# Integrating Non-Animal Test Information into an Adaptive Testing Strategy – Skin Sensitization Proof of Concept Case

*Joanna Jaworska [1], Artsiom Harol [1], Petra S. Kern [1], and G. Frank Gerberick [2]*

[1]Procter & Gamble Eurocor, Strombeek-Bever, Belgium; [2]The Procter & Gamble Company, Miami Valley Innovation Center, Cincinnati, OH, USA

**Summary**

*There is an urgent need to develop data integration and testing strategy frameworks allowing interpretation of results from animal alternative test batteries. To this end, we developed a Bayesian Network Integrated Testing Strategy (BN ITS) with the goal to estimate skin sensitization hazard as a test case of previously developed concepts (Jaworska et al., 2010). The BN ITS combines in silico, in chemico, and in vitro data related to skin penetration, peptide reactivity, and dendritic cell activation, and guides testing strategy by Value of Information (VoI). The approach offers novel insights into testing strategies: there is no one best testing strategy, but the optimal sequence of tests depends on information at hand, and is chemical-specific. Thus, a single generic set of tests as a replacement strategy is unlikely to be most effective. BN ITS offers the possibility of evaluating the impact of generating additional data on the target information uncertainty reduction before testing is commenced.*

*Keywords: integrated testing strategy, skin sensitization, Bayesian network, LLNA potency*

## 1 Introduction

There is a pressing need for non-animal test methods, driven by the forthcoming ban on animal testing for cosmetic ingredients in Europe, the large number of tests potentially required to fill in data gaps for the REACH legislation, and animal welfare concerns. Skin sensitization was identified as the hazard endpoint for which most animal tests would need to be conducted and which required a large number of animals (van der Jagt et al., 2004).

The individual steps involved in the skin sensitization process are illustrated in File A (Supplementary Data at www.altex-edition.org). In short, the chemical must penetrate the skin to react with endogenous proteins, either directly or after activation through enzymatic or oxidative processes. Next, epidermal Langerhans cells (LC) and immature dendritic cells (DC) take up and process haptenated proteins. LC cells mature into antigen presenting cells, which after migration to the lymph nodes present haptenized protein fragments to T-cells.

Many research groups are working on the development of alternative tests for skin sensitization (Vandebriel and van Loveren, 2010; Aeby et al., 2010). As a chemical's reactivity towards proteins is deemed a key determining factor in its ability to act as a skin sensitizer, a lot of research has focused on *in chemico* measurements of reactivity with model nucleophiles. Various nucleophiles are used and experiments are either done by direct peptide reactivity measurements (Gerberick et al., 2008), semi-kinetic (Aptula et al., 2006) or more complex kinetics (Aleksic et al., 2009). In addition, the induction of an antioxidant response element (ARE) dependent gene activity in a recombinant cell line (Natsch et al., 2008) can be used to indirectly characterize reactivity. To further elucidate the skin sensitization induction process, various measures of dendritic cell activation are considered. Recent advances in the *in vitro* generation of immature dendritic cells and the availability of cell line surrogates with various DC-like characteristics has led to the development of *in vitro* tests based on the measurement of various cell surface markers or secretion of cytokines modulated upon exposure to chemicals (Aeby et al., 2010; Ryan et al., 2005; Lambrechts et al., 2010). Numerous attempts were also made to predict *in silico* the skin sensitization potential *in vivo* (Roberts et al., 2007; Patlewicz et al., 2007, 2008; Patlewicz and Worth, 2008; Karlberg et al., 2008).

Many authors share the opinion that a single test method cannot replace the *in vivo* skin sensitization animal testing; however, it remains open which tests are actually needed. To address this point of view, several data integration frameworks have been developed. Jowsey et al. (2006) proposed a conceptual scoring system based on Structure Activity Relationships (SAR), penetration, peptide reactivity, and dendritic and T-cell activation

Received March 17, 2011; accepted in revised form June 9, 2011.

to obtain a prediction of skin sensitization potential. Following Jowsey et al., Maxwell and Mackay (2008) developed a mechanistic model of skin allergy using a systems biology approach. While the model provided a valuable mechanistic hypothesis, its use in risk assessment is limited due to its need for a large amount of experimental data. Natsch et al. (2009) combined two *in vitro* measurements with *in silico* predictions into a yes/no classification model. Recently, Nukada et al. (2010) combined data from a dendritic cell activation assay with peptide reactivity data using a rule-based scoring system.

Basketter and Kimber (2009) reviewed the current state of the art of *in vitro* alternatives for skin sensitization and updated the Jowsey et al. (2006) proposal. In parallel, several groups are pursuing a different route to explain skin sensitization effects *in vivo* (e.g., Roberts and Patlewicz, 2009). Roberts et al. (2008) explore the concept of a molecular initiating event that is represented by covalent chemical binding with "protein" and focus on interpretation of this step to explain sensitization, considering cell-based assays only for stages downstream of the reactivity step. There is no consensus on the relative merits of different proposed frameworks.

A data integration framework is already a goal on its own and useful in risk assessment (Maxwell and Mackay, 2008). In this paper we are pursuing the closely related, but broader in scope, goal of constructing an Integrated Testing Strategy (ITS). ITS requires developing a data integration framework allowing for the synthesis of information in a cumulative manner and guiding testing in such a way that information gain in a testing sequence is maximized. In narrative terms, ITS can be described as combinations of tests in a battery covering relevant mechanistic steps and organized in a logical, hypothesis-driven decision scheme, which is required to make efficient use of generated data and to gain a comprehensive information basis for making decisions regarding hazard or risk (Jaworska and Hoffmann, 2010). The importance the information target formulation is discussed in Jaworska and Hoffmann (2010). ITS for several human health and environmental safety endpoints were outlined in the REACH Technical Guidance Document (TGD). Grindon et al. (2007) further customized the REACH ITS for skin sensitization potential. Analysis of existing ITS approaches towards the objective to optimize chemical testing can be found in Jaworska et al. (2010). In short, these authors identified the following shortcomings: 1) The use of flow charts as the ITS' underlying structure may lead to inconsistent decisions; 2) There is no underlying methodology to derive consistent and transparent inferences about the information target (e.g., a well-defined toxicological endpoint serving to address hazard), taking into account all available evidence and its interdependence; 3) Moreover, there is no objective guidance, or only purely expert-driven guidance, regarding the choice of the subsequent tests that would maximize information gain in predicting the information target.

The aim of the study was to put the previously developed concepts on data integration and ITS (Jaworska et al., 2010; Jaworska and Hoffmann, 2010), into practice and evaluate their utility in a proof-of-concept case. For this test case we chose skin sensitization potential as the ITS target because of a relatively good mechanistic understanding of the underlying steps involved in skin sensitization induction, the availability of several non-animal tests characterizing these steps, as well as the overall importance of skin sensitization in the context of safety evaluation (Basketter and Kimber, 2009). To this end, we developed an ITS for skin sensitization potential with the specific goal of estimating potency in the mouse LLNA (Local Lymph Node Assay).

## 2 Materials and methods

### 2.1 Dataset

Representative assays for each of the steps in the induction of skin sensitization (File A in supplementary data; www.altex-edition.org) except the T-cell recognition step, for which no assay data are available, were chosen as inputs to the integrated testing strategy. The data set consisted of responses of 142 chemicals in the following tests: epidermal bioavailability data, peptide reactivity assays, dendritic cell activation, and TIMES predictions. The complete data set, together with LLNA data, is available in File B (supplementary data; www.altex-edition.org).

*Information target: Murine Local Lymph Node Assay*

LLNA data (OECD testing guideline 429, http://www.oecd-ilibrary.org/content/book/9789264071100-en) were compiled from multiple sources, which included the published literature (Gerberick et al., 2005; Kern et al., 2010) and previously unpublished data from several laboratories. The chemicals were chosen based on quality of LLNA studies and availability of data. The data comprise a variety of chemical classes, including fragrances, preservatives, dyes, dye-precursors, halogenated alkanes, and solvents, and cover a wide range of physico-chemical properties. Out of the 142 chemicals, 37 were non-sensitizers and 105 were sensitizers. The four-way classification scheme – non-sensitizing (NS), weak (W), moderate (M), and strong (S) (which also included extreme sensitizers) (Kimber et al., 2003) – was used to characterize potency in the LLNA.

*Epidermal bioavailability*

Finite and infinite dose variables were considered. Finite dose-related variables were calculated using a transdermal transport model (Kasting et al., 2008). During simulation of a single exposure, free and total maximum mid-epidermal concentrations, *C*free and *C*max ($\mu$mol/cm$^3$), as well as % systemically absorbed, were calculated. The infinite dose related variables, Kow and Kp (permeability coefficient), were estimated using KOWWIN and EPIWIN (v.1.6.7) software. Bioavailability data were generated for two exposures (V1=1 mg/cm$^2$), V2=10 mg/cm$^2$) that cover a range of chemical exposures relevant to typical consumer products.

*Direct Peptide Reactivity Assay (DPRA)*

Peptide reactivity data were generated using a method to measure reactivity of a test chemical with model hepta-peptides containing lysine (Lys) or cysteine (Cys) (Gerberick et al., 2004). Peptide reactivity is reported as a percent depletion based on the decrease in free peptide concentration in the sample.

*Cell-based ARE assay*
ARE data were taken from Natsch et al. (2009). AREc32 is a stable cell line derived from the human MCF7 breast carcinoma cell line (Wang et al., 2006). The average $I_{max}$ (maximal induction of gene activity reported as fold-induction vs. untreated cells) and the average concentration inducing 1.5-fold enhanced gene activity (EC 1.5) are determined. For the analyses in this paper EC 1.5 values were used and reported as ARE luciferase (Luc). (Data reported as >1000 are listed with 2000 $\mu$M in our dataset.)

*Dendritic cell activation*
The data were generated using the U937 Activation Test, an *in vitro* cell-based skin sensitization screening test which uses the human myeloid cell line U937 (Python et al., 2007). Cell surface CD86 expression and IL-8 secretion are measured as activation markers.

*TIMES*
The TIMES software (V.2.25.7) (Dimitrov et al., 2005) was run to predict the skin sensitization potential. Predictions based on the parent molecule (TIMES-P), as well as considering potential skin metabolism (TIMES-M), were investigated in the study.

Out of the 142 x12 records, 14.2% were missing. Specifically, only 70 chemicals had dendritic cell activation data (i.e. CD86 and IL-8 data), thus 72*2/284=51% records were missing. For Reactivity data 93; for Luc 75, for Cys 8 and for Lys 10 records were missing. There were no missing records for Bioavailability as all data were generated *in silico*. Only 45 out of 142 chemicals had complete data records for all tests. The abbreviated input variables' names and their units are presented in Table 1.

## 2.2 Bayesian Network construction

Prior to the network construction, the tests considered as inputs were mapped onto a mechanistic scheme of the skin sensitization induction process as described in Basketter and Kimber (2009) and described in File A (supplementary data; www.altex-edition.org). Next, the structure of the BN and the probabilistic relationships between the variables were extracted directly from the data. The network development consisted of the following steps: 1) Transforming the training set into discrete variables; 2) Latent variables structure learning; 3) Missing data imputation; 4) Final model structure learning; 5) Elucidation of the conditional probabilities, i.e. parameter learning. The network was constructed using the BayesiaLab software (www.bayesia.com).

BN Structure learning's objective is to build a graph representing dependence between data, achieving the best fit of data with minimal structural complexity of the net. In the BN language, the variable for which we develop a hypothesis (LLNA potency, in this study) is a target variable, while the variables providing evidence (all three of the above listed types of tests with 12 readouts) are referred to as manifest variables. In addition to manifest variables, the latent variables Bioavailability, Dendritic cells and Reactivity were introduced to the network structure. The latent variables are not observable and are conceptual. They allow the communication of summary results obtained from the network, simplifying the structure of the network by reducing the number of arcs between conditionally dependent variables as well as simplifying numerical computations for the joint probabilities.

Removing chemicals with incomplete records would leave only 45 chemicals for which the full record is available. Hence, 97 chemicals would be ignored and not analyzed further and valuable information would be lost. Skipping such a big portion of data may result in biased estimates, especially in cases when missing data contain information considerably different from the rest of the dataset. Hence, to maximize use of available information, the data gaps in the training set were filled in by imputation. Details of each BN construction step are described in File C (supplementary data; www.altex-edition.org).

The performance of the network in terms of classification performance was evaluated on the test set of 12 chemicals provided in the Supplementary Data. However, we would like

**Tab. 1: Tests used in BN ITS as input variables and the abbreviations used in the text and figures**

| Manifest variables | Unit | Abbreviation |
|---|---|---|
| IL-8 activation | $\mu$M | IL-8 |
| CD86 expression | $\mu$M | CD86 |
| Free epidermal concentration | $\mu$mol/cm$^3$ | Cfree |
| Molecular weight | g/mol | MW |
| Octanol/water coefficient (log) | – | Kow |
| Mid epidermal concentration | $\mu$mol/cm$^3$ | Epi conc |
| Permeation coefficient Kp | cm/hr | Kp |
| Systemically absorbed dose | % | Dose abs |
| TIMES prediction considering parent only/metabolites | – | TIMES-P / TIMES-M |
| Lysine reactivity | % | Lysine (Lys) |
| Cysteine reactivity | % | Cysteine (Cys) |
| ARE Luciferase activity | $\mu$M | Luciferase (Luc) |

to emphasise that the value of using the network is far more than a prediction framework. The network represents key steps of the skin sensitization process and it can be queried to find a variety of options to develop a mechanistically interpretable testing strategy. Finding equivalent tests, assessing the value of adding an additional test when a related one is known, and demonstrating the evolution of the testing strategy based on the amounting evidence are useful features of the network approach.

### 2.3 Methodology to guide testing

Value of Information (VoI) measures and one-step look-ahead hypothesis were used as the methodology to guide testing. The one-step look-ahead hypothesis calculates the VoI from all possible individual information sources and chooses the one for which the information gain about the target variable is maximized. The foundation of this reasoning is the analysis of the changes in the probability distribution of the information target given a set of existing data versus generation of new data. In this study, we use relative mutual information MI (X, Y) between variable X and Y to measure VoI. MI measures the amount of uncertainty in Y, which is removed by knowing X. This corroborates the intuitive meaning of mutual information as the amount of information (that is, reduction in uncertainty) that knowing X variable provides about the Y. The relativity refers to % of entropy of the parent node Y, H(Y), reduced by knowledge of X. Thus, relative MI amounts to MI(X,Y)/H(Y) and is expressed in %. In the remainder of the paper relative MI is abbreviated as MI. For more technical information, see File C in supplementary data; www.altex-edition.org.

## 3 Results

### 3.1 Bayesian Network construction

*Input data transformation to discrete values*

The histograms representing the discretized training data set according to the process described in the Supplementary Data are shown in Figure 1.

*Elucidation of the latent variables*

The unsupervised clustering algorithm identified 2 clusters that were biologically meaningful. The first cluster contained variables associated with the Bioavailability variables (1st latent variable): MW, Epi conc., Cfree, Kp and Kow. The second cluster contained variables associated with the Reactivity variables (2nd latent variable): Lys, Cys, Luc and TIMES-M or TIMES-P. For CD86 and IL-8 variables no meaningful clustering was found. Failing of the automatic clustering of these variables is a result of large data gaps in the original data set. Aiming for the mechanistic interpretation of the latent variables, we manually added Dose abs to the bioavailability cluster, while the 3rd cluster was manually created with Dendritic cells data (3rd latent variable): CD86 and IL-8.

The structure learning algorithm identified local networks for each latent variable in the form of Naïve Bayes (Fig. 2). A joint probability distribution was calculated for each cluster to represent a latent variable.

*Missing data imputation*

Chemicals with no missing data were selected for the Bioavailability and Dendritic cells clusters. For the Reactivity cluster the filtering was done considering only Cys and Lys, and not Luc,

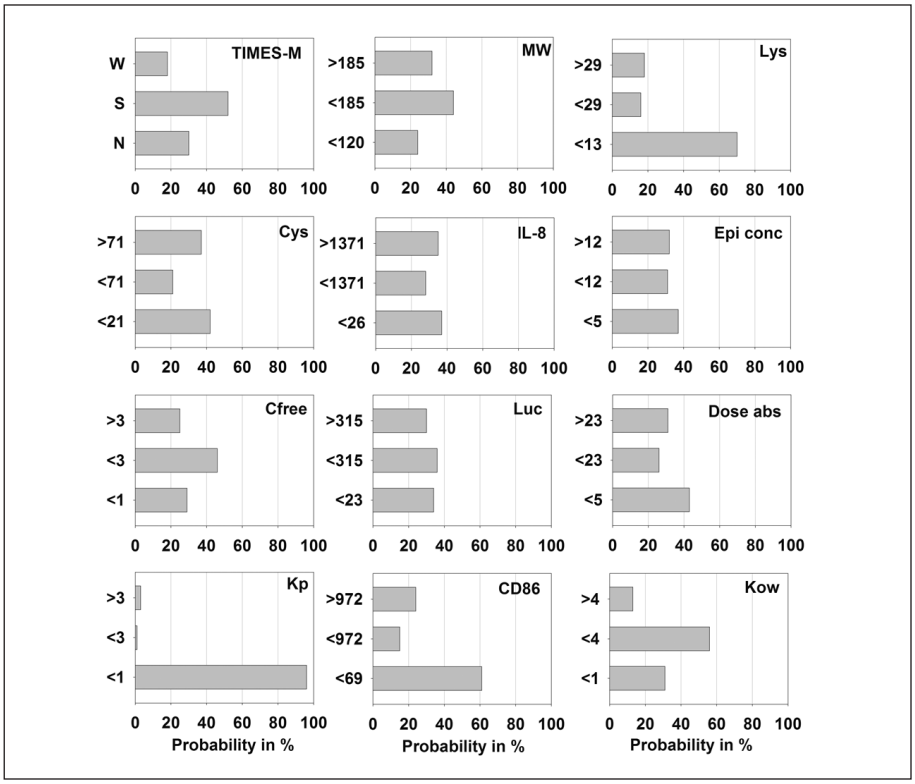


**Fig. 1: Histograms representing the discretized training data set.** This is the initial state, from which all further network analyses are conducted.

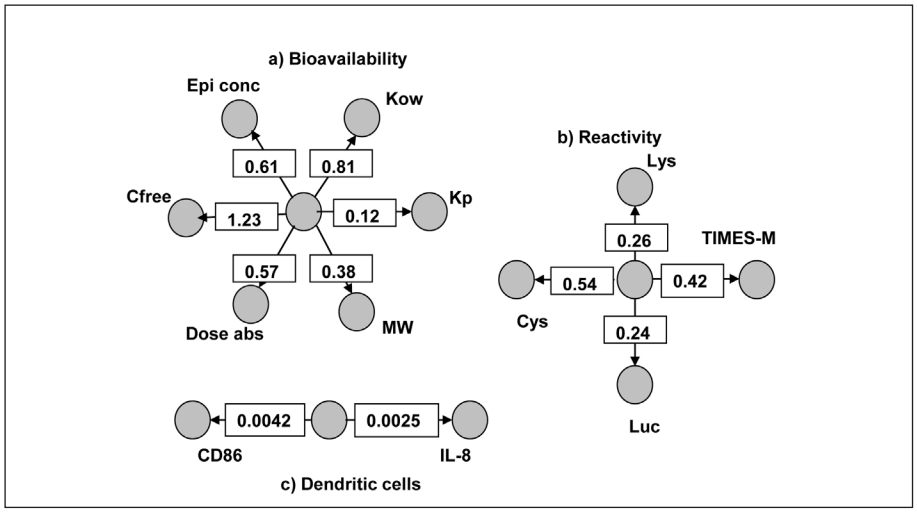


**Fig. 2: Latent variables: a) Bioavailability; b) Reactivity; c) Dendritic cells local networks**
Each arc is tagged with an MI value between the nodes it connects.

because there were no chemicals containing full records. Based on the resulting local data sets, all missing values were filled using the EM algorithm (Meng and Rubin, 1993) that allowed for an amended but complete new training set. Now the data were prepared for final model structure and parameters learning.

*The final Bayesian Network structure*
The network is able to follow the skin sensitization process by choosing a test sequence representing individual steps in the process. The BN ITS structure represents a Hierarchical Naïve Bayes classifier (Langseth and Nielsen, 2006) except that the TIMES node is connected to both Reactivity cluster and directly to the hypothesis variable LLNA (Fig. 3). It represents advancement over a popular Naïve Bayes (NB) classifier that assumes independence among manifest variables and ignores dependence between tests that translates to information duplication. HBN models have been shown to improve classification accuracy over NB by introducing latent variables to account for conditional dependence between manifest variables (Langseth and Nielsen, 2006), as well as for data heterogeneity between the clusters representing latent variables (Demichelis et al., 2006).

## 3.2 Value of Information analysis

*Test Ranking based on Mutual Information (MI) with LLNA*
The global ranking ordered all tests regardless of the LLNA potency group or state. The local ranking information ranked the tests differentiating possible LLNA states and can be used to advise on the next test to further refine a particular hypothesis, for example, that a chemical is a weak sensitizer. We can evaluate which test is the most informative globally as a starting point and afterwards refine the hypothesis suggested by local ranking.

From the global view the Reactivity latent variable was the most informative and contained more information explaining LLNA activity than the Bioavailability and the Dendritic cells variables together. On the level of latent variables, we observed that inclusion of the TIMES-M model in the network improved the mutual information of the Reactivity latent variable by 35%. Presence of the TIMES-M model corrected the joint probability distribution of the Reactivity latent variable so it can better predict LLNA potency in the global and per LLNA state analysis (except for weak sensitizers). From the local ranking we observed different patterns of importance for different potency classes (Tab. 2). Interestingly, the Bioavailability profoundly dominated the ability to explain weak sensitizers. Dendritic cells were always the least informative latent variable except in the case of strong sensitizers, for which they came in as second most informative (latent) variable. We emphasize that results for

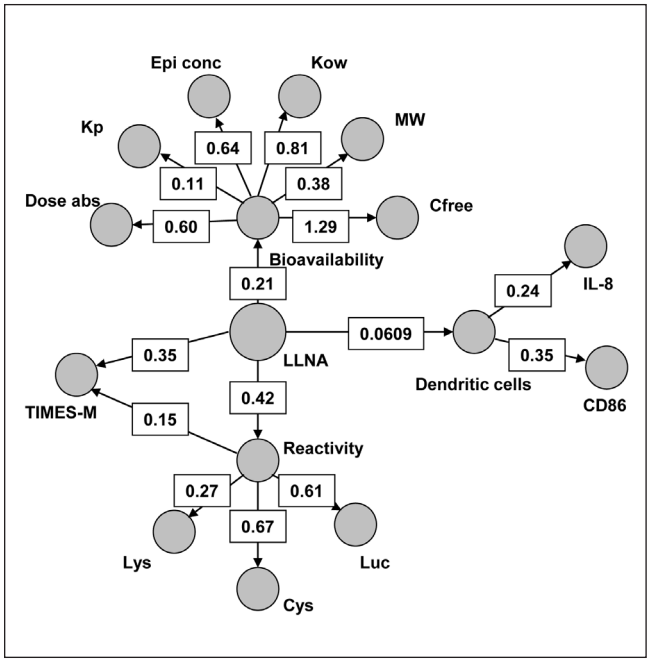


**Fig. 3: Final BN ITS for LLNA potency**
To calculate LLNA potency probability distribution one needs to compute Bioavailability, Reactivity, Dendritic cells and TIMES probability distributions. Every latent variable explains cumulative influence of the manifest variables attached to it with respect to the LLNA node. Each arc is tagged with an MI value between the nodes it connects.

**Tab. 2: Mutual Information (MI) between latent variables and LLNA both global and categorized using a 4-way classification scheme: non-sensitizing (NS), weak (W), moderate (M), and strong (which also included extreme sensitizers) (S)**

**With TIMES-M**

| Global | | NS | | W | | M | | S | |
|---|---|---|---|---|---|---|---|---|---|
| | MI | | MI | | MI | | MI | | MI |
| Reactivity | 0.42 | Reactivity | 0.36 | Bioavailability | 0.13 | Reactivity | 0.08 | Reactivity | 0.11 |
| Bioavailability | 0.21 | Bioavailability | 0.10 | Reactivity | 0.00 | Bioavailability | 0.05 | Dendritic cells | 0.03 |
| Dendritic cells | 0.06 | Dendritic cells | 0.05 | Dendritic cells | 0.00 | Dendritic cells | 0.01 | Bioavailability | 0.01 |

**Without TIMES-M**

| Global | NS | | W | | M | | S | | |
|---|---|---|---|---|---|---|---|---|---|
| | MI | | MI | | MI | | MI | | MI |
| Reactivity | 0.31 | Reactivity | 0.26 | Reactivity | 0.15 | Reactivity | 0.06 | Reactivity | 0.07 |
| Bioavailability | 0.23 | Bioavailability | 0.10 | Bioavailability | 0.00 | Bioavailability | 0.04 | Dendritic cells | 0.03 |
| Dendritic cells | 0.06 | Dendritic cells | 0.05 | Dendritic cells | 0.00 | Dendritic cells | 0.01 | Bioavailability | 0.01 |

**Tab. 3: Mutual Information (MI) between manifest variables and LLNA both global and categorized using a 4-way classification scheme: non-sensitizing (NS), weak (W), moderate (M), and strong (which also included extreme sensitizers) (S)**

**With TIMES-M**

| Global | | NS | | W | | M | | S | |
|---|---|---|---|---|---|---|---|---|---|
| TIMES-M | 0.61 | TIMES-M | 0.39 | TIMES-M | 0.20 | TIMES-M | 0.14 | TIMES-M | 0.13 |
| Cysteine | 0.31 | Cysteine | 0.27 | Cfree | 0.10 | Cysteine | 0.06 | Cysteine | 0.08 |
| Luciferase | 0.29 | Luciferase | 0.25 | Dose abs | 0.06 | Luciferase | 0.06 | Luciferase | 0.07 |
| Cfree | 0.16 | Lysine | 0.12 | Kow | 0.06 | Cfree | 0.03 | Lysine | 0.03 |
| Lysine | 0.13 | Cfree | 0.08 | MW | 0.04 | Lysine | 0.02 | CD86 | 0.02 |
| Dose abs | 0.08 | Dose abs | 0.04 | Kp | 0.02 | Kow | 0.02 | IL-8 | 0.01 |
| Kow | 0.06 | CD86 | 0.03 | Epi conc | 0.01 | Epi conc | 0.01 | Cfree | 0.01 |
| CD86 | 0.04 | Epi conc | 0.03 | Cysteine | 0.00 | CD86 | 0.01 | Dose abs | 0.00 |
| MW | 0.04 | IL-8 | 0.02 | Luciferase | 0.00 | Dose abs | 0.01 | Kow | 0.00 |
| Epi conc | 0.04 | Kow | 0.01 | Lysine | 0.00 | MW | 0.01 | Epi conc | 0.00 |
| IL-8 | 0.03 | MW | 0.01 | CD86 | 0.00 | IL-8 | 0.00 | MW | 0.00 |
| Kp | 0.02 | Kp | 0.00 | IL-8 | 0.00 | Kp | 0.00 | Kp | 0.00 |

**Without TIMES-M**

| Global | | NS | | W | | M | | S | |
|---|---|---|---|---|---|---|---|---|---|
| Cysteine | 0.26 | Cysteine | 0.22 | Cfree | 0.12 | Cysteine | 0.06 | Cysteine | 0.06 |
| Luciferase | 0.24 | Luciferase | 0.21 | Dose abs | 0.07 | Luciferase | 0.05 | Luciferase | 0.06 |
| Cfree | 0.18 | Cfree | 0.08 | Kow | 0.06 | Cfree | 0.03 | CD86 | 0.02 |
| Dose abs | 0.09 | Lysine | 0.07 | MW | 0.04 | Kow | 0.02 | Lysine | 0.02 |
| Lysine | 0.08 | Dose abs | 0.04 | Kp | 0.02 | Lysine | 0.01 | IL-8 | 0.01 |
| Kow | 0.06 | CD86 | 0.03 | Epi conc | 0.01 | Epi conc | 0.01 | Cfree | 0.00 |
| CD86 | 0.04 | Epi conc | 0.03 | Cysteine | 0.00 | Dose abs | 0.01 | Dose abs | 0.00 |
| MW | 0.04 | IL-8 | 0.02 | Luciferase | 0.00 | CD86 | 0.01 | Kow | 0.00 |
| Epi conc | 0.04 | Kow | 0.01 | Lysine | 0.00 | MW | 0.01 | MW | 0.00 |
| IL-8 | 0.03 | MW | 0.01 | CD86 | 0.00 | Kp | 0.00 | Epi conc | 0.00 |
| Kp | 0.02 | Kp | 0.00 | IL-8 | 0.00 | IL-8 | 0.00 | Kp | 0.00 |

Dendritic cells may be biased as a result of so many data gaps and need confirmation with more data.

Next, we analyzed manifest variables MIs with LLNA both globally and per state (Tab. 3). Based on MI values, all types of variables carried more VoI for NS class than for other classes. We studied the rankings with and without TIMES-M in the network. Inclusion of TIMES-M increased the MIs between all manifest variables belonging to the Reactivity in the network, meaning that the TIMES-M model corrected the joint probability distribution of Cys, Lys and Luc. The MI index and high rankings of TIMES-M compared to other Reactivity variables are inflated because of the 72% overlap between chemicals in the TIMES training set and the training set used in this study. In other words TIMES already had "seen" 72% of LLNA data and learned rules from these data. In contrast, the experimental methods are entirely unbiased. As a consequence, comparison of VoI carried by TIMES with VoI of experimental data is not fair based only on these values. To address this, we investigated performance of variants of the network with and without TIMES later in the paper.

Among Bioavailability manifest variables, the most informative variable was the Cfree and the next was Dose abs. However, due to empirical formulation of partitioning in the trans-dermal transport model equations (Kasting et al., 2008), it is premature to ascribe Cfree as the key bioavailability-related driver for skin sensitization. Nevertheless, both Cfree and Dose abs carried more information than Kow and MW and demonstrated the value of including finite dose exposure calculations in BN ITS for LLNA potency assessment.

*Test may carry equivalent information towards explaining the target*

MI can be used to determine whether two different tests carry equivalent information towards explaining the target and whether there is added value to conducting a second test when one of the tests is available. We illustrated this with the peptide reactivity tests. Cys and Luc had very similar MI with LLNA both globally and per state. This means that based on the available evidence, both tests can be used interchangeably to learn about the LLNA potency. In addition, generating evidence on both tests did not advance our knowledge about LLNA potency, e.g., MI (LLNA, Cys or Luc) = MI (LLNA, Cys and Luc) (data not shown). Given that this conclusion was reached with many data gaps for Luc, confirmation with more data is needed.

*Value of adding Lys test if Cys result is known*

In BN one can study not only the value of adding additional evidence, as discussed above, but also how an additional test *result* changes the hypothesis about the target distribution by directly observing changes in the posterior distribution. We illustrated this by studying LLNA posterior distribution changes after combining evidence from Cys reactivity with Lys reactivity. First, the impact of providing the network Cys results was examined. According to the discretization (Fig. 2) 3 ranges, i.e., C1 ≤21%, C2 [21-70]%, C3 >71% depletion in the reactivity assay, covered all possible Cys results. If result C1 was obtained, the chance that the evaluated chemical is a non-sensitizer increased from 26% to 56% (Fig. 4), while it decreased to 15% to be a moderate, and to 7% to be a strong sensitizer. Note that the chance of obtaining a Lys ≤13% depletion increased to 94%, indicating a strong dependence between C1 and L1 results.

Subsequently information on Lys was added. Three simulations were carried out, one for each possible state to study differences in resulting distributions for LLNA (Fig. 5). If Lys was L1 (≤13%) then very little further refinement was obtained regarding the LLNA potency. The probability of the hypothesis that a chemical is a non-sensitizer changes from 56% to 59%. This was a consequence of a high conditional dependence between Cys=C1 and Lys=L1, Pr(L1|C1)=0.93. However if the result for Lys was L2 [13-29% depletion] or L3 >29% depletion), the LLNA distributions shifted from non-sensitizer centered towards moderate (35% for both L2 and L3) or strong sensitizer (26% for L2 and 27% for L3).

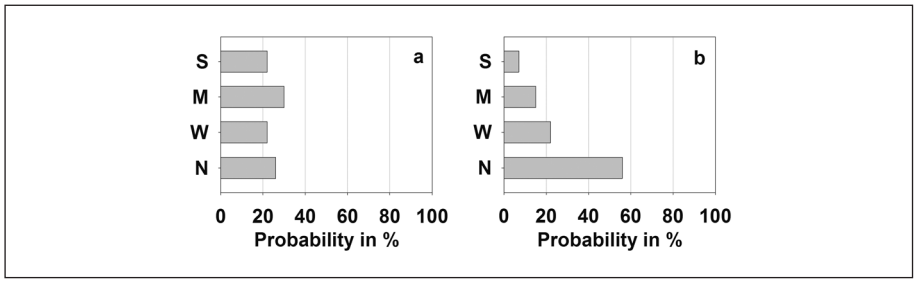


**Fig. 4: LLNA probability distributions a) before evidence for Cys was provided; b) after evidence for Cys equal C1 (i.e. we are 100% sure that it was C1) was provided**

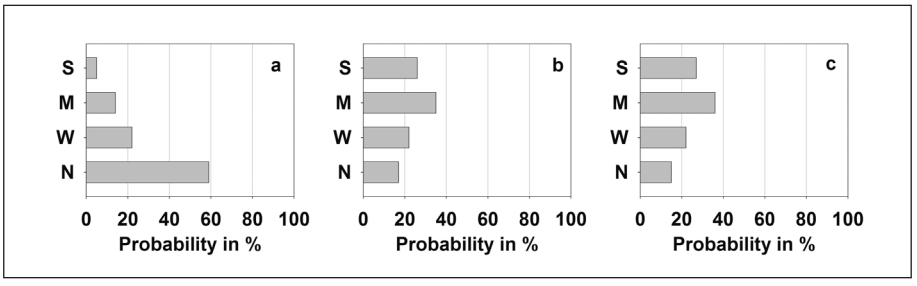


**Fig. 5: LLNA probability distribution after adding Lys L1 (a), L2 (b), L3 (c) given that Cys was C1 ≤21%**

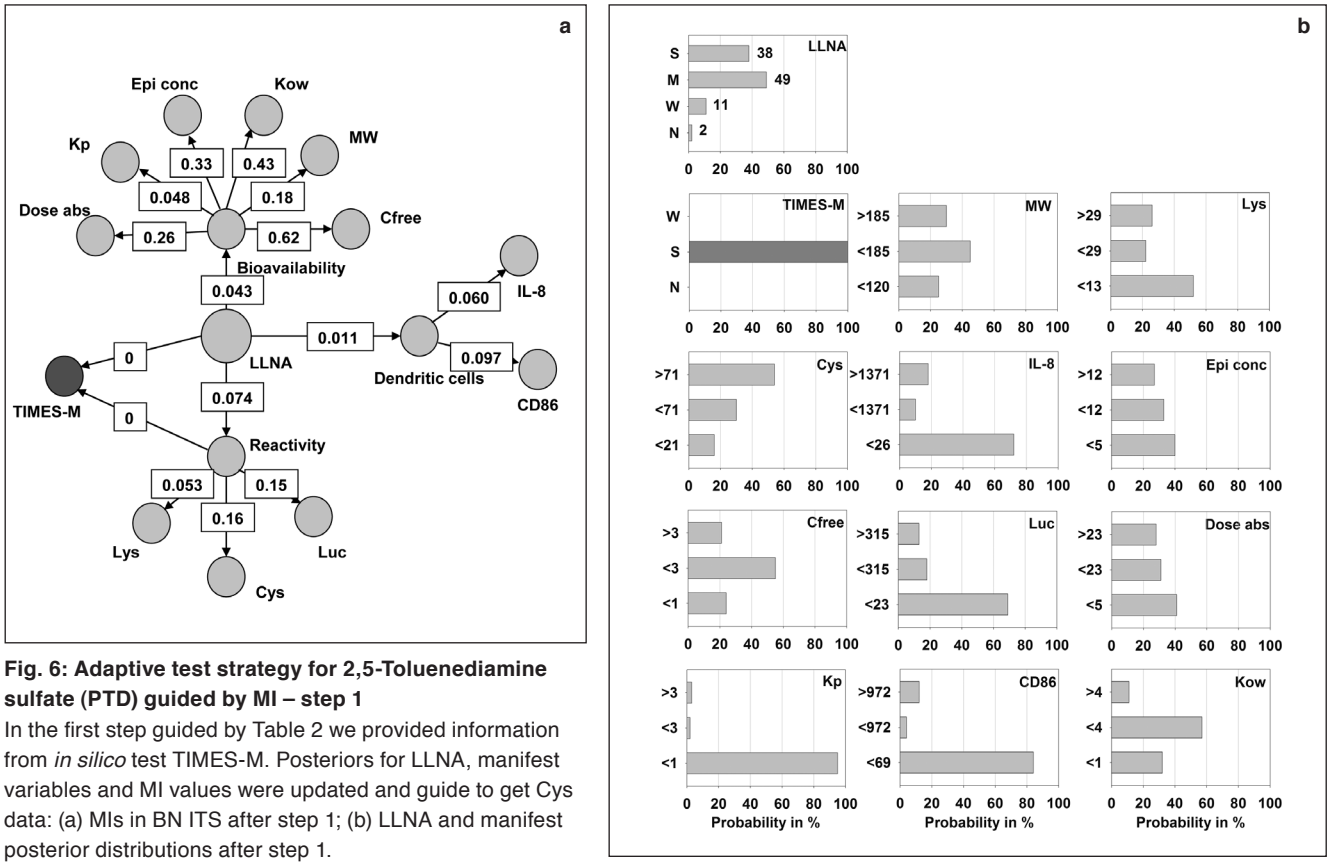


**Fig. 6: Adaptive test strategy for 2,5-Toluenediamine sulfate (PTD) guided by MI – step 1**

In the first step guided by Table 2 we provided information from *in silico* test TIMES-M. Posteriors for LLNA, manifest variables and MI values were updated and guide to get Cys data: (a) MIs in BN ITS after step 1; (b) LLNA and manifest posterior distributions after step 1.

If the result for Cys was larger than 21%, i.e., either the C2 or C3 state was observed, the results were similar to those for C1. The C2 result alone yielded NS=3%, W=22%, M=42%, S=34%, while the C3 result alone yielded NS= 4%, W=22%, M=41%, S=33%, respectively. Further, changes of the LLNA distribution after evidence on Lys was added are smaller than 2% for all the states. The above analyses suggest no value in conducting the Lys test if Cys reactivity is greater than 21%. However, there is a value in conducting the Lys test when Cys reactivity is smaller than 21%. Out of 142 chemicals in the training set, 57 have Cys reactivity values <21%, the majority being non-sensitizers.

*Testing strategy depends on the initial information and changes based on incoming new information in an adaptive manner*

Frequently a full record for the assessed chemical is not available. In a BN setting, an initial hypothesis can always be generated based on incomplete evidence. Testing should start with a test having the highest MI with the target among all available tests. After obtaining the result from one test (or several, if one chooses so) the hypothesis about the target can be revised and the calculation of MI repeated. Figures 6, 7, and 8 illustrate a sequential testing strategy for 2,5-Toluenediamine sulfate (PTD) guided by MI. Figures 6, 7, and 8 are integral to understanding how the network works and its potential to alleviate unnecessary animal testing. They aim to show that the process is iterative and that different parts of the network can be interrogated on the fly as the whole network will update itself. By this, we mean that all the probability distributions for all the nodes of the network, not only the target node, are updated.

*Toxicity signatures*

We generated toxicity signatures for each LLNA state. A toxicity signature is a fingerprint consisting of manifest variables with values that maximize probability for a particular LLNA state (Tab. 4).

## 3.3 BN ITS classification performance

BN ITS classification performance was assessed via AUC of ROC (Area Under the Curve of Receiver Operating Characteristic) curves (Tab. 5). Due to the fact that we have a four-way classification model, the AUC indices larger than 25% are better than a random guess. Three variants of the BN ITS were examined: a) with TIMES-P; b) with TIMES-M; and c) without TIMES. These comparisons were completed for two exposures: V1 and V2. In all cases, while the overall structure of the network was the same, changes in the performance were noted. The networks with and without TIMES performed similarly well when predicting NS class. The network without TIMES, however, performed worse for W, M, and S classes. This suggests that TIMES and experimental reactivity data are about equivalent to predict NS class. This also suggests that TIMES is a very valuable component of ITS for predictions of

a b

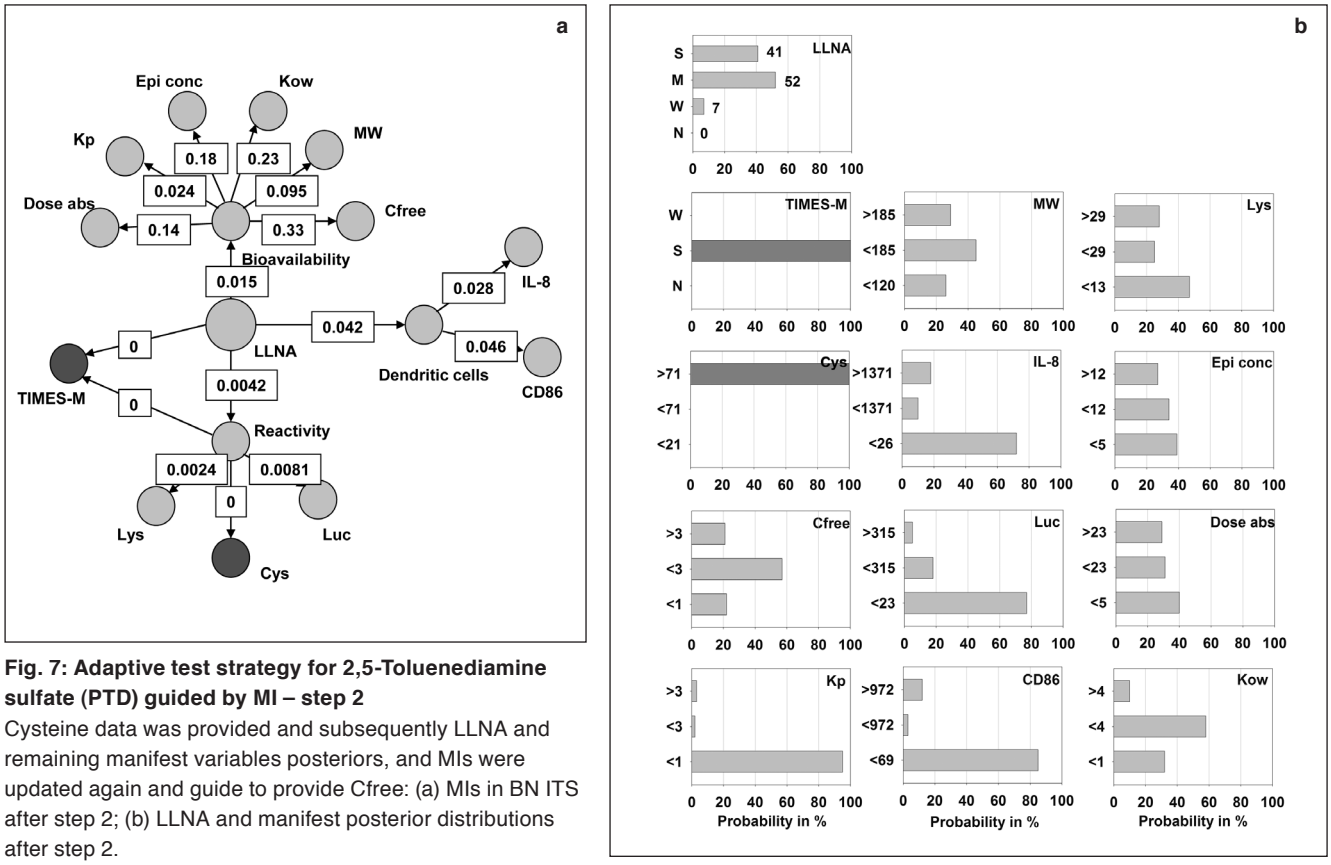


**Fig. 7: Adaptive test strategy for 2,5-Toluenediamine sulfate (PTD) guided by MI – step 2**

Cysteine data was provided and subsequently LLNA and remaining manifest variables posteriors, and MIs were updated again and guide to provide Cfree: (a) MIs in BN ITS after step 2; (b) LLNA and manifest posterior distributions after step 2.

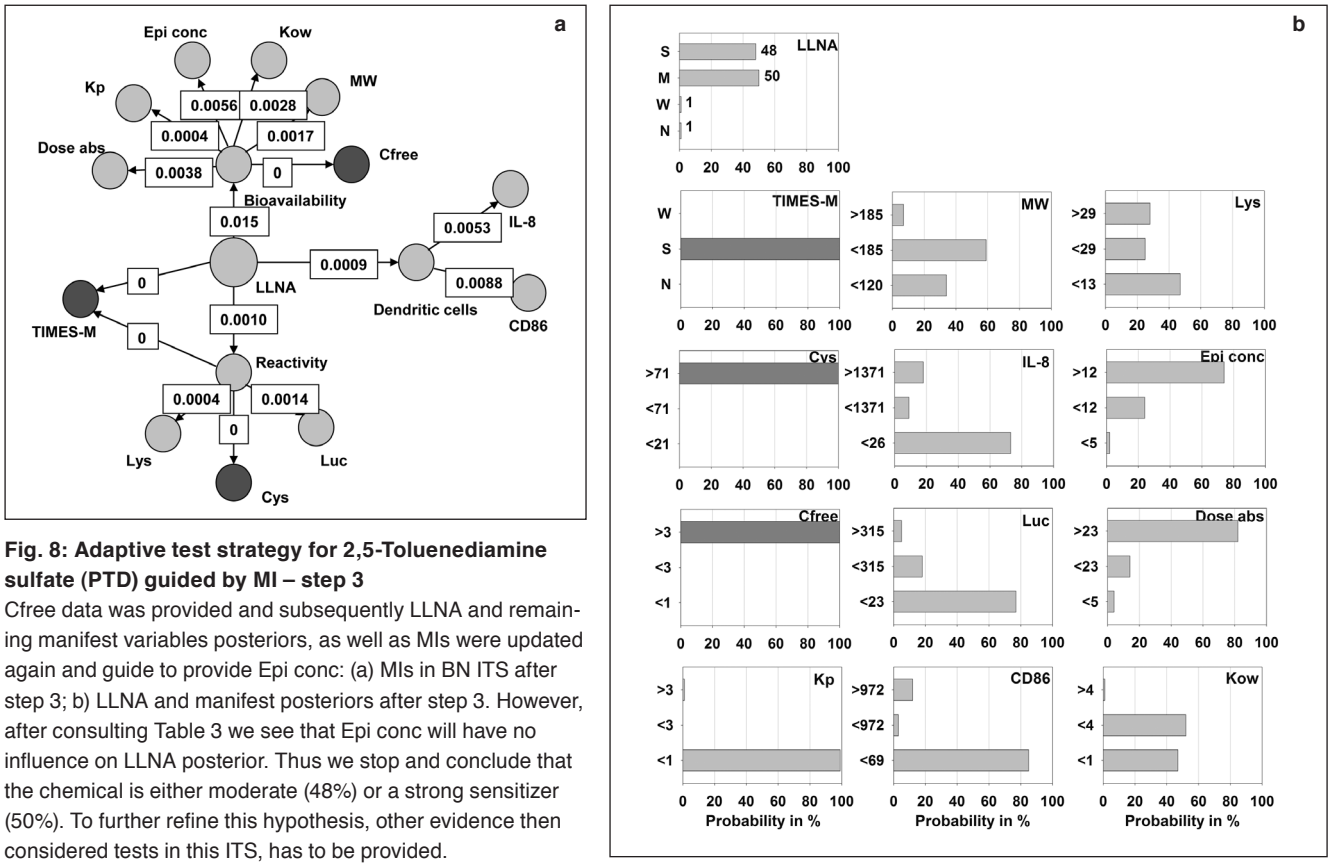


**Fig. 8: Adaptive test strategy for 2,5-Toluenediamine sulfate (PTD) guided by MI – step 3**

Cfree data was provided and subsequently LLNA and remaining manifest variables posteriors, as well as MIs were updated again and guide to provide Epi conc: (a) MIs in BN ITS after step 3; b) LLNA and manifest posteriors after step 3. However, after consulting Table 3 we see that Epi conc will have no influence on LLNA posterior. Thus we stop and conclude that the chemical is either moderate (48%) or a strong sensitizer (50%). To further refine this hypothesis, other evidence then considered tests in this ITS, has to be provided.

**Tab. 4: Toxicity signatures based on the manifest variables for each of the LLNA states**
Units for the variables are provided in Table 1. These numbers should be treated in a qualitative manner as they reflect data in the training set only.

| | LLNA state | | | |
|---|---|---|---|---|
| Test | NS | W | M | S |
| TIMES-M | Non-sensitizer | Weak | Strong | Strong |
| Cysteine | <21 | <71 | <71 | <71 |
| Luciferase | >314 | <23 | <23 | <23 |
| Cfree | >3 | <1 | <3 | >3 |
| Lysine | <13 | >29 | >29 | >29 |
| Dose abs. | >23 | <23 | >23 | >23 |
| Kow | <1 | >4 | <1 | <1 |
| CD86 | <971 | <971 | <971 | <69 |
| MW | <185 | >185 | <119 | >185 |
| Epi conc | >12 | >12 | <5 | >12 |
| IL-8 | <1371 | <1371 | <1371 | <26 |
| Kp | <1 | >3 | <3 | <1 |

individual W, M, and S classes for which experimental reactivity data seems less informative. This observation is in line with excellent 88% correct predictions for W and M+S in TIMES in this study. The analyses of experimental reactivity data by Gerberick et al. (2007) were restricted to binary classifications (Se=88%, Sp=90% on the training set), so their performance to predict individual W, M, and S classes separately was not available.

### 3.4 Different TIMES reference predictions: TIMES-M and TIMES-P

Both TIMES-M and -P showed strong connection with the Reactivity cluster in addition to the LLNA node. In the V1 version, the MI between TIMES-M and the Reactivity latent variable was 0.42 and TIMES-P model was 0.29. A stronger link for TIMES-M likely reflects the fact that TIMES-M assesses skin sensitization potency based on the most potent molecule from parent chemical and metabolites, and that reactivity is positively correlated with potency.

The network with TIMES-M gave a better fit for weak and moderate sensitizers than the network with TIMES-P. Both networks performed similarly for moderate and strong sensitizers. Clearly, consideration of metabolism for the non-sensitizers and weak classes seemed to be important. Further, the network performance with both TIMES predictions was worse for strong and moderate classes compared to non-sensitizers and weak sensitizers. This result can be partially explained by the fact that TIMES predicts only three states (none, weak, and moderate/strong/extreme together), while our network has four states for LLNA potency, thus moderate and strong sensitizers are less precisely predicted.

For some chemical management decisions, for example Regulation EC No 1272/2008 as required for REACH, distinction between non-sensitizers and sensitizers suffices. Our network can easily be adopted for this purpose by pooling weak, moderate, and strong into one class: sensitizers. The advantage of pooling together at the decision-making stage and not the modeling stage allows for a more precise identification of non-sensitizers. Results suggested that TIMES-M performs slightly better in this setup (Tab. 6).

**Tab. 5: BN ITS performance expressed as AUC of the ROC curve considering 2 different exposures V1 and V2 and 3 versions of ITS: 1) With TIMES-M; 2) with TIMES-P; and 3) without TIMES**

**V1**

| | NS | W | M | S |
|---|---|---|---|---|
| TIMES-M | 96% | 81% | 58% | 65% |
| TIMES-P | 89% | 80% | 63% | 67% |
| w/o TIMES | 84% | 59% | 43% | 49% |

**V2**

| | NS | W | M | S |
|---|---|---|---|---|
| TIMES-M | 95% | 85% | 69% | 70% |
| TIMES-P | 88% | 82% | 65% | 72% |
| w/o TIMES | 79% | 59% | 35% | 60% |

**Tab. 6: BN ITS performance to differentiate between non-sensitizers and sensitizers using TIMES-P and TIMES-M as *in silico* priors on the training set**
Chemicals predicted Weak, Moderate, Strong were pooled into 1 class-sensitizers.

| | TIMES-P | TIMES-M |
|---|---|---|
| sensitivity | 89% | 92% |
| specificity | 90% | 95% |
| PPV | 77% | 87% |
| NPV | 96% | 97% |
| accuracy | 90% | 94% |

### 3.5 BN ITS performance on the test set

We assessed network performance on a small set of 12 chemicals. They represent different chemical classes, e.g., haloalkanes, amines, and acids, overlapping with the training set. However, it is a challenging test set to predict, as some of them are prohaptens (e.g., para-toluene diamine (PTD), aniline) or have unclear reaction mechanisms. For the test chemicals only TIMES, Lys, Cys and Bioavailability-related inputs were available, except for PTD, for which Dendritic cell results were also available. Performance of the networks with TIMES-M, TIMES-P and without TIMES was compared recognizing that the variability in EC3 is in the range 0.5 to 2 times (0.5EC3 to 2EC3) and therefore potency class assignment can be off by one class (Tab. 7) due to test variability. Therefore, we considered as correct exact correct predictions, defined as a match between

**Tab. 7: BN ITS predictions on the test set on the test set using TIMES-M and TIMES-P and without TIMES**

| | Observed | | TIMES-M | | | | TIMES-P | | | | w/o TIMES | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Chemical** | **class** | **EC 3%** | **NS** | **W** | **M** | **S** | **NS** | **W** | **M** | **S** | **NS** | **W** | **M** | **S** |
| 2,5-Toluenediamine sulfate (PTD) | S | 0.4 | 0.01 | 0.01 | 0.49 | 0.49 | 0.07 | 0 | 0.21 | 0.72 | 0.18 | 0.04 | 0.40 | 0.38 |
| Aniline | W | 89 | 0.31 | 0.08 | 0.35 | 0.26 | 0.72 | 0 | 0.19 | 0.09 | 0.86 | 0.02 | 0.08 | 0.04 |
| Benzoic acid | NS | | 0.17 | 0.16 | 0.06 | 0.61 | 0.3 | 0.05 | 0.12 | 0.53 | 0.05 | 0.04 | 0.46 | 0.46 |
| Diethyl sulfate | M | 3.3 | 0 | 0.12 | 0.51 | 0.38 | 0 | 0.05 | 0.67 | 0.28 | 0.03 | 0.35 | 0.35 | 0.26 |
| Dimethyl sulfate | S | 0.19 | 0 | 0.14 | 0.49 | 0.36 | 0 | 0.05 | 0.67 | 0.28 | 0.07 | 0.36 | 0.33 | 0.24 |
| Dimethylsulfoxide | W | 72 | 0.4 | 0.09 | 0.44 | 0.06 | 0.59 | 0.1 | 0.24 | 0.07 | 0.24 | 0.29 | 0.38 | 0.09 |
| N-Ethyl-N-nitrosourea | M | 1.1 | 0 | 0.12 | 0.51 | 0.37 | 0 | 0.11 | 0.73 | 0.15 | 0.02 | 0.36 | 0.36 | 0.27 |
| 1-Iododododecane | W | 13 | 0.07 | 0.92 | 0 | 0.01 | 0.06 | 0.94 | 0 | 0 | 0.29 | 0.61 | 0.06 | 0.04 |
| 1-Iodononane | W | 24 | 0.09 | 0.89 | 0 | 0.01 | 0.06 | 0.94 | 0 | 0 | 0.35 | 0.54 | 0.07 | 0.04 |
| N-Methyl-N-nitrosourea | S | 0.05 | 0 | 0.04 | 0.64 | 0.32 | 0.01 | 0.02 | 0.59 | 0.39 | 0.00 | 0.14 | 0.57 | 0.29 |
| Pyridine | W | 72 | 0.9 | 0.01 | 0.06 | 0.03 | 0.6 | 0 | 0.19 | 0.21 | 0.86 | 0.02 | 0.08 | 0.04 |
| Undec-10-enal | M | 6.8 | 0.18 | 0.81 | 0.01 | 0.01 | 0.06 | 0.94 | 0 | 0 | 0.50 | 0.37 | 0.10 | 0.04 |

experimental and the class prediction with the highest probability, as well as a number of predictions off by one class. The network with TIMES-P performed similarly to the network with TIMES-M. It predicted 6 exact matches and 5 one class off matches (92% total correct), while the network with TIMES-M had five exact predictions and five one-class off predictions (83% total correct). The network without Times had three exact matches, six one-class off predictions and two two-classes off predictions (75% total correct). The more detailed analysis revealed differences in LLNA posterior probability distributions for the TIMES-P and TIMES-M networks for the chemicals for which TIMES-P and TIMES-M class predictions were the same (Tab. 7). All test set molecules were part of the TIMES training set. The prediction of potencies of pro-haptens, i.e., aniline, PTD, and pyridine, improved with TIMES-M. It is most clear for the potent prohapten sensitizer PTD, for which prediction without TIMES did not result in a clear identification as a potent sensitizer. For weak sensitizers (aniline, pyridine), consideration of metabolism was less impactful as expected.

Most LLNA predictions (e.g., dimethyl sulfate) have a unimodal pattern, whereas some (e.g., benzoic acid) reveal a bimodal pattern. Bimodality is a sign that the chemical input data reveals a pattern unseen in the training dataset, thus likely to be outside the model domain. The other possibility is that the input data are in conflict with each other due to some error. The potential errors can be experimental errors or prediction errors for *in silico* tests. For the benzoic acid both TIMES and reactivity data suggested that the chemical is a non-sensitizer, but bioavailability data were in conflict with strong evidence for a strong class. In this case, bioavailability data were unreliable because the epidermal bioavailability model (Kasting et al., 2008) is not suitable for acids. In general, data conflict can be used for the purpose of quality assurance of the overall prediction.

## 4 Discussion

This study presented an attempt to construct and test an integrated testing strategy following up on the concepts laid out in earlier works. The goals of the study were achieved and the BN ITS for LLNA potency was constructed despite a challenging data set with many data gaps. The developed BN ITS combined prior biological knowledge with heterogeneous experimental *in silico*, *in chemico* and *in vitro* evidence and generated a probabilistic hypothesis about potency of a chemical in the LLNA assay. It can be used purely for data integration and combined inference, as well as an adaptive testing strategy guiding tool. Bessems (2009) and many other authors acknowledge limitations of alternative assays to provide replacement for an *in vivo* study and recommend shifting focus towards reduction and refinement. The BN ITS framework can be viewed as a reduction strategy, as chemicals with clear potency can be separated from chemicals for which more evidence needs to be generated. The approach carries resemblance to current trends in clinical trial design that strive towards optimizing efficiency and increasingly rely on adaptive Bayesian design (Berry et al., 2010).

The BN ITS for LLNA potency provided better predictions compared to earlier approaches assessing LLNA and at the same time offered new insights to testing strategies. The framework formulated a flexible, adaptive testing strategy. It offers objective guidance on how to identify situations in which generating additional data would not reduce uncertainty about the target. The results clearly showed that there is no one best test sequence, but rather that testing strategy depends on chemical structure, exposure, and initial information. Value of Information analysis demonstrated that differences in VoI rankings depend on potency of a chemical. The section *Testing strategy depends on the initial information and changes based on incoming new information in an adaptive manner* under 3.2 and Table 7 further demonstrate that the BN ITS not

only calculates different VoI based on the potency class but it eventually is chemical-specific as the individual chemical is associated with its unique biological fingerprint, resulting in a particular unique LLNA probability distribution. Therefore, mandating a single, generic set of tests as a replacement strategy is unlikely to be efficient.

Suitability of BNs as underpinning methodology to ITS development has been discussed previously (Jaworska et al., 2010). Use of a Bayesian network, as a formal framework, provides a basis for consistent and transparent reasoning when integrating different, incomplete, and conflicting data. The network is able to follow the skin sensitization process by choosing a test sequence representing individual steps in the process. Because it is a probabilistic approach it allows 1) addressing uncertainty in the biological knowledge, 2) combining heterogeneous evidence, and 3) quantifying uncertainty about target and relationships. Uncertainty in relationships is characterized in probabilistic terms as Conditional Probability Tables (CPTs). Differentiating between strong and weak evidence can be accomplished by different shapes of evidence distributions. In this study we used the simplest form of evidence, allocating it always to one class, but in general, it is possible to allocate evidence to more than one class. Only by quantifying uncertainty about target and relationships can we develop strategies to objectively and effectively reduce it.

In addition to prediction functionality, BNs allow different analyses (e.g., evidence sensitivity, Value of Information analysis) that can be used to guide testing. Further, this framework yields a refined model, as new data become available without discarding old data. As such, it provides functionality for reassessment of predictions in light of additional evidence or to reason with incomplete data. We inferred the potency in the LLNA given full and partial input data and showed testing strategy guided by Value of Information (VoI) calculations. Our specific case study results showed that the integration of biological knowledge with data in the form of a BN ITS is a step forward in making efficient use of alternative data and has potential to become a practical part of a toxicologist's toolbox.

Determining the causal structure is a key for mechanistic interpretation capability of ITS. The structure of the developed network reflected the current knowledge about skin sensitization and included key processes, such as dermal penetration, reaction with proteins, and dendritic cell activation (Jowsey et al., 2006). BNs are interpreted as *causal models* where causal model is understood as a model that conveys causal assumptions, not necessarily a model that produces validated causal conclusions (Pearl, 1988). However, the BN ITS framework also requires quantification of the relationships between nodes. In this study, a data-driven approach was used to quantify the relationships. Thus we regard the used training set as the applicability domain of the constructed ITS. Since its outcomes strongly depend on the quality and the appropriateness of the input information, the choice of tests, and the underlying training set of chemicals, expert knowledge plays an important role in assessing the quality and relevance of input information. BN ITS framework cannot suggest tests that are not a part of the network. Such problems can only be solved with approaches such as that of Maxwell and Mackay (2008). In addition, BN ITS in its current form, by the virtue of the underlying input information, only considers metabolism through TIMES-M.

In the VoI we analysed the MI rankings for latent and manifest variables to compare their relative importance in explaining LLNA potency. As postulated in previous studies (Roberts et al., 2007) reactivity characterization is very important in predicting sensitization. The MI rankings not only confirmed this postulate but in addition quantified the importance. However, the bioavailability, and not reactivity, appeared as the major driver determining that a chemical is a weak sensitizer. This exception for chemicals with Kow >3.9 and MW >185 Da suggested poor dermal penetration irrespective of reactivity profile. Our results regarding lack of dendritic cell importance came as a surprise. In other studies, dendritic cell data correlated well with LLNA data (Nukada et al., 2010; Lambrechts et al., 2010). These results should, therefore, be interpreted with caution, especially since 50% of records were missing dendritic cell data. Among manifest variables TIMES-M was the most informative. This result is biased due to TIMES being already trained on a part of the ITS training set. TIMES was followed by Cys and Luc tests that carried similar VoI, with respect to LLNA. As said earlier, all variables carried more VoI for NS class than for the sensitizing classes. This is not surprising, as it shows that the net is best suited to discriminate NS from sensitizers, which is a biological distinction. The split between W, M, and S is based on an arbitrary cutoff based on the LLNA data. It is more difficult for a model to separate different potency classes, as compared to discriminating non-sensitizers from sensitizers. However, while separating NS from sensitizers is often sufficient for hazard identification within regulatory requirements, the discrimination of potency classes remains a very important aspect in model development, as it is critical for conducting skin sensitization risk assessments. These conclusions are limited to the analyzed data set and require further analysis with more data.

Further, we showed how MI can be used in sequential testing to determine when a follow-up test may add value. Thus, Lys data generated under defined conditions adds value to existing Cys data. Our results suggested no value in conducting the Lys test if Cys reactivity was greater than 21% and the important contribution of Lys data in explaining LLNA potency when Cys reactivity was smaller than 21%. These findings confirm Alvarez-Sanchez et al. (2003) and Eilstein et al. (2006) observations regarding importance of considering various amino acid nucleophiles to understand skin sensitization. The importance of studying reactivity towards Lys was already discussed by Gerberick et al. (2004) and further investigated by Troutman et al. (2011) who concluded that Lys was important for molecules with specific reactivity towards $NH_2$ groups (e.g., anhydrides, isocyanates). Since this study evaluated data in the context of potency, further work is needed to link these two different views – potency oriented and chemistry oriented results.

Among bioavailability-related tests the higher rankings of Cfree and Dose abs versus Kow and Kp suggested that quantifying finite dose exposure conditions is more suitable compared to infinite dose in explaining LLNA potency.

The mutual information analysis allowed us also to answer the question of how many tests are useful before we start generating information that does not further improve our knowledge about the target variable. In contrast to the classic variable selection performed globally on a training set which generates one set of results, our framework allows variable adaptive selection on various levels of detail: globally, per potency class, and for an individual chemical. The answer is related to both target and manifest test variability, and in this study four tests were about the maximum.

While constructing the network all available data were efficiently harvested. The training set contained data for many chemicals with one or more missing records. Instead of deleting chemicals with incomplete records (a typical procedure in data processing among toxicologists), missing data were filled in by imputation. Missing data are a common problem in analyzing toxicological data sets. Little and Rubin (1987), among others, have demonstrated the dangers of simply deleting cases. Case deletion strategies can appreciably diminish the statistical power of the analysis and introduce substantial bias into the study. Without the imputation step we would not be able to construct the network with such a complex structure. Good results in the testing phase confirmed utility of imputation.

Several variants of the network were generated to evaluate its performance using AUC of ROC. ROC curves allow a comprehensive assessment and comparison of classification model accuracy among different studies as they do not depend on the prevalence of actives in the training set (Pepe, 2003). In contrast, sensitivity, specificity, and accuracy cannot be interpreted correctly without knowing the prevalence of active chemicals in the training set. Therefore, comparisons with other studies on the basis of these indexes, albeit frequent in toxicology literature, should be done with caution. Using smaller data sets to train their models (Natsch and Emter, 2008) reported 83% accuracy and Gerberick et al. (2007) reported 89% accuracy on the training sets. Since this BN ITS (V1+ TIMES-M) network achieved 94% accuracy on a larger training set and 92% on a test set, it likely classifies more robustly. BN ITS configured with TIMES-M predicted better than the one with TIMES-P. This result demonstrated the value of considering metabolism when assessing skin sensitization potency. Further, TIMES-M had a large impact on improved BN ITS performance for M and S classes compared with BN ITS configured without TIMES.

The developed toxicity signatures can be used in multiple ways. First, they can guide screening criteria. Toxicity signatures can be considered as biological fingerprints and bases for SAR development and read-across. These signatures can be also used as a simple look-up table to classify a chemical if running a network would not be possible or calculating posterior distribution would not be useful.

To build upon results from this proof-of-concept work, a follow up BN ITS will need to be constructed with a comprehensive data set to establish more certain relationships among individual inputs and between inputs and target information. As the amount of data on originally used tests grows, other alternative tests and data relevant to skin sensitization assessment are becoming available (e.g., additional *in chemico* reactivity data as reviewed in Schwoebel et al. (2011) and HCLAT dendritic cell activation assays (Nukada et al., 2010)) which could be used in an ITS. As the consequence of adding new tests, the follow up BN ITS may result in a revised structure. To further generalize practicality of the BN ITS approach and develop more mechanistic insights, there is a need investigate how BN ITS results, geared currently to assess potency, translate back to chemical structural information. In general, it is to be expected that the ITS will continue to evolve, taking into account novel mechanistic insights and new tests, and will eventually transform to be able to assess skin sensitization potential in humans in a dose-response manner based on clinical biomarkers that remain to be established.

While at this point we focus on the scientific credibility of ITS, efforts are needed to make this type of systematic approach more accessible, viable, and feasible. Specifically, there are needs to build publicly accessible infrastructure of quality databases allowing for storing structural and experimental data and workflows to simultaneously mine these databases.

## Supplementary data description

The supplementary data available online (www.altex-edition.org) consists of 3 files: File A: figure illustrating the skin sensitization induction process, File B: training and test data sets, and File C: an annex with details of the BN construction as well as mathematical formulations of VoI and relative mutual information (MI).

## References

Aeby, P., Ashikaga, T., Bessou-Touya, S., et al. (2010). Identifying and characterizing chemical skin sensitizers without animal testing: Colipa's research and method development program. *Toxicol. In Vitro 24*, 1465-1473.

Aleksic, M., Thain, E., Roger, D., et al. (2009). Reactivity profiling: Covalent modification of single nucleophile peptides for skin sensitization risk assessment. *Toxicol. Sci. 108*, 401-411.

Alvarez-Sanchez, R., Basketter, D., Pease, C., and Lepoittevin, J.-P. (2003). Studies of chemical selectivity of hapten, reactivity, and skin sensitization potency. 3. Synthesis and studies on the reactivity toward model nucleophiles of the 13C-labeled skin sensitizers, 5-Chloro-2-methylisothiazol-3-one (MCI) and 2-Methylisothiazol-3-one (MI). *Chem. Res. Toxicol. 16*, 627-636.

Aptula, A. O., Patlewicz, G., Roberts, D. W., et al. (2006). Non-enzymatic glutathione reactivity and in vitro toxicity: A non-animal approach to skin sensitization. *Toxicol. In Vitro 20*, 239-247.

Basketter, D., and Kimber, I. (2009). Updating the skin sensitization in vitro data assessment paradigm in 2009. *J. Appl. Toxicol. 29*, 545-550.

Berry, S., Bradley, P., Carlin, P., et al. (2010). *Bayesian adaptive methods for clinical trials*. Chapman & Hall/CRC Biostatistics Series.

Bessems, J. G. M. (2009). Opinion on the usefulness of in vitro data for human risk assessment. Suggestions for better use of non-testing approaches. *RIVM report # 320016002*.

Demichelis, F., Magni, P., Piergiorgi, P., et al. (2006). A Hierarchical Naïve Bayes model for handling sample heterogeneity in classification problems: An application to tissue micro arrays. *BMC Bioinformatics 7*, 514-526.

Dimitrov, S. D., Low, L. K., Patlewicz, G. Y., et al. (2005). Skin sensitization: modeling based on skin metabolism simulation and formation of protein conjugates. *Int. J. Toxicol. 24*, 189-204.

Eilstein, J., Gimenez-Arnau, E., Duche, D., et al. (2006). Synthesis and reactivity toward nucleophilic amino acids of 2,5-[13C]-Dimethyl-p-benzoquinonediimine. *Chem. Res. Toxicol. 19*, 1248-1256.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2003). *Bayesian data analysis* (2nd ed.). Chapman & Hall/CRC.

Gerberick, G. F., Vasallo, J. D., Bailey, R. E., et al. (2004). Development of peptide reactivity assay for screening contact allergens. *Toxicol. Sci. 81*, 332-343.

Gerberick, G. F., Ryan, C. A., Kern, P. S., et al. (2005). Compilation of historical lymph node data for evaluation of skin sensitization alternative methods. *Dermatitis 16*, 157-202.

Gerberick, G. F., Vasallo, J. D., Foertsch, L. M., et al. (2007). Quantification of chemical peptide reactivity for screening contact allergens: A classification tree model approach. *Toxicol. Sci. 97*, 417-427.

Gerberick, G. F., Aleksic, M., Basketter, D., et al. (2008). Chemical reactivity measurement and the predictive identification of skin sensitizers. *ATLA 36*, 215-242.

Grindon, C., Combes, R., Cronin, M. T. D., et al. (2007). An integrated decision-tree testing strategy for skin sensitization with respect to the requirements of the EU REACH legislation. *ATLA 35*, 683-697.

Jaworska, J., Gabbert, S., and Aldenberg, T. (2010). Towards optimization of chemical testing under REACH: a Bayesian network approach to Integrated Testing Strategies. *Regul. Toxicol. Pharmacol. 57*, 157-167.

Jaworska, J., and Hoffmann, S. (2010). Integrated Testing Strategy (ITS) – opportunities to better use existing data and guide future testing in toxicology. *ALTEX 27*, 231-242.

Jowsey, I. R., Basketter, D., Westmoreland, C., and Kimber, I. (2006). A future approach to measuring relative skin sensitizing potency: a proposal. *J. Appl. Toxicol. 26*, 341-350.

Karlberg, A.-T., Bergstroem, M. A., and Boerje, A. (2008). Allergic contact dermatitis – formation, structural requirements, and reactivity of skin sensitizers. *Chem. Res. Toxicol. 21*, 53-69.

Kasting, G. B., Miller, M. A., and Nitsche, J. M. (2008). Absorption and evaporation of volatile compounds applied to skin. In K. A. Walters, and M. S. Roberts (eds.), *Dermatologic, cosmeceutic and cosmetic development* (385-400). New York: Informa Healthcare USA.

Kern, P. S., Gerberick, G. F., Ryan, C. A., et al. (2010). Historical local lymph node data for the evaluation of skin sensitization alternatives: a second compilation. *Dermatitis 21*, 8-32.

Kimber, I., Basketter, D., Butler, M., et al. (2003). Classification of contact allergens according to potency: Proposals. *Food Chem. Toxicol. 41*, 1799-1809.

Lambrechts, N., Vanheel, H., Nelissen, I., et al. (2010). Assessment of chemical skin-sensitizing potency by an in vitro assay based on human dendritic cells. *Toxicol. Sci. 116*, 122-120.

Langseth, H., and Nielsen, T. D. (2006). Classification using Hierarchical Naïve Bayes models. *Machine Learning 63*, 135-159.

Little, R. J. A., and Rubin, D. B. (1987). *Statistical analysis with missing data*. New York: John Wiley & Sons.

Maxwell, G., Aleksic, M., Aptula, A., et al. (2008). Assuring consumer safety without animal testing: A feasibility case study for skin sensitisation. *ATLA 36*, 557-568.

Maxwell, G., and Mackay, C. (2008). Application of a systems biology approach to skin allergy risk assessment. *ATLA 36*, 521-556.

Meng, X.-L., and Rubin, D. B. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika 80*, 267-278.

Munteanu, P., and Bendou, M. (2001). The EQ framework for learning equivalence classes of Bayesian networks, Proceedings First IEEE International Conference on Data Mining (IEEE ICDM), San José, USA.

Natsch, A., and Emter, R. (2008). Skin sensitizers induce antioxidant response element dependent genes: Application to the in vitro testing of the sensitization potential of chemicals. *Toxicol. Sci. 102*, 110-119.

Natsch, A., Emter, R., and Ellis, G. (2009). Filling the concept with data: Integrating data from different in vitro and in silico assays on skin sensitizers to explore the battery approach for animal-free skin sensitization testing. *Toxicol. Sci. 107*, 106-121.

Nukada, Y., Foertsch, L., Ashikaga, T., et al. (2010). Predicting allergy potential by battery evaluation system using two in vitro skin sensitization tests; Direct Peptide Reactivity Assay (DPRA) and human Cell Line Activation Test (h-CLAT). *Contact Dermatitis 63, Suppl. 1*, 79-80.

Patlewicz, G., Aptula, A. O., Uriarte, E., et al. (2007). An evaluation of selected global (Q)SARs/expert systems for the prediction of skin sensitisation potential. *SAR QSAR Environ. Res. 18*, 515-541.

Patlewicz, G., Aptula, A. O., and Roberts, D. W. (2008). A minireview of available skin sensitization (Q)SAR/expert systems. *QSAR Comb. Sci. 27*, 60-76.

Patlewicz, G., and Worth, A. (2008). Review of data sources, QSARs and Integrated Testing Strategies for skin sensitization. *JRC Scientific and Technical Reports EUR 23225 EN - 2008*.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems:*

*Networks of plausible inference*. San Francisco, CA: Morgan Kaufmann.

Pepe, M. (2003). The statistical evaluation of medical tests for classification and prediction. *Oxford Univ. Press*, 302.

Python, F., Goebel, C., and Aeby, P. (2007). Assessment of the U937 cell line for the detection of contact allergens. *Toxicol. Appl. Pharm. 220*, 113-124.

Roberts, D. W., Aptula, A. O., Cronin, M. T. D., et al. (2007). Global (Q)SARs for skin sensitization-assessments against OECD principles. *SAR QSAR Environ. Res. 18*, 343-365.

Roberts, D. W., Aptula, A. O., Patlewicz, G., et al. (2008). Chemical reactivity indices and mechanism-based read-across for non animal based assessment of skin sensitization potential. *J. Appl. Toxicol. 28*, 443-454.

Roberts, D. W., and Patlewicz, G. (2009). Updating the skin sensitization in vitro data assessment paradigm in 2009 – a chemistry and QSAR perspective. *J. Appl. Toxicol. 30*, 286-288.

Ryan, C. A., Gerberick, G. F., Gildea, L. A., et al. (2005). Interactions of contact allergens with dendritic cells: Opportunities and challenges for the development of novel approaches to hazard assessment. *Toxicol. Sci. 88*, 4011.

Schwoebel, J. A. H., Koleva, Y. K., Enoch, S. J., et al. (2011). Measurement and estimation of electrophilic reactivity for predictive toxicology. *Chem. Rev. 111*, 2562-2596.

Troutman, J. A., Foertsch, L. M., Kern, P. S., et al. (2011). The incorporation of lysine into the peroxidase peptide reactivity assay for skin sensitization risk assessments. *Toxicol. Sci. 122*, 422-436.

Vandebriel, R. J., and van Loveren, H. (2010). Non-animal sensitization testing: State-of-the-art. *Crit. Rev. Toxicol. 40*, 389-404.

van der Jagt, K., Munn, S., Torslov, J., et al. (2004). Alternative approaches can reduce the use of test animals under REACH. Addendum to: Assessment of additional testing needs under REACH effects of (Q)SARS, risk based testing and voluntary industry initiatives. *IHCP report EUR 21405 EN*.

Wang, X. J., Hayes, J. D., and Wolf, C. R. (2006). Generation of a stable antioxidant response element-driven reporter gene cell line and its use to show redox-dependent activation of Nrf2 by cancer chemotherapeutic agents. *Cancer Res. 66*, 10983-10994.

Yuan, Y., and Shaw, M. J. (1995). Induction of fuzzy decision trees. *Fuzzy Sets and Systems 69*, 125-139.

## Acknowledgements

We thank J. Kasting and M. Miller for generating bioavailability data and K. Blackburn and C. Ryan and 2 reviewers for their thorough reviews. The work was supported in part (JJ, AH) by funding of the European Union 6th Framework OSIRIS Integrated Project (GOCE-037017-OSIRIS).

## Correspondence to

Joanna Jaworska, PhD
Procter & Gamble Eurocor,
Temselaan 100
1853 Strombeek-Bever
Belgium
Phone: +32 2 456 2076
Fax: +32 2 4563098
e-mail: Jaworska.J@pg.com,