



Predicting skin sensitizers with confidence — Using conformal prediction to determine applicability domain of GARD

Andy Forreryd^{a,*}, Ulf Norinder^{b,c}, Tim Lindberg^a, Malin Lindstedt^{a,*}

^a Department of Immunotechnology, Lund University, Medicin Village (406), 22381 Lund, Sweden

^b Swetox, Karolinska Institutet, Unit of Toxicology Sciences, Forskargatan 20, SE-151 36 Södertälje, Sweden

^c Department of Computer and Systems Sciences, Stockholm University, Box 7003, SE-164 07 Kista, Sweden

ARTICLE INFO

Keywords:

GARD
In vitro assay
Skin sensitization
Conformal prediction
Applicability domain

ABSTRACT

GARD — Genomic Allergen Rapid Detection is a cell based alternative to animal testing for identification of skin sensitizers. The assay is based on a biomarker signature comprising 200 genes measured in an *in vitro* model of dendritic cells following chemical stimulations, and consistently reports predictive performances ~90% for classification of external test sets. Within the field of *in vitro* skin sensitization testing, definition of applicability domain is often neglected by test developers, and assays are often considered applicable across the entire chemical space.

This study complements previous assessments of model performance with an estimate of confidence in individual classifications, as well as a statistically valid determination of the applicability domain for the GARD assay. Conformal prediction was implemented into current GARD protocols, and a large external test dataset ($n = 70$) was classified at a confidence level of 85%, to generate a valid model with a balanced accuracy of 88%, with none of the tested chemical reactivity domains identified as outside the applicability domain of the assay.

In conclusion, results presented in this study complement previously reported predictive performances of GARD with a statistically valid assessment of uncertainty in each individual prediction, thus allowing for classification of skin sensitizers with confidence.

1. Introduction

Skin sensitizers are chemical compounds with an intrinsic ability to activate specific adaptive immune responses in susceptible individuals following skin contact and penetration into dermis. The adverse health effects arising from such exposure originate from the initial priming of chemical specific skin homing Th1 and CD8⁺ T-cells, which is followed upon re-exposure by the development of clinical symptoms, such as allergic contact dermatitis (ACD) (Akhavan and Cohen, 2003; Fonacier et al., 2010). Once acquired, sensitization it is often a life-long condition, and elicitation of dermatitis can only be prevented by avoidance of exposure (Karlberg et al., 2008). Thus, identification of chemicals with sensitizing properties is a key toxicological endpoint within chemical hazard testing.

To further prevent potentially harmful chemicals from reaching the consumer market or working environment, different legal frameworks within the European Union, such as REACH (EU, 2006) and the GHS/CLP (CLP, 2016) now require information on skin sensitizing hazard. At the same time, ethical considerations associated with animal welfare

has also been implemented in current legislations, and the Cosmetic Directive REGULATION(EC) No 1223/2009 (EU, 2009), now prohibits *in vivo* testing of cosmetic products. Thus, development of non-animal based methods for chemical hazard assessment has recently gained momentum, and to date, three such models, the direct peptide reactivity assay (DRPA) (Gerberick et al., 2004), the ARE-Nrf2 (KeratoSens) (Natsch, 2010), and the human cell line activation test (h-CLAT) (Ashikaga et al., 2006; Nukada et al., 2011) have been formally validated and assigned OECD test guideline numbers (OECD 442C, OECD 442D, and OECD 442E, respectively). These assays are all anchored to the key events (KE) in the adverse outcome pathway (AOP) for skin sensitization (OECD, 2012), where DRPA measures *in chemico* peptide reactivity (KE1), KeratoSens measures keratinocyte activation (KE2), and h-CLAT measures dendritic cell (DC) activation (KE3). The above-mentioned assays all monitor single or few biomarkers, and thus, provide limited mechanistic information.

To meet regulatory requirements, a consensus has emerged within the scientific community that no single *in vitro* assay will in the near future be able to fully replace animal testing. Thus, attempts have been

* Corresponding authors.

E-mail addresses: andy.forreryd@immun.lth.se (A. Forreryd), ulf.norinder@swetox.se (U. Norinder), tim.lindberg@immun.lth.se (T. Lindberg), malin.lindstedt@immun.lth.se (M. Lindstedt).

<https://doi.org/10.1016/j.tiv.2018.01.021>

Received 18 August 2017; Received in revised form 18 January 2018; Accepted 23 January 2018

Available online 31 January 2018

0887-2333/ © 2018 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

made to combine information from different assays targeting the various steps in the AOP into an Integrated Testing Strategy (ITS) based on majority voting from individual assays (2 out of 3 ITS) (Bauch et al., 2012; Natsch et al., 2013). Recently, more advanced ITS models based on machine learning (Tsujita-Inoue et al., 2015; Tsujita-Inoue et al., 2014), or Bayesian network statistics (Jaworska et al., 2015) have also been proposed. These are still far from regulatory acceptance, and ultimately, the benefit of an ITS will depend on the individual components capacity to provide reliable predictions. Importantly, no assay, regardless of being used as a stand-alone test, or within an ITS is perfect, and the confidence on individual classifications, as well as the predictive boundaries *i.e.* the chemical space on which each model is applicable, needs to be thoroughly evaluated in order to ultimately arrive at reliable predictions.

The Genomic Allergen Rapid Detection (GARD) assay (Johansson et al., 2011) was developed as a multiparametric alternative to current *in vitro* assays as an attempt to generate highly accurate mechanism-of-action based hazard classifications, in line with current recommendations by the National Research Council (NRC, 2007). GARD is a highly advanced assay based on multiple measurements of transcriptional changes induced in a myeloid cell line following exposure to chemicals. By monitoring 200 genes associated with sensitization, commonly referred to as the GARD prediction signature (GPS), the assay use supervised machine learning to compare expression profiles induced by unknown testing compounds to the expression profiles of reference samples, and consistently reports accuracies close to 90% (Johansson et al., 2017; Johansson et al., 2014; Zeller et al., 2017). The assay has been taken up by the OECD Test Guideline Program (TGP 4.106), and are currently undergoing the final stages in a formal validation procedure under the supervision of the European Reference Laboratory for alternatives to animal testing (EURL ECVAM),

In this study, we introduce the conformal prediction framework (Vovk et al., 2005) to establish a transparent assessment of the confidence in individual predictions, as well as a statistically valid determination of the predictive boundaries of the GARD assay, for which prediction of novel chemicals are valid. The latter is commonly referred to as the applicability domain (AD) and is informative to ensure proper usage of the derived model and its current limits with respect to extrapolation outside of the chemical space occupied by the reference samples. In contrast to *in silico* methods, such as quantitative structure–activity relationships (QSAR) assays (ECHA, 2008; OECD, 2007; OECD, 2004), this information is not yet required for regulatory purposes of *in vitro* assays. Such methods are often instead considered universally applicable to all chemicals, an assumption we consider to be incorrect. As a proof of concept, we here apply the conformal prediction framework to the GARD model by classifying a total of 70 chemicals, using a Support Vector Machine (SVM) (Cortes and Vapnik, 1995) as the statistical model and the genes in GPS as descriptors. Results presented in this study illustrates that the conformal prediction framework was useful for providing an instance based measure of confidence for each individual prediction, which resulted in a rigorous definition of the predictive boundaries of GARD. In addition, we show that it was also possible to control the error rate of the model by adjusting the user-defined confidence threshold by internal validation of the training set, and thereby optimize the efficiency in predictions also for corresponding test set classifications.

In conclusion, this study demonstrates and discusses the usefulness of conformal prediction for estimating applicability domains of *in vitro* assays, and illustrates how the framework could be directly implemented into the GARD settings to generate a measure of uncertainty in each prediction, with the ambition to ultimately achieve confidence in final assignments as sensitizers or non-sensitizers.

2. Material and methods

2.1. Chemical compounds

The analysis was performed using a dataset comprising a total of 108 chemicals divided into a training dataset ($n = 38$) and an external test dataset ($n = 70$). Compounds included in the training dataset were identical to those used for initial model development (Johansson et al., 2011), and comprised a balanced set of well characterized skin sensitizing ($n = 18$) and non-sensitizing ($n = 20$) chemicals, apart from the vehicle controls dimethyl sulfoxide and water. Compounds included in the test dataset comprised a more unbalanced mixture of skin sensitizing ($n = 54$) and non-sensitizing ($n = 16$) chemicals, and were either obtained from recent studies (Johansson et al., 2014; Zeller et al., 2017), or acquired from our in house database of chemical stimulations which have not previously been published. These chemicals included the skin sensitizing compounds diethylthiocarbamyl benzothiazole sulfide, dimethylthiocarbamyl benzothiazole sulfide, glyceryl monothioglycolate, hydrazine, kanamycin, neomycin sulphate, and the non-sensitizing compounds hydrocortisone, phenoxyethanol, triethanolamine, and vanillin. Chemical compounds were assigned into protein reactivity domains using the Toxtree v 2.13.6 application (Enoch et al., 2008), according to the reaction mechanistic domains proposed by Aptula and Roberts (2006). All five reactivity domains were represented in the dataset of 108 chemical compounds included in this study with the following distribution: Nucleophilic aromatic substitution (SNAr, 4 compounds), Schiff base formation (SB, 18 compounds), Michael acceptor (MA, 23 compounds), Acyl transfer agent (AT, 7 compounds), Substitution nucleophilic bi-molecular (SN2, 6 compounds). Among these, the two compounds taxifolin and methyl-di-bromo glutaronitrile contained reactivity domains for both MA and SN2 reactions. Seven compounds (kathon CG/ICP, potassium permanganate, CAPE, dimethylthiocarbamyl, kanamycin, dimethyl CAPE, and α -glucosylrutin) could not be assigned into any of the reactivity domains. In addition, the dataset also contained several pre- and pro-haptens (annotated according to the information provided in Patlewicz et al. (2016)), as well as 45 compounds where no protein reactivity could be identified. Information on chemical compounds included in this study are summarized along with reactivity class and concentrations used for chemical stimulations of cells (referred to as GARD input concentrations) in Supplementary Table S1.

2.2. Description of data set for model construction

Gene expression data for model construction were generated using protocols described in (Johansson et al., 2012; Johansson et al., 2011; Zeller et al., 2017). Readers are referred to these articles for a detailed description of cell maintenance procedures, establishment of GARD input concentrations, cellular stimulations, phenotypic control of cells prior to stimulations, RNA isolation, and subsequent microarray hybridizations. The majority of the compounds were assayed in triplicate reactions (for details on number of replicate samples for each stimulation, see Supplementary Table S1), using separate cell batches for each replicate, generating a total data set comprising 324 individual stimulations, distributed across the training dataset ($n = 122$) and the test dataset ($n = 202$). In the subsequent study, raw data from the resulting Affymetrix HuGene 1.0 ST microarray measurements (33,296 transcripts/sample) were imported into R statistical environment (R Development Core Team, 2014) and normalized using the SCANfast algorithm (Piccolo et al., 2012). Since several experimental campaigns were combined into one data set in this study, potential batch effects were eliminated using the ComBat algorithm (Johnson et al., 2007; Leek et al., 2012). A filtering process was then initiated to isolate the 200 genes in the GARD prediction signature (GPS), which were subsequently used for model constructions and predictions. For information on how predictors in GPS were selected, and the identity of the

genes in the signature, the reader is referred to the original papers (Johansson et al., 2012; Johansson et al., 2011; Zeller et al., 2017).

2.3. Conformal prediction

A conformal predictor is a confidence predictor which always makes valid predictions according to a confidence level determined by the user (Vovk et al., 2005). In contrast to single class predictions provided by standard supervised machine learning algorithms, a conformal predictor outputs a prediction region. In the subsequent binary classification setting, this is accomplished by assigning class labels to external test compounds by comparison to a calibration set with known experimental classes. If the prediction for a new test compound is “similar enough” (determined by the user defined confidence level) to a certain class label in comparison to the predictions on the calibration set, the new test compound is assigned to that specific class. This assignment process is repeated for each of the two classes in the test dataset (Mondrian Conformal Prediction), thus, there are four possible outcomes for an external test compound: it can be assigned to either of the two classes (*single* classification), it can be assigned to both classes (*both* classification), or it can be assigned to neither one of the two classes (*empty* classification). A prediction is correct if it includes the correct class, i.e. *both* classifications are always correct, while *empty* classifications are always incorrect. Performance of the conformal predictor is measured by validity and efficiency. A model is valid if the number of prediction errors does not exceed the significance level ε ($1 - \text{user defined confidence level}$), i.e. the maximum percentage of errors that can be tolerated for the problem at hand, while efficiency corresponds to the percentage of single class predictions.

2.4. Statistical approach, software and machine learning algorithm

With the aim to implement the conformal prediction framework to the GARD platform and to introduce its usefulness for predictive *in vitro* assays, a classification algorithm based on SVM predictions (Cortes and Vapnik, 1995) was used as the underlying model building method for the conformal predictions in this study. Gene expression data from GPS was used as variables to explain the descriptor space, and similar to the classification algorithm used in GARD, the binary SVM model was constructed based on samples in training dataset, using default parameters and a linear kernel as decision boundary. The distance to the separating hyperplane was used to generate a conformal prediction p -value for each class for the test dataset compounds, as a measure of similarity (conformity) between the two class labels of calibration compounds (sensitizers and non-sensitizers) and each compound in the test dataset. In Fig. 1, we provide an illustrative example of this prediction process using simulated data, with the aim to clarify the interpretation of the link between the SVM model, conformity values, and conformal prediction p -values. In the figure, the distribution of samples in the calibration set (sensitizers ($n = 15$, in black circles), non-sensitizers ($n = 12$, in open circles)) are shown in the SVM space generated by samples in the training set (not shown). The resulting conformity values (distances from the hyperplane) are shown in the upper right corner in the same figure (sorted in decreasing order) for each of the two classes. A conformity value was then generated for each compound in the external test set (test ($n = 3$, in grey circles), and compared to corresponding values for each class for the compounds in the calibration set, taking the sign, i.e. the side of the hyperplane, into account. Final ranking and classification of samples in the test dataset were performed using conformal prediction p -values, the fraction of compounds in the calibration set with a lower conformity value compared to the test compound. The p -value must be greater than the user defined significance level ε for final assignment to a certain class. Thus, in Fig. 1, test compound 1 (conformity value = 0.94), located further from the hyperplane than any of the sensitizing compounds in the calibration set, on the opposite side of the hyperplane, will be positioned at the end

of the sorted conformity list for the sensitizer class. With respect to the non-sensitizer class, the compound will instead be sorted as number 9 from the bottom, since there are 8 non-sensitizers in the calibration set with smaller conformity values. Resulting p -values for compound 1 will be 0/16 (0) and 8/13 (0.615) for sensitizer class and non-sensitizer class, respectively. Similarly, p -values for compound 2 (conformity value = 0.555) will be 6/16 (0.375) and 1/13 (0.0769), and for compound 3 (conformity value = 1.008) 10/16 (0.625) and 0/13 (0) for sensitizer and non-sensitizer classes, respectively. At the significance level $\varepsilon = 0.15$, compounds 1–3 will thus be classified as non-sensitizer (1), sensitizer (2) and sensitizer (3).

2.5. Model building and validation

Similar to Carlsson et al. (2014) we applied an aggregated conformal prediction method for internal validation of the training dataset. The training dataset was initially randomly divided into an internal training dataset (80%), and an internal test dataset (20%). An ensemble comprising 100 SVM models was subsequently compiled by iteratively and randomly dividing the internal training dataset into *proper* training (70%) and calibration sets (30%). During each iteration, a certain SVM model was initially derived from the randomly selected stimulations selected in the *proper* training set, and subsequently applied to predict the stimulations in the associated calibration set, i.e. to assign a conformity (similarity) value to each compound. After predicting samples in the calibration set, each SVM model was applied to predict conformity values and, subsequently, conformal prediction p -values for all the compounds in the internal test dataset. The median conformal prediction p -value from the 100 predictions for each compound and for each class in the internal test dataset was calculated and used for class assignment in accordance with the set confidence level. The aggregated conformal prediction framework for prediction of the internal test dataset is illustrated in Fig. 2A. Additionally, a similar strategy was applied to perform validation of an external test dataset, however, during this procedure, no prior subdivision of the training dataset was performed. Thus, the complete training dataset was randomly divided into the ensemble of 100 *proper* training and calibration sets, which were subsequently applied for class assignment of the stimulations in the external test dataset in accordance with the set confidence level, similar as described above for the internal validation of the training set. The aggregated conformal prediction framework for prediction of the external test dataset is illustrated in Fig. 2B. All models were built and predicted on individual stimulations (training dataset: 122 and test dataset: 202) and aggregated, as described above, over each compound for final class assignments. All modeling were performed within the Anaconda2 package version 4.3.1 (<https://www.continuum.io/>) using Python (<http://www.python.org>) and Scikit-learn (Pedregosa et al., 2011) version 0.18.1. SVM models were derived using the SVC module. Illustrative examples on how conformal predictions are carried out have been described previously (Norinder et al., 2014).

3. Results

3.1. Conformal prediction — internal validation of the training dataset

In a first step to implement the conformal prediction framework to the GARD assay and to transform the SVM point predictions into confidence classifications with prediction regions, an internal validation of the 38 compounds in the training dataset was performed with the purpose to provide an estimate of how well the model can be expected to perform on the external test dataset. Since conformal prediction by definition allows the user to determine a reasonable and acceptable confidence level to guarantee a maximum error rate of the predictions, the performance of a model is evaluated by estimating efficiency, i.e. the frequency of *single* class labels in the prediction set. Statistics from conformal prediction classifications for the training dataset are shown

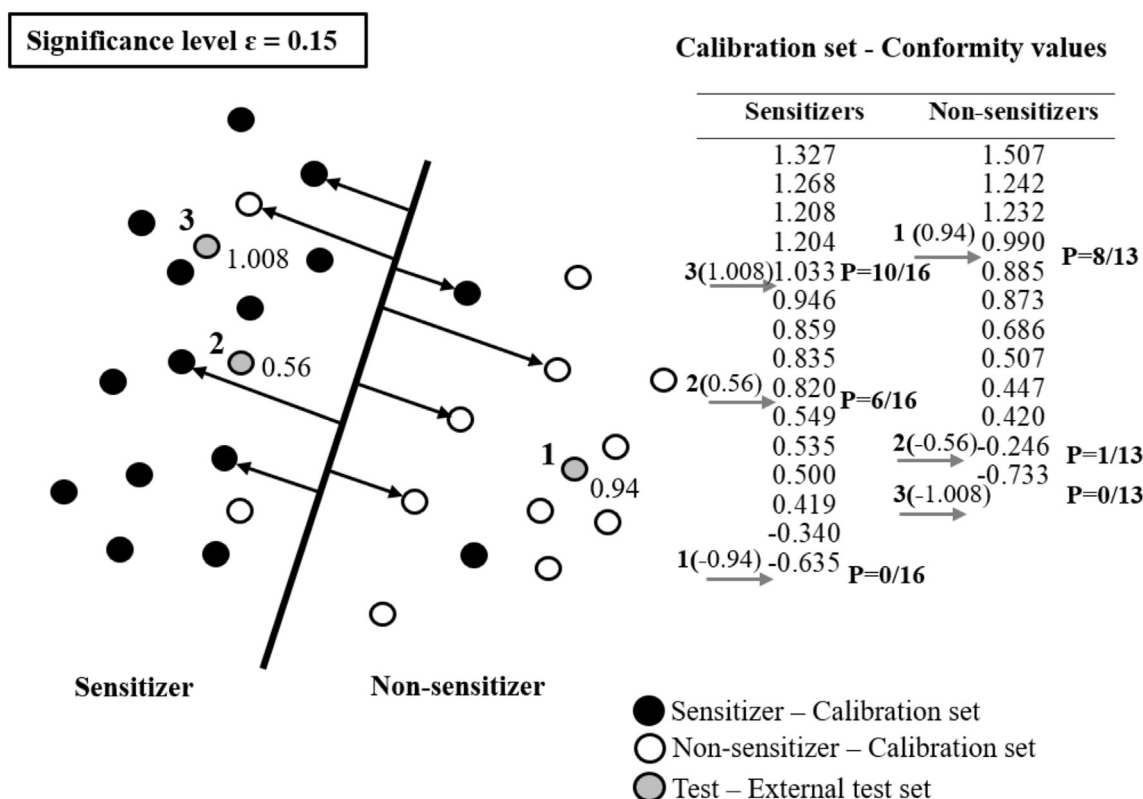


Fig. 1. Conformal predictions using Support Vector Machines (SVM).

Illustration of the relationship between conformity values (CV) and p -values when conformal predictions are used in conjunction with SVM. The SVM model was trained on samples in training set (not shown in figure), and applied to predict the samples in the calibration set (sensitizers ($n = 15$, in black circles), non-sensitizers ($n = 12$, in open circles)). The output from the SVM, i.e. distances to the separating hyperplane were used to generate the two sets of CV (table in upper right in figure). CVs were then generated for samples in test set (test ($n = 3$, in grey circles)). As a measure of similarity to samples in calibration set, p -values for each compound in the test set were subsequently calculated for each class by determining the fraction of compounds in the calibration set with a lower conformity value in comparison to the specific test compound. Resulting p -values (P) for compound 1 (CV = 0.94), compound 2 (CV = 0.56), and compound 3 (CV = 1.008) were calculated to 0/16, 6/16 and 10/16 for sensitizer class, and 8/13, 1/13, and 0/13 for non-sensitizing class, respectively. Thus, at significance level (ϵ) = 0.15, compounds in external test set were classified as non-sensitizer (1), sensitizer (2) and sensitizer (3).

in Table 1, and visualized in Fig. 3. Starting with a confidence level of 0.8, thus accepting 20% prediction errors, valid models for both the sensitizer and the non-sensitizer class could be constructed (Fig. 3A), when taking random fluctuations into account (as described in Vovk et al. (2005)). In addition, the framework also allows users to adjust the desired level of validity and readily inspect the consequences of the actions, and increasing the confidence level to 0.85 also resulted in a valid model for both classes of compounds. Moreover, the consequence of increasing the confidence level from 0.8 to 0.85 is an increase in the percentage of *single* class labels (Fig. 3B–C). Thus, the result of using the higher confidence level is a slightly more efficient model with more *single* class predictions (92% and 90% for confidence levels 0.85 and 0.80, respectively) based on internal validation of the training dataset. Conformal predictions for each chemical from the internal validation of the training dataset are summarized in Supplementary Table S2.

3.2. Conformal prediction — prediction of the external test dataset

The internal validation exercise provided an estimate on the performance of the conformal prediction framework applied to the GARD assay, and as illustrated above, the procedure resulted in highly effective and statistically valid models at both the 0.8 as well as 0.85 confidence level. However, in order to rule out the possibility that the results observed with the training dataset were simply due to over-training, or to overestimation of the performance, an external test dataset containing 70 samples previously unseen by the model was subsequently classified within the framework (Supplementary Table S1). Statistics from conformal prediction classifications for the external test

dataset are shown in Table 1, and visualized in Fig. 4. Starting with the confidence level 0.8, the validity of the derived conformal predictor was compromised for both class labels, and far from being valid for any of the classes. However, increasing the confidence level to 0.85 resulted in a valid model (Fig. 4A). Furthermore, as indicated in Fig. 4B and C, similar to the training dataset, the consequence of increasing confidence level from 0.8 to 0.85 was an increase in the percentage of *single* class labels. At confidence level 0.8, 55/70 of compounds were assigned to a *single* class, corresponding to an efficiency of only 79%. Increasing the confidence level to 0.85 resulted in much better performance with respect to efficiency (99%, 69/70), with only one compound classified as *both*. Moreover, another consequence of increasing the confidence level was that 12 of the 14 compounds previously classified as *empty* at the confidence level 0.8 was now instead assigned to the correct *single* class (Fig. 4B–C, and Table 2). Accuracy for predictions was calculated on *single* class predictions, thus provided information on how well the model performed within the applicability domain. Individual accuracies, as well as balanced accuracies were high for both models, ranging between 83% and 100% (Table 1), demonstrating that also the model at confidence level 0.8 were able to provide accurate predictions for compounds assigned to an unique class, and hence determined to be within the applicability domain of the model.

3.3. Conformal predictions — instance based uncertainties

Until now, we have illustrated the use of the conformal prediction framework and how it can be incorporated into current GARD settings to generate statistically valid prediction regions, and how the training

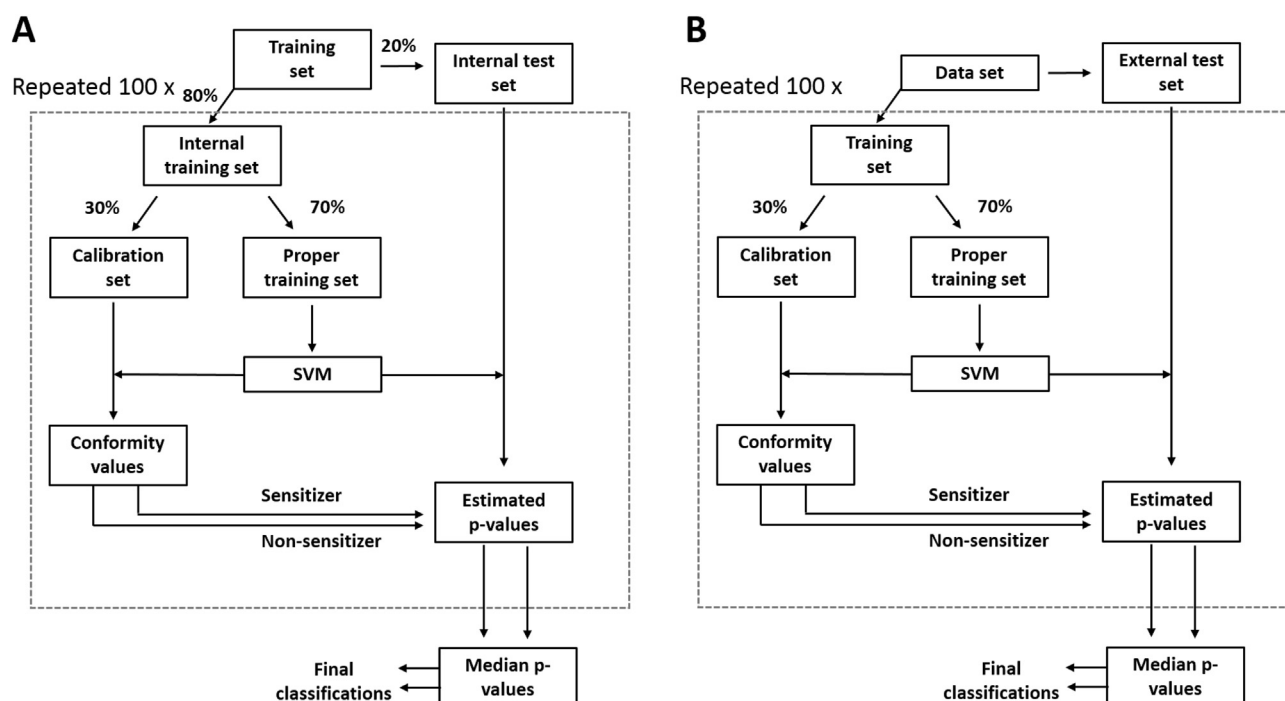


Fig. 2. Aggregated conformal prediction (ACP).

Class assignment of a certain compound was performed using median calculated conformal prediction p -values from an ensemble of 100 SVM models. (A) For the internal validation procedure, the training dataset was initially randomly divided into an internal training set (80%) and an internal test dataset (20%). SVM models were generated by repeatedly and randomly divide internal training dataset into *proper* training (70%) and calibration sets (30%), and the resulting conformity values from the 100 SVM models were summarized. SVM models were subsequently applied to assign conformity values also to samples in the internal test set, and final class assignments were performed by comparing the median conformal prediction p -value from the ensemble of SVM models for each compound and each class to corresponding p -values for the calibration set. (B) Similar strategy as described above was applied to perform validation of the external test dataset, except no subdivision of training dataset prior to model construction. Training dataset was divided into the ensemble of 100 *proper* training and calibration sets, which were subsequently applied for class assignment of the stimulations in the external test dataset, similar as described for internal validation of the training set.

Table 1

Conformal predictions on internal validation of training dataset and external test dataset at confidence levels 0.8 and 0.85.

Set	Conf- level	Chemicals (#)	Chemicals [empty] (#)	Chemicals [both] (#)	Efficiency (%)	Validity sens class	Validity non-sens class	Accuracy sens class	Accuracy non-sens class	Accuracy (balanced)
Training	0.8	38	4	0	0.895	0.778	0.900	0.875	1.000	0.938
Training	0.85	38	1	2	0.921	0.833	0.950	0.875	0.947	0.911
Test	0.8	70	15	0	0.786	0.685	0.625	0.841	0.909	0.875
Test	0.85	70	0	1	0.986	0.833	0.938	0.830	0.938	0.884

dataset can be applied to select and control the level of confidence to generate a valid and highly effective model where performance on the training dataset served as an indicator to future performance of test substances. Additionally, in contrast to conventional machine learning algorithms, the conformal prediction framework is instance based, thus providing a measure of uncertainty (p -value) for each classified compound in a test dataset. Table 2 summarizes the list of conformal prediction p -values, obtained by relating conformity values for compounds in the calibration set to the calculated conformity value for each individual compound in the test dataset. Statistical hypothesis testing was used to compare p -values to the user defined significance value. Thus, if the p -value for a certain prediction was $< \epsilon$, the classification was rejected. Three compounds with various outcomes were selected from Table 2: citral, pyridine and benzocaine, as a proof of concept of the interpretation of instance based measures of uncertainties. Starting with the 0.8 confidence level, only citral could be accurately classified as a sensitizer by the model, while pyridine and benzocaine were classified as *empty*, and were thus outside the applicability domain of the model. When the confidence level was adjusted to 0.85, benzocaine and pyridine ended up within the applicability domain of the model, and classified as *both* and sensitizer, respectively. Consequently, with

85% probability, pyridine and citral are most conforming to the sensitizers, and classified accordingly, while classification of benzocaine was still ambiguous (*both* class). Furthermore, conformal prediction p -values can almost always, if not identical for the two classes, be used to provide a single class prediction, although not statistically significant in relation to the validity of the model. The results of such a forced classification of compounds in above mentioned example was that all samples were (in this case correctly) classified as sensitizers.

4. Discussion

Development of accurate and reliable *in vitro* assays for replacement of animal toxicity testing for the endpoint of skin sensitization is a main priority within chemical risk assessment. Although literature describes several different strategies (recently reviewed in Ezendam et al. (2016)), the determination of applicability domain, *i.e.* under which circumstances predictions from a certain model are likely to be accurate is often neglected by test developers, as discussed by Roberts and Patlewicz (2017). Instead, predictive performances of such models are often only evaluated on how well they perform on external test sets, or through cross-validation procedures. Such strategies are based on the

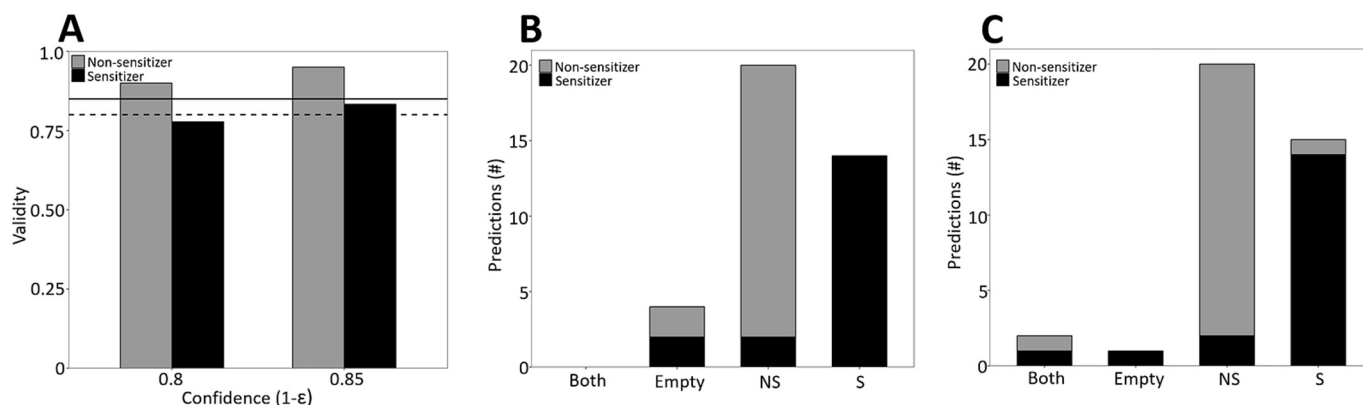


Fig. 3. Conformal predictions from internal validation of samples in training dataset.

Internal validation of the compounds in the training dataset ($n = 38$, sensitizers = 18, non-sensitizers = 20) was performed by initially dividing samples into an internal training dataset (80%), and an internal test dataset (20%). The internal training dataset was then iteratively divided into 100 *proper* training (70%) and calibration sets (30%), and samples in *proper* training sets were subsequently used to generate an ensemble of SVM models, which were in turn applied to assign conformity values, and for generating p -values for samples in internal test dataset. Median conformal prediction p -value from the 100 SVM models was used for class assignment in accordance with the set confidence level for each compound. (A) Valid models (prediction errors $\leq \epsilon$) were generated at confidence levels 0.8 (filled line) and 0.85 (dotted line) for both sensitizers (black) and non-sensitizers (grey) when taking statistical fluctuations into account. (B) At confidence level 0.8, 34/38 (4 *empty*, 0 *both*) compounds in training dataset were assigned to a single class, resulting in an efficiency of 90%. Accuracies were calculated based on *single* class predictions: sensitizers (S, colored in black) = 0.88 (14/16), non-sensitizers (NS, colored in grey) = 1.00 (18/18). (C) Efficiency was slightly higher at the 0.85 confidence level: 92% (1 *empty*, 2 *both*, 35 *single*) while accuracy was similar for sensitizers (0.88, 14/16), and slightly lower for non-sensitizers (0.95, 18/19).

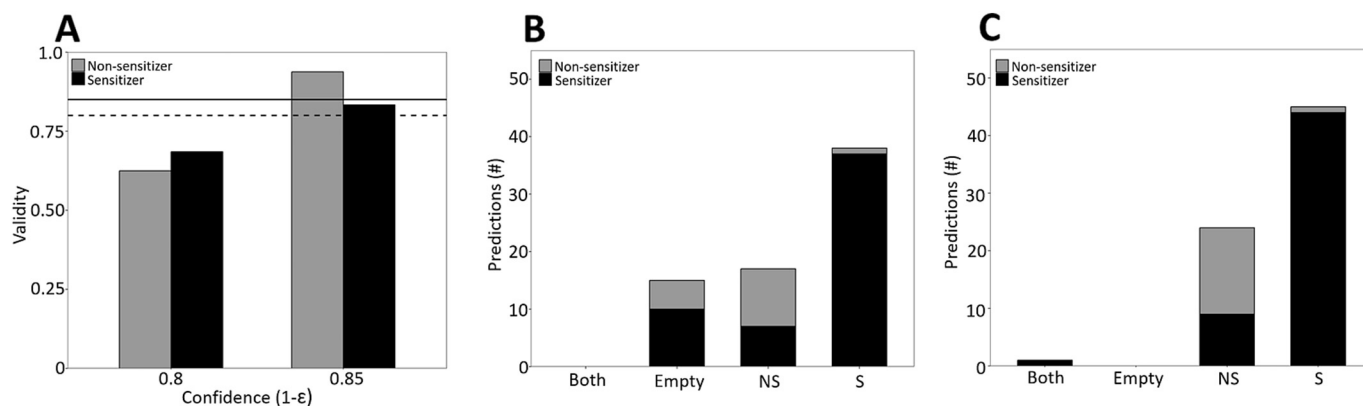


Fig. 4. Conformal predictions from external test set.

Prediction of the compounds in the external test dataset ($n = 70$, sensitizers = 54, non-sensitizers = 16) was performed by randomly divide samples in the training dataset into a *proper* training set (70%) and calibration sets (30%). An SVM model was trained on samples in the *proper* training set, and then first applied to assign conformity values to samples in calibration set, and then subsequently to samples in external test dataset. p -Values for each compound in the test dataset was generated by comparing percentage of compounds in the calibration set with a lower conformity value compared to the test compound. The process was iterated 100 times, to generate an ensemble of SVM models, and the median p -value from all models were used for final classification. (A) A valid model (prediction errors $\leq \epsilon$) was generated at confidence level 0.85 (validity sensitizer = 0.83, validity non-sensitizer = 0.94), while the model generated at confidence level 0.80 was not valid (validity sensitizer = 0.69, validity non-sensitizer = 0.63). (B) The percentage of single class predictions (model efficiency) was low at confidence level 0.80 (55/70), resulting in an efficiency of 79% (15 *empty*, 0 *both*, 55 *single*). Accuracies were calculated based on *single* class predictions: sensitizers (S, colored in black) = 0.84 (37/44) non-sensitizers (NS, colored in grey) = 0.91 (10/11). (C) Increasing confidence level to 0.85 resulted in more single class predictions, and consequently a more effective model (99%, 0 *empty*, 1 *both*, 69 *single*). Accuracy was similar for both sensitizers (0.83, 44/53) and for non-sensitizer (0.94, 15/16).

assumption that past performance is a true indicator of future performance, and although providing estimates on how well a model performs in general, it will not inform on the uncertainty involved in the predictions of individual instances. As a consequence, although chemical sensitizers are highly heterogeneous in terms of reaction mechanisms, and the majority of available *in vitro* assays are reductionist models providing little mechanistic information, they are often assumed to be applicable across the entire chemical space.

In contrast to the majority of *in vitro* strategies for identification of skin sensitizers, which are based on single or few markers, the GARD platform measure chemically induced changes across a genomic biomarker signature of 200 genes in a cellular model of dendritic cells, to ultimately arrive at more mechanistically relevant classifications. In this study, we introduce and demonstrate the application of the conformal prediction framework within GARD to complement current classifications with confidence predictions, and thus addressing the critical concerns of assessing uncertainty for each prediction, and for

estimating predictive boundaries of the assay. Conformal prediction is a framework based on the concept of validity, and a conformal predictor will always make valid predictions according to a set significance level. The framework outputs a set of labels (*single*, *both*, *empty*) where the probability of excluding correct label is guaranteed to be smaller than the significance level provided by the user. Thus, the guaranteed error level of a model is known up front, but it comes at the price that all predictions are not assigned a single class label prediction.

In a first attempt to implement the conformal prediction framework into the GARD assay, an internal validation procedure was performed on samples in the training dataset. The result showed that it was possible to construct valid models at acceptable levels of significance, indicating that the conformal prediction framework could be directly implemented into current GARD workflow. Since both confidence levels resulted in statistically valid models, this raised the question on how to select the most useful level of significance for future external test datasets. Increasing the confidence level has important consequences for

Table 2
Conformal predictions on external test dataset.

Treatment	Non-sensitizer (p-value)	Sensitizer (p-value)	Conformal classifications: significance level (ϵ)	
			$\epsilon = 0.15$	$\epsilon = 0.20$
Non-sensitizers				
1-Brombutane	0.227	0.114	NS	NS
α -Glucosylrutin	0.400	0.000	NS	NS
Benzoic acid	0.190	0.125	NS	Empty
Benzyl alcohol	0.355	0.000	NS	NS
Citric acid	0.209	0.133	NS	NS
Dextran	0.167	0.125	NS	Empty
Dimethyl CAPE	0.364	0.000	NS	NS
Dimethyl isophthalate	0.400	0.000	NS	NS
Hexane	0.190	0.133	NS	Empty
Hydrocortisone	0.000	0.667	S	S
Kanamycin A	0.238	0.105	NS	NS
Menthol	0.212	0.125	NS	NS
Phenoxyethanol	0.191	0.143	NS	Empty
Tartaric acid	0.167	0.143	NS	Empty
Vanillin	0.474	0.000	NS	NS
Xylene	0.250	0.114	NS	NS
Sensitizers				
2,4-Dinitrofluorobenzene	0.000	0.529	S	S
3-Methylcatechol	0.000	0.696	S	S
4-Methylaminophenol sulfate	0.000	0.273	S	S
4-Nitrobenzylbromide	0.000	0.308	S	S
Abietic acid	0.087	0.211	S	S
Amylcinnamyl alcohol	0.000	0.550	S	S
Anethole	0.000	0.294	S	S
Aniline	0.275	0.105	NS	NS
Anisyl alcohol	0.227	0.111	NS	NS
Benzalkonium chloride	0.000	0.313	S	S
Benzocaine	0.150	0.158	Both	Empty
Benzyl benzoate	0.100	0.200	S	S
Bisphenol A-diglycidyl ether	0.000	0.737	S	S
Butyl glycidyl ether	0.000	0.289	S	S
CAPE	0.000	0.250	S	S
Chloroanilin	0.400	0.000	NS	NS
Chlorpromazine	0.000	0.429	S	S
Cinnamaldehyde	0.000	0.938	S	S
Citral	0.000	0.750	S	S
Citronellol	0.000	0.474	S	S
Cyanuric chloride	0.136	0.158	S	Empty
Diethanolamine	0.238	0.111	NS	NS
Diethyl maleate	0.000	0.636	S	S
Diethylthiocarbamyl benzothiazole sulfide	0.000	0.824	S	S
Dimethylthiocarbamyl benzothiazole sulfide	0.000	1.000	S	S
Diphenylcyclopropanone	0.000	0.531	S	S
Glutaraldehyde	0.000	0.529	S	S
Glyceryl monothioglycolate	0.143	0.150	S	Empty
Hexyl salicylate	0.000	0.353	S	S
Hydrazine	0.000	0.294	S	S
Hydroquinone	0.000	0.706	S	S
Hydroxycitronellal	0.000	0.462	S	S
Imidazolidinyl urea	0.000	0.556	S	S
Iodopropynyl butylcarbamate	0.000	0.455	S	S
Isopropyl myristate	0.190	0.118	NS	Empty
Kanamycin	0.136	0.167	S	Empty
Lauryl gallate	0.105	0.188	S	Empty
Lilial	0.000	0.556	S	S
Limonene	0.167	0.143	NS	Empty
Linalool	0.095	0.200	S	S
Lylal	0.000	0.500	S	S
Methyl heptene carbonate	0.000	0.591	S	S
Methyldibromo glutaronitrile	0.143	0.158	S	Empty
Neomycin sulphate	0.130	0.158	S	Empty
p-Benzochinone	0.000	0.700	S	S
Pentachlorophenol	0.333	0.000	NS	NS
Phenyl benzoate	0.075	0.214	S	S

Table 2 (continued)

Treatment	Non-sensitizer (p-value)	Sensitizer (p-value)	Conformal classifications: significance level (ϵ)	
			$\epsilon = 0.15$	$\epsilon = 0.20$
Phenylacetaldehyde	0.000	0.391	S	S
Pinanediol	0.250	0.111	NS	NS
Propyl gallate	0.000	0.818	S	S
Pyridine	0.125	0.176	S	<i>Empty</i>
Taxifolin	0.095	0.200	S	S
Tetramethylthiuram disulfide	0.000	0.824	S	S
Triethanolamine	0.316	0.000	NS	NS

Compounds indicated in bold were falsely classified with incorrect single label by the conformal prediction model. Empty and both classifications are indicated in italic. NS = non-sensitizer, S = sensitizer

the model. Fewer compounds are allowed to fall outside the predictive boundaries in order for the model to still be valid, thus fewer compounds are designated to the *empty* category, while the opposite trend is observed for the *both* category. Thus, there is a balance between the number of compounds in the *empty* and *both* domain with respect to model validity, but from an efficiency point of view (*i.e.* the single class predictions), *both* and *empty* are equally uninformative. Consequently, there is also a trade-off between validity and efficiency, and a model with more *single* class predictions is considered more informative. In our case, increasing the confidence level from 0.8 to 0.85 resulted in a slightly more effective model, but since both confidence levels resulted in valid models, neither of the two models could be discarded at this point. These results demonstrate the flexibility of conformal prediction to adapt to the problem at hand, and the possibility to adjust confidence levels in order to increase efficiency, and to generate valid and usable models which are key concepts of the framework.

For a reliable *in vitro* classification model, it is important that internal validation of training data reflects future performance also on external test sets. In the presented study, classification of samples in the external test set was even more challenging due to the slightly imbalanced class distribution between sensitizers ($n = 54$) and non-sensitizers ($n = 16$). Thus, the model are built on evenly distributed classes, but applied to predict a test set with a different ratio, a situation referred to as the “sample selection bias” (Cortes et al., 2008), which often cause complications for building models with good forecasting capabilities. Indeed, the model at confidence level 0.8 was not valid and produced *single* class predictions for only 79% of the compounds, with a large number designated to the *empty* category. At this point, we had the option to increase the confidence level to 0.85, which was also indicated by the internal validation procedure to generate a more effective model. Indeed, when increasing the confidence level, a valid model with significantly more *single* class predictions was generated, and the accuracy for the *single* class predictions in the test dataset was similar to the accuracy for the training dataset for both sensitizers and non-sensitizers. These results highlights two important characteristics of conformal predictions: First, almost no decrease in model performance is observed between training dataset and test dataset, and second, the framework handles imbalanced dataset well, which is a consequence of the Mondrian scoring applied within the framework *i.e.* where different distributions are used for the binary classes, as previously discussed by Norinder and Boyer (2016).

Furthermore, if an *in vitro* model is to be useful for risk assessment, it is important to provide a measure on statistical reliability in predictions, and provide a warning if such reliability is not reached. As an integrated part of model development, conformal prediction produces an instance based measure of uncertainty for each classified compound, where the generated *p*-values serves as a measure on how different a test sample is to existing compounds in the model. In addition, *p*-values can be compared to a user defined confidence level for each binary class

label, and testing compounds failing to reach this level for any of the two classes are subsequently assigned to the *empty* class, thus considered nonconforming (dissimilar) in relation to existing compounds. Therefore, the mathematical framework applied within conformal predictions defines the predictive boundaries of the model, hence the definition of the applicability domain, *i.e.* the region where reliable predictions can be made, is also implicitly defined as an integrated part of model development.

Moreover, while *both* and *empty* classifications do not affect the efficiency of the model, these compounds may still provide important information. Since samples in the *empty* domain are by definition different from existing compounds in the model, these compounds can contribute with new important information. Thus, the model should be retrained to include these compounds, for future reliable predictions of similar compounds. In contrast, compounds in *both* domain are within the AD of the model, but predictors are not sufficiently informative to distinguish the two class labels. The solution to this problem is to include these samples in a new feature selection process, to identify and implement complementary predictors (*i.e.* genes in the GPS) able of achieving a better separation of the two class labels. Indeed, for the test set classified in this study, only one compound was classified in the *both* domain, and none in the *empty* domain (at confidence level 0.85), thus no immediate action needs to be taken at this time. As a final remark, these results also demonstrate that the multivariate genomic biomarker signature used as predictor variables in GARD are applicable across all tested chemical reactivity domains.

In conclusion, this study address the important issue of establishing applicability domain for *in vitro* assays by illustrating how the conformal prediction framework can be implemented directly to the GARD assay for identification of skin sensitizing chemicals. The benefits of using conformal predictions in this settings includes a straightforward interpretation of confidence: a confidence level of 0.85 means that no more than 15% errors are acceptable for the model to be valid, as well as a possibility to modify the confidence level and immediately inspect consequences of such actions. But most important, the framework transforms the SVM classifier into an instance based confidence predictor with an unambiguous definition for the applicability domain of GARD.

Conflicts of interest

All authors state no conflict of interest.

Transparency document

The <http://dx.doi.org/10.1016/j.tiv.2018.01.021> associated with this article can be found, in online version.

Acknowledgements

The work at LTH (AF, TL, ML) was supported by grants from the Swedish Research Council VR (521-2014-2286), Swedish Foundation for Strategic Research (IB13-0118), and the Faculty of Engineering (LTH). The research at Swetox (UN) was supported by Knut & Alice Wallenberg Foundation (2013.0253) and Swedish Research Council FORMAS (2016-02031).

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.tiv.2018.01.021>.

References

Akhavan, A., Cohen, S.R., 2003. The relationship between atopic dermatitis and contact dermatitis. *Clin. Dermatol.* 21 (2), 158–162.

- Aptula, A.O., Roberts, D.W., 2006. Mechanistic applicability domains for nonanimal-based prediction of toxicological end points: general principles and application to reactive toxicity. *Chem. Res. Toxicol.* 19, 1097–1105.
- Ashikaga, T., Yoshida, Y., Hirota, M., Yoneyama, K., Itagaki, H., Sakaguchi, H., Miyazawa, M., Ito, Y., Suzuki, H., Toyoda, H., 2006. Development of an *in vitro* skin sensitization test using human cell lines: the human Cell Line Activation Test (h-CLAT). I. Optimization of the h-CLAT protocol. *Toxicol. in Vitro* 20 (5), 767–773. <http://dx.doi.org/10.1016/j.tiv.2005.10.012>.
- Bauch, C., Kolle, S.N., Ramirez, T., Eltze, T., Fabian, E., Mehling, A., Teubner, W., van Ravenzwaay, B., Landsiedel, R., 2012. Putting the parts together: combining *in vitro* methods to test for skin sensitizing potentials. *Regul. Toxicol. Pharmacol.* 63 (3), 489–504. <http://dx.doi.org/10.1016/j.yrtph.2012.05.013>.
- Carlsson, L., Eklund, M., U., N., 2014. Aggregated conformal prediction. In: Lladis, L., Maglogiannis, I., Papadopoulos, H., Sioutas, Makris, C. (Eds.), *Artificial Intelligence Applications and Innovations*. Springer, Berlin Heidelberg, pp. 231–240.
- CLP (2016). CLP Regulation. <http://echa.europa.eu/sv/regulations/clp/> (accessed 16. 07.13).
- Cortes, C., Vapnik, V., 1995. Support-vector networks. *Mach. Learn.* 20, 273–297.
- Cortes, C., Mohri, M., Riley, M., Rostamizadeh, A., 2008. Sample selection bias correction theory. In: Freund, Y., Györfi, L., Turán, G., Zeugmann, T. (Eds.), *Algorithmic Learning Theory*. Springer, Berlin, Heidelberg, pp. 38–53. http://dx.doi.org/10.1007/978-3-540-87987-9_8.
- ECHA, 2008. Guidance on Information Requirements and Chemical Safety Assessment. Chapter R.6: QSARS and Grouping of Chemicals. Available at: https://echa.europa.eu/documents/10162/13632/information_requirements_r6_en.pdf/77f49f81-b76d-40ab-8513-4f3a533b6ac9.
- Enoch, S.J., Madden, J.C., Cronin, M.T., 2008. Identification of Mechanisms of Toxic Action for Skin Sensitization Using a SMARTS Pattern Based Approach.
- EU, 2006. Regulation (EC) No 1907/2006 of the European Parliament and of the Council of 18 December 2006 concerning the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH), establishing a European Chemicals Agency, amending Directive 1999/45/EC and repealing Council Regulation (EEC) No 793/93 and Commission Regulation (EC) No 1488/94 as well as Council Directive 76/769/EEC and Commission Directives 91/155/EEC, 93/67/EEC, 93/105/EC and 2000/21/EC. *Off. J. Eur. Union* L 396, 1–1355.
- EU, 2009. Regulation (EC) No 1223/2009 of the European Parliament and of the Council of 30 November 2009 on cosmetic products. *Off. J. Eur. Union* L 342, 1–59.
- Ezendam, J., Braakhuis, H.M., Vandebriel, R.J., 2016. State of the art in non-animal approaches for skin sensitization testing: from individual test methods towards testing strategies. *Arch. Toxicol.* 90 (12), 2861–2883. <http://dx.doi.org/10.1007/s00204-016-1842-4>.
- Fonacier, L.S., Dreskin, S.C., Leung, D.Y., 2010. Allergic skin diseases. *J. Allergy Clin. Immunol.* 125 (2 Suppl 2), S138–49. <http://dx.doi.org/10.1016/j.jaci.2009.05.039>.
- Gerberick, G.F., Vassallo, J.D., Bailey, R.E., Chaney, J.G., Morrall, S.W., Lepoittevin, J.P., 2004. Development of a peptide reactivity assay for screening contact allergens. *Toxicol. Sci.* 81 (2), 332–343. <http://dx.doi.org/10.1093/toxsci/kfh213>.
- Jaworska, J.S., Natsch, A., Ryan, C., Strickland, J., Ashikaga, T., Miyazawa, M., 2015. Bayesian integrated testing strategy (ITS) for skin sensitization potency assessment: a decision support system for quantitative weight of evidence and adaptive testing strategy. *Arch. Toxicol.* 89 (12), 2355–2383. <http://dx.doi.org/10.1007/s00204-015-1634-2>.
- Johansson, H., Lindstedt, M., Albrekt, A.S., Borrebaeck, C.A., 2011. A genomic biomarker signature can predict skin sensitizers using a cell-based *in vitro* alternative to animal tests. *BMC Genomics* 12, 399. <http://dx.doi.org/10.1186/1471-2164-12-399>.
- Johansson, H., Albrekt, A.S., Borrebaeck, C.A., Lindstedt, M., 2012. The GARD assay for assessment of chemical skin sensitizers. *Toxicol. in Vitro* 27 (3), 1163–1169. <http://dx.doi.org/10.1016/j.tiv.2012.05.019>.
- Johansson, H., Rydner, F., Kuhn, J., Schepky, A., Borrebaeck, C., Lindstedt, M., 2014. Genomic allergen rapid detection in-house validation—a proof of concept. *Toxicol. Sci.* 139 (2), 362–370. <http://dx.doi.org/10.1093/toxsci/kfu046>.
- Johansson, H., Gradin, R., Forreryd, A., Agemark, M., Zeller, K., Johansson, A., Larne, O., van Vliet, E., Borrebaeck, C., Lindstedt, M., 2017. Evaluation of the GARD assay in a blind cosmetics Europe study. *ALTEX*. <http://dx.doi.org/10.14573/altex.1701121>.
- Johnson, W.E., Li, C., Rabinovic, A., 2007. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* 8 (1), 118–127. <http://dx.doi.org/10.1093/biostatistics/kxj037>.
- Karlberg, A.T., Bergstrom, M.A., Borje, A., Luthman, K., Nilsson, J.L., 2008. Allergic contact dermatitis—formation, structural requirements, and reactivity of skin sensitizers. *Chem. Res. Toxicol.* 21 (1), 53–69. <http://dx.doi.org/10.1021/tx7002239>.
- Leek, J.T., Johnson, W.E., Parker, H.S., Jaffe, A.E., Storey, J.D., 2012. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* 28 (6), 882–883. <http://dx.doi.org/10.1093/bioinformatics/bts034>.
- Natsch, A., 2010. The Nrf2-Keap1-ARE toxicity pathway as a cellular sensor for skin sensitizers—functional relevance and a hypothesis on innate reactions to skin sensitizers. *Toxicol. Sci.* 113 (2), 284–292. <http://dx.doi.org/10.1093/toxsci/kfp228>.
- Natsch, A., Ryan, C.A., Foertsch, L., Emter, R., Jaworska, J., Gerberick, F., Kern, P., 2013. A dataset on 145 chemicals tested in alternative assays for skin sensitization undergoing prevalidation. *J. Appl. Toxicol.* 33 (11), 1337–1352. <http://dx.doi.org/10.1002/jat.2868>.
- Norinder, U., Boyer, S., 2016. Conformal prediction classification of a large data set of environmental chemicals from ToxCast and Tox21 estrogen receptor assays. *Chem. Res. Toxicol.* 29 (6), 1003–1010. <http://dx.doi.org/10.1021/acs.chemrestox.6b00037>.
- Norinder, U., Carlsson, L., Boyer, S., Eklund, M., 2014. Introducing conformal prediction in predictive modeling. A transparent and flexible alternative to applicability domain

- determination. *J. Chem. Inf. Model.* 54 (6), 1596–1603. <http://dx.doi.org/10.1021/ci5001168>.
- NRC, 2007. *Toxicity Testing in the 21st Century: A Vision and a Strategy*. National Academies Press, Washington DC.
- Nukada, Y., Ashikaga, T., Sakaguchi, H., Sono, S., Mugita, N., Hirota, M., Miyazawa, M., Ito, Y., Sasa, H., Nishiyama, N., 2011. Predictive performance for human skin sensitizing potential of the human cell line activation test (h-CLAT). *Contact Dermatitis* 65 (6), 343–353. <http://dx.doi.org/10.1111/j.1600-0536.2011.01952.x>.
- OECD, 2004. OECD Principles for the Validation for Regulatory Purposes of (quantitative) Structure-Activity Relationship Models. Available at: <http://www.oecd.org/chemicalsafety/risk-assessment/37849783.pdf>.
- OECD, 2007. Guidance document on the validation of (quantitative) structure-activity relationship models. In: OECD Series on Testing and Assessment 69 ENV/JM/MONO (2007)2.
- OECD, 2012. The adverse outcome pathway for skin sensitisation initiated by covalent binding to proteins, part 1: scientific evidence. In: Environment, Health and Safety Publications. Series on Testing and Assessment 168.
- Patlewicz, G., Casati, S., Basketter, D.A., Asturiol, D., Roberts, D.W., Lepoittevin, J., Worth, A.P., Aschberger, K., 2016. Can currently available non-animal methods detect pre and pro-haptens relevant for skin sensitization? *Regul. Toxicol. Pharmacol.* 82, 147–155.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: machine learning in python. *J. Mach. Learn. Res.* 12, 2825–2833.
- Piccolo, S.R., Sun, Y., Campbell, J.D., Lenburg, M.E., Bild, A.H., Johnson, W.E., 2012. A single-sample microarray normalization method to facilitate personalized-medicine workflows. *Genomics* 100 (6), 337–344. <http://dx.doi.org/10.1016/j.ygeno.2012.08.003>.
- R Development Core Team, 2014. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria URL. <http://www.R-project.org/> Accessed 17.08.16.
- Roberts, D.W., Patlewicz, G., 2017. Non-animal assessment of skin sensitization hazard: is an integrated testing strategy needed, and if so what should be integrated? *J. Appl. Toxicol.* <http://dx.doi.org/10.1002/jat.3479>.
- Tsujita-Inoue, K., Hirota, M., Ashikaga, T., Atobe, T., Kouzuki, H., Aiba, S., 2014. Skin sensitization risk assessment model using artificial neural network analysis of data from multiple in vitro assays. *Toxicol. in Vitro* 28 (4), 626–639. <http://dx.doi.org/10.1016/j.tiv.2014.01.003>.
- Tsujita-Inoue, K., Atobe, T., Hirota, M., Ashikaga, T., Kouzuki, H., 2015. In silico risk assessment for skin sensitization using artificial neural network analysis. *J. Toxicol. Sci.* 40 (2), 193–209. <http://dx.doi.org/10.2131/jts.40.193>.
- Vovk, V., Gammerman, A., Shafer, G., 2005. *Algorithmic Learning in a Random World*. Springer, New York. http://dx.doi.org/10.1007/0-387-25061-1_2.
- Zeller, K.S., Forreryd, A., Lindberg, T., Albrekt, A.S., Chawade, A., Lindstedt, M., 2017. The GARD platform for potency assessment of skin sensitizing chemicals. *ALTEX* 34 (4), 539–559. online first. [10.14573/altex.1701101](https://doi.org/10.14573/altex.1701101).