., 2001.
84-74-2	dibutylphtalate	500 mg/kg/day	10 µg/ml	WEC	Rhee et al., 2002.
50-28-2	estradiol	0,225 mg/kg/day	30 ug/ml	WEC	Bowden et al., 1993.
4759-48-2	13-cis-retinoic acid	50 mg/kg/day	0,08 µg/ml	MM	Renault et al., 1989.
127-07-1	hydroxycarbamide	200 mg/kg/day	14,3 µg/ml	MM	Renault et al., 1989.
50-78-2	aspirin	250 mg/kg/day	1436 µg/ml	MM	Renault et al., 1989
50-02-2	dexamethasone	0,08 mg/kg/day	30,5 µg/ml	MM	Renault et al., 1989
53-86-1	methazine	1 mg/kg/day	4 µg/ml	MM	Flint et al., 1984.
147-24-0	diphenhydramine	100 mg/kg/day	48,8 µg/ml	MM	Renault et al., 1989
84-74-2	dibutylphtalate	500 mg/kg/day	27,47 µg/ml	MM	Rhee et al., 2002.
85-68-7	benzylbutyl hthalate	270 mg/kg/day	412,24 µg/ml	MM	Rhee et al., 2002
305-03-3	chlorambucil	6 mg/kg/day	2,6 µg/ml	MM	Renault et al., 1989
56-75-7	chloramphenicol	2500 mg/kg/day	230 µg/ml	MM	Flint et al., 1984.
2921-88-2	chlorpyrifos	3 mg/kg/day	16 µg/ml	MM	Cosenza et al., 1995.
50-18-0	cyclophosphamide	10 mg/kg/day	325 µg/ml	MM	Flint et al., 1984.
439-14-5	diazepam	500 mg/kg/day	150 µg/ml	MM	Flint et al., 1984.
60-00-4	Ethylenediaminetetra acetic acid	954 mg/kg/day	2,8 µg/ml	MM	Flint et al., 1984.
50-00-0	Formaldehyde	8 mg/kg/day	5,3 µg/ml	MM	Renault et al., 1989.
302-79-4	all-trans retinoic acid	5 mg/kg/day	0,000105 µg/ml	EST	Spielmann et al., 1997.
127-07-1	hydroxycarbamide	200 mg/kg/day	1,7 µg/ml	EST	Spielmann et al., 1997.
50-78-2	aspirin	250 mg/kg/day	248 µg/ml	EST	Spielmann et al., 1997.
50-02-2	dexamethasone	30,5 µg/ml	18,3 µg/ml	EST	Spielmann et al., 1997.
53-86-1	methazine	1 µg/ml	66 µg/ml	EST	Spielmann et al., 1997.
147-24-0	diphenhydramine	1 µg/ml	6,7 µg/ml	EST	Spielmann et al., 1997.
58-08-2	caffeine	100 µg/ml	185 µg/ml	EST	Spielmann et al., 1997.
50-18-0	cyclophosphamide	325 µg/ml	21 µg/ml	EST	Newall et al., 1996.
55-98-1	busulfan	100 mg/kg/day	4,6 µg/ml	EST	Spielmann et al., 1997.
51-21-8	fluorouracil	0,6 µg/ml	0,029 µg/ml	EST	Spielmann et al., 1997.
50-35-1	thalidomide	1000 µg/ml	67 µg/ml	EST	Zur Nieden et al., 2001.
57-41-0	phenytoin	100 µg/ml	5,8 µg/ml	EST	Newall et al., 1996.

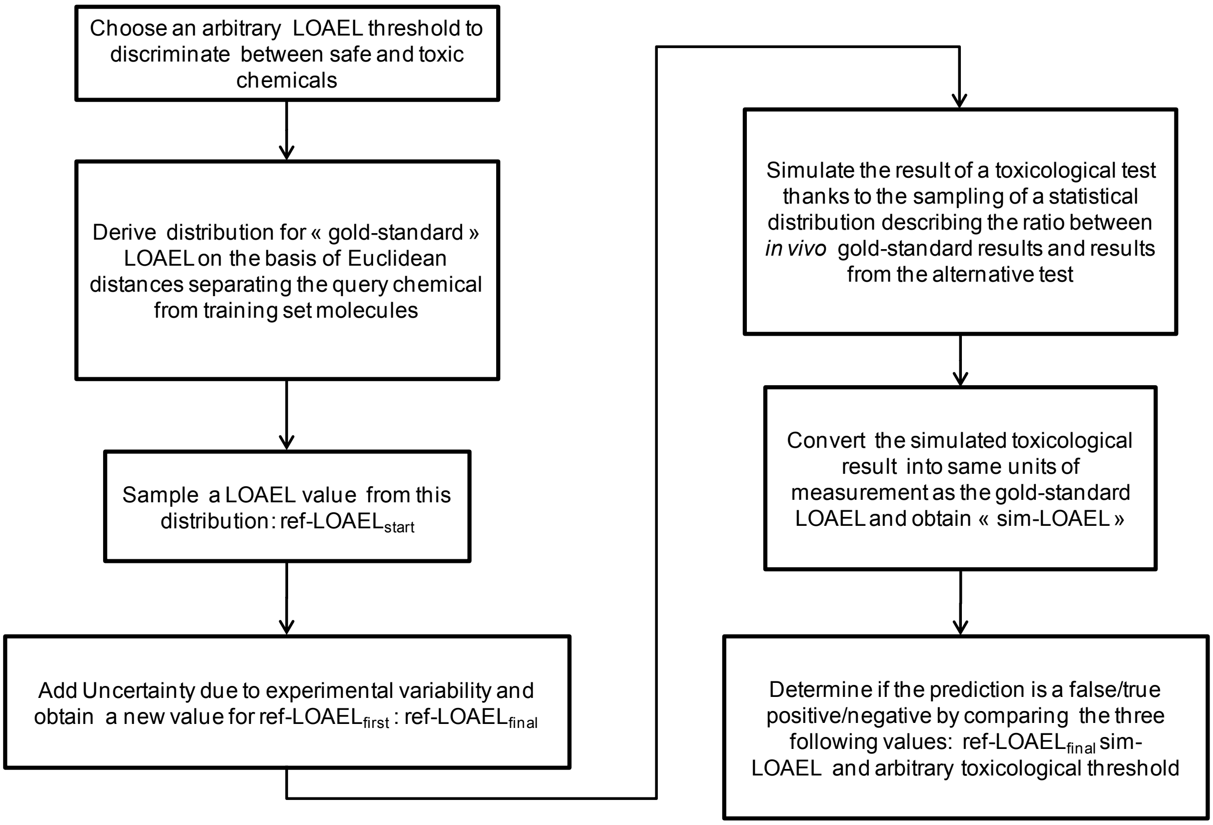
Choose an arbitrary LOAEL threshold to discriminate between safe and toxic chemicals
Derive distribution for « gold-standard » LOAEL on the basis of Euclidean distances separating the query chemical from training set molecules
Sample a LOAEL value from this distribution: ref-LOAEL$_{start}$
Add Uncertainty due to experimental variability and obtain a new value for ref-LOAEL$_{first}$: ref-LOAEL$_{final}$
Simulate the result of a toxicological test thanks to the sampling of a statistical distribution describing the ratio between *in vivo* gold-standard results and results from the alternative test
Convert the simulated toxicological result into same units of measurement as the gold-standard LOAEL and obtain « sim-LOAEL »
Determine if the prediction is a false/true positive/negative by comparing the three following values: ref-LOAEL$_{final}$ sim-LOAEL and arbitrary toxicological threshold

Fig. 1 Flowchart representing the essential steps of the simulation of a battery of toxicological tests.

Fig. 2. Plot of the orthogonal latent variables LV1 and LV2. Training set chemicals are represented by black diamonds, DEHP and Indomethacin are represented as a white square and a white circle respectively. The isolated point on the left represents TCDD and the dashed circle represents σ=1.24. A query chemical located at a distance equal at σ (i.e. on the dashed lines) from TCDD will be characterized by a predictive certainty equal at 1/e ≈ 0.37 where e is the Neperian number. The grey dots represent virtual query chemical probing empty spaces: for all of them the predictive certainty is lower than 1. The dotted ellipse represents the 95% confidence region of the model according to Hotelling's T2.

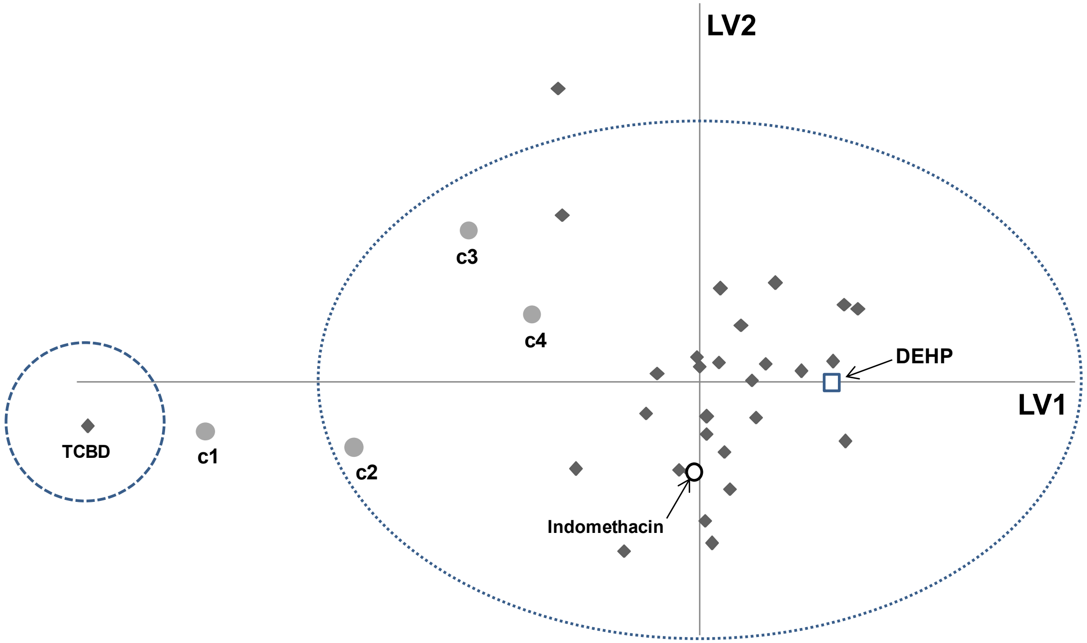

Table 3 LOAEL thresholds and estimated sigma values for the two DEHP and Indomethacin.

DEHP		**Indomethacin**	
Threshold for Log LOAEL	σ	Threshold for Log LOAEL	σ
-1.52	2.7	-0.3	0.7
-0.22	0.7	0.6	1.97
0.48	1.97	1.6	1.24
1.48	1.24	2.6	0.77

Table 4 Toxicological probabilities for the test set corresponding to the different scenarios for DEHP and Indomethacin. Chemicals for which the LOAEL is below the threshold (i.e. toxic chemicals) are indicated in bold.

CAS	Probability	Certainty	CAS	Probability	Certainty
DEHP Threshold=-1.52			Indomethacin Threshold=-0.3		
107534-96-3	0	1	107534-96-3	0	1
117-81-7	0	1	117-81-7	0	1
13710-19-5	0	1	13710-19-5	0	1
32774-16-6	0.04	0.14	**32774-16-6**	**0.07**	**0**
53-86-1	0	1	53-86-1	0	1
54-31-9	0	1	54-31-9	0	1
58138-08-2	0	1	58138-08-2	0	1
69806-50-4	0	1	69806-50-4	0	1
76-44-8	0	1	76-44-8	0.04	0.48
84-66-2	0	1	84-66-2	0	1
85-68-7	0	1	85-68-7	0	1
DEHP Threshold=-0.22			Indomethacin Threshold=0.6		
107534-96-3	0	1	107534-96-3	0.04	1
117-81-7	0	1	117-81-7	0.05	1
13710-19-5	0	1	**13710-19-5**	**0.12**	**1**
03274-16-6	**0.07**	**0**	**03274-16-6**	**0.15**	**0**
53-86-1	0	1	**53-86-1**	**0.24**	**1**
54-31-9	0	1	54-31-9	0.16	1
58138-08-2	0	1	58138-08-2	0.01	1
69806-50-4	0	1	69806-50-4	0.07	1
76-44-8	0.04	0.48	76-44-8	0.01	0.93
84-66-2	0	1	84-66-2	0.03	1
85-68-7	0	1	85-68-7	0.06	1
DEHP Threshold=0.48			Indomethacin Threshold=1.6		
107534-96-3	0.04	1	107534-96-3	0.18	1
117-81-7	0.05	1	117-81-7	0.10	1
13710-19-5	**0.12**	**1**	**13710-19-5**	**0.56**	**1**
32774-16-6	**0.15**	**0**	**32774-16-6**	**0.48**	**0**
53-86-1	**0.24**	**1**	**53-86-1**	**0.76**	**1**
54-31-9	0.16	1	54-31-9	0.73	1
58138-08-2	0.01	1	58138-08-2	0.1	1
69806-50-4	0.07	1	69806-50-4	0.3	1
76-44-8	0.01	0.93	**76-44-8**	**0.91**	**0.83**
84-66-2	0.03	1	84-66-2	0.19	1
85-68-7	0.06	1	85-68-7	0.22	1
DEHP Threshold=1.48			Indomethacin Threshold=2.6		
107534-96-3	0.18	1	**107534-96-3**	**0.88**	**1**
117-81-7	0.10	1	**117-81-7**	**0.42**	**1**
13710-19-5	**0.56**	**1**	**13710-19-5**	**1**	**1**
32774-16-6	**0.48**	**0**	**32774-16-6**	**0.85**	**0**
53-86-1	**0.76**	**1**	**53-86-1**	**1**	**1**
54-31-9	0.73	1	**54-31-9**	**1**	**1**
58138-08-2	0.1	1	**58138-08-2**	**1**	**1**
69806-50-4	0.3	1	**69806-50-4**	**1**	**1**
76-44-8	**0.91**	**0.83**	**76-44-8**	**0.93**	**0.55**
84-66-2	0.19	1	84-66-2	0.99	1
85-68-7	0.22	1	**85-68-7**	**0.82**	**1**

Table 5 Predictive certainty for virtual query chemicals computed for $\sigma = 1.24$. All the certainties are lower than 1.

Chemical	**Certainty**
C1	< 0.1
C2	< 0.1
C3	0.23
C5	0.27

Table 6 Retained descriptors and regression coefficients of the retained QSAR model. Coefficients refer to scaled and centered descriptors and scaled response values.

Descriptor	Category	Coefficient
MATS4m	2D-autocorrelation	0.19377
E3e	WHIM	0.17081
JhetZ	Topological	0.158696
RDF050m	RDF	-0.149818
G2m	WHIM	-0.144327
GATS5v	2D-autocorrelation	-0.138099
G2u	WHIM	-0.135932
RDF080m	RDF	-0.11942
G(O..Cl)	Geometrical	-0.114555
GATS4v	2D-autocorrelation	-0.110849
Mor31u	3d-Morse	0.10955
E1v	WHIM	-0.0825062
Mor19v	3d-Morse	0.0607
G1m	WHIM	-0.0532846
R1e	GETAWAY	0.0464452
BELe7	Burden Eigenvalues	-0.0253866
H3m	GETAWAY	-0.0112353

Table 7 Expected false positive and false negative rates for DEHP and Indomethacin for different toxicological thresholds.

	In vivo		***In vitro***					
			WEC		**MM**		**EST**	
DEHP [mg/kg/day]	False positive	False negative	False positive	False negative	False positive	False negative	False positive	False negative
3-30 10^{-3}	0	0	0	0.01	0	0.02	0	0.02
0.6	0	0	0	0.02	0	0.05	0	0.05
3	0.003	0.002	0.2	0.01	0.23	0.01	0.24	0.01
30	0.06	0.04	0.24	0.07	0.27	0.07	0.27	0.07
Indomethacin [mg/kg]	False positive	False negative	False positive	False negative	False positive	False negative	False positive	False negative
0.5	0.04	0.02	0.21	0.03	0.24	0.03	0.25	0.03
4	0.04	0.03	0.22	0.10	0.24	0.11	0.24	0.11
40	0.04	0.06	0.09	0.24	0.09	0.26	0.09	0.26
400	0	0.001	0	0.13	0	0.17	0	0.17