EXTERNAL SCIENTIFIC REPORT

APPROVED: 12 March 2019

doi:10.2903/sp.efsa.2019.EN-1598

Evaluation of the applicability of existing (Q)SAR models for predicting the genotoxicity of pesticides and similarity analysis related with genotoxicity of pesticides for facilitating of grouping and read across

Romualdo Benigni*, Chiara Laura Battistelli**, Cecilia Bossa**, Alessandro Giuliani**, Elena Fioravanzo***, Arianna Bassan***, Mojca Fuart Gatnik***, James Rathman****, Chihae Yang****and Olga Tcheremenskaia**

***Alpha-Pretox (www.alpha-pretox.com), Rome, Italy**

****Istituto Superiore di Sanità (ISS), Environmental and Health Department, Rome, Italy**

*****S-IN Soluzioni Informatiche Srl, Vicenza, Italy**

******Altamira, LLC, USA**

Abstract

To facilitate the practical implementation of the guidance on the residue definition for dietary risk assessment, EFSA has organized an evaluation of applicability of existing in silico models for predicting the genotoxicity of pesticides and their metabolites, including analysis of the impact on genotoxicity of the metabolic structural changes. The prediction ability of (Q)SARs for in vitro and in vivo tests were evaluated. For the Ames test, all (Q)SAR models generated statistically significant predictions, comparable with the experimental variability of the test; instead, the reliability of the (Q)SAR models for assays / endpoints different from in vitro bacterial mutagenicity appears to be quite far from optimality. Secondly, two new Read Across approaches were applied to predict Ames mutagenicity and in vitro Chromosomal Aberrations: Read Across was largely successful for predicting the Ames test results, but much less for in vitro Chromosomal Aberrations. The worse results for endpoints different from Ames may be attributable to the several revisions of experimental protocols and evaluation criteria of results that have made the databases qualitatively non-homogeneous, and poorly suitable for modelling. A third dimension of this research is the evaluation of the impact of the structural changes-in result of metabolic or degradation processes-on the genotoxic potential of the substances. Parent/Metabolite structural differences (beyond the known Structural Alerts) that may, or may not cause changes in the Ames mutagenicity were identified and catalogued. In addition, Structural Alerts analysis applied under human expert supervision permitted the rationalization of the large majority of the changes of patterns of genotoxicity. The findings from this work are suitable for being integrated into Weight-of-Evidence and Tiered evaluation schemes. The importance of the human expert knowledge is particularly emphasized.

Key words: dietary risk assessment, genotoxicity, in silico, metabolite, pesticide active substance, (Q)SAR, Read Across

Question number: EFSA-Q-2016-00340

Correspondence: pesticides.peerreview@efsa.europa.eu

Acknowledgements: Paola Leopardi is acknowledged for the continued scientific support. The generous cooperation of developers / vendors of QSAR software is gratefully acknowledged, namely Lhasa Limited, MultiCASE Inc., Leadscope Inc, Molecular Networks GmbH, ACD/Labs Inc., In Silico Toxicology GmbH. The support and collaboration of EFSA staff members Rositsa Serafimova and Juan Manuel Parra Morte is thankfully acknowledged.

Suggested citation: Romualdo Benigni, Chiara Laura Battistelli, Cecilia Bossa, Alessandro Giuliani, Elena Fioravanzo, Arianna Bassan, Mojca Fuart Gatnik, James Rathman, Chihae Yang, Olga Tcheremenskaia, 2019. Evaluation of the applicability of existing (Q)SAR models for predicting the genotoxicity of pesticides and similarity analysis related with genotoxicity of pesticides for facilitating of grouping and read across. EFSA supporting publication 2019:EN-1598. 221 pp. doi:10.2903/sp.efsa.2019.EN-1598

ISSN: 2397-8325

Summary

The use of pesticides on food and feed crops may lead to residues in edible parts of plants or animal products and hence results in exposure of the consumer to a mixture of compounds including the active substance and/or its metabolites, breakdown or reaction products. The terminal residues of pesticide active substances in food and feed commodities should be taken into account for risk assessment and need to be duly identified following the requirements of Commission Regulation (EU) No 283/2013. While a comprehensive toxicological dossier is developed for active substances, often none or only limited information about toxicological properties of their metabolites is available. Because of this and following the recommendations of the PPR Panel Scientific opinion, use of (Quantitative) Structure-Activity Relationship ((Q)SAR) models and read across is proposed for the assessment of genotoxic potential of all metabolites as a first step in the residue definition procedure.

To this aim, EFSA has implemented a procurement procedure to evaluate the different facets of the use of *in silico* models for predicting the genotoxicity of pesticide active substances and their metabolites. The results of this work are reported here.

A first dimension was the evaluation of the (Q)SAR models, based on a literature search and on the application of models to the new EFSA genotoxicity database on pesticides. This evaluation pointed to the existence of a very rich literature, including comparative prediction exercises, as well as more specific topics like combination of QSAR models, effect of Applicability Domain on the predictions, integration between models and expert knowledge. However, almost all studies in the literature focused on modelling the Ames test, and very little was found on the other genotoxicity assays or endpoints.

Based on recent literature (2010 – 2016 period), the abilities of software tools to predict the Ames test mutagenicity were comparable to evaluations previously published. Sensitivity ranged 0.72 to 0.96, and Specificity ranged 0.65 to 0.86, when the systems were applied to predict Ames mutagenicity results in the public dataset (retro-fitting, with a defined but variable percentage of the test chemicals also present in the model training sets). At the same time, the (Q)SARs showed a quite high variability when validated with different external test sets. These results indicate that the QSAR technology is good enough to fit existing data, but the coverage / representation of the chemical space needs further improvement. Other studies have considered the difference in predictivity within and outside the Applicability Domains of the models: the available evidence seems to indicate quite limited performance improvements within the Applicability Domain. Thus the predictions outside the Applicability Domain should not be dismissed as insignificant. Regarding combinations of tools, the analysis of the literature indicated that Sensitivity can be remarkably increased, but at the expense of a decrease in Specificity.

In parallel to the literature search, this work developed a large scale comparative prediction exercise. A wide range of commercial and publicly available (Q)SAR models were applied to the recently published EFSA genotoxicity database on pesticides and their metabolites. Five experimental assays were selected for the exercise: Bacterial Reverse Mutation Assay (Ames test), Mammalian Bone Marrow Chromosome Aberration Test, Mammalian Erythrocyte Micronucleus Test, *In vitro* Mammalian Chromosome Aberration Test*, In vitro* Mammalian Cell Gene Mutation Test.

Overall, the results of the comparative exercise pointed to a substantial difference between the performance in the prediction of the Ames test on one hand, and that of the other experimental assays on the other hand. For the Ames test, all (Q)SAR models generated statistically significant predictions: Sensitivity ranged between 46% and 71%, Specificity ranged between 66% and 98%. These results with the EFSA genotoxicity database confirm the statistically significant predictions reported in previous exercises in the literature. These values are comparable with the intrinsic experimental variability of the Ames data, and indicate the satisfactory level of the (Q)SARs for the Ames test.

On the opposite, the reliability of the (Q)SAR models for assays / endpoints different from in vitro bacterial mutagenicity (Ames) appears to be still quite far from optimality. As a matter of fact, the inspection of the ROC graphs of predictions shows that most models tend to be close to the diagonal line of random responses. There is no possibility of comparing these results obtained with the pesticides and metabolites database, with previous studies in the literature, since an extensive literature search

did not retrieve similar prediction exercises. Thus, this EFSA projects contributes with original information to the research on the predictivity of QSARs for genotoxicity endpoints different from bacterial mutagenicity.

Combinations of QSAR predictions -based on the EFSA pesticides genotoxicity results- confirmed the evidence from literature: as a general trend, the combination of QSARs increases Sensitivity, but at the expense of Specificity. On the other hand, predictions within and outside the Applicability Domain of the models do not seem to be drastically different.

Whereas QSARs have undergone during the years many performance evaluations, with special emphasis on comparative prospective exercises, nothing analogous can be found in the literature for Read Across. The literature is rich in proposals for general workflows and criteria, but the published examples of applications –even though often quite detailed- are limited in number and do not provide sufficient material for assessing the real predictive value of the proposed workflows.

Using the EFSA genotoxicity database, we performed around sixty Read Across exercises aimed at predicting the genotoxicity of metabolites from the information on the parent pesticide. The properties of the active substances are systematically documented in the dossiers submitted to the Regulatory Authorities, thus are the primary source of information on which to base the Read Across for their metabolites. When necessary, further information from a wider range of analogues was used.

Read Across was applied to both Ames and in vitro Chromosomal Aberrations assays, with two new strategies, based on different approaches and integrating different sets of information. A common result is that Read Across appears to be largely successful for predicting the Ames test results. The performance of the two strategies was partially different with in vitro Chromosomal Aberrations, but overall it was lower than that obtained with the Ames test.

The discrepancy between applications to the Ames test on one hand, and to the other assays / endpoints is in agreement with the results obtained in the evaluation of (Q)SARs. Since the structure-activity approaches used are identical, it is hypothesized that the poor performance with the non-Ames data depends on the different type and quality of biological data. Unlike the Ames assay, other in vitro genotoxicity assays may be subject to artifactual positive response, whose recognition by the scientific community has stimulated several revisions of protocols and evaluation criteria. Thus the available database of experimental results for studying QSARs and Read Across on these assays is not only remarkably smaller of that for the Ames test, but probably also of a lower quality since include data obtained under different conditions in different periods of time.

A third dimension of this research was the evaluation of the impact of the structural changes -in result of metabolic or degradation processes- to the genotoxic potential of the substances. One line of research concentrated on changes in the pattern of Structural Alerts. The information on Structural Alerts was not applied in an automatic way, but was filtered through human expert knowledge: this permitted the rationalization of the large majority of the patterns of genotoxicity in the subgroups of substances in which parent and (some) metabolites have different Ames outcomes. As a matter of fact, the supervision by the human expert permitted a better predictive performance in respect to automatic applications of rules. In addition, an extensive analysis of Parent / Metabolite structural differences -beyond the known Structural Alerts- was performed with chemoinformatics tools. This resulted in a list of structural changes that were catalogued into those related to changes in the Ames mutagenicity, and others which are neutral in this respect. The knowledge on these structural factors complements the knowledge on Structural Alerts, and may be used in combination in the assessment of the genotoxicity of metabolites.

Several remarkable take-home lessons were provided by this project. Among others, it provided the first comparative exercise on the QSAR modelability of assays / endpoints different from the Ames test, and indicated that further development is necessary, including better data curation. Critical areas that need further progress are also that of the Applicability Domain and models combinations. Read Across approaches in particular need the development of more objective criteria of predictivity, similarly to what happens with the QSARs.

Since both experiments and in silico methods are probabilistic in nature, the combined use of a wide array of tools within integrated testing strategies can surely increase the reliability and confidence in the assessment. In this project, we have evaluated several types of evidence (QSAR, Structural Alerts, Read Across, Structural Factors, Chemical Similarity) that can be used to assess the potential toxicity of metabolites. Integration of evidence –at the best of professional judgement -must take place in order to reach conclusions on the individual pesticides and metabolites. In this work, we provide examples of how the above factors can be combined for identifying categories of chemicals (e.g., potential negatives, potential positives, uncertain or border-line) that: a) may permit efficient evaluations of large numbers of chemicals for prioritization, through e.g., tiered approaches; or b) may provide an assessment of the individual chemicals (suitable also for further in-depth evaluations, if necessary).

Finally, a prominent result is that the integration of the (Q)SAR (including Structural Alerts) predictions with expert knowledge is a way to generate better predictions; the supervision role for the human expert knowledge in all evaluations has to be strongly emphasized.

Table of contents

1. Introduction

1.1. Background and Terms of Reference as provided by the requestor

This contract was awarded by EFSA to a consortium with Istituto Superiore di Sanità in the lead.

Members of the consortium are:

Istituto Superiore di Sanità

Romualdo Benigni

Soluzioni Informatiche (S-IN)

Altamira, LLC, USA was engaged as a sub-contractor of the consortium.

Contract title: Evaluation of the applicability of existing (Q)SAR models for predicting the genotoxicity of pesticides and similarity analysis related with genotoxicity of pesticides for facilitating of grouping and read across

Contract number: OC/EFSA/PRAS/2016/01

1.2. Interpretation of the Terms of Reference

Background as provided by EFSA

In order to facilitate the practical implementation of the guidance on the residue definition for dietary risk assessment (EFSA, 2016), an evaluation of applicability of existing (Q)SAR models and Read Across approaches for prediction of genotoxicity of pesticides and their metabolites, is needed. It would lead to the compilation of a list of recommended (Q)SAR models with the best performance and with the most reliable predictions of genotoxicity of pesticides active substances and their metabolites. This will be beneficial for the work of the risk assessors when applying the guidance for residue definition as well as in other areas of risk assessment of pesticides.

Read across is widely recognised as an alternative approach for toxicological data gaps filling. In the last years intensive efforts were done for development of different methodology for read across and to overcome the obstacles which are preventing the widespread use of the approach for regulatory purpose. Although some guidance documents have been published, a specific guidance for the application of read across for genotoxicity prediction is missing (ECHA, 2017a; b; OECD, 2017a; b). The molecular initiating events leading to genotoxicity i.e. covalent binding of substances to DNA and/or proteins are well studied and described. Many lists exist with structural alerts (SA) recognised as responsible for interaction of the chemicals with the biological macromolecules. Usually these lists are used for grouping of substances and read across, however it is also known that the remaining part of the molecule could affect the reactivity of these moieties and therefore should be taken into account when the similarity is evaluated as crucial part of the read across approach. Practical guidance on the assessment of the similarity, including (non) common SA, (non)common functional group, "core" structure and their relevance for the purpose of read across for prediction of genotoxicity is needed and would facilitate the systematic, objective and automatic evaluation of the similarity and consequently the justification and acceptability of the read across approach.

A procurement procedure to conclude a direct Contract for the execution of specific tasks has been implemented by EFSA. The Pesticides Unit requires:

- the performance of a critical review of existing (Q)SAR models for prediction of genotoxicity and evaluation of their predictability for pesticide active substances and their metabolites.
- the performance of a critical review of the existing methodologies of Read Across for prediction of genotoxicity.

- to investigate the combination of the resulting best methodologies for read across and (Q)SAR model(s) for the improvement of the overall predictability of genotoxicity
- the performance of an analysis of the impact of structural changes resulting from metabolic or degradation processes on the genotoxic potential of the substances.

1.3. Additional information

The activities of the procurement procedure are articulated in a number of specific Objectives, namely:

Objective 1:
To perform a comprehensive and critical review of the state of the art of (Q)SAR models for prediction of *in vitro* and *in vivo* genotoxicity. The review should cover free available and commercial models as well as literature models for the period from 1 January 2009 to 31 December 2016;

Objective 2:
To critically evaluate the performance (e.g. sensitivity, specificity, concordance, positive predictability, etc..) of existing (Q)SAR models (identified by carrying out objective 1) for prediction of genotoxicity (gene mutation, aneugenicity and clastogenicity) of pesticide active substances and their metabolites;

Objective 3:
To perform a comprehensive and critical review of the state of the art of the available methodologies and tools for performing Read-Across for the prediction of genotoxicity. The review should cover the period from 1 January 2006 to 31 December 2016;

Objective 4:
To critically evaluate the applicability and reliability of different methodologies (identified by carrying out Objective 3) for grouping and Read-Across for prediction of genotoxicity of pesticide active substances and their metabolites.

Objective 5:
Evaluation of the impact of the structural changes in the molecule in result of metabolic or degradation processes to the genotoxic potential of the substances.

2. Data and Methodologies

2.1. Scientific background: the present state of the art

This section presents the present state of the art in the fields of QSAR and Read Across, as evidenced from the literature of the recent years.

This literature review follows, and is complementary to the external scientific report done by the Joint Research Center (JRC) in 2010 (Worth et al., 2010). The work by JRC was comprehensive of the state-of-art situation, and surveyed a wide range of aspects, including not only the (Q)SAR models and studies, but also an historical perspective on the development of frameworks for assessing the usefulness of QSAR models in terms of the practical applicability of the models and the adequacy of the predictions; a survey of how QSAR analysis is used by national regulatory bodies and international advisory organizations in the field of food safety; case studies (research investigations) on the potential use of QSARs for genotoxicity and carcinogenicity, with a view to developing a conceptual framework for QSAR analysis that can be integrated with the application of the Threshold of Toxicological Concern (TTC) concept; identification of research and development needs, leading to recommendations for further activities aimed at promoting the uptake of computational methods in the food safety area. In the present review, we do not duplicate the content of the JRC review.

2.1.1. QSAR literature survey

A systematic search of literature on QSARs for the period 1 January 2009 to 31 December 2016 was performed. We have carefully selected only papers providing new evidence, and we have not included even excellent general presentations that do not present such novelties. We will only quote a general review on the application of QSAR to pesticides by NAFTA (NAFTA, 2012), that includes also the genotoxicity aspects. It presents general principles, together with the regulatory perspective. Other general papers are: (Barber et al., 2015; Barber et al., 2016; Modi et al., 2012; Patlewicz and Fitzpatrick, 2016; Rybacka et al., 2014; Wichard, 2017). However, the above papers do not add substantial information that was not already in the JRC work. In addition, we have not included papers focusing on very limited technical aspects, with no follow up or independent evaluation.

The search was carried out according to the principles of systematic reviews (see the EFSA 'Guidance on Application of systematic review methodology to food and feed assessments to support decision making'); a detailed documentation is provided in Appendix A. In particular, information regarding the identity of the databases/information sources (e.g. website), criteria for inclusion and exclusion of literature/model, are provided. In addition, the retrieved publications are documented as EndNote database (including the publications) and provided to EFSA.

This review of the literature is articulated as follows: a) Evaluations of QSAR methods through predictive exercises: b) Studies on specific research aspects.

List and description of Software tools available for genotoxicity assessment is provided in Appendix B.

2.1.2. Literature survey: Evaluations of QSAR methods through comparative exercises

Of particular value are evaluations based on comparative studies, where the systems are applied to predict the mutagenicity / genotoxicity of the same set of chemicals. For the period covered by this review, a number of well-designed exercises are available and they provide crucial information on the systems performance. It should be remarked however that these comparative studies virtually address only the prediction of the Ames test mutagenicity, whereas nothing comparable exists for the prediction of other genetic endpoints.

It should be remarked as well that development of the algorithms is a continuous process and so newer versions of applications may now be available. Thus, the present review and analysis represents a snapshot in time, with the data available in the literature today.

To provide a more understandable picture of the state-of-art of prediction models, we extracted quantitative data relative to crucial topics from different papers and put them into perspective. The presentation as Receiver Operating Characteristics (ROC) graphs provides an easy visualization. It plots True Positive Rate (Sensitivity) versus False Positive rate (1 – Specificity). The diagonal line indicates random results (Swets, 1988).

2.1.2.1. QSARs applied to publicly available Ames data sets

Ames data in the public domain range 6000 to 7000 chemicals tested. These data have been used as main source of information for training the QSAR predictive systems, and for evaluations of their performance. For example, (Jolly et al., 2015) compiled and curated an extended dataset of Ames results from publicly available sources such as the Vitic Nexus database by Lhasa Ltd. (Lhasa, 2014) and several literature sources (Feng et al., 2003; Hansen et al., 2009; Kazius et al., 2005), the Carcinogenic Potency Database (Fitzpatrick, 2008), and the Physician's Desk Reference (PDR) http://www.pdr.net/. On its turn, the Hansen's database contains data gathered from several sources, such as Chemical Carcinogenesis Research Information (CCRIS) https://toxnet.nlm.nih.gov/newtoxnet/ccris.htm (Feng et al., 2003; Helma et al., 2004; Kazius et al., 2005), VITIC (Judson et al., 2005), and the GeneTox databases https://toxnet.nlm.nih.gov/cgi-bin/sis/htmlgen?GENETOX. It appears that there is a large overlapping between the databases in the public domain.

Using the above data, virtually all important QSAR models were assessed and reported in three studies (Bakhtyari et al., 2013; Jolly et al., 2015; Valencia et al., 2013). The results on the performance of the systems are combined and presented in Figure 1. It should remarked that this can be defined as a case of internal validation (retro-fitting): the systems were applied to variable portions of the public data, which were also the original source of information for setting the QSAR models. For the statistical systems, it can be said that QSAR models predicted the same chemicals used as training sets; for the expert system, the public data were, in non-formalized ways, at the origin of the rules.

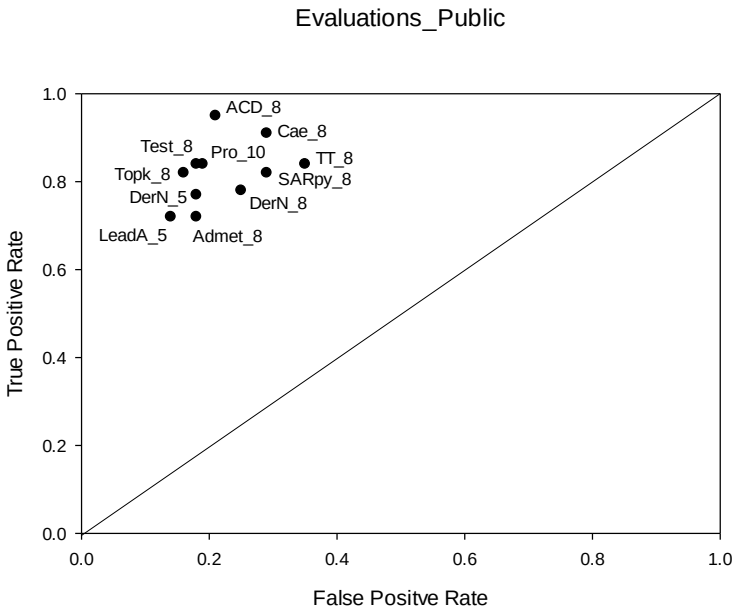


Figure 1: Different QSAR models were used to predict Ames mutagenicity using the publicly available data set (consisting of about 6000 chemicals tested)

Derek Nexus (derN_5) and Leadscope Model Applier (leadA_5) were assessed by (Jolly et al., 2015). ACD/ToxSuite 2.95 (acd_8), ADMET 6.06.0007 (admet_8), CAESAR VEGA 2.1.10 (cae_8), Derek Nexus 2.0 (derN_8), SARpy VEGA (sarpy_8), T.E.S.T. 4.0.1 (test_8), Topkat 3.1 (topk_8), and ToxTree 2.5.0 (tt_8) were assessed by (Bakhtyari et al., 2013). Prous Institute's Symmetry (pro_10) was assessed by (Valencia et al., 2013)

The inspection of the ROC graph shows that there are differences in the reported performance of the systems, with Sensitivity ranging 0.72 (ADMET) to 0.96 (ACD), and Specificity ranging 0.65 (ToxTree) to 0.86 (Leadscope Model Applier).

In spite of these differences, all the predictive systems are grouped in a well-defined area of the ROC space, irrespective of the technology used (either statistically-based, or based on expert knowledge). The area of the ROC space spanned by the predictive systems is well distanced from the diagonal, random performance line, thus pointing to statistically significant performance. The relative similarity of performances, irrespective of the method used, may be related to the fact that the same large body of publicly available data played a central role in the development of the systems.

2.1.2.2. QSARs applied to different databases: intra-system variability

In the case study shown above, the whole range of QSAR systems were applied to predict the Ames mutagenicity of chemicals belonging to the same database, i.e., the data in the public domain that have been largely exploited to train the QSAR models. Of special importance is the study of the variability of systems performance when applied to a range of different databases, particularly in-house databases not included in the collection of data in the public domain.

Figure 2 summarizes the results of a number of QSAR applications to in-house databases, retrieved from different papers (see references in the legend). The evidence is related to three popular QSAR systems (Derek Nexus, LeadScope Model Applier, MultiCase). It should be emphasized that in this case study: a) most of the databases are not in the public domain (mostly in house databases from industry), and b) the databases had not been used for training the systems. Thus, these results are examples of

external validation, and not retro-fitting as in the first case study shown above. For a comparison, we included also the retro-fitting performance of Derek Nexus (DerN5p) and LeadScope Model Applier (LeadA5p) with the large public database of the previous analysis.

A striking evidence is that, for the same QSAR model, the performance variability when applied to different databases may be much larger than that of different models when applied to the same database: see, for example, the extreme difference of two applications of MultiCase (MCase2 and MCase4), or those of LeadScope Model Applier (LeadA5h and LeadA5p).

It appears that the impact of the composition in chemicals of the test set is much stronger than that of the differences between systems when applied to the same database; this points to a problem of generalizability of the models.

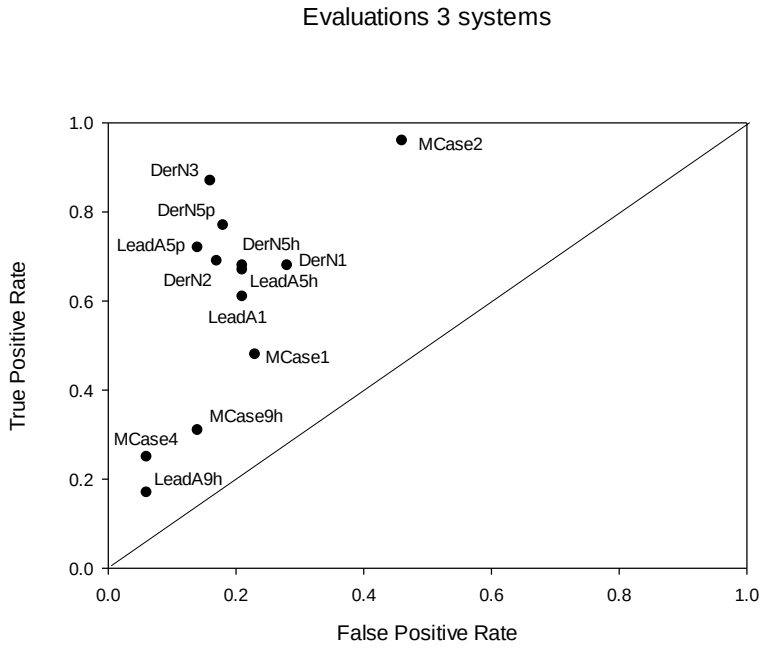


Figure 2: QSAR models were applied to different data sets to study their variability

In-house, not publicly available data sets were studied with Derek Nexus, Leadscope Model Applier, and MultiCase.
The specific models are:
a) Derek Nexus 3.01 (DerN_1), LeadScope Model Applier 1.6.0 (LeadA_1), and Case Ultra (Mcase_1) (Greene et al., 2015);
b) Derek Nexus 4.0.5 (DerN_2), and Case Ultra 1.4.6.6 (Mcase_2) (Araya et al., 2015);
c) Derek Nexus 3.0.1 (DerN_3) (Aiba nee Kaneko et al., 2015);
d) MultiCase 1.90 (Mcase_4) (Ono et al., 2012) ;
e) Derek Nexus 3.01 (DerN_5 h) and LeadScope Model Applier 3.1.1 (LeadA_5 h) (Jolly et al., 2015);
f) LeadScope Model Applier 1.2 (LeadA_9 h) and MultiCase MC4PC 21.0.99 (Mcase_9 h) (Hillebrecht et al., 2011).
For a comparison, applications of Derek Nexus 3.01 (DerN_5 p) and LeadScope Model Applier 3.1.1 (LeadA_5 p) (Jolly et al., 2015) are shown.

In another study on the variability of performance (Barber et al., 2016), the statistical prediction system Sarah Nexus (version 1.1) was validated against 14 private Ames test data sets supplied by nine pharmaceutical companies (Figure 3). The data sets sizes ranged 100 to 2100 chemicals. In agreement with the evidence above, the study pointed to a large variation of performance, with sensitivity ranging 0.3 to 0.7, and specificity ranging 0.7 to 0.9. The Authors argued that, in order to obviate the uncertainty factors, the predictions obtained from software tools should eventually be better supervised by expert judgement.

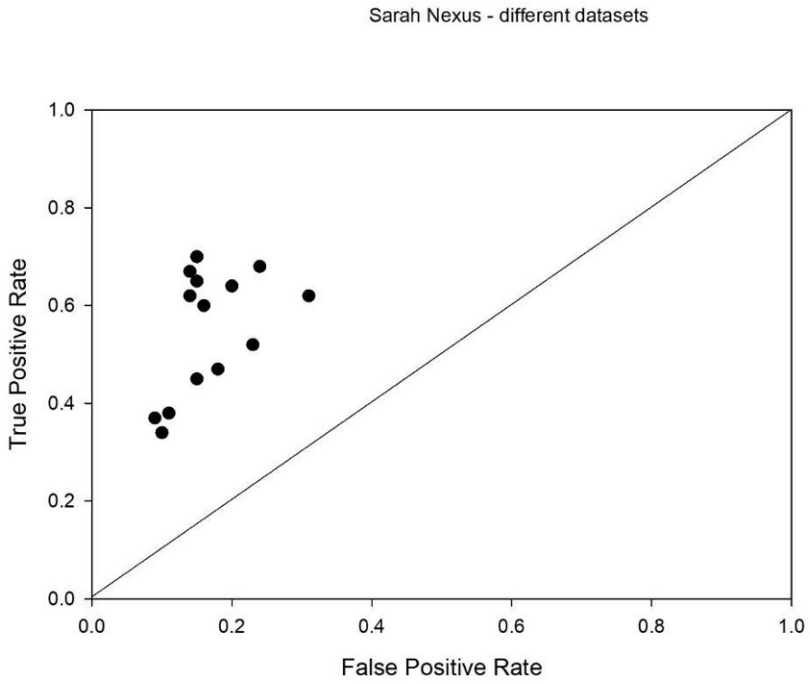


Figure 3: Sarah Nexus 1.1 was applied to different in-house data sets to study its variability (see details in the text)

2.1.2.3. Combinations of QSAR systems

Attempts have been made to see if the combination of results from more than one QSAR system improves the predictive performance. (Greene et al., 2015) have applied four QSAR systems (Derek Nexus, Sarah Nexus, LeadScope Model Applier, and MultiCase) to 801 intermediates and starting material for pharmaceutical synthesis, likely not in the training sets. In addition, three binary combinations of systems have been explored: Derek Nexus and Sarah Nexus, Derek Nexus and Leadscope, Derek Nexus and MultiCase (Figure 4). In the combinations, a conservative worst-case approach was adopted: a prediction is positive when either prediction is positive, whereas a prediction is negative when all predictions are negative. The Applicability Domain issue is automatically dealt with by the systems.

A first result is that the four systems, when applied to the same chemicals set, are in the same area of the ROC space and have relatively similar performance (thus confirming the results of the case study in Figure 1).

Regarding combinations, Figure 4 shows that the combination of QSARs predictions with the rule of the worst case increases remarkably the sensitivity, but at the expense of a decrease in specificity.

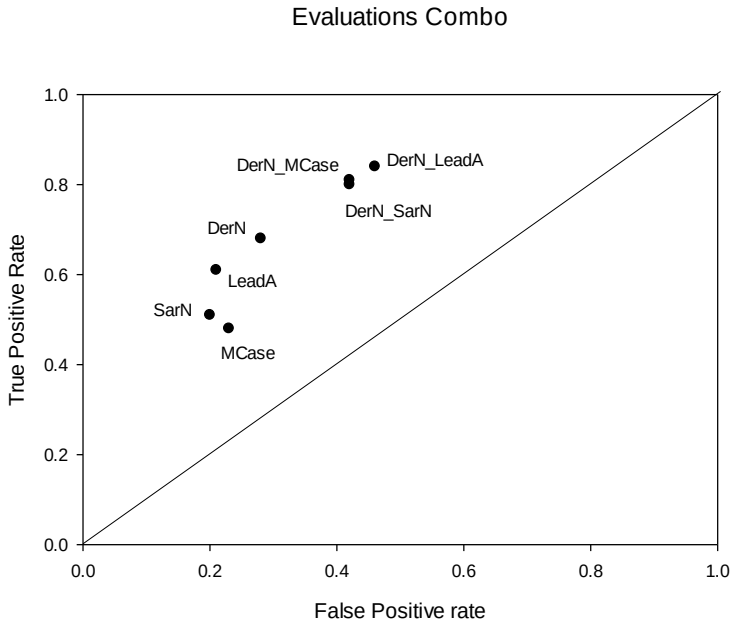


Figure 4: Derek Nexus 3.01 (DerN), Sarah Nexus 1.0 (SarN), LeadScope Model Applier 1.6.0 (LeadA), and Case Ultra (MCase_U) were applied to 801 in-house chemicals. In a following step, binary combinations of QSARs (DerN_MCase, DerN_SarN, DerN_LeadA) were applied to the same set of chemicals (Greene et al., 2015)

Other studies considered combinations of models, and arrived to similar conclusions on the balance between sensitivity and specificity. In particular, Contrera (Contrera, 2013), in a study within the scope of ICH M7 guideline (Assessment and control of DNA-reactive impurities in pharmaceuticals to limit potential carcinogenic risk (Amberg et al., 2016)), used an expert knowledge-based system (ToxTree) and an *ad hoc* developed statistical model (SciQSAR). The public nonproprietary 6489 compound Hansen benchmark mutagenicity data set was used as a validation data. The Toxtree validation specificity, sensitivity, concordance and false negative rate for this mutagenicity data set was 66%, 80%, 74% and 20%, respectively. This mutagenicity data set was also used to create a statistically-based SciQSAR-Hansen mutagenicity model. In a 10% leave-group-out internal cross validation study the specificity, sensitivity, concordance and false negative rate for the SciQSAR mutagenicity model was 71%, 83%, 77% and 17%, respectively. Combining Toxtree and SciQSAR predictions and scoring a positive finding in either software as a positive mutagenicity finding reduced the false negative rate to 7% and increased sensitivity to 93%, at the expense of specificity which decreased to 53%.

Combinations of models were explored also by Modi et al. (Modi et al., 2012). Three in house QSAR models were built using three different modelling techniques: (1) an in-house alert model; (2) a kNN approach (k-Nearest Neighbours); (3) a naïve Bayesian model (NB) using chemical features (e.g., physico-chemical, structural descriptors). A benchmark set of 6718 compounds from public sources, as well as in house data were used. The in house models were compared against two well-known alert models (DEREK and ToxTree), and against different combination methods: Categorical Bayesian Integration Approach (CBI), Partial Least Squares Discriminate Analysis (PLS-DA), and various simple majority vote combinations. Consensus predictions made by PLS-DA and Bayesian classification integration methods for integration of all five methods offers better predictivity and confidence as compared to simple voting integration. However, the predictivity of the consensus methods (around 90% accuracy) was no better than some of the individual QSARs composing the battery (e.g., 90% accuracy of the NB model).

2.1.2.4. Combinations of QSAR systems, plus Expert judgement

In another study (Araya et al., 2015) assessed 188 substances used in pharmaceutical production with Derek Nexus, Sarah Nexus, and MultiCase (Ultra version). In this case study, two systems (Derek Nexus and Sarah Nexus) showed mediocre sensitivity, whereas MultiCase had high sensitivity but mediocre specificity (Figure 5).

In a next step, the Authors complemented the outcomes of the combination of three systems with expert knowledge. Evaluation of the predictions followed a conservative approach, i.e., a prediction was regarded negative when the query molecule is devoid of a structural concern for mutagenicity using all three systems. Mutagenic potential was assumed when at least one positive prediction result was obtained, and the positive prediction could not be overruled by expert knowledge. Expert knowledge included the full exploitation of the predictive systems and also other approaches, as for example comparison to structurally similar substances.

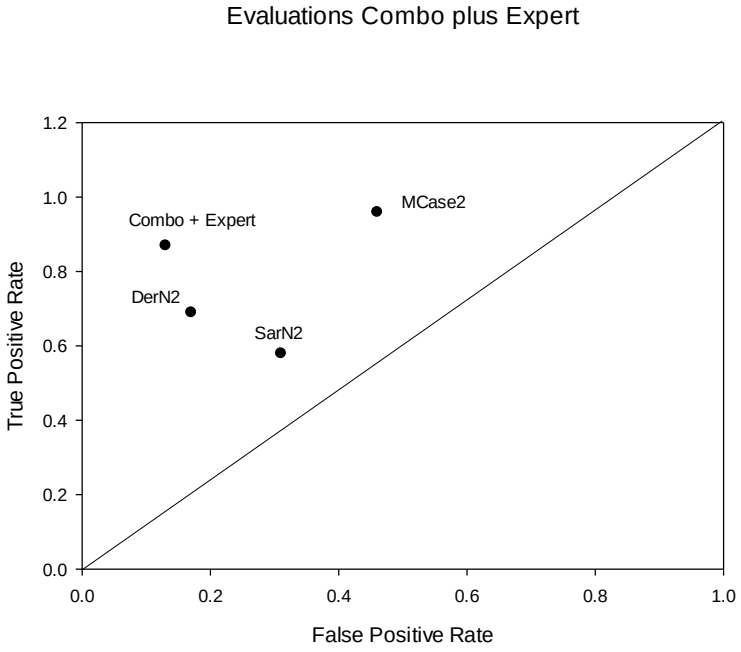


Figure 5: Derek Nexus 4.0.5 (DerN2), Sarah Nexus 1.1.2 (SarN2), and Case Ultra (MCase 2) were applied to 188 in-house chemicals. In a subsequent test, the three systems were combined together and with expert judgement (Combo+Expert) (Araya et al., 2015)

Figure 5 shows that the combination of the prediction systems, plus expert review of the predictions, generated a more equilibrated increase of both sensitivity and specificity.

2.1.2.5. Influence of the Applicability Domain, and of training / test set status

In a study, (Mombelli et al., 2016) used three models included in the VEGA system (Caesar, SARpy, and ToxTree as implemented in VEGA) to study differences in performance between retro-fitting application, and application only to chemicals simultaneously outside the training set and inside a rigorously defined Applicability Domain (Figure 6).

The models were first applied to the usual, large Ames database in the public domain (composed of around 6000 chemicals), used as training set for the various QSAR models.

In another step, chemicals outside the training sets were identified, and, out of them, were selected only those in the Applicability Domain. The Applicability Domain Index (ADI) implemented in VEGA is composed of a number of indexes (e.g., similarity index, concordance index, accuracy index, and atom-centered fragments index). The application of ADI also included a final supervision by experts.

Figure 6 shows that the consideration of ADI and of training / test set affects performance: however the differences are not particularly large, and are not always in the same direction. For example, the performance of ToxTree is improved both as sensitivity and specificity, whereas the other two systems have increased sensitivity but decreased specificity within the AD.

In the specific case of ToxTree, the ISSCAN database was assumed to be the training set, and the AD was calculated on this basis. Rigorously speaking, this is not entirely correct, since ToxTree rules are not derived from a specific database, but from expert knowledge.

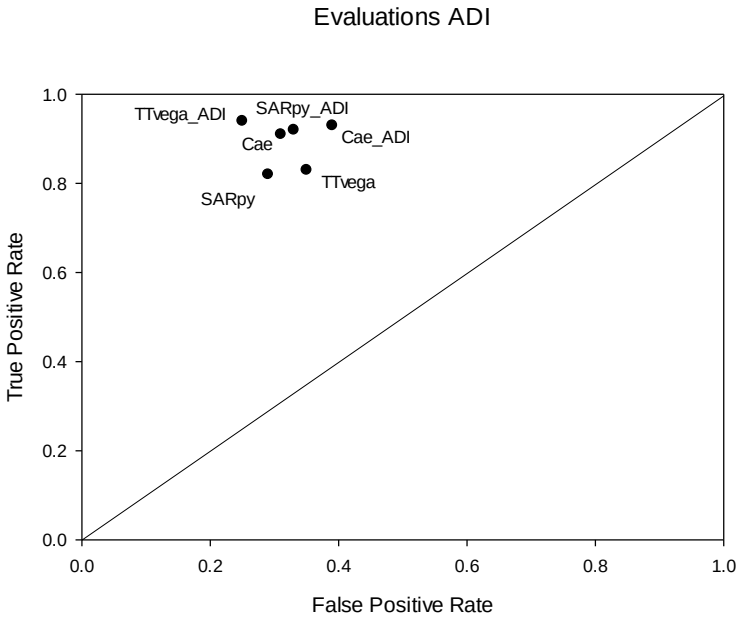


Figure 6: Caesar (Cae), SARpy, and ToxTree (TTVega) as implemented in VEGA were applied to the public Ames data base. In a subsequent step, the systems were applied only to the chemicals in the Applicability Domain defined by the VEGA algorithm (_ADI) (Mombelli et al., 2016)

2.1.3. Literature on specific research topics

This section considers papers presenting methodological developments of (Q)SAR approaches.

In the field of statistical models, Xu et al. (Xu et al., 2012) reported on the development of five new machine learning methods, namely support vector machine (SVM), C4.5 decision tree (C4.5 DT), artificial neural network (ANN), k-nearest neighbors (kNN), and naïv e Bayes (NB), along with five fingerprints, namely CDK fingerprint (FP), Estate fingerprint (Estate), MACCS keys (MACCS), PubChem fingerprint (PubChem), and substructure fingerprint (SubFP) aimed at predicting Ames mutagenicity. The training set consisted of 7617 diverse compounds, including 4252 mutagens and 3365 nonmutagens. On the basis of this data set, high predictive models were then built. Performances were measured by cross validation and an external test set containing 831 diverse chemicals. Information gain and substructure analysis were used to interpret the models. The accuracies of fivefold cross validation were from 0.808 to 0.841 for top five models. The range of accuracy for the external validation

set was from 0.904 to 0.980. Three models (PubChem-kNN, MACCS-kNN, and PubChem-SVM) showed the highest predictive accuracy.

Several papers reported studies aimed at improving the Structural Alerts.

Ahlberg et al. (Ahlberg et al., 2014) presented a fully automatic method that highlights significant substructures for toxicologically important data sets. The method identifies important substructures by computationally breaking chemical structures into fragments and analyzing those fragments for their contribution to the given activity by the calculation of a p-value and a substructure accuracy. The method is intended to aid the expert in locating and analyzing alerts by automatic retrieval of alerts or by enhancing existing alerts. The method has been applied to a data set of Ames mutagenicity results and compared to the substructures generated by manual curation of this same data set as well as another computationally based substructure identification method. The authors reported that this method can retrieve significant substructures quickly, that the substructures are comparable and in some cases superior to those derived from manual curation, that the substructures found covers all previously known substructures, and that they can be used to make reasonably accurate predictions of Ames activity.

Another approach to identify automatically Structure Alerts was reported by Ferrari et al. (Ferrari et al., 2013). The algorithm (called SARpy) generates substructures of arbitrary complexity, and the fragment candidates to become SAs are automatically selected on the basis of their prediction performance on a training set. In SARpy, fragmentation is done directly on the SMILES notation of structures. This approach has been tested on a large public database of Ames results, and showed marked prediction skills and pointed to the knowledge already collected in the literature as well as to new alerts. This suite of alerts was implemented in the VEGA software platform.

Liew et al. (Liew et al., 2012) used ensemble modelling of mixed structural / physical chemical features to develop a model able to classify the metabolic activation of chemicals into covalently reactive species (starting from a training set of chemicals with known effects). Compounds which produce reactive metabolites that form GSH-, protein-, or DNA-adducts were included in this study. A total of 663 1D and 2D molecular descriptors were calculated. An ensemble model of 13 naive Bayes classifiers was built from a diverse set of 1,479 compounds. The ensemble model was validated internally with five-fold cross validation and it has achieved sensitivity of 67.4% and specificity of 93.4% when tested on the training set. The final ensemble model was made available for public use. In the study, it was found that there was only a small overlap between the structural alerts for metabolic activation of chemicals into covalently reactive species when compared with the structural alerts for covalent DNA binding.

Another study incorporating knowledge on metabolic pathways was reported by (Kamath et al., 2015). A knowledge-based approach, combining information from Structural Alerts of SARpy with metabolic triggers generated based on simulation with the CRAFT software, was developed. The new model was externally validated to predict mutagenicity *in vitro* of chemicals, which were predicted unknown by SARpy. This model has a higher accuracy than the SARpy model, with an accuracy of 89% for the training set and 75% for an external validation set.

A novel method that automatically extracts potential structural alerts from a data set of molecules, through the identification of emerging graph patterns, was reported by Metivier et al. (Metivier et al., 2015). The method automatically outputs a manageable number of structural patterns that are strongly related to mutagenicity. A part of the resulting structures corresponds to already known structural alerts. The reported accuracy is around 72%.

A series of works were aimed at improving models through the use of sophisticated computer chemistry approaches.

Ford et al. (Ford et al., 2017) performed a comparative analysis on how accurately 11 routinely used *in silico* programs correctly predicted the mutagenicity of selected compounds (n=20) that contained either bulky or electron-withdrawing substituents. The eleven *in silico* programs were evaluated and compared: Derek for Windows, Derek Nexus, Leadscope Model Applier (LSMA), LSMA featuring the *in vitro* microbial Escherichia coli–Salmonella typhimurium TA102 A-T Suite (LSMAþ), TOPKAT, CAESAR,

TEST, ChemSilico (±S9 suites), MC4PC and a novel in house DNA docking model. The presence of bulky or electron-withdrawing functional groups in the vicinity of a mutagenic toxicophore in the test compounds clearly affected the ability of each *in silico* model to predict non-mutagenicity correctly. This was because of an over reliance on the part of the programs to provide mutagenicity alerts when a particular toxicophore is present irrespective of the structural environment surrounding the toxicophore. The DNA docking model was the most sensitive model evaluated, suggesting that this approach could usefully complement other approaches.

Snyder (Snyder, 2010) focused on the prediction of clastogenicity. Whereas Ames mutagenicity assay is highly reproducible and, for the most part, transparent, the same is not necessarily true of *in vitro* mammalian chromosome aberration assays, since there is a fairly large and growing number of molecules without clear structural genotoxicity alerts (DEREK, MCASE), which are negative in Ames testing but positive in aberration studies, often only at high concentrations and / or cytotoxicity. The Snyder's paper suggests that non covalent drug/ DNA interactions, which are not adequately modeled in computational programs, may help explain some of these unexpected positive results. In particular, it is suggested that N-dimethyl groups and certain pyridine/piperidine aryl ketones may contribute to genotoxicity, perhaps via DNA intercalation and topo-isomerase inhibition. Clastogenicity arising from topo-isomerase inhibition would be expected to be a threshold phenomenon and to have a different risk relative to clastogenicity associated with covalent drug/DNA interactions.

The above work had a follow-up by Snyder et al. (Snyder et al., 2013). The earlier studies were extended by examining a series of over 1,350 drugs for their ability lo noncovalently bind to different DNA sequences using two computational programs: Autodock and SurAex. These drugs were also evaluated for binding to the crystallographic ATP-binding site of human topoisomerase II. The results obtained point to multiple series of noncovalent DNA binding structure activity relationships which would not have been predicted based on cursory structural examination. Many drugs within these series are genotoxic although not via any commonly recognized structural covalent alerts. The study confirmed previously implicated features such as N-dialkyl groups and specific N-aryl ketones as potential genotoxic chemical moieties acting through noncovalent mechanisms.

2.1.4. Recapitulation of the main literature results on QSARs applications

The present literature review continues and updates the previous one performed for EFSA by the Joint Research Center (JRC) (Worth et al., 2010), that exhaustively addressed the genotoxicity QSAR subject up to 2010. The research performed for the present review (2010 to 2016) has shown that no substantial, groundbreaking novelties have been proposed since then. This period of time has mainly witnessed refinements of the predictive systems, based on collection of larger training sets and on continued fine-tuning (with the introduction of few new predictive QSAR systems), and an increased interest for the use of QSAR by regulatory authorities.

A major new fact in the area is the recently developed ICH-M7 guideline (Assessment and control of DNA-reactive impurities in pharmaceuticals to limit potential carcinogenic risk (Amberg et al., 2016)), that allows the use of the *in silico* approach to predict Ames mutagenicity for initially assessing impurities in pharmaceuticals. This is the first international guideline addressing the use of QSAR models *in lieu* of an actual toxicological study for human health assessment. The guideline requires the use of two complementary approaches, an expert rule-based method and a statistical algorithm. In addition, the guidance states that the output from these computer-based assessments can be reviewed using expert knowledge to provide additional support or resolve conflicting predictions. This approach is designed to maximize the sensitivity for correctly identifying DNA reactive compounds while providing a framework to reduce the number of compounds that need to be synthesized, purified and subsequently tested in an Ames assay (Amberg et al., 2016; Greene et al., 2015).

The attention by Regulatory Authorities has strongly stimulated improvements of the QSAR systems. In particular, many attempts have been made to derive (possibly new) Structural Alerts in an "automatic" way, and to fine-tune the existing Alerts with more sophisticated modeling approaches. Particular efforts have been devoted to model non-DNA reactive interaction with the genetic material, that might give

rise to chromosomal damage and for which reliable Alerts do not exist yet. All these interesting attempts may contribute to future improvements of the predictivity of the (Q)SAR models.

In parallel with the above refinement activity, several new comparative prediction exercises have evaluated the presently available (Q)SAR models, and thus have provided an updated snapshot of the present state-of-art in the field.

It should be emphasized that virtually all the reported predictive exercises focus on the Ames test, whereas information is lacking for other assays / endpoints.

Overall, the abilities of software tools to predict Ames test mutagenicity were comparable to previously published evaluations. Sensitivity ranged 0.72 (ADMET) to 0.96 (ACD), and Specificity ranged 0.65 (ToxTree) to 0.86 (Leadscope Model Applier), when the systems were applied to predict Ames mutagenicity results in the public dataset (retro-fitting, with a defined but variable percentage of the test chemicals also present in the model training sets) (Figure 1).

However, the systems showed a quite high variability when validated with different external test sets: Sensitivity ranged 0.17 to 0.96, and Specificity 0.54 to 0.94 (Figures 2 and 3). This variability of response affects –to different degrees- all the (Q)SAR models studied. Taken together, these results indicate that the QSAR technology is good enough to fit existing data, but the coverage / representation of the chemical space is still to be improved.

Other studies have considered the difference in predictivity within and outside the formalized Applicability Domain of the models. However, the available evidence seems to indicate quite limited performance improvements within the Applicability Domain (Figure 6). Thus the predictions outside the Applicability Domain should not be dismissed as insignificant.

Combinations of tools have been explored as well. When a simple conservative, worst case approach was adopted (a prediction is positive when either prediction is positive, whereas a prediction is negative when both predictions are negative), Sensitivity was remarkably increased, but at the expense of a parallel decrease in Specificity (Figure 4). On the other hand, Figure 5 shows that the combination of the prediction systems, plus expert review of the predictions, generated an equilibrated increase of both Sensitivity and Specificity. Expert review included consideration of similar chemicals, critical evaluation of experimental mutagenicity data of identified analogues, and Read-across (Araya et al., 2015; Greene et al., 2015; Mombelli et al., 2016)).

Even though outside of the temporal limits of this literature survey, it should be mentioned that in 2014 the Division of Genetics and Mutagenesis, National Institute of Health Sciences (DGM/NIHS) of Japan has launched the Ames/QSAR international collaborative project. DGM/NIHS has the largest Ames mutagenicity database, containing approximately 12,000 new chemicals that have not been previously used for developing QSAR models. These Ames data were provided to developers for a prospective prediction exercise, aimed at finally improving the QSAR models. The exercise was recently completed (Honma et al., 2018). It appears that all tools were considerably improved with this extended database.

2.1.5. Read Across Literature Survey

In this section, firstly we shortly summarize the literature on the general principles of Read Across, and then we illustrate a number of specific applications to genotoxicity. However, while the literature on QSAR applications to genotoxicity is quite extended, very few Read Across applications are reported. Since Read Across is still an evolving area, methodology is the primary interest, and papers deal mostly with how it should be conducted. QSAR is a science already well developed, so the main interest is not on principles but in applications. Accordingly, we adopt here a different style of presentation and give more emphasis on the methodology of Read Across.

Read Across is a method that estimates the potential toxicity of an untested substance based on structurally or functionally similar substances with known toxicity information. The principle of the Read Across technique is that endpoint or test information for one or more substances (called analogue or

source) is used to predict the same endpoint or test for another / other substance(s) (called target), the latter being considered to be similar by scientific justification. In principle, Read Across can be used to estimate physicochemical properties, toxicity, environmental fate, and ecotoxicity.

Read-Across can be performed in different ways to fill data gaps: a) One-to-one (one analogue used to make an estimation for a single chemical); b) One-to-many (one analogue used to make estimations for two or more chemicals); c) Many-to-one (two or more analogues used to make an estimation for a single chemical); d) Many-to-many (two or more analogues used to make estimations for two or more chemicals). In cases c) and d), the analogues are said to form a category.

The similarity or analogy (to be intended in a broad sense, not only as structural similarity) between target and analogue(s) may be based on the following: a) a common functional group (e.g. aldehyde, epoxide, ester, specific metal ion); b) common constituents or chemical classes; c) similar carbon range numbers; d) commonality in positions of double bonds within the same basic molecular skeleton; e) an incremental and constant change across the category (e.g. a chain-length category); f) similar values of physical chemical parameters, e.g., LogP, Energy of the Lowest Unoccupied Molecular Orbital (LUMO), Energy of the Highest Occupied Molecular Orbital (HOMO); g) the likelihood of common precursors and/or breakdown products, via physical or biological processes, which result in structurally similar chemicals (e.g. the metabolic pathway approach of examining related chemicals such as acid/ester/salt). The observation of a quantitative trend (increasing, decreasing, or constant) in the experimental data for a given endpoint across chemicals in a category can also be used as the basis for interpolation or extrapolation (i.e., trend analysis). Similarity from more than of the criteria above obviously strengthen the validity of the analysis.

Although there has been a number of technical guidance developed (ECHA, 2017a; b; OECD, 2017a; b) which describe the workflow of category/analogue development and associated read-across, many challenges still remain. Uncertainties on the consistency in how Read Across predictions are made, and the level of evidence required to substantiate a read-across prediction and document its justification persist, thus thwarting greater acceptance of Read Across for regulatory purposes

Many researchers are working to address these challenges and to provide practical indications on how to perform Read Across, and to permit the implementation of the above quoted guidance documents. These efforts have led to a vision which is largely accepted in its general principles; however, the hands-on practice can vary very much on a case-by-case basis, to reflect differences in data availability, types of chemicals, mechanisms of toxic action, as well as regulatory requirements. The illustration of the following generic workflow for Read Across is freely based on the following excellent papers: (Madden, 2013; Patlewicz and Fitzpatrick, 2016; Patlewicz et al., 2017; Wu et al., 2010).

2.1.5.1. The Category/analogue workflow

According to the literature quoted in the previous paragraph, key steps in the development of a category or analogue approach are as follows:

1. Decision context
2. Data gap analysis
3. Overarching similarity rationale
4. Analogue identification
5. Analogue evaluation
6. Data gap filling
7. Uncertainty assessment

Decision context

The first step is a consideration of the decision context. Decision contexts can take many forms including prioritisation, screening level hazard assessment, or risk assessment. The type of decision will dictate the level of uncertainty that can be tolerated with the Read Across prediction being made. For example, a prioritisation decision for a target chemical can tolerate more uncertainty than a risk assessment outcome considering the downstream consequences of the decision context.

Data gap analysis

This step refers to a data collection exercise for the target chemical to understand what is known from a hazard perspective in order to be able to prioritise next steps and to determine whether any Read Across approach should be broadly based in scope, or limited to a specific endpoint.

Overarching similarity rationale for the category/analogue approach

The data gap analysis for a target chemical should inform the most practical and pragmatic means of identifying source analogues. For example, according to the OECD guidance for what a category represents, if the overarching rationale is a common functional group or structural similarity, this will focus the tactical approach of identifying analogues. If, for example, the data gap analysis shows that the only gap is for a single endpoint, such as genotoxicity, then a more targeted search strategy might be applied to identify analogues on the basis of their common reaction mechanistic domains.

Analogue identification (Analogue searching)

Analogue identification is the process of searching for analogues similar to the target chemical. The overarching similarity rationale dictates how this search is conducted practically. A search on the basis of structural similarity where a similarity index such as the Jaccard distance (Tanimoto coefficient) is used as a convenient threshold to limit the number of source analogues retrieved would be categorised as an 'unsupervised' approach. A search that is informed by parameters relevant to the endpoint (e.g. a specific structural alert) on the other hand, would be categorised as a supervised approach. Combination of both approaches can be envisaged as well.

Analogue evaluation

After a search of source analogues has been performed, a critical step is to evaluate the validity and relevance of these analogues. Source analogues with limited data, and particularly for the endpoint(s) of interest required for the target chemical, are not viable candidates for further consideration. Source analogues should be evaluated in terms of their similarity relative to the target chemical specifically with respect to their general physicochemical characteristics, metabolic profile and reactivity. A preliminary indication of the relative similarity can also be made by reference to existing (Q)SAR tools. QSAR tools can be particularly helpful to provide an estimate of physicochemical characteristics such as LogKow, molecular weight (MW) and vapour pressure, all of which will be informative in assessing bioavailability. Tools that can identify structural alerts will be helpful to judge whether the toxicity profile of the source analogues relative to the target chemical are likely to be similar. Other tools exist that are able to make predictions of likely metabolites which provide an indication of whether metabolic pathways diverge or converge to any extent.

Data gap filling

This step requires an evaluation of the validity of the analogues with respect to their actual experimental data, and judging the concordance and consistency of their effects across the members and across the

endpoints. The data evaluation itself is largely expert driven. In a category approach, the prediction made is either based on expert judgement using one or more of the source analogues, or objectively estimated by mathematical calculation using the source analogues. Depending on the type of property data under consideration, the Read Across prediction could be qualitative or quantitative.

Uncertainty assessment

Although this step has not been systematically or consistently performed in practice, there are two main approaches: expert-driven or data-driven. Expert-driven approaches rely on the judgement of domain scientists/experts to evaluate the relevance of the analogues as well as their underlying data. A framework was proposed by Blackburn and Stuard (Blackburn and Stuard, 2014) which describes potential areas of uncertainty, and provides a questionnaire to help assign a level of uncertainty using qualitative scores. This framework was adapted and extended by Schultz et al. (Schultz and Cronin, 2017) whereby templates were proposed to assist in assessing similarity in the context of chemistry, toxicokinetics and toxicodynamics as well as to guide the systematic characterisation of uncertainty both in the context of the similarity rationale, the Read Across data, and overall approach and conclusion.

The scientific challenges concern the preparation of scientifically valid and robust read-across justifications that build on the knowledge of the presumed Mode of Action (MOA) driving the endpoint(s) under consideration. If the justification for Read Across is not robust or poorly characterised, there is a risk that hazards will be mis-represented either too conservatively, or not conservatively enough. For some endpoints, such as Ames mutagenicity, skin/eye irritation or skin sensitisation, the presumed MOA has been reasonably established and structural rules/profilers have been encoded in (Q)SAR models or in tools such as the OECD Toolbox. Electrophilicity is well known to be an important factor in driving mutagenicity and carcinogenicity (Miller and Miller, 1981a; b). On this basis, a wide range of tools (SAs and QSARs) have been developed. These provide many options to address data gaps for mutagenicity. Models for Ames are to an extent well developed and could be used to make estimates of mutagenicity without recourse to experimental testing. However, their utility is even more pronounced in substantiating a Read Across by providing the alert information to demonstrate commonality in reaction mechanism. For other endpoints, particularly repeated dose toxicity, adequate mechanistic information may be unavailable. In these cases Absorption, Distribution, Metabolism and Excretion (ADME) information as well as information on other endpoints can be helpful in substantiating the Read Across justification developed (Patlewicz et al., 2013; Patlewicz et al., 2017).

The following section describes case studies found in the literature. Criteria for the literature search are given in Appendix C.

List and description of available software for Read Across are given in Appendix D.

2.1.6. Main literature results on genotoxicity Read Across case studies

An extensive literature search of Read Across applications to genotoxicity was performed for this project, however it generated a very limited number of published individual case studies, and almost no comparative study such as those available for the assessment of QSARs.

A number of studies were reported by the Research Institute for Fragrance Materials (RIFM). Several assessed fragrance ingredients lacked mutagenicity studies, so suitable analogs were looked for (Api et al., 2016a; b; c; Api et al., 2016d; Api et al., 2016e; Api et al., 2017). They were all 1:1 Read Across analyses. The analogy was established based on structural similarity, reactivity, metabolism data, physico-chemical properties. The QSARs and profilers in the OECD QSAR Toolbox were extensively used, including metabolism prediction. The targets and analogs showed similar alerts for DNA binding, mutagenicity, genotoxicity and Oncologic classification. The identified Read Across analogs were confirmed by using expert judgment, and their mutagenicity data were used to fill the data gaps. The above publications refer to cases in which the similarity was quite un-questionable.

Another case study on the Polyethylene Glycol (PEG) cocamines was reported (Skare et al., 2015). These are tertiary amines with an alkyl group derived from coconut fatty acids and two PEG chains of varying length. Toxicology (genotoxicity, and systemic or developmental/reproductive toxicity with use in cosmetics) data gaps for the PEG cocamines was addressed by Read Across based on structure–activity relationship using the framework described by (Wu et al., 2010) for identifying suitable structural analogs. Analogs with toxicological data were identified by searching an in-house database developed by the Procter & Gamble Company with more than 800,000 chemicals linked to toxicological data. SciFinder, ToxNet, Scopus and Google searches were used as well. For each of the PEG cocamines selected as potential analogs, information to justify the categorization was evaluated. This evaluation included chemical structure and structural alert identification using Derek for Windows™ and TIMES prediction models, chemical reactivity assessed by expert judgment and a comparison of physicochemical properties. The potential metabolic transformations for the PEG cocamines in comparison to the analogs were also evaluated. This process involved medicinal chemistry expert judgment along with a literature and database search for related chemicals that could assist in understanding the expected metabolic pathways. Genotoxicity data for structural analogs supported the conclusion that the PEG cocamines of interest are non-genotoxic.

A small comparative exercise on Read Across was conducted within the CALEIDOS LIFE project (Benfenati et al., 2016). The participants were invited to assess the hazard posed chemicals, applying in silico methods and Read Across approaches. The exercise focused on three endpoints: mutagenicity, bioconcentration factor and fish acute toxicity. Nine chemicals were assigned for each endpoint and the participants were invited to complete a specific questionnaire communicating their conclusions. The platforms used were very different, sometimes in combination: OECD QSAR Toolbox, VEGA, Toxread, T.E.S.T., Toxtree, ChemID Plus +eChemportal, Leadscope. When only one program was used, the OECD QSAR Toolbox was the most used. For the analysis of the same chemical, there was a higher rate of disagreement between different users of the Toolbox than between different users of ToxRead. Only one chemical was positive, and the predictions were remarkably skewed towards false positives. Given the limitation in number of case studies and skewed distribution of positives / negatives, the comparative exercise was more informative on the use of the different tools by the users, than on the value of the Read Across approaches employed.

In this section, also an OECD manual for applying the OECD QSAR Toolbox workflow to the Read Across analysis of mutagenic / genotoxic compounds: "Strategies for grouping chemicals to fill data gaps to assess genetic toxicity and genotoxic carcinogenicity" http://www.oecd.org/chemicalsafety/risk-assessment/46985336.pdf should be mentioned The usefulness of the manual is that it applies the workflow to a number of specific cases studies, and proposes in detail how the tools in the Toolbox can be used in different situations starting from an initial characterization of chemicals and continuing through a series of refinements of the analysis, based on the extensive experience of the Toolbox Authors / Developers. It should be emphasized as well that different strategies may be applied, depending on the specific characteristics of the case study.

2.2. The EFSA genotoxicity database

In 2014, EFSA has commissioned the compilation of a database specific for the pesticide residues including active substances and their metabolites, which comprises different genotoxicity endpoints, i.e. point mutations, structural and numerical chromosome aberrations, and DNA damage.

Data collection on individual genotoxicity studies has been retrieved from regulatory toxicological reports (Draft or Renewal Assessment Reports, i.e. DARs or RARs, respectively) as provided by the Rapporteur Member State (RMS) during the pesticide peer review process at European Level. The final EFSA conclusion on the overall genotoxic potential of active substance or metabolites taking into account all available information is not included in the database.

The database contains identity and genotoxicity information on more than 290 active substances and a large number of their metabolites (around 600).

The database represents a practical tool to complement *in silico* tools i.e. QSAR (Quantitative structure–activity relationship models), grouping and read across for prediction of the genotoxicity hazard of the pesticides residues, and it supposes to enlarge the chemical domains for their application.

The database can be freely downloaded: https://data.europa.eu/euodp/data/dataset/database-pesticide-genotoxicity-endpoints. Recently, it has been implemented in the OECD QSAR Toolbox as well.

Since the EFSA genotoxicity database is central to the analyses required for this project, it has been the subject of curation and preliminary analyses by the Consortium. These include the characterization of the chemical typology of the compounds, especially in relation to other existing databases, and of the patterns of genotoxicity data.

After curation and characterization, the EFSA genotoxicity database was used for Objectives 2, 4, and 5 analyses, i.e., prediction exercises with commercial and publicly available QSAR systems, Read Across exercises, and investigations on the impact of chemical features on the genotoxicity outcomes.

The curated data used for the analyses are provided in Annex 1.

The version of the database used by us for the analysis is different from the publicly available one: because of confidentiality reasons, some of the data available to us are not included in the public version. In some of the following sections, confidential data in the tables are undisclosed.

2.2.1. Chemical typology of the EFSA genotoxicity database

The first step was the curation of the EFSA genotoxicity database (EFSA DB).

From the initial dataset containing 1,109 compounds, the following compounds were excluded: 41 inorganics and organometallics, 39 compounds where no structural representation was available, 3 compounds where the compound was listed as "representative compound" in the column title "structure shown". Next, the entries in the database with duplicated identifiers were grouped, first by COM_ID (91 COM_IDs grouped into 34) and second, by SMILE notations (22 duplicated SMILES grouped into 9), Thus, the removal and arrangement steps resulted in the removal of 153 structures, thereby leaving 956 chemicals with unique SMILEs notations in the EFSA DB. This dataset was ready for subsequent analyses.

2.2.1.1. Physicochemical space analysis, and comparison with a reference dataset of pesticides

An analysis of chemical space was performed to evaluate the representativeness of the EFSA genotoxicity dataset in respect to a broader "pesticides" chemical space, and to assess the chemical space described by the pesticide active compounds and their metabolites. For the purpose of this analysis -in addition to a dataset of pesticide active substances (DS1) and metabolites (DS2) from the EFSA genotoxicity database- a separate "Reference dataset" (DS3), including a vast number of pesticides derived from different sources (e.g. OpeFood Tox DB and http://www.alanwood.net/ (available as an excel file supplementary to (Richard et al., 2016)) was compiled.

Error! Reference source not found.1 gives the number of structures belonging to the dataset of a ctive substances (DS1) and to the dataset of metabolites (DS2) in the genotoxicity EFSA DB, and the number of structures in the "Reference dataset" (DS3) representing pesticides, are given. Additionally, the number of structures in common between different datasets is reported.

Table 1: Number of compounds, and number of overlapping compounds present in the analysed datasets

	DS1	DS2	DS3	

DS1	349	0	340	Dataset of active substances
DS2		607	37	Dataset of metabolites
DS3			1667	Reference dataset

In short, the chemical space of the datasets was defined as ranges of selected physico-chemical descriptors and was characterised by the following, predicted, physical chemical properties i.e. HDonors, HAcceptors, Freely Rotatable Bonds (FRB), Molecular Weight (MW), Polar surface Area (PSA), Octnol-water partitioning coeficient (LogP), Dissociative Partition Coefficient (LogD). The DS1, DS2 and DS3 datasets were investigated also by means of Principal Component Analysis (PCA).

The analysis demonstrated that the three datasets overlap to a considerable extent, thus confirming that the pesticide EFSA genotoxicity dataset is representative of a broader "pesticides" chemical space, having the selected physical chemical / structural characteristics consistent with those of the reference pesticide dataset.

The details of this analysis are presented in Appendix E.

In conclusion, the chemical spaces of the three datasets overlap, thus confirming the representativeness of the EFSA genotoxicity DB of the pesticides and emphasising its wider representativity in the evaluation of the applicability of QSARs to pesticides.

2.2.2. Congenericity of the EFSA genotoxicity database

In addition to the chemical characterization of the EFSA DB presented in the previous section, the database was also characterized for its congenericity. This aspect is particularly important in the use of the database for extracting structure-activity rules.

The general experience on structure-activity relationships indicates that the optimal condition for identifying structural factors that influence a biological activity is when different structural changes are observed in a set of congeneric chemicals sharing the same basic skeleton, with different substituents attached to it. This applies to QSARs, that are most informative and sound when are derived for congeneric series of chemicals. This applies as well to the identification of the modulating factors of Structural Alerts (SA)(Benigni et al., 2007). It is important that there is a sufficient number of chemicals to represent the different structural motifs, or different values of physical chemical properties. For example, in (Benigni et al., 2009) rodent carcinogenicity data were available for around 70 aromatic amines; this large representation of structures permitted a detailed analysis of the modulating factors of the activity, that thus complemented the knowledge on structural alerts.

A systematic analysis of congenericity in the EFSA DB was performed by studying the existence of classes of very similar compounds. In addition, a comparison was made between EFSA and ISSSTY databases for this aspect. ISSSTY is an implementation of the publicly available Ames test database (7367 chemicals (Benigni et al., 2008)) It is available also in the OECD QSAR Toolbox.

In practice, the analysis was performed with the OECD QSAR Toolbox that provides different options for grouping chemicals within the Category Formation functionality. The Clustering option with a cut-off of 70% Dice chemical similarity was applied.

It appears that 55% of EFSA chemicals were in clusters with at least 70% chemical similarity, giving rise to relatively small clusters. The maximum numerosity was n = 9, with 10 clusters of up to 7 elements (Figure 7).

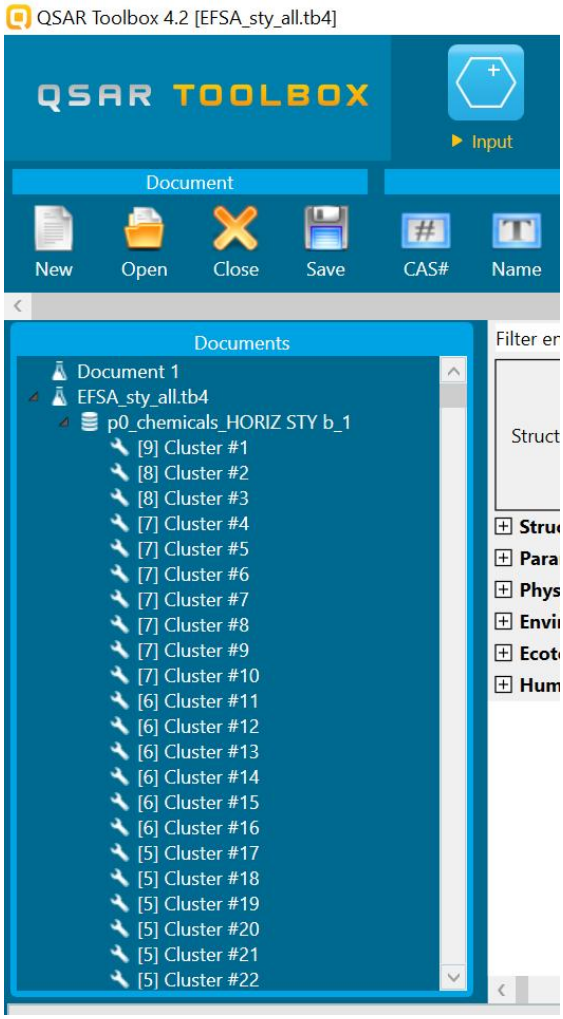


Figure 7: Snapshot of the clustering results for the EFSA genotoxicity database, performed with the OECD QSAR Toolbox

For a comparison, Figure 8 shows the clustering results for the ISSSTY database (in the interval 275 – 300 Molecular Weight (MW)). Here the percentage of chemicals not classified in clusters is similar (42%), but the cluster numerosity is remarkably higher (up to 32 elements, with many clusters with more than 20 elements). Similar results were obtained in the other MW intervals (results not shown).

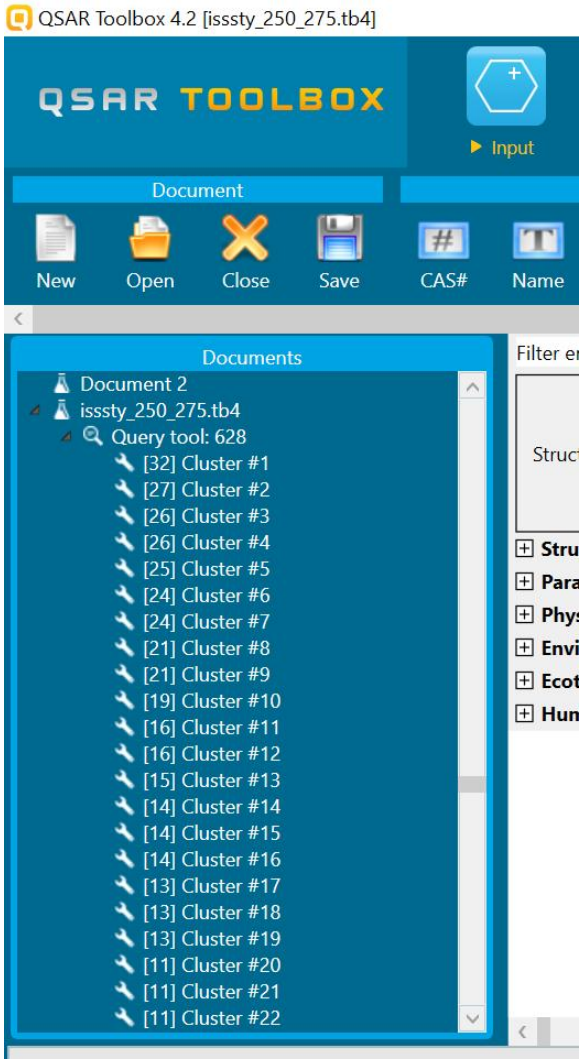


Figure 8: Snapshot of the clustering results for the ISSSTY database, performed with the OECD QSAR Toolbox

Overall, the above comparison with the historical Ames database (in ISSSTY) emphasizes the fact that the EFSA database may provide a more limited possibility for a systematic study of the modulating effects of substituents on basic chemical skeletons, and is a caveat for what can be expected from such a study.

2.2.3. The genotoxicity assays results: Overall outcomes generation, and descriptive analysis

The EFSA Genotoxicity Database reports the genotoxicity results in a granular form, i.e., all the experimental results for every e.g. strain, test repetition, etc.. without generating final summary outcomes for the individual chemicals. In order to provide a descriptive statistics, and to construct a basis for assessing the predictive ability of the (Q)SAR models and Read Across procedures, we have preliminarily generated Overall Outcomes for every chemical in every test system from the granular results in the EFSA DB.

The criteria followed for building the Overall Outcomes was a conservative, worst case approach.

In particular:

In vitro assays, without and with metabolic activation: positive if either result is positive (for Ames test, consisting of several strains without and with metabolic activation: positive if one strain result is positive.

In vivo assays: no metabolic activation is applicable. A special case is that of the Mammalian erythrocyte micronucleus test: negative results without evidence of interaction with target cells were not considered (based on the variable INVIVOTISSUEEXP in the EFSA DB).

Experiments repeated with more systems (e.g., *in vitro* mammalian cells gene mutation), or repeated with the same system(s), or contradictory results: judgement by Expert inspection according to a conservative, worst case approach. Decisions were taken on a case-by-case basis taking into account all available information, including the variables ReportAuthor , and ReportTitle in the EFSA database. The latter information helped to evaluate the tradition and expertise of the authors and laboratories that had generated the experiments.

The Overall Outcomes are compiled in Annex1.

Table 2 reports the descriptive statistics of the Overall Outcomes for the ten most represented assays. It appears that the numbers of assayed chemicals vary considerably from test to tests, and that the proportion of positives to negatives is highly unbalanced, with a prevalence of negatives for every assay.

Table 2: Overall Outcomes for the ten most represented assays: the number of chemicals with results (N) and the proportion of positive results (% Pos) are reported

Assay	N	% Pos
in vivo chromosome aberration assay	128	0.18
in vitro unscheduled DNA synthesis in mammalian cells	247	0.06
in vivo unscheduled DNA synthesis	132	0.003
in vitro mammalian cell micronucleus test	39	0.46
in vitro mammalian chromosome aberration test	592	0.25
in vitro mammalian cell gene mutation assay	608	0.11
in vivo Mammalian erythrocyte micronucleus test	236	0.09
in vitro single cell gel/comet assay in mammalian cells	37	0.46
in vitro sister chromatid exchange assay in mammalian cells	95	0.51
in vitro bacterial reverse mutation assay	990	0.04

Five assays were selected for the subsequent predictivity analysis (Table 3) based on the combination of the following criteria:

- the assays evaluated had an OECD technical guidance reported in the respective EFSA dataset;
- the OECD TG reported for the specific assay does not have a status "deleted" in the OECD *"Guidance document on Revision to OECD Genetic Toxicology Test Guidelines (2015)"*;
- a software/model for the prediction of the specific assay is available;
- the assay is relevant for the assessment of gene mutation and chromosome aberrations endpoints;
- there were experimental records available for more than 100 substances.

The selection –based on the combination of the above criteria, together with practical experience in regulatory work- permitted to focus the analysis on QSARs that model assays representative of the range of genotoxicity endpoints and that are highly employed in genotoxicity profiling of chemicals.

Table 3: Gene mutation and chromosome aberration assays selected for further analysis.

Guideline	Method	Test Type
OECD TG 471 and TG 472	*in vitro*	Bacterial Reverse Mutation Assay
OECD TG 475	*in vivo*	Mammalian Bone Marrow Chromosome Aberration Test
OECD TG 474	*in vivo*	Mammalian Erythrocyte Micronucleus Test
OECD TG 473	*in vitro*	*in vitro* Mammalian Chromosome Aberration Test
OECD TG 476	*in vitro*	*in vitro* Mammalian Cell Gene Mutation Test

For the five selected assays, Table 4 reports the number of chemicals with defined SMILES notation, hence suitable for being treated with the QSAR software systems.

Table 4: Number of chemicals with SMILES and experimental data available for five selected assays

TEST TYPE	No of components with exp data	Negative	Positive	In-conclusive	Data not reliable
Bacterial Reverse Mutation Assay	921	879	39 (4%)	3	3
Mammalian Bone Marrow Chromosome Aberration Test	116	99	14 (12%)	2	1
Mammalian Erythrocyte Micronucleus Test	452	200	19 (4%)	1	232
***in vitro* Mammalian Chromosome Aberration Test**	574	427	139 (24%)	1	7
***in vitro* Mammalian Cell Gene Mutation Test**	592	515	65 (11%)	4	8

A pictorial representation of Table 4 is shown in Figure 9.

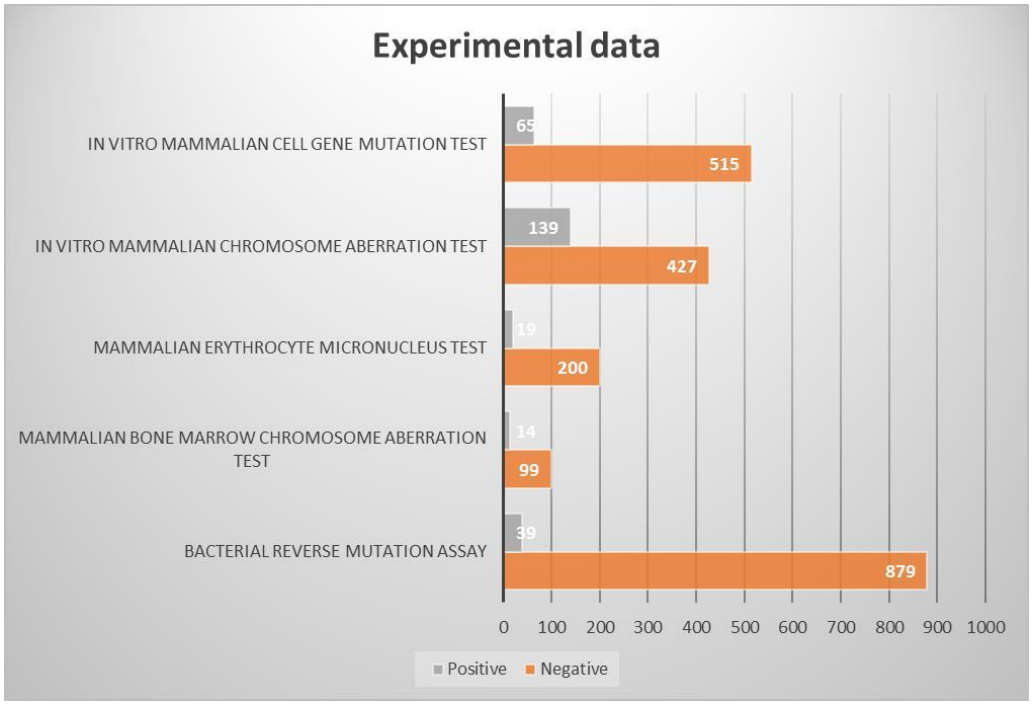


Figure 9: Chemicals with SMILES and experimental data available for five assays selected for the prediction exercise

3. Assessment/Results

The assessment included the evaluation of QSARs predictions performed with commercial and publicly available software (**Objective 2**), Read Across predictions (**Objective 4**), and investigation on substructures that impact on genotoxicity (**Objective 5**).

3.1. QSAR predictions of the genotoxicity of chemicals in the EFSA database

3.1.1. Software applied for the analysis

The Consortium has been granted access by developers / owner companies, to the commercial software tools listed below. For the purpose of the project and in collaboration with software houses, developers run the predictions in the first person in order to obtain the best possible results. The Consortium collected the predictions and performed the assessment of the predictive performance.

The selection of freely available (Q)SARs for the evaluation of predictivity was based on the consideration of the ease of use i.e. model implemented in a user-friendly software and batch mode is available to run the predictions. Freely available tools were run by the Consortium, apart from Lazar which was run by the developer.

The commercial software included:

Derek Nexus v.5 and Sarah Nexus v.2.0.1 by Lhasa Limited

CASE Ultra 1.6.2.1 by MultiCASE Inc.

Leadscope Model Applier v2.2.1.1 by Leadscope Inc.

ChemTunes ToxGPS by Molecular Networks GmbH

Percepta 2016 (Build 2911) by ACD/Labs Inc.

The free software included:

Lazar v. 1.1.0 by *In Silico* Toxicology GmbH

ToxTree v. 2.6.13 by Ideaconsult Ltd and JRC

Vega v.1.1.4 by IRCCS

Some software tools include models for predicting a range of assays / endpoints, as well as more models for predicting a single endpoint. The number of models available for the prediction of specific genotoxicity endpoint is provided in Table 5.

Table 5: Number of models available for the prediction of selected assays

TEST TYPE	GUIDELINE	METHOD	Models	statistical based	expert-rule based
Bacterial Reverse Mutation Assay	OECD Guideline 471 and 472	in vitro	18	12	6
Mammalian Bone Marrow Chromosome Aberration Test	OECD Guideline 475	in vivo	3	2	1
Mammalian Erythrocyte Micronucleus Test	OECD Guideline 474	in vivo	6	4	2
In vitro Mammalian Chromosome Aberration Test	OECD Guideline 473	in vitro	7	6	1
In vitro Mammalian Cell Gene Mutation Test	OECD Guideline 476	in vitro	6	5	1

The following tables (Tables 6-10) detail the QSAR models and specific assays predicted.

Table 6: List of software and their models for predicting the Bacterial Reverse Mutation Assay

Software	Model Name
Commercial statistical based	
Percepta	Mutagenicity/Procaryote/Bacterial composite
Percepta	Mutagenicity/Procaryote/Salmonella composite
Model Applier	Ames consensus from two models: Leadscope Salmonella statistical-based QSAR model v3 and Leadscope E.coli/TA102 statistical-based QSAR model v1
Sarah	Mutagenicity Endpoint *in vitro*
ChemTunes	Bacterial reverse mutagenesis (Ames mutagenicity)
MultiCASE	GT1_A7B, Salmonella G:C mutation
MultiCASE	BMUT_PHARMA, OECD471 bacterial mutagenicity
Free statistical based	
Lazar	Salmonella typhimurium mutagenicity
Vega	Mutagenicity on Salmonella typhimurium (Ames test)
Commercial expert rule-based	
Model Applier	Leadscope genetox expert alerts v4
Derek	Mutagenicity Endpoint *in vitro*
MultiCASE	GT_Expert, expert rule system for bacterial mutagenicity
Free expert rule-based	
ToxTree	Mutagenicity on Salmonella typhimurium (Ames test)
Vega	Mutagenicity on Salmonella typhimurium (Ames test)

Table 7: List of software and their models for predicting the Mammalian Bone Marrow Chromosome Aberration Test *in vivo*

Software	Model Name
Commercial statistical based	
Percepta	Clastogenicity/Chromosome aberrations/ Chromosome aberrations *In vivo* composite
Model Applier	Leadscope *In Vivo* Chrom Ab Comp statistical-based QSAR model v1
Commercial expert rule-based	
Derek	Chromosome Damage Endpoint *in vivo*

Table 8: List of software and their models for predicting the Mammalian Erythrocyte Micronucleus Test

Software	Model Name
Commercial expert rule-based	
ChemTunes	*in vivo* micronucleus
MultiCASE	GT3_MNT_MOUSE (Micronucleus test, *in vivo*, mouse)
Percepta	Clastogenicity/Micronucleus/ Micronucleus *In Vivo* composite
Model Applier	Leadscope *In Vivo* Micronuc Mouse statistical-based QSAR model v2
Commercial expert rule-based	
Derek	Chromosome Damage Endpoint *in vivo*
Free expert rule-based	
ToxTree	Structural alerts for the *in vivo* micronucleus assay in rodents

Table 9: List of software and their models for predicting the *In vitro* Mammalian Chromosome Aberration Test

Software	Model Name
Commercial statistical based	
ChemTunes	*in vitro* chromosome aberration
MultiCASE	GT2_CHROM_CHL (Chromosomal aberrations, *in vitro*, CHL cell line)
MultiCASE	GT2_CHROM_CHO (Chromosomal aberrations, *in vitro*, CHO cell line)
Percepta	Clastogenicity/Chromosome aberrations/ Chromosome aberrations *In vitro* composite
Model Applier	Leadscope *In Vitro* Chrom Ab CHL statistical-based QSAR model v2
Model Applier	Leadscope *In Vitro* Chrom Ab CHO statistical-based QSAR model v2
Commercial expert rule-based	
Derek	Chromosome Damage Endpoint *in vitro*

Table 10: List of software and their models for predicting the *In vitro* Mammalian Cell Gene Mutation Test

Software	Model Name
Commercial statistical based	
Percepta	Mutagenicity/Eucaryote/Mouse lymphoma (MLA) composite
Percepta	Mutagenicity/Eucaryote/CHO/CHL all loci composite
MultiCASE	GT4_L5178Y (Gene mutation, *in vitro* mouse lymphoma L5178Y, TK loci)
MultiCASE	GT4_ML_ACT (Mouse Lymphoma, activated)
MultiCASE	GT4_ML_UNACT (Mouse Lymphoma, unactivated)
Commercial expert rule-based	
Derek	Mutagenicity Endpoint *in vitro*

3.1.2. Training sets of the QSAR predictive systems

In the assessment of QSAR predictions, a critical issue is that of distinguishing between the chemicals included in the training set of the model, and the really external chemicals. The Consortium had no direct access to the training sets of the specific models, however several software tools vendors / developers provides the information whether the predicted compound is present in the model's training set. Thus, it has been possible to analyse compounds of the EFSA genotoxicity database that are included in training sets of the applied models. The training set concept is in general not applicable to expert rule-based system, with the exception of expert models from Leadscope, MultiCase, and Vega (see Table 6).

In Figures 10 and 11, the training set coverage for the models evaluated is presented as percentage of compounds from the genotoxicity EFSA DB part of the model's training sets. As a software may include more than one model for a certain test / endpoint, the same software may be repeated on the ordinate scale. For a detailed review of models per software available for a specific assay please refer to Tables 6 to 10.

Figure 10: Training set coverage for the QSAR models evaluated

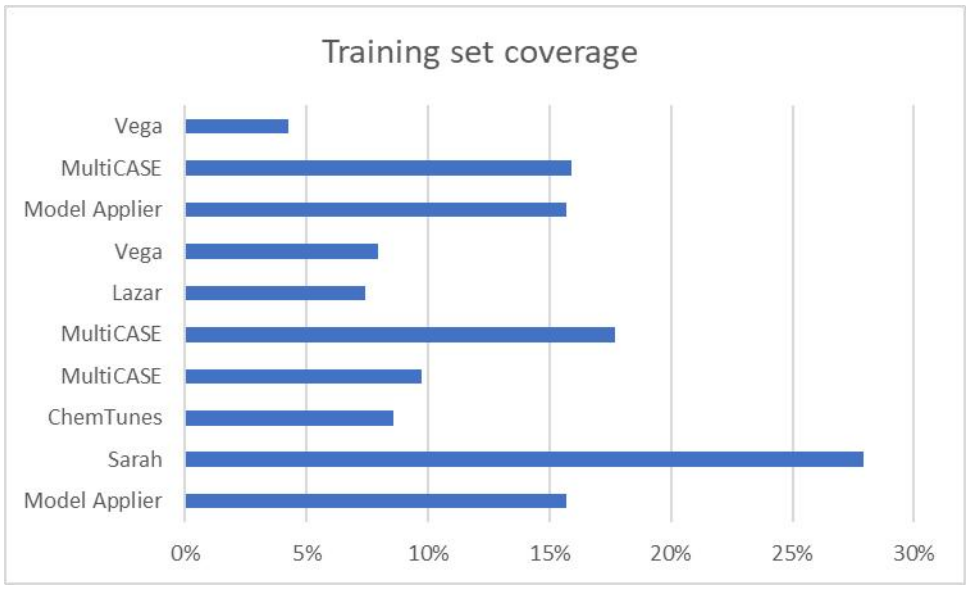


Figure 11: Training set coverage for the models predicting bacterial reverse mutation

Training set coverage is also represented for the models in the intersection mode in Figures 12 - 15. The intersections of models (row/column) are showing the number of same compounds , from the EFSA genotoxicity database, that are common for training sets of two models. The colour scale ranks the number of common compounds, going from green (highest) to red (lowest).

In general, it is observed that -for the evaluated assays- a low percentage, always less than 30%, of compounds from the EFSA genotoxicity database is included in the training sets of the evaluated models. On the average, the coverage is around 10%. A maximum coverage of the training set was observed for Sarah, a statistical tool from Lhasa, predicting bacterial reverse mutation, where 267 (28%) compounds are included in the model's training set. Figure 12 also shows that, for example for Vega and Lazar, all EFSA genotoxicity pesticides from their training sets, are included in the training set of most models.

For other assays the number of compounds included in the training sets range between 16 compounds (a ChemTunes model predicting mammalian erythrocyte micronucleus and a MultiCASE model predicting *in vitro* mammalian gene mutation) and 62 compounds (included in the ChemTunes training set of a model predicting *in vitro* mammalian chromosome aberration).

		Lazar	Sarah	ChemTunes	MultiCASE	MultiCASE	MultiCASE	Vega	Vega	Model Applier	Model Applier
Lazar	Salmonella typhimurium mutagenicity	71	55	41	69	55	68	25	70	71	71
Sarah	Mutagenicity Endpoint in vitro		267	52	107	66	117	29	59	99	99
ChemTunes	Bacterial reverse mutagenesis (Ames mutagenicity)			82	72	70	78	36	43	81	81
MultiCASE	GT_Expert, expert rule system for bacterial mutagenicity				152	93	148	40	75	129	129
MultiCASE	GT1_A7B, Salmonella G:C mutation					93	89	39	59	92	92
MultiCASE	BMUT_PHARMA, OECD471 bacterial mutagenicity						169	40	73	139	139
Vega	Mutagenicity on Salmonella typhimurium (Ames test)							41	27	41	41
Vega	Mutagenicity on Salmonella typhimurium (Ames test)								76	76	76
Model Applier	Leadscope genetox expert alerts v4									150	150
Model Applier	Ames consensus from two models: Leadscope Salmonella statistical-based QSAR model v3 and Leadscope E.coli/TA102 statistical-based										150

Figure 12: Intersection graph of overlapping training sets for models predicting bacterial reverse mutation

		Chem Tunes	Multi CASE	Multi CASE	Model Applier	Model Applier
Chem Tunes	in vitro CA	62	30	35	32	38
Multicase	GT2_CHROM_CHL (Chromosomal aberrations, in vitro, CHL cell line)		36	17	36	17
Multicase	GT2_CHROM_CHO (Chromosomal aberrations, in vitro, CHO cell line)			48	17	47
Leadscope	Leadscope In Vitro Chrom Ab CHL statistical-based QSAR model v2				38	18
Leadscope	Leadscope In Vitro Chrom Ab CHO statistical-based QSAR model v2					52

Figure 13: Intersection graph of overlapping training sets for models predicting *in vitro* mammalian chromosome aberration test

		ChemTunes	MultiCASE	Model Applier
ChemTunes	in vivo MN	16	7	10
MultiCASE	GT3_MNT_MOUSE (Micronucleus test, in vivo, mouse)		32	32
Model Applier	Leadscope In Vivo Micronuc Mouse statistical-based QSAR model v2			49

Figure 14: Intersection graph of overlapping training sets for models predicting mammalian erythrocyte micronucleus test

		MultiCASE	MultiCASE	MultiCASE
MultiCASE	GT4_L5178Y (Gene mutation, in vitro mouse lymphoma L5178Y, TK loci)	16	12	11
MultiCASE	GT4_ML_ACT (Mouse Lymphoma, activated)		47	11
MultiCASE	GT4_ML_UNACT (Mouse Lymphoma, unactivated)			16

Figure 15: Intersection graph of overlapping training sets for models predicting *in vitro* mammalian gene mutation assay

In conclusion, it can be said that the statistically based QSAR models have a defined training set on which the model was built. On the other hand, expert knowledge-based models, which are based on human knowledge, do not necessarily have a clearly identified training set. Under the current project, the partners did not have direct access to training sets of the models applied in order to assess how independent the training sets are. In addition, access to training sets may be limited also due to the existence of proprietary data in the training sets. However, it was possible to analyse the number of compounds from the EFSA genotoxicity database present in the training sets of the applied models. In general, it was observed that a low number of pesticide/metabolites from EFSA genotoxicity DB is included in training sets of models predicting genotoxicity endpoints. Based on the analysis of training sets only, it is difficult to conclude on how the models are independent from each other; however it is possible to say that the evaluations performed are a sound exercise of external validation of the models.

3.1.3. Applicability Domain, and Prediction Rules of the QSAR software

The Applicability Domain

The principle of Applicability Domain obliges the users to specify the scope of their proposed models thus, defining the model limitations with respect to its structural domain and response space. If an external compound is beyond the defined scope of a given model, it is considered outside that model's Applicability Domain (AD) and its reliability should be considered with caution.

REACH (ECHA, 2017a; b) and ICH M7 (Fioravanzo et al., 2012) guidelines on the use of in silico models for regulatory purposes ask for predictions within the AD of the model.

For the purpose of this project firstly only the predictions within the applicability domain were considered. For expert rule-based models, the concept for AD is generally not applicable except for Leadscope, Multicase and Vega. On the other hand, all predictions from a statistical model were associated with an assessment of the AD provided by the model itself. Each model has its own method to assess the AD.

Based on the assessment of applicability domain automatically generated by some models, it was concluded that the highest number of compounds within the AD was determined for most models predicting bacterial reverse mutation and ranging between 80% and 90% (Figure 16). The only models for bacterial reverse mutation, with the percentage of compounds within AD below the above figures, were the two models from Vega (a statistical and expert-rule based model with 53% and 28% of compounds within AD respectively), and a model from Percepta (with 64% compounds within the AD).

Considering other assays, reported in Figures 17 - 20, the percentage of predictions assessed to be within AD was around 50% for the majority of models, except for the ChemTunes models predicting Mammalian Erythrocyte Micronucleus Test and In vitro Mammalian Chromosome Aberration Test, with 100% and 96% of compounds within AD, respectively. In the figures, the histograms relative for the models that provide the AD assessment are provided and the following abbreviation are used: "com"

for a commercial software; "free" for a freely available tool; "stat" for a statistically based and "rule" for an expert rule based tool.

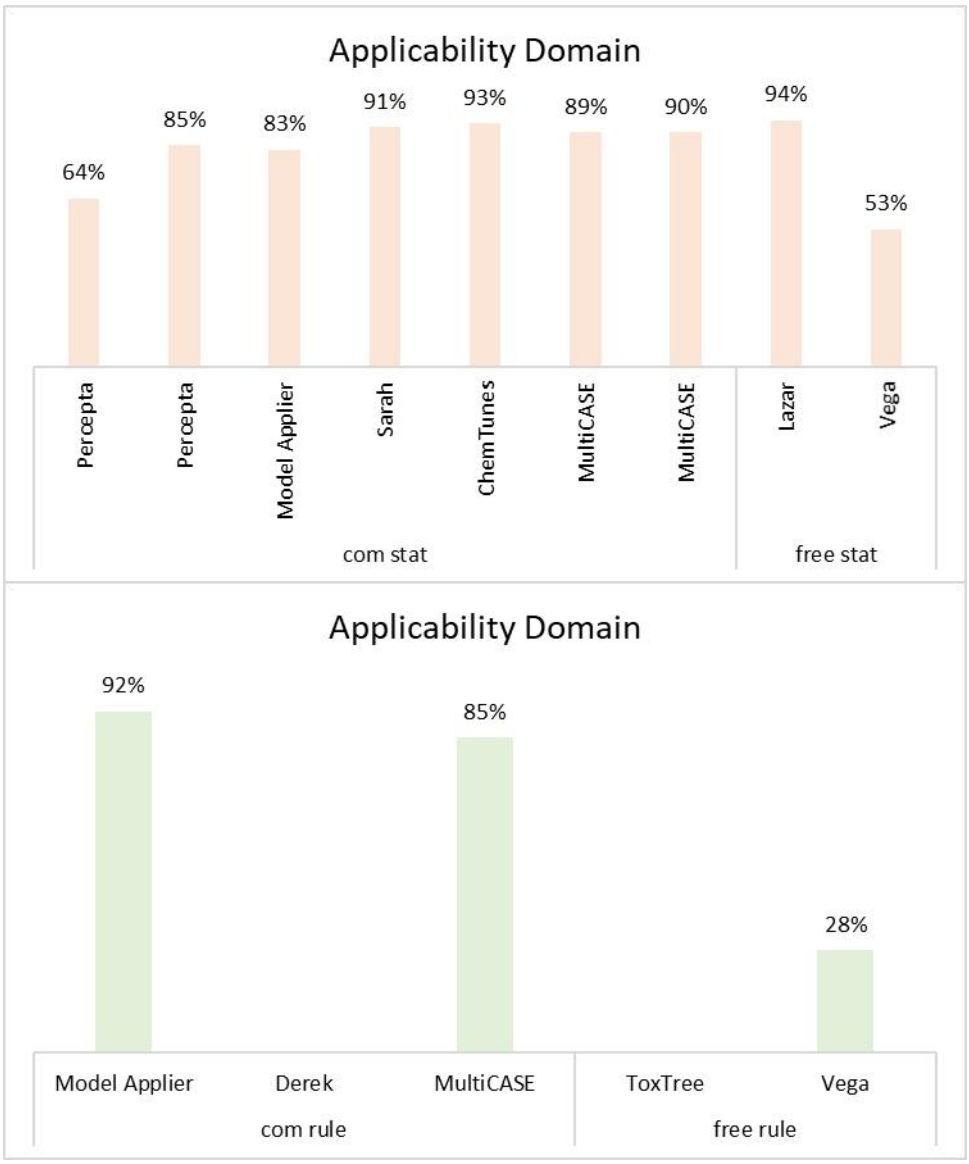


Figure 16: Percentage of compounds within the applicability domain of models for predictions of the bacterial reverse mutation. Note that the concept of Applicability Domain does not apply to the DEREK and ToxTree rule-based systems.

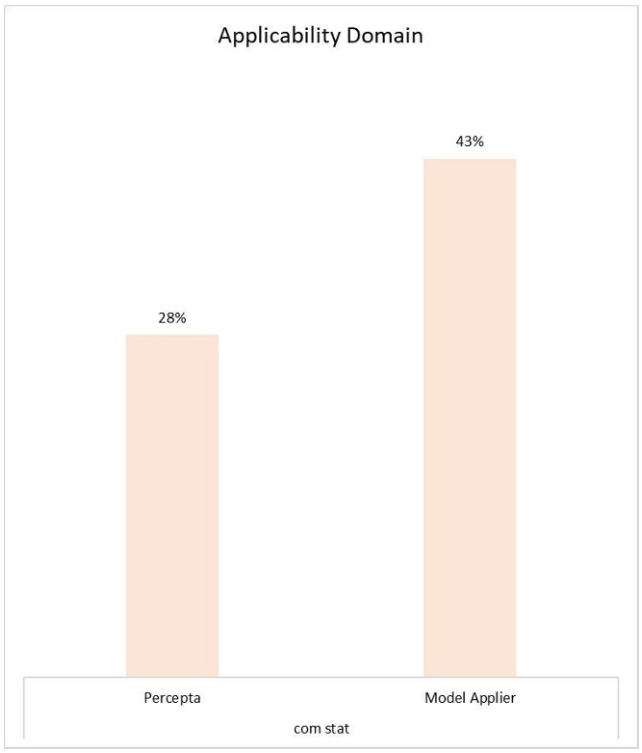


Figure 17: Percentage of compounds within the applicability domain of models for predictions of mammalian bone marrow chromosome aberration test

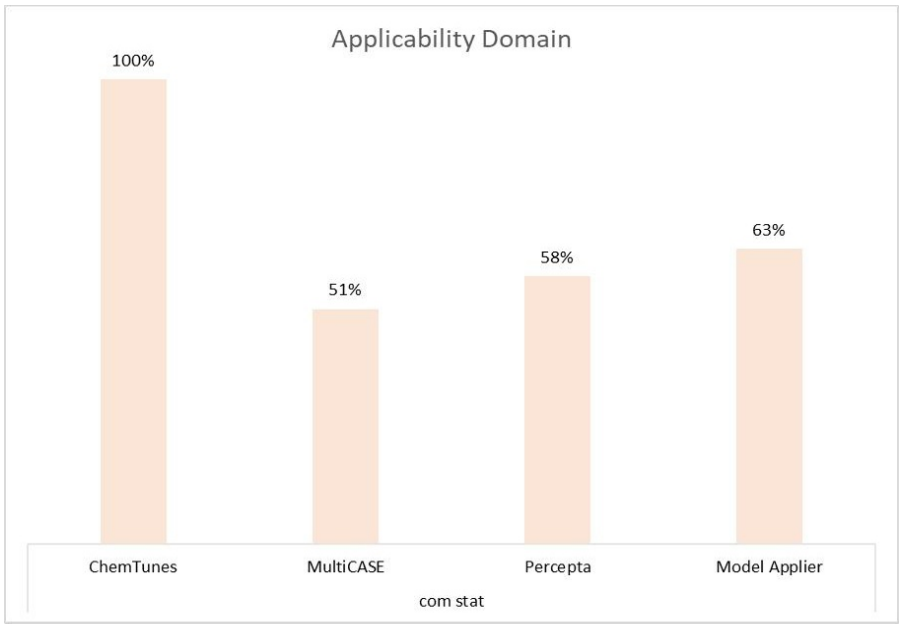


Figure 18: Percentage of compounds within the applicability domain of models for predictions of Mammalian Erythrocyte Micronucleus Test

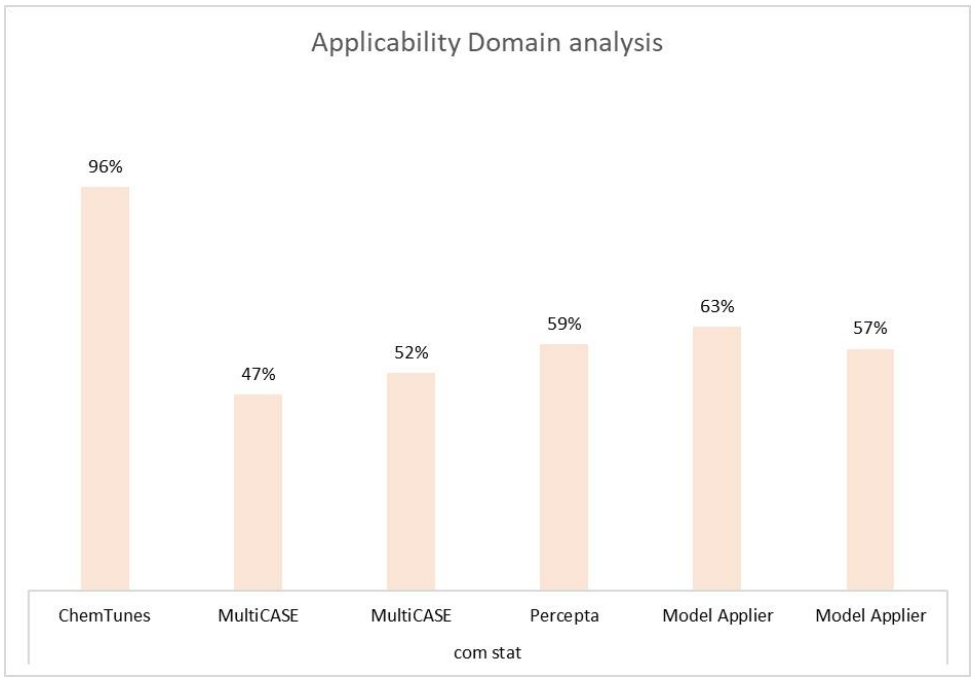


Figure 19: Percentage of compounds within the applicability domain of models for predictions of *in vitro* mammalian chromosome aberration test

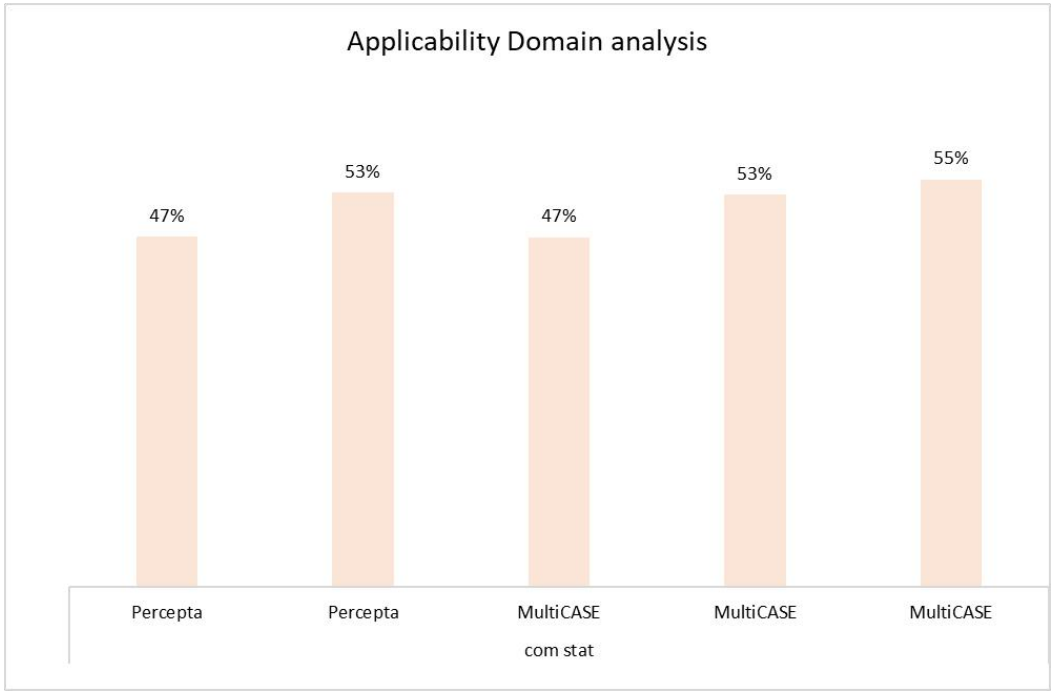


Figure 20: Percentage of compounds within the applicability domain of models for predictions of *In vitro* Mammalian Cell Gene Mutation Test

Rules applied to interpret the predictions

Table 11 reports the general rules that were applied to assess the predictions as generated by the models. Predictions not providing a negative or positive outcome were in general reported as equivocal and were therefore considered as uncertain, therefore not reliable.

The predictions were associated with the assessment of the AD, where the estimations are considered to be associated with an acceptable level of accuracy (reliability). Therefore, if the chemical was within the applicability domain, the results of the model predictions were considered to be reliable. Uncertainty is in general assessed considering a probability which is associated to each prediction. The probability was assessed according to rules either agreed with the commercial developers of the models or derived from experience. No fine tuning of these rules is possible in a screening mode.

The commercial developers have analysed their results, and provided comments useful to guide the application of their models at best.

Table 11: Rules applied to the predictions of the software systems

Software	Com/free	General Rules	
Case Ultra	Commercial	Negative, positive, known negative and known positive as given	Inconclusive were considered inconclusive
ChemTunes ToxGPS	Commercial	Negative and positive as given	Equivocal were considered inconclusive
Derek Nexus	Commercial	Probable, plausible and equivocal were considered positive; Improbable, inactive and blank as negative	-
Model Applier	Commercial	Negative and positive as given	Inconclusive, Indeterminate and Missing Descriptors were considered inconclusive.
Percepta	Commercial	positive or negative according to the following algorithm: if(ln([p]/(1 - [p])) + 3 - 2.5 * [RI] < 0, -1, if(ln([p]/(1 - [p])) - 3 + 2.5 * [RI] > 0, 1, 0)); *this produces:* 1 - for Positive; -1 - for Negative	0 (Undefined) - were considered inconclusive.
Sarah Nexus	Commercial	Negative and positive as given	Equivocal were considered inconclusive
Lazar	Free	Mutagenic was considered as positive non-mutagenic was considered as negative	Blank was considered as inconclusive.
ToxTree	Free	Structural alert(s) was considered as positive NO alerts was considered as negative	
Vega	Free	non-Mutagenic and mutagenic were considered as negative and positive	Suspect Mutagenic was considered as inconclusive.

3.1.4. QSAR predictions of the genotoxicity of the chemicals in the EFSA database: results

To visualise and compare the performance of the models, the relationship between sensitivity and false positive rate were plotted in a Receiver Operating Characteristic (ROC) graph. In a ROC graph, a model located in the top left corner is the ideal model, having a perfect (100%) prediction of positives and a perfect (0%) false positive rate. The diagonal line is associated with random results (Swets, 1988).

The colour scheme applied in the ROC graphs represents the percentage of predicted compounds within AD and with available experimental data. It follows the following code: red is for less than 60% compounds, orange is for 60%-80% compounds, and green is for more than 80% compounds.

Overall, this investigation points to a substantial difference between the satisfactory level of accuracy of the QSAR predictions for the Ames test, and the much lower reliability of predictions obtained with the other experimental assays.

Detailed results from evaluations are available in Annex 2.

3.1.4.1. Ames test predictions

The predictions of the Ames test results obtained are displayed in Figure 21. The figure shows that there are differences among the performances of the different (Q)SARs. However all (Q)SAR models are in the top left ROC area, thus pointing to statistically significant predictions. This is clearly apparent even though the EFSA genotoxicity sample is strongly skewed towards negative results, very far from an ideal 50 / 50 % distribution of positives and negatives.

Sensitivity ranges between 46% for Toxtree and 71% for a model from Leadscope, namely the Ames consensus from two models: Leadscope Salmonella statistical-based QSAR model v3 and Leadscope E.coli/TA102 statistical-based QSAR model v1. Specificity ranges between 66% for Lazar and 98% for the Percepta model (namely Mutagenicity/Procaryote/Salmonella composite).

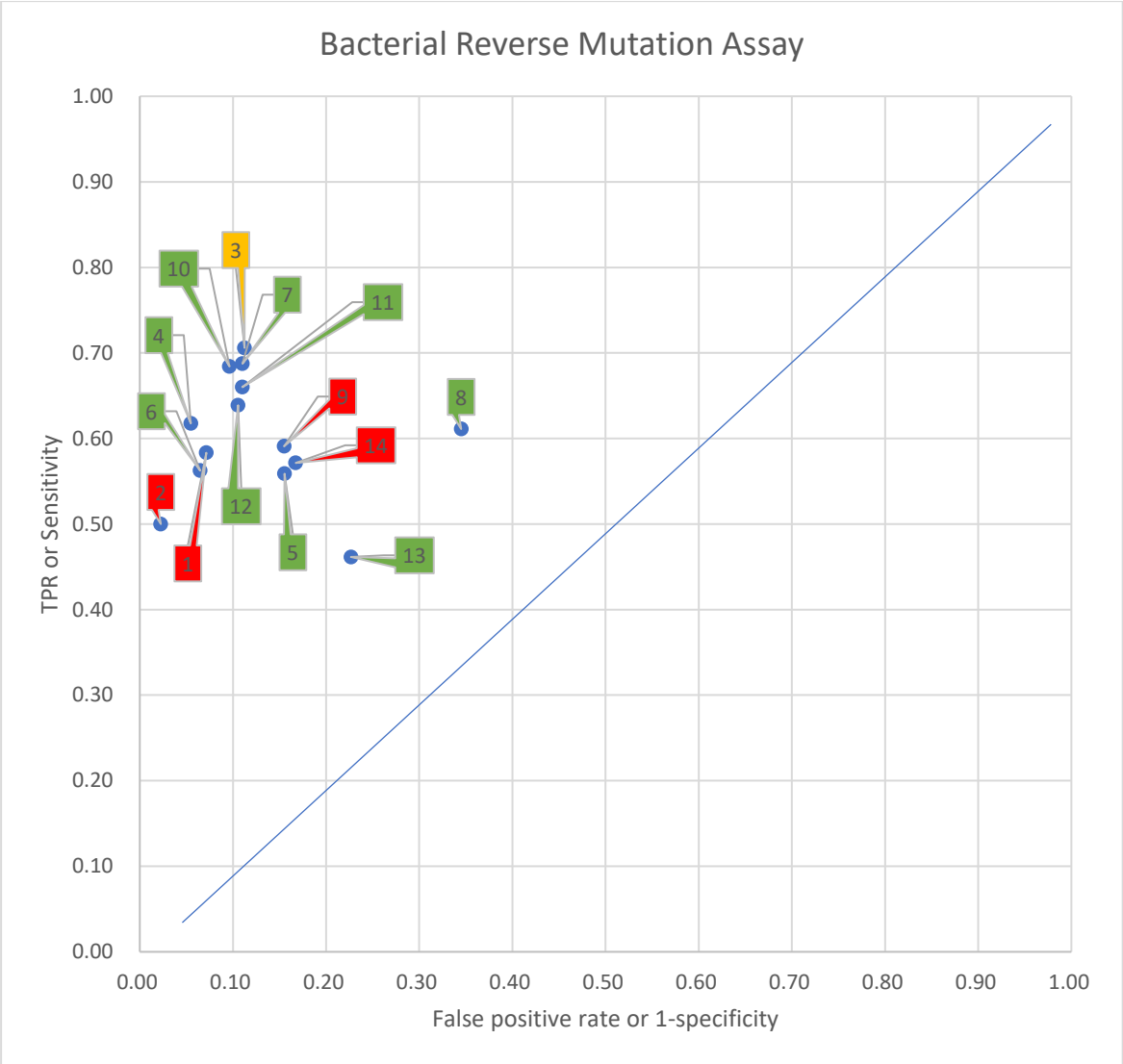


Figure 21: ROC graph for the Ames test predictions

The codes of the (Q)SAR models are:

1 Percepta	Mutagenicity/Procaryote/Bacterial composite
2 Percepta	Mutagenicity/Procaryote/Salmonella composite
3 Model Applier	Ames consensus from two models: Leadscope Salmonella statistical-based QSAR model v3 and Leadscope E.coli/TA102 statistical-based QSAR model v1
4 Sarah	Mutagenicity Endpoint *in vitro*
5 ChemTunes	AMES
6 MultiCASE	GT1_A7B, Salmonella G:C mutation
7 MultiCASE	BMUT_PHARMA, OECD471 bacterial mutagenicity
8 Lazar	Salmonella typhimurium mutagenicity
9 Vega	Mutagenicity on Salmonella typhimurium (Ames test)
10 Model Applier	Leadscope genetox expert alerts v4
11 Derek	Mutagenicity Endpoint *in vitro*
12 MultiCASE	GT_Expert, expert rule system for bacterial mutagenicity
13 ToxTree	Mutagenicity on Salmonella typhimurium (Ames test)
14 Vega	Mutagenicity on Salmonella typhimurium (Ames test)

It is interesting to compare the pattern in Figure 21 with that shown by similar analyses in Objective 1 based on literature data

Figure 1 reports the performance of different (Q)SAR models in retro-fitting exercises on the classical database of Ames mutagenicity results available in the public domain (i.e., how well the models fit the training sets). In both cases (Figure 1 and Figure 21), the predictive models –collectively- are in the ROC area of statistically significant results. At the same time, the patterns are different. This is expected, and is explained by the very different study designs: a) the compositions of chemicals of the two databases were largely different. The public domain database consists mainly of industrial chemicals, with a minor proportion of pesticides; b) the proportion of negatives / positives is strongly skewed towards negatives in the EFSA database, whereas the public database has a majority of positives; c) the predictions in Objective 1 were retrieved from retro-fit studies in literature (systems applied to the training set), whereas the predictions in Objective 2 were prospective ones on a database largely different from the training sets. The latter explains why the predictivity is higher in the literature data (Figure 1) in respect to the exercise on the EFSA data (Figure 21).

Another comparison is possible between the results of Figure 21, and those of Figures 2 and 3. In the latter figures, different QSARs were challenged to predict external test sets (like in the case of the EFSA database). It should be noticed that the variability of responses is quite high, larger than with the EFSA database. A conclusion is that the Ames mutagenicity of pesticides can be predicted with good confidence, since the patterns of predictions by the different models were quite consistent with each other.The performance of the models for the Ames test was further assessed by considering also the predictions out of AD. For the majority of models, it was observed that the percentage of predicted compounds changes only slightly and the performance matrices are very similar (results not shown). Only for the two Vega models it was noted that by considering also the compounds out the AD the number of predicted compounds increased significantly, but there was almost no change observed in the performance matrices (Figure 22).

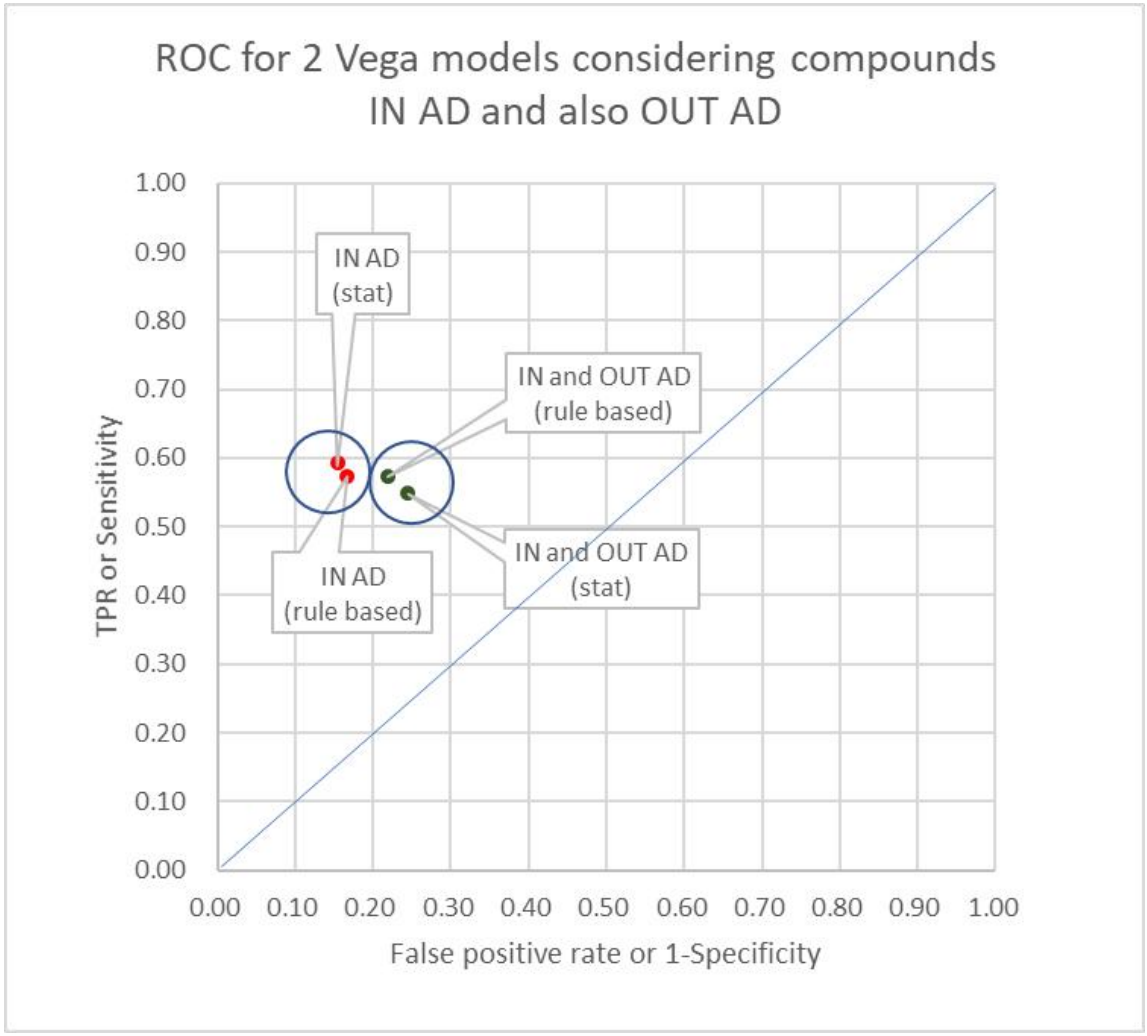


Figure 22: ROC graph for Ames test predictions by one statistical and one rule-based VEGA model, with considerations of compounds: a) within; and b) within and outside the Applicability Domain

3.1.4.2. Predictions for assays different from the Ames test

Figures 23, 24, 25 report the ROC graphs for assay systems / endpoints different from *in vitro* bacterial mutagenicity (Ames), i.e., *in vitro* Micronucleus, *in vitro* mammalian cells gene mutation and *in vitro* mammalian cells chromosomal aberrations (Objective 2 of this project).

Overall, the study indicated that the reliability of these (Q)SAR models is still far from optimality. In fact, the inspection of the figures shows that the predictions –collectively- are quite close to the diagonal line, that represents random results in the ROC graphs. It should be noted that Percepta models showed a promising performance in two cases (Figures 24 and 25): however this was obtained at the expenses of reducing the number of chemicals actually predicted to around 30%, due to very strict Applicability Domain rules.

There is no possibility of comparing these results for tests different from the Ames test with previous studies in the literature, because similar prediction exercises have not been published. Thus this EFSA projects contributes with original information to the research on the predictivity of QSARs for genotoxicity assays different from the Ames test.

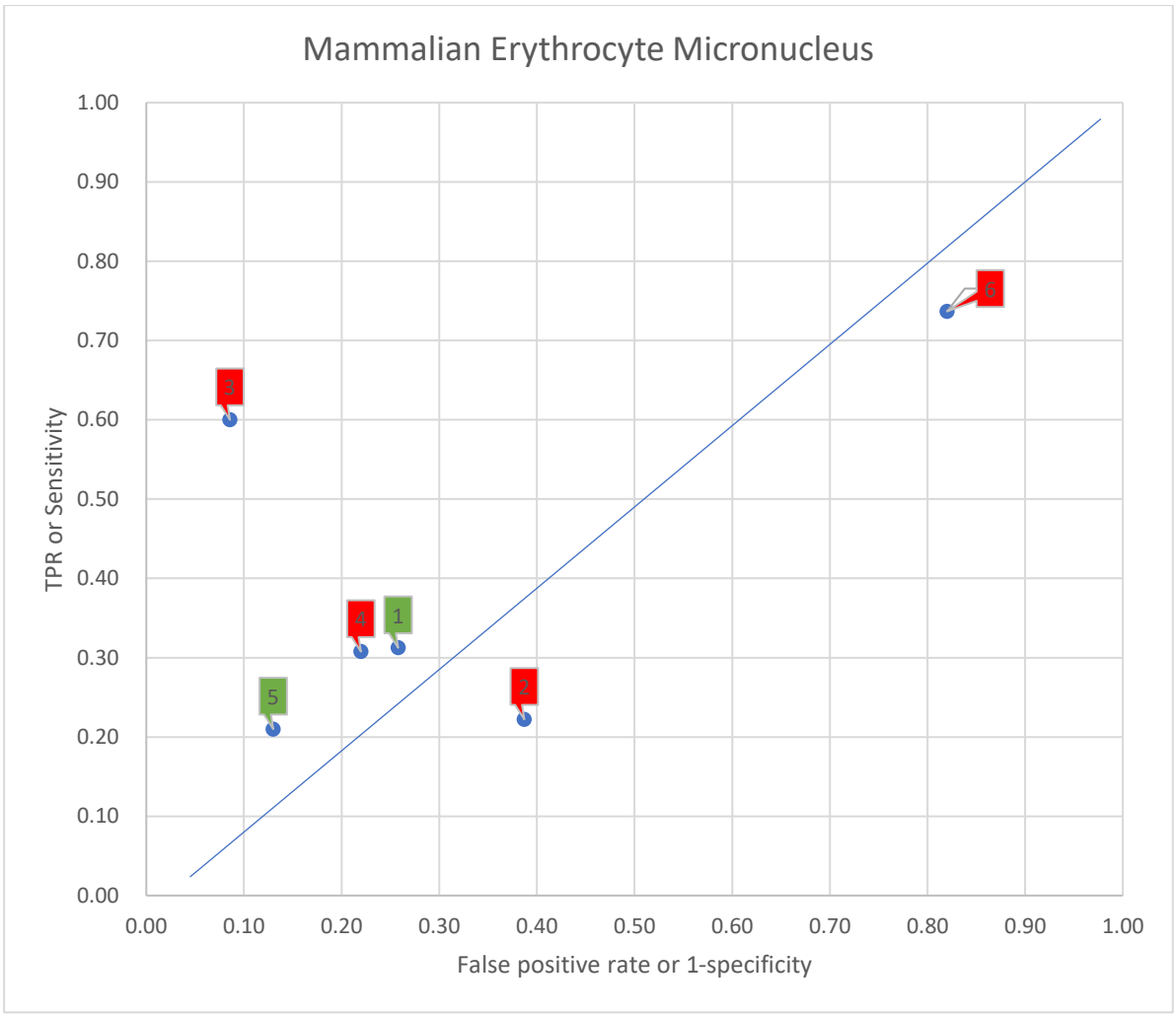


Figure 23: ROC graph for the predictions of *in vivo* micronucleus assay

1 ChemTunes	*in vivo* MN
2 MultiCASE	GT3_MNT_MOUSE (Micronucleus test, *in vivo*, mouse)
3 Percepta	Clastogenicity/Micronucleus/ Micronucleus *In Vivo* composite
4 Model Applier	Leadscope *In Vivo* Micronuc Mouse statistical-based QSAR model v2
5 Derek	Chromosome Damage Endpoint *in vivo*
6 ToxTree	Structural alerts for the *in vivo* micronucleus assay in rodents

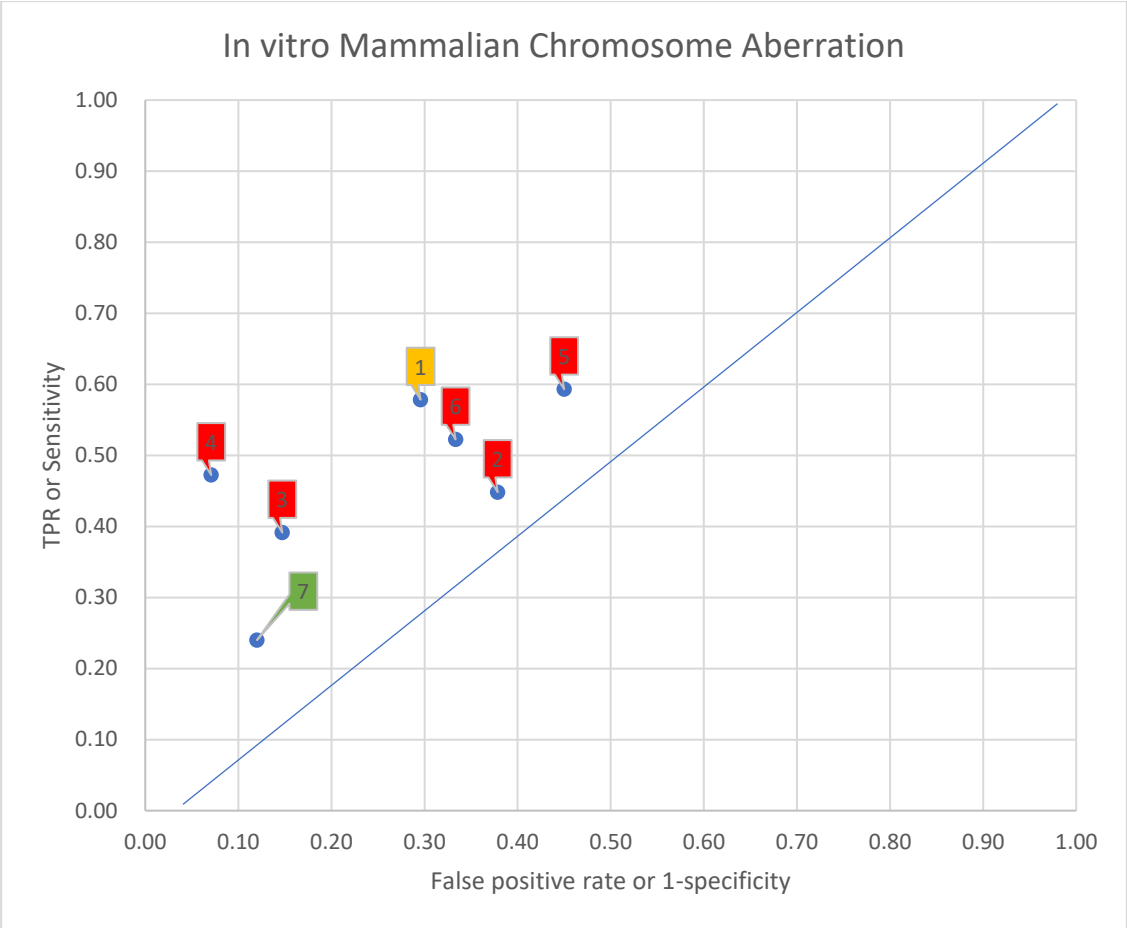


Figure 24: ROC graph for the predictions of *in vitro* mammalian chromosomal aberration test

1 ChemTunes	*in vitro* CA
2 MultiCASE	GT2_CHROM_CHL (Chromosomal aberrations, *in vitro*, CHL cell line)
3 MultiCASE	GT2_CHROM_CHO (Chromosomal aberrations, *in vitro*, CHO cell line)
4 Percepta	Clastogenicity/Chromosome aberrations/ Chromosome aberrations *In vitro* composite
5 Model Applier	Leadscope *In Vitro* Chrom Ab CHL statistical-based QSAR model v2
6 Model Applier	Leadscope *In Vitro* Chrom Ab CHO statistical-based QSAR model v2
7 Derek	Chromosome Damage Endpoint *in vitro* com rule

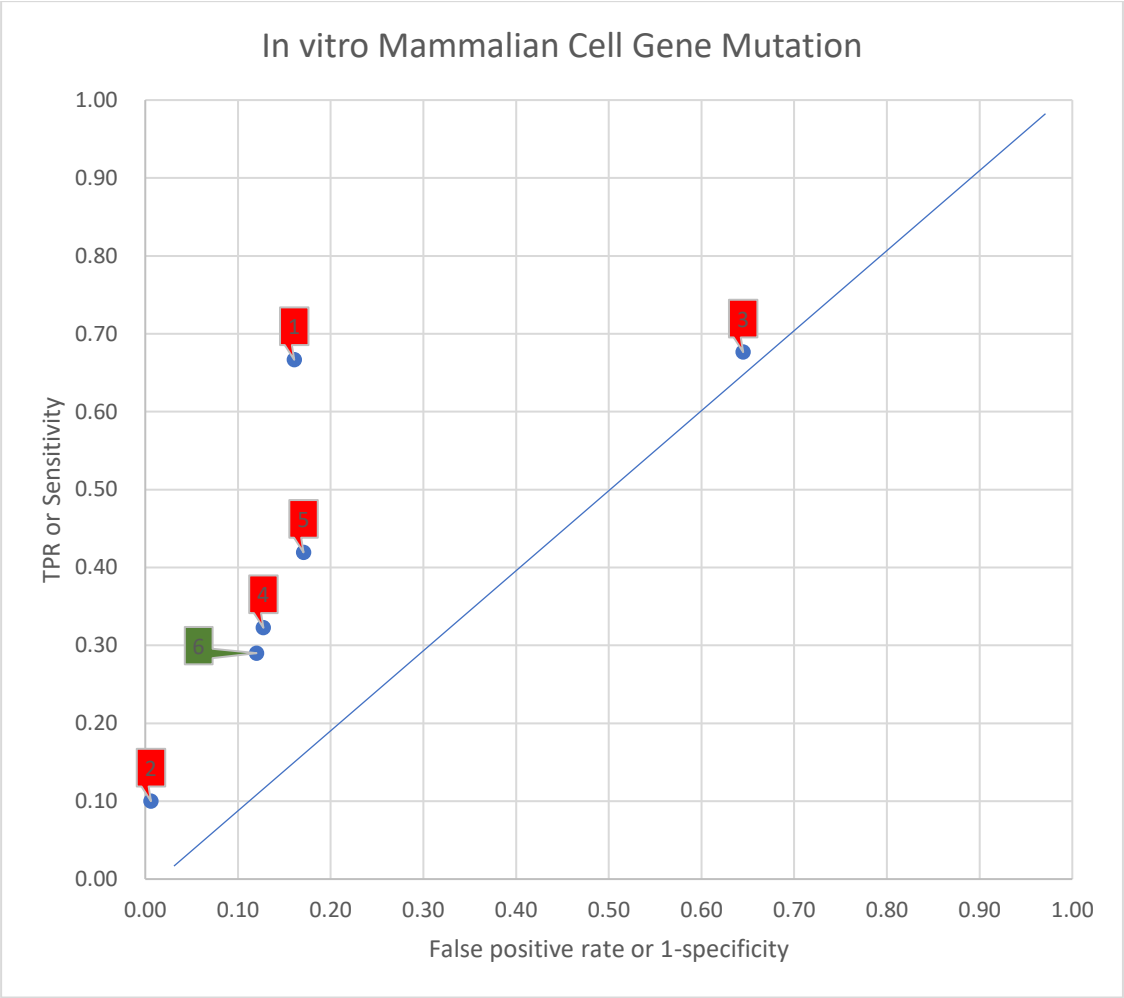


Figure 25: ROC graph for the predictions of *in vitro* mammalian cells gene mutations

1 Percepta	Mutagenicity/Eucaryote/Mouse lymphoma (MLA) composite
2 Percepta	Mutagenicity/Eucaryote/CHO/CHL all loci composite
3 MultiCASE	GT4_L5178Y (Gene mutation, *in vitro* mouse lymphoma L5178Y, TK loci)
4 MultiCASE	GT4_ML_ACT (Mouse Lymphoma, activated)
5 MultiCASE	GT4_ML_UNACT (Mouse Lymphoma, unactivated)
6 Derek	Mutagenicity Endpoint *in vitro*

Thus, at present the (Q)SAR models do not seem to be able to provide reliable genotoxicity predictions for assays / endpoints different from Ames, and need to be further developed.

3.1.5. Combination of QSAR predictions

The analysis of literature (reported above) has shown that much attention focuses on the issue of combining different QSAR models with the aim of attempting to improve performance.

In Objective 2, we have tested a number of combinations of the (Q)SARs assessed in this Project, using the EFSA genotoxicity database as a probe. We used two approaches: a) simple combinations of (Q)SAR predictions; b) combinations that weight the (Q)SAR predictions based on parameters related to the reliability of each model (Weight-of-Evidence (WoE) approach).

3.1.5.1. Simple combinations

In this section, a number of simple combinations of pairs of models predictions were studied without giving weights to either of them. General criteria in the selection of models were: a) the maximum coverage of the number of substances predicted; and b) the combination of independent models, as suggested in ICH and EFSA guidance.

The predictions were combined according to a conservative approach: if either prediction is positive, the combined prediction is positive as well. We considered only the substances for which the prediction of both systems is available, and we included also predictions of substances nominally out of the applicability domain.

Figure 26 displays the combination of predictions of Ames mutagenicity obtained with Derek Nexus and ToxTree (both expert systems). It appears that the combination attains an increased Sensitivity, with some loss of Specificity.

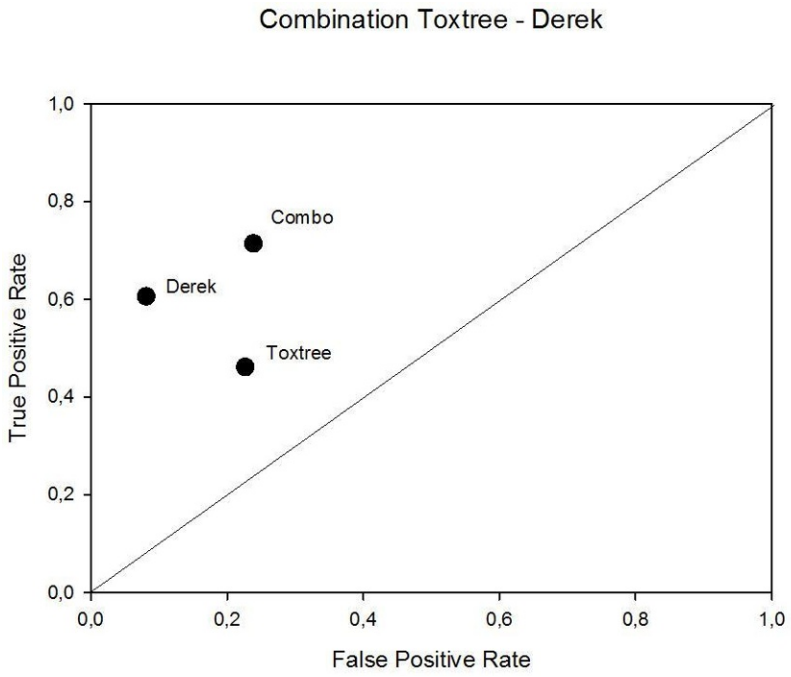


Figure 26: Predictions of Ames mutagenicity obtained within this project, by combining the two expert systems ToxTree and Derek Nexus (see details in the text)

Table 12 reports details for the combination shown in the figure above, and for additional combinations of models for Ames test mutagenicity (implemented according to the same conservative approach as above). In every case, Sensitivity is increased with a correspondent loss of Specificity. This is irrespective of the nature of the models combined (statistical or expert system). The total number of substances predicted decreases as a consequence of the combination of results (i.e., for some chemicals, the predictions are not available for all the systems in the combination).

Table 12: Performance of combinations of QSAR models for Ames test prediction, together with the performance of the composing models (see details in the text)

		Sensitivity (%)	**Specificity (%)**	**False positive Rate (%)**	**Number of predicted substances (%)**
statistically based model	Sarah	61.76	94.52	5.48	82.95
rule based model	Leadscope MA	66.67	91.05	8.95	96.42
	mean	64.22	92.79	7.22	89.69
Combination		**82.35**	**89.27**	**10.73**	**80.83**
	Difference	**18.14**	**-3.52**	**3.52**	**-8.86**
statistically based model	MultiCase (BMUT_PHARMA)	68.75	88.98	11.02	80.56
rule based model	Derek	60.61	91.89	8.11	94.57
	mean	64.68	90.44	9.57	87.57
Combination		**69.23**	**84.93**	**15.07**	**78.00**
	Difference	**4.55**	**-5.51**	**5.51**	**-9.57**
statistically based model	Sarah	61.76	94.52	5.48	82.95
rule based model	Derek	60.61	91.89	8.11	94.57
	mean	61.19	93.21	6.80	88.76
Combination		**75.86**	**88.97**	**11.03**	**79.19**
	Difference	**14.68**	**-4.24**	**4.24**	**-9.57**
statistically based model	MultiCase (BMUT_PHARMA)	68.75	88.98	11.02	80.56
statistically based model	Sarah	61.76	94.52	5.48	82.95
	mean	65.26	91.75	8.25	81.76
Combination		**78.57**	**87.35**	**12.65**	**68.52**
	Difference	**13.32**	**-4.40**	**4.40**	**-13.24**

		Sensitivity (%)	Specificity (%)	False positive Rate (%)	Number of predicted substances (%)
rule based model	Derek	60.61	91.89	8.11	94.57
rule based model	Toxtree	46.15	77.33	22.67	99.57
	mean	53.38	84.61	15.39	97.07
Combination		**71.43**	**76.13**	**23.87**	**95.10**
	Difference	**18.05**	**-8.48**	**8.48**	**-1.97**

Combination of models for the prediction of *in vitro* Mammalian Chromosome Aberration and Mammalian Erythrocyte Micronucleus Tests were studied as well. Table 13 reports data relative to the combination of a statistically-based model (ChemTunes) and an expert system (Derek). In this case, the models where selected in order to ensure the maximum coverage of substances predicted. A conservative approach, as for Ames test, was applied. The performances show the same trend of the other combinations.

Table 13: Performances of combination of QSAR models for *in vitro* Mammalian Chromosome Aberration, and *in vivo* Mammalian Erythrocyte Micronucleus Test (see details in the text)

			Sensitivity (%)	Specificity (%)	False positive Rate (%)	Number of predicted substances (%)
in vitro Mammalian Chromosome Aberration	rule based model	Derek	17.44	93.84	6.16	99.82
	statistically based model	Chemtunes	55.81	71.23	28.77	67.49
		mean	36.63	82.54	17.47	83.66
	Combination		**58.14**	**69.18**	**30.82**	**67.26**
		Difference	**21.52**	**-13.36**	**13.36**	**-16.40**
in vivo Mammalian Erythrocyte Micronucleus Test	rule based model	Derek	15.79	94.47	5.53	99.54
	statistically based model	Chemtunes	31.25	74.21	25.79	94.06
		mean	23.52	84.34	15.66	96.80
	Combination		**50.00**	**72.49**	**27.51**	**93.60**
		Difference	**26.48**	**-11.85**	**11.85**	**-3.20**

The results of Table 13 are displayed as ROC graphs in Figures 27 and 28.

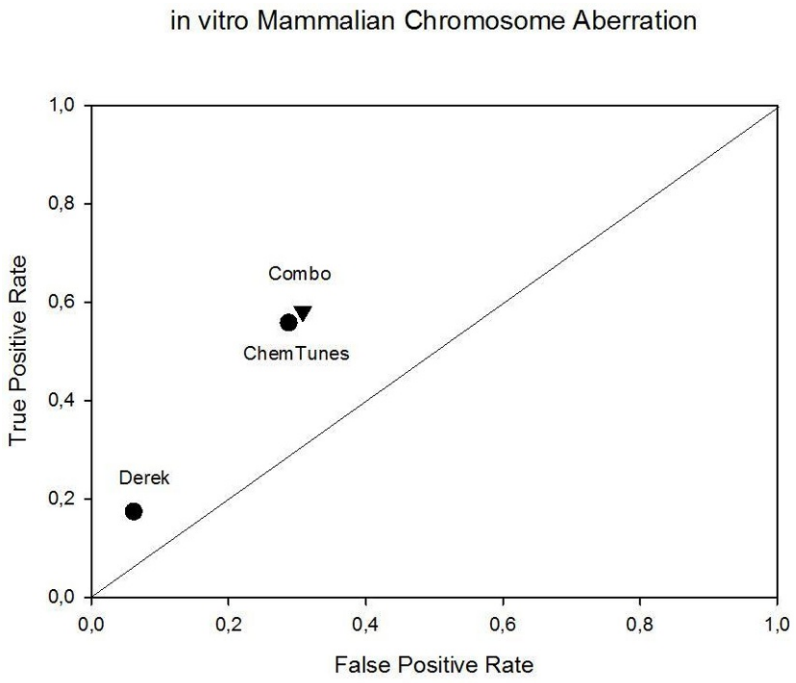


Figure 27: The predictions of *in vitro* Mammalian Chromosome Aberration obtained within this project by combining ChemTunes (statistically based system) and Derek (expert system) (see details in the text)

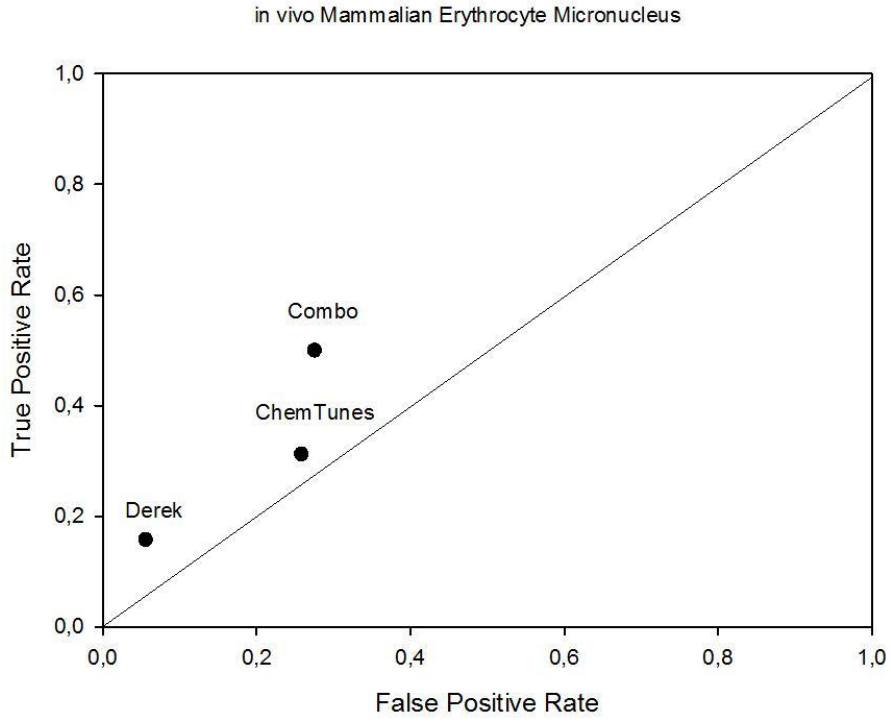


Figure 28: The predictions of *in vitro* Mammalian Erythrocyte Micronucleus test obtained within this project by combining ChemTunes (statistically based system) and Derek (expert system)(see details in the text)

3.1.5.2. Weight-of-Evidence combinations

The second approach to the combination of QSAR predictions considers weighting the evidence (WoE) with factors related to the reliability of models. It is an application of Dempster-Shafer theory (DST) to binary classification QSAR models for chemical toxicity prediction (http://fitelson.org/topics/shafer.pdf) (Beynon et al., 2000). Key features of this approach are:

1) The reliability of any given model is quantitatively incorporated into determining the prediction for a new query (test) compound. While performance statistics (e.g., AUC, sensitivity, specificity, positive and negative predictivity, etc.) are commonly cited for computational models based on model-validation exercises, this information is typically not used in any way in generating predictions for query compounds. The DST-based approach does take this information into account, so that each prediction can be reported with an associated level of uncertainty. For binary classifiers, the DST-based approach may therefore generate an "equivocal" prediction, indicating a level of uncertainty too high to confidently predict either of the two possible outcomes (e.g., positive or negative).

2) Results from multiple sources (e.g., multiple QSAR models) can be combined in an approach that is quantitative and statistically rigorous.

The results presented here should only be considered illustrative, because it was not possible to properly apply the rigorous DST-based combination approach to this case study. Reasons for this are the following:

- The original plan was to obtain reliability metrics by applying each model to a subset of the EFSA test set; however, this was not possible due to the extremely low proportion of positives (e.g., 4% for Ames, only 39 positive compounds in the set of 921) in the EFSA test set. Model reliability cannot be accurately assessed from validation sets that are too highly skewed.

- Model providers were asked to self-report reliability metrics based on their own model validation calculations. Unfortunately, only a few provided this information. We considered extracting reliability information from published papers documenting performance of these models but decided against this because it would require some interpretation of the modeling method, and also because quite different results are often reported for the same model in different studies.

- Given the importance of the above information, it is recommended that model developers provide/publish self-report reliability metrics to properly assess uncertainties. This information could be provided in published papers, and submitted with the model to QMFR database.

- For the purposes of comparing and evaluating QSAR models, it is important that the model reports the QSAR prediction even for cases where a test set compound happens to be present in the model's training set or underlying knowledgebase. However, in such cases, several of the models instead report the experimental result. Obviously, the reliability of such a prediction is unrelated to the reliability of the QSAR model, so application of the DST-based combination approach is ambiguous. (Note that this is not only an issue for weight-of-evidence combination approaches, but more generally for evaluating any given QSAR model. A model that reports experimental results, when available, is simply executing a data lookup, which is substantially different from predictions based on QSAR.)

- Ideally, each prediction generated by a QSAR model should include a measure of the uncertainty associated with that prediction. For example, rather than simply reporting a particular compound to be Ames positive, it is desirable to have a probabilistic prediction that captures the uncertainty (e.g., 80% probability of being Ames positive). This quantitatively accounts for the fact that a given model may work much better for certain molecules than for others. Unfortunately, most model providers did not provide this level of detail for their predictions.

Weight-of-evidence combination for Ames models

Given the issues and limitations described above, in a first instance a much simplified DST-based analysis was performed on the Ames results (Table 14). Lacking model reliability metrics for most models, we first assumed all models to be equally reliable. Lacking probabilistic estimates for individual predictions, we used a probability of 1 for each positive prediction and 0 for each negative prediction. Finally, model predictions were used as reported, regardless of whether they are computational results from the QSAR model or experimental results from the model's underlying training data. Table 14 below summarizes this preliminary analysis. Figure 29 displays the results as a ROC graph.

Table 14: Preliminary results for Ames predictions for the 921-compound EFSA tests set using simplified weight-of-evidence approach

models combined	sensitivity	specificity
all 19 models	67%	91%
7 selected models*	69%	91%
2 models with highest sensitivity	58%	83%

* Selection criteria: sensitivity ≥ 55%, specificity ≥ 85%, number of compounds in domain of applicability ≥ 80%

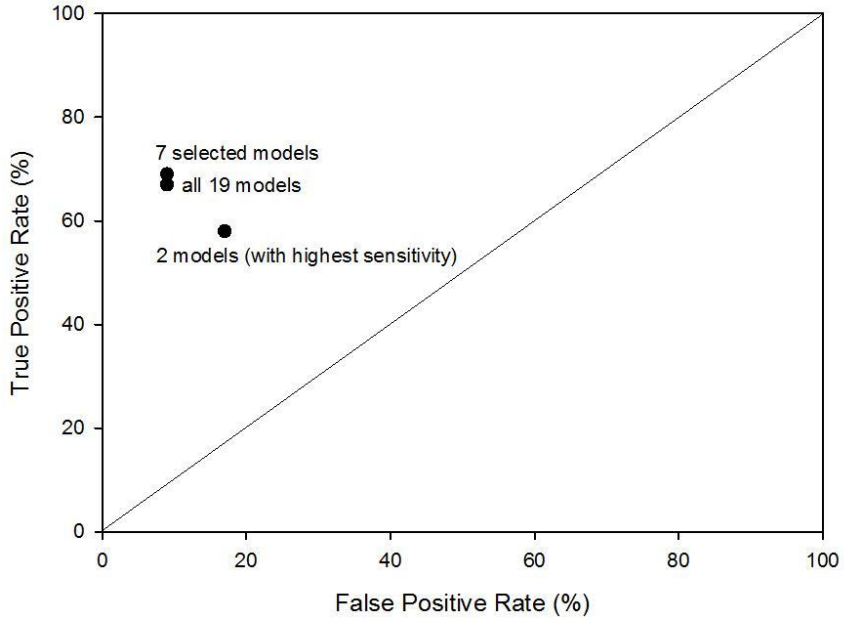


Figure 29: The results of Table 14 are displayed as ROC graph. The models are:

The 7 selected models:

1. Derek (Model name: Mutagenicity Endpoint *in vitro*, Assay: *In vitro* Mammalian Cell Gene Mutation Test)
2. Sarah (Model name: Mutagenicity Endpoint *in vitro*, Assay: Bacterial reverse mutation)
3. ChemTunes (Model name: AMES, Assay: Bacterial reverse mutation)
4. MultiCASE (Model name: GT_Expert Bacterial mutagenicity model, Assay: Bacterial reverse mutation)

5. MultiCASE (Model name: BMUT_PHARMA, OECD471 bacterial mutagenicity, Assay: Bacterial reverse mutation)
6. Leadscope Model applier (Model name: Leadscope genetox expert alerts v4, Assay: Bacterial reverse mutation)
7. Leadscope Model applier (Model name: Ames consensus from two models: Leadscope Salmonella statistical-based QSAR model v3 and Leadscope E.coli/TA102 statistical-based QSAR model v1, Assay: Bacterial reverse mutation)

The 2 models with highest sensitivity:

1. MultiCASE (Model name: BMUT_PHARMA, OECD471 bacterial mutagenicity, Assay: Bacterial reverse mutation)
2. Leadscope Model applier (Model name: Ames consensus from two models: Leadscope Salmonella statistical-based QSAR model v3 and Leadscope E.coli/TA102 statistical-based QSAR model v1, Assay: Bacterial reverse mutation)

For the individual QSARs, sensitivity ranged between 29% and 69% across all 19 models, while specificity values ranged from 65% to 99%. As discussed previously, only 4% of the compounds in the 918-compound EFSA test set are classified as Ames positive based on experimental data, so high prediction specificity is easily achieved. The advantage of selectively combining multiple sources of evidence is best illustrated by the results for the 7 selected models, which together do slightly better than any single model alone or all 19 models together. Another interesting result is that the observation that a combination using only the 2 models having the high sensitivities (slightly less than 70% for both) results in a substantial decrease in sensitivity (58%) after combination, in contrast to what it is usually observed. This indicates a fair amount of disagreement between these two apparently good models with regard to which particular compounds are predicted to be positive.

Four of the model providers did report reliability metrics for their Ames models, so we could perform a slightly more rigorous DST-based combination for these four models. Results are summarized in Table 15 and Figure 30.

Table 15: DST-based combination of 4 models for Ames

models	number of predictions	sensitivity	specificity
Ames-A	865	61%	66%
Ames-B	885	56%	85%
Ames-C	845	65%	88%
Ames-D	373	58%	94%
Combination of all 4 models	708	74%	83%

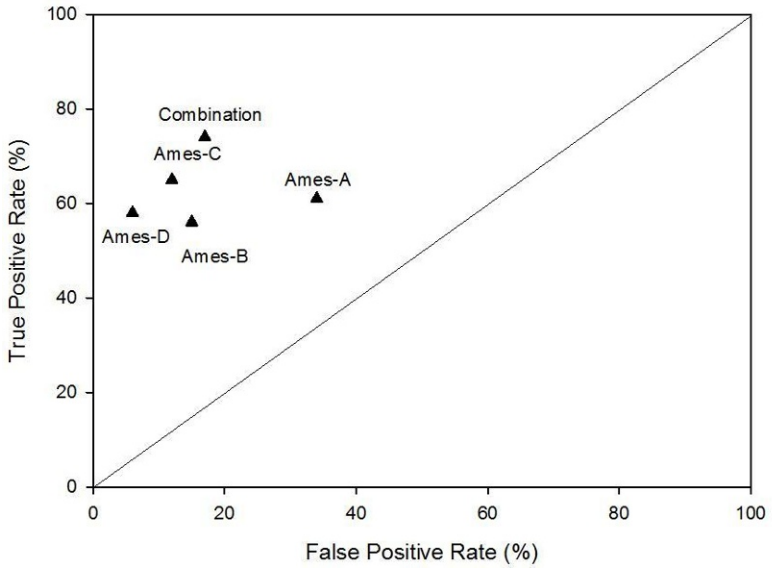


Figure 30: The results of Table 15 are displayed as ROC graph.

The codes are:
AMES-A: Lazar - Salmonella typhimurium mutagenicity;
AMES-B: ChemTunes - AMES
AMES-C: MultiCASE - BMUT_PHARMA, OECD471 bacterial mutagenicity
AMES-D: Percepta - Mutagenicity/Procaryote/Bacterial composite

Compounds for which models tend to disagree result in equivocal predictions when combined, and so the number of predictions (positive or negative) is somewhat lower for the combination; however, this approach attains a sensitivity appreciably higher than any of the individual models alone, in line with what usually happens when models are combined.

WoE combination for mammalian erythrocyte micronucleus (ivvMN) models

The EFSA test set includes experimental data for the mammalian erythrocyte micronucleus test for 219 compounds. Only 3 QSAR models were provided for this endpoint and results for a DST-based combination of the 2 models with relatively high sensitivities are summarized in Table 16 and Figure 31. It's worth noting that these 2 models generate predictions for less than 40% of the test set compounds. In this case, combination of the evidence does not increase the domain of applicability of the combined prediction, but there is a significant improvement in sensitivity.

Table 16: DST-based combination of 2 QSAR models for the mammalian erythrocyte micronucleus assay

models	number of predictions	sensitivity	specificity
ivvMN-A	86	42%	51%
ivvMN-B	78	60%	92%

Combination of both models	79	67%	84%

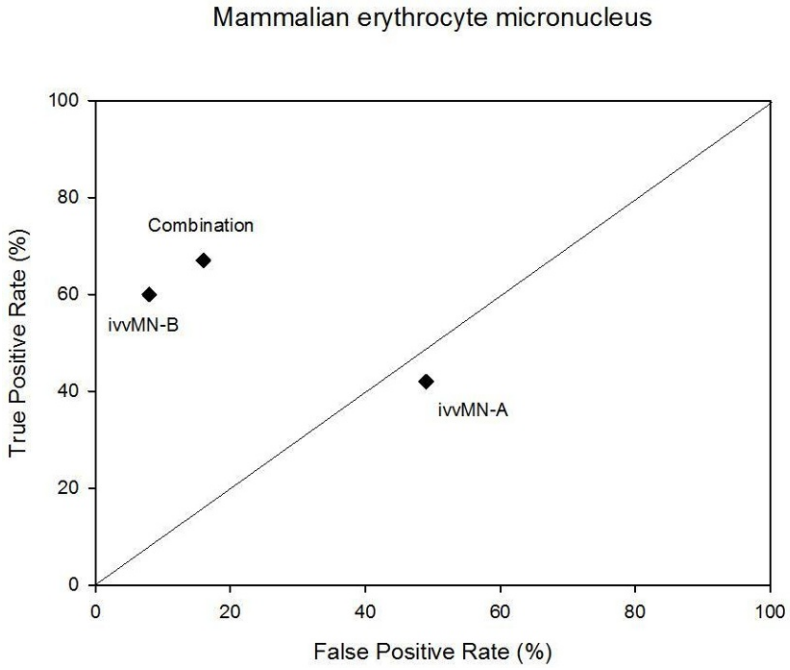


Figure 31: The results of Table 16 are displayed as ROC graph. The codes are:

ivvMN-A: (MultiCASE) GT3_MNT_MOUSE (Micronucleus test, *in vivo*, mouse)
ivvMN-B: (Percepta) Clastogenicity/Micronucleus/ Micronucleus *In Vivo* composite

WoE combination for in vitro mammalian chromosome aberration models

The EFSA test set includes experimental data for the *in vitro* mammalian chromosome aberration test for 566 compounds. Only 2 QSAR models for this endpoint reported reliability metrics, and results for these 2 models and their combinations are summarized in Table 17 and Figure 32.

Table 17: DST-based combination of 2 models for the *in vitro* mammalian chromosome aberration assay

models	**number of predictions**	**sensitivity**	**specificity**
ivtCA-A	382	56%	71%
ivtCA-B	166	47%	93%
Combination of both models	241	58%	80%

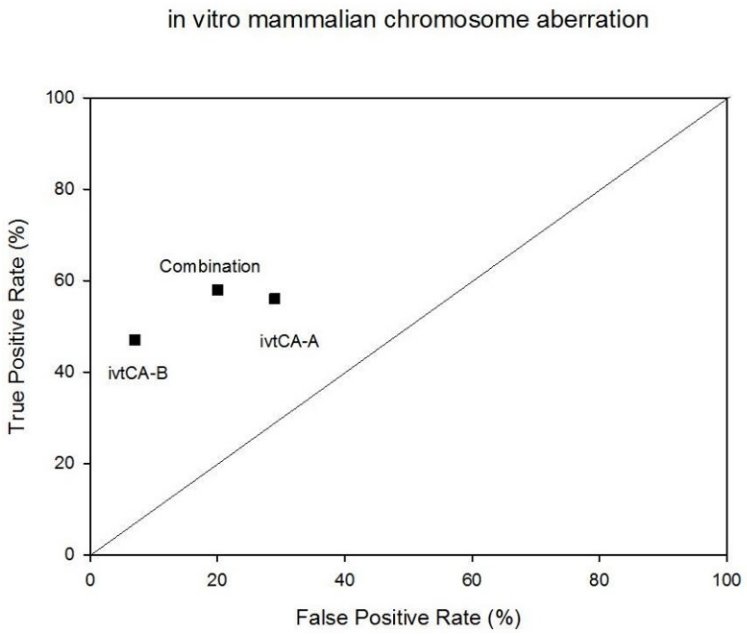


Figure 32: The results of Table 17 are displayed as ROC graph. The codes are:

ivtCA-A: (ChemTunes) *in vitro* chromosome aberration
ivtCA-B: (Percepta) Clastogenicity/Chromosome aberrations/ Chromosome aberrations *In vitro* composite

Overall, Tables 14 to 17 indicate that combinations of QSARs under a DST approach tend to increase Sensitivity at the expense of Specificity, thus supporting the results obtained with the simple combinations.

3.1.5.3. Concluding remarks on combinations

The evidence from the exercises on the EFSA database substantially agrees with the evidence from literature. As a general trend, the combination of QSARs increases Sensitivity at the expense of Specificity. This is valid both for simple combinations, and for the DST approach (even if a rigorous application of the DST approach has not been possible due to the partial lack of important information). This result may have an important role in pre-screening or prioritization: highly conservative (very sensitive) (Q)SARs can permit the conclusion that no further testing is necessary for chemicals with negative results. In addition, the literature shows that the integration of combinations of QSARs with expert knowledge has the potential to substantially improve QSAR screening. In principle, combinations of independent models (e.g., systems trained on different databases; statistical and expert systems; systems using different structural / physical chemical descriptors) should be preferred.

3.1.6. QSAR predictions: Summary/ Conclusions

Objective 2 focuses on a large scale exercise in which several commercial and publicly available (Q)SAR models were challenged in the prediction of the genotoxicity of pesticides and their metabolites, whose data are in a database provided by EFSA.

The exercise was organized by the Consortium that contacted, and collaborated with the owners / developers of the models. The commerciale software included: Derek Nexus v.5 and Sarah Nexus v.2.0.1 by Lhasa Limited, CASE Ultra 1.6.2.1 by MultiCASE Inc., Leadscope Model Applier v2.2.1.1 by Leadscope Inc., ChemTunes ToxGPS by Molecular Networks GmbH, Percepta 2016 (Build 2911) by ACD/Labs Inc. In order to obtain the best possible results, the predictions were kindly run directly by the developers, that have always offered the highest cooperation. The free software included Lazar v. 1.1.0 by *In Silico* Toxicology GmbH, ToxTree v. 2.6.13 by Ideaconsult Ltd and JRC, and Vega v.1.1.4 by IRCCS. Among them, Lazar predictions were kindly run by the developer.

A preliminary step of the analysis was the transformation of the granular genotoxicity data in the EFSA database (including details for all experiments) into Overall Outcomes for each chemical in each assay, in a format compatible with the outcomes of the (Q)SAR models. The numbers of genotoxicity results ranged from almost 1000 for the Ames test, to some dozens for other assays. It also appears that the experimental results are largely unbalanced, with a strong prevalence of negative genotoxicity results.

In another preliminary step, the general "chemistry" of the EFSA database (i.e., physical chemical properties and substructures composition) was compared with that of another database of pesticides collected *ad hoc* from the literature. No remarkable differences among the two databases were observed.

For the prediction exercise, five experimental assays were selected: Bacterial Reverse Mutation Assay (Ames test), Mammalian Bone Marrow Chromosome Aberration Test, Mammalian Erythrocyte Micronucleus Test, *In vitro* Mammalian Chromosome Aberration Test, *In vitro* Mammalian Cell Gene Mutation Test. The above assays were selected because: a) they represent different genotoxicity endpoints, b) they have a prominent regulatory role; c) (Q)SAR models are available; d) the number of chemicals tested is considered high enough as to permit reliable conclusions ($n > 100$).

Overall, the results of Objective 2 point to a substantial difference between the prediction of the Ames test on one hand, and that of the other experimental assays on the other hand.

The predictions of the Ames obtained in Objective 2 are displayed in Figure 21. The figure shows that there are differences among the performances of the different (Q)SARs. However all (Q)SAR models are in the top left ROC area, thus pointing to statistically significant predictions. Sensitivity ranges between 46% (Toxtree) and 71% (a model from Leadscope), Specificity between 66% (Lazar) and 98% (Percepta). This confirms the statistically significant predictions reported in previous exercises available in the literature (Figure 1).

The difference between QSAR applications to data in the public domain (Figure 1), and to the EFSA database (Figure 21) can be explained by the very different study designs: a) the compositions of chemicals of the two databases were largely different. The public domain database consists mainly of industrial chemicals, with a minor proportion of pesticides; b) the proportion of negatives / positives is strongly skewed towards negatives in the EFSA database, whereas the public database has a majority of positives; c) the predictions in Objective 1 were retrieved from retro-fit studies in literature (systems applied to the training set), whereas the predictions in Objective 2 were prospective ones on a database largely different from the training sets. The latter explains why the predictivity is higher in the literature data (Figure 1) in respect to the exercise on the EFSA data (Figure 21).

Regarding assays / endpoints different from *in vitro* bacterial mutagenicity (Ames), Objective 2 indicated that the reliability of the (Q)SAR models is still far from optimality. In fact, Figures 23 - 25 show that the predictions –collectively- are quite close to the diagonal line that represents random results in the ROC graphs. There is no possibility of comparing Objective 2 results for tests different from the Ames test with previous studies in the literature, since Objective 1 did not retrieve similar prediction exercises. Thus, this EFSA projects contributes with original information to the research on the predictivity of QSARs for genotoxicity endpoints different from bacterial mutagenicity.

As a matter of fact, the structure-activity approaches used for the various endpoints / tests are identical, whereas the type and quality of biological data is different. The Ames test has a clear scientific basis (each strain has been designed and constructed as to be able to respond to certain types of potentially mutagenic chemical structures), and has repeatedly been shown to have a high positive predictivity towards carcinogens. Unlike the Ames assay, other *in vitro* genotoxicity assays are subject to artifactual positive response and are not as effective in distinguishing carcinogens from noncarcinogens (Zeiger 2004). Thus, the Ames test appears to be a "cleaner" (less noisy) tool to identify DNA-reactive chemicals; as a consequence, the relationship between the biological effect and the causative chemical features are expected to be more easily identified, and consistent structure-activity rules established.

The high frequency of false positives of *in vitro* assays different from Ames have stimulated several revisions of protocols and evaluation criteria (Kirkland et al. 2007): the available database of experimental results for studying QSARs and Read Across on these assays is not only remarkably smaller of that for the Ames test, but probably also of a lower quality since include data obtained under different conditions. For example, a very recent work (Schisler et al. 2018) re-examined critically the database relative to the mouse lymphoma assay. The Authors found that, out of more than 1900 experiments representing 342 chemicals examined against updated acceptance criteria for background mutant frequency (MF), cloning efficiency (CE), positive control values, appropriate dose selection, and data consistency, only 17% of the evaluated experiments met all acceptance criteria. The Authors concluded that a similar curation should be done for other widely used genetic toxicology assays; however, it will be more difficult for certain assays (e.g., *in vitro* chromosomal aberrations) because important parameters such as level of cytotoxicity were often not evaluated/reported (see also (Honda et al. 2018)).

Finally, combinations of QSAR predictions of the EFSA genotoxicity results were explored, confirming evidence from literature. As a general trend, the combination of QSARs increases Sensitivity at the expense of Specificity. This may have an important role in pre-screening or prioritization: the application of highly conservative (very sensitive) (Q)SARs, especially in combination, may permit the conclusion that no further testing is necessary for chemicals with homogeneously negative results.

3.2. Read Across application to the EFSA genotoxicity database

Whereas during the years QSARs have undergone many performance evaluations, with special emphasis on comparative prospective exercises, nothing analogous exists for Read Across. The literature is rich in proposals for general workflows and criteria, but the published examples of applications –even though often quite detailed- are limited in number and do not provide sufficient material for assessing the real value of the proposed workflows. This is also related to the fact that Read Across is –by definition- the

distillation and use of information which is tailored on individual cases. However, a sufficient degree of generalization is necessary to draw conclusions on the most appropriate workflows.

This current project from EFSA concerns the evaluation of pesticides and metabolites and focuses on establishing the conceptual frameworks to identify and document sources of evidence, estimate uncertainty, and combine all information for the final outcome. Rigorous and objective metrics to evaluate the performance or quantify uncertainties of RA results have not been the main focus of many current computational tools. To address these deficiencies, there is a clear need to develop RA approaches that are more robust and reproducible whilst maintaining the original rationales.

In most conventional Read Across, the initial step usually starts with identifying the best analogues for a given target compound with a specific endpoint in mind. In the present case, analogues are pre-determined by the selection of metabolites, since a primary need of EFSA is that of predicting the metabolites genotoxicity by exploiting all the available information on the parent compound (for which full dossiers are submitted to EFSA). Whenever possible, this is performed with 1:1 Read Across applications.

To explore better the field, the analysis is performed with two different strategies. Then results are compared and discussed.

3.2.1. Read Across exercises: Strategy I

This section presents **twenty-eight Read Across exercises**, performed *ad hoc* for this project for the two endpoints **Ames test** and ***in vitro* Chromosomal Aberrations**.

As said above, whereas QSARs have undergone many performance evaluations, with special emphasis on comparative prospective exercises, nothing analogous exists for Read Across. However, a sufficient degree of generalization is necessary to draw conclusions on the most appropriate workflows. With this in mind, in this section we present a transparent, clearly structured RA workflow. The workflow is platform independent. It was kept as simple and transparent as possible, so that it is easy to highlight strength and weakness of the approach, which can be easily replicated by other investigators. The simplicity of the approach allows for the control and intervention of the human expert at every stage of the process.

Together with this, we have focused on the primary need of EFSA of predicting the metabolites genotoxicity by exploiting all the available information on the parent compound (for which full dossiers are submitted to EFSA). Typically, this is performed with 1:1 Read Across applications. The assessment of the feasibility –or not- of this step is crucial.

3.2.1.1. Study design

To maximize the information gained from the RA application, a careful selection of chemicals was performed: a) parent / metabolite pairs that have experimental results for both Ames and *in vitro* Chromosomal Aberrations (CHA) tests; b) presence of both positive and negative experimental results; c) inclusion of a number of metabolites erroneously predicted by QSARs for the Ames test in Objective 2 of this project. The total number of parent / metabolite pairs is 14 (Tables 18 and 19). In the tables, Source is the Pesticide, or Parent chemical, and Target is the metabolite whose genotoxicity has to be predicted.

Appendix F displays the structures of the selected Pesticide / Metabolite pairs. The structures were depicted with the OECD QSAR Toolbox. The inspection of the structures shows that: a) the "metabolic" transformations (including all sorts of interactions, degradations, etc…) did not produce large structural changes in a number of pairs (1 to 6), whereas the changes are more dramatic for other pairs. This fact contributes to the diversity and representativity of the chemical pairs selected for the exercise.

Table 18: Experimental genotoxicity results of the molecules investigated

RA	SUB_ID	Source COM_ID	Target COM_ID	Source STY	Target STY	Source CHA	Target CHA
1	1170	1653	50021	0	0	1	0
2	1347	1698	1895	0	0	0	0
3	1133	1624	6028	1	1		
4	35058	50616	1689	0	0	1	0
5	1166	1668	50576	0	0	0	1
6	1347	1698	1897	0	0	0	0
7	15043	15674	15678	0	0	0	0
8	85027	75507	75509	0	0	0	1
9	85027	75507	75511	0	0	0	1
10	3842	6185	15493	0	0	0	1
11	35031	50309	50554	1	0		1
12	1416	1509	2098	0	1	0	0
13	1166	1668	15576	0	0	0	1
14	85018	75368	75367	1	0	0	0

STY = Salmonella typhimurium Ames test; CHA = *in vitro* Chromosomal Aberrations test. Colors point to chemicals erroneously predicted by the QSARs in Objective 2: Blue = erroneous prediction by all QSARs; Light Blue = erroneous prediction by the majority of QSARs.

Table 18 gives in tabular form the experimental genotoxicity data for the Salmonella assay (STY) and the *in vitro* Chromosomal Aberrations assay (CHA) (0 = Negative; 1 = Positive). The outcomes reported are Overall scores generated by the genotoxicity experts of the Team upon inspection of the granular data in the EFSA database. The chemicals are identified by the codes used in the EFSA Genotoxicity Database: COM_ID for both the source (Parent) and target (Metabolite), together with SUB_ID of the group they belong to.

Table 19: physical chemical data of molecules, and structural similarity between source and target

RA	SUB_ID	Source COM_ID	Target COM_ID	Source MW	Target MW	Source logKoW	Target logKoW	Similarity
1	1170	1653	50021	256	240	0.56	0.44	0.91
2	1347	1698	1895	230	211	3.27	2.55	0.87
3	1133	1624	6028	213	229	-0.79	0.72	0.87
4	35058	50616	1689	223	199	3.46	3.19	0.82
5	1166	1668	50576	216	271	2.92	3.33	0.78
6	1347	1698	1897	230	202	3.27	2.33	0.72
7	15043	15674	15678	412	248	4.26	3.09	0.73
8	85027	75507	75509	407	300	1.08	3.33	0.83
9	85027	75507	75511	407	226	1.08	0.14	0.58
10	3842	6185	15493	450	243	6.85	3.85	0.56
11	35031	50309	50554	339	215	1.49	-0.07	0.54
12	1416	1509	2098	447	189	4.69	0.97	0.49
13	1166	1668	15576	174	271	1.22	3.33	0.48
14	85018	75368	75367	240	153	1.7	-3.14	0.29

The color indicates when the parameter values of parent and metabolite were considered "too" different (see further explanation, and use below).

It should be emphasized that the selection of erroneous QSAR predictions was intentional only for the Ames test; the many more erroneous QSARs for Chromosomal Aberrations were just a consequence of the above primary choice. Even with a limited number of chemicals, the pattern of erroneous QSAR predictions in the table points to the different development levels of QSARs for the two endpoints (and reflects the conclusions of Objective 2): satisfactory for the Ames test, and mediocre for the *in vitro* Chromosomal Aberrations assay.

Table 19 reports the values of the following properties of the chemicals: Molecular Weight, Log KoW, Dice / Atom Centered Structural Similarity between Target and Source. These were calculated with the OECD QSAR Toolbox.

3.2.1.2. The Read Across approach in Strategy I

For every chemical, we calculated the following basic properties: Molecular Weight (MW), Log KoW, Dice / Atom Centered Structural Similarity (Table 19); these were used to assess the level of similarity (or analogy) between parents and metabolites. Regarding the selected parameters: a) differences in MW code for coarse-grain changes between the structures of parent and metabolite; b) differences in

Log KoW indicate how the structural changes are reflected in a property that is crucial for ADME effects; and c) the level of Structural Similarity is related to the similarity in the types of atoms composing the molecules. Since the three parameters are poorly inter-correlated (this can be easily perceived by inspecting Table 19), taken in combination they offer a thorough tool to assess the chemical analogy between parent and metabolite.

In this approach, for chemicals in the pair similar for the entire profile of the three parameters, we attributed the genotoxicity of the parent pesticide to the metabolite. Based on our experience and of other investigators, the criteria for accepting the similarity of the three parameters were set as follows: a) MW ± 20%; b) Log KoW ± 1 unit; c) Structural Similarity higher than 70%. Based on these stringent criteria, for 5 out 14 pairs the 1:1 RA was applied directly for Ames, and 4 out 14 for Chromosomal Aberrations (since one data point was missing). The color in Table 19 points to cases where the above similarity criteria are not fulfilled.

The metabolites for which the similarity of the overall pattern with the parent compound was not judged sufficient for directly reading across its genotoxicity, were assessed with a "one-target / many-source chemicals" approach, in four different ways.

First, we looked for analogues in the EFSA genotoxicity database, by setting:

1) Structural Similarity higher than 70%; and

2) Structural Similarity higher than 60%.

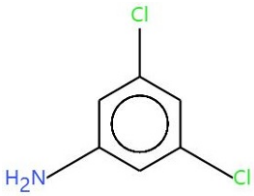
In another type of search, we described the basic skeleton of the metabolite with SMARTS, and we looked for chemicals sharing the same SMARTS in:

3) the EFSA genotoxicity database; and

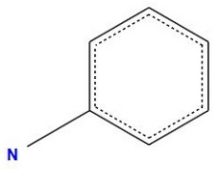
4) all the genotoxicity databases contained in the OECD QSAR Toolbox.

SMILES and SMARTS: short introduction

To explain in short the difference between SMILES and SMARTS, the following structure refers to a specific molecule (3,5 Dichloro aniline, not included in this study), and can be coded with the SMILES linear notation: Nc1cc(Cl)cc(Cl)c1.



The basic skeleton of aniline, without substituents, is common to all aromatic amines and can be coded with another linear notation called SMARTS: Nc1ccccc1, corresponding to the following:



The use of SMARTS in software tools (like the Toolbox and many others), allows one to retrieve in a database all chemicals sharing the some basic skeleton (in the example above, all aromatic amines irrespective of substituents). This approach was used in Options 3) and 4) above to identify analogues of the targets, by selecting every time the appropriate basic skeleton (the example shown is not among the EFSA db chemicals, but was selected just for the sake of a simple illustration of the approach).

A consensus prediction from the expert inspection of the four groups of analogues, largely based on the majority of positives / negatives, was obtained (see details below).

All the above steps were performed with the OECD QSAR Toolbox; however they are platform independent, and do not follow the main workflow suggested in the structure of the Toolbox itself.

In the search for analogues, the EFSA genotoxicity database (transformed into Overall genotoxicity outcomes) was implemented in the Toolbox and was used for the above Options 1), 2) and 3).

Option 4) in addition used: the Bacterial Mutagenicity ISSSTY, the Genotoxicity OASIS, and the Toxicity Japan MHLW databases as implemented in the Toolbox.

3.2.1.3. Strategy I: The case studies analyses: prediction of the Ames test

Table 20 presents the results for the Ames test Read Across. The table reports the original experimental data and the predictions for the target chemical.

Table 20: Read Across analysis for Ames test case studies: results.

RA	Source STY	Target STY	RA 1:1	POS_Anal. >70% sim EFSA db	POS_Anal. >60% sim EFSA db	POS_Anal. SMARTS EFSA db	POS_Anal. SMARTS All db	RA many to 1
1	0	0	0	0__1	0__4	0__1	0__1	
2	0	0	0	0__4	0__4	0__14	0__17	
3	1	1	1	1__1	1__1		0__1	
4	0	0	0	0__7	0__11	3__13	3__13	
5	0	0	0	0__1	0__2			
6	0	0	0	0__2	0__5	0__2	0__9	
7	0	0	NA	0__9	0__10			0
8	0	0	NA	0__5	0__11	0__3	0__3	0
9	0	0	NA	0__3	0__11	0__3	0__3	0
10	0	0	NA		0__1	0__4	0__4	0
11	1	0	NA	0__3	0__5	0__1	0__1	0
12	0	1	NA	0__1	0__13	2__15	10__38	0

13	0	0	NA		0__2		0__1	0
14	1	0	NA		0__1			0

RA = code of the chemical pair (see identifiers in Tables 18 and 19);
Source STY = Ames test experimental result for the Source chemical;
Target STY = Ames test experimental result for the Target chemical;
RA 1:1 = direct Read Across prediction from Source to Target (NA = Not Applicable);
POS_Anal. >70% sim EFSA db = Number of Positive Analogues _out of _ Total number of analogues with at least 70% similarity with the target (search in the EFSA Genotoxicity DB);
POS_Anal. >60% sim EFSA db = Number of Positive Analogues _out of _ Total number of analogues with at least 60% similarity with the target (search in the EFSA Genotoxicity DB);
POS anal SMARTS EFSA db = Number of Positive Analogues _out of _ Total number of analogues sharing the same basic skeleton (SMARTS) of the target (search in the EFSA Genotoxicity DB);
POS anal SMARTS All db = Number of Positive Analogues _out of_ Total number of analogues in the Toolbox genotoxicity databases sharing the same basic skeleton (SMARTS) of the target (search in all genotoxicity DBs in the Toolbox);
RA many to 1= Read Across predictions based on the majority of positive analogues.
NA = Not Applicable
Color codes: Blue = Ames outcomes incorrectly predicted by all QSARs;
Yellow = Correct Read Across predictions.

As shown in Table 20, Source and Target in pairs 1 to 6 have very similar profiles of MW, Log KoW and Structural Similarity: thus, 1:1 RA was considered acceptable. They also correspond to cases in which the "metabolic" transformations produced minor structural changes (see structures in Appendix F). The Ames test values of the Source (Parent) were attributed to the Target. The correct 1:1 RA predictions were 6 out 6.

The results of the 1:1 RA were confirmed by the analysis of additional analogues. The table reports the mutagenicity of additional analogues (retrieved by the 70% or 60% similarity, or by SMARTS search). It appears that, for each target, the majority of Ames results were concordant with those of both the target and source.

For the remaining pairs, 1:1 RA was not considered reliable because of insufficient similarity between Target and Source. The Read Across predictions were performed by searching for analogues in the four ways described above. With the exception of RA 12, the analogues were all negative, in agreement with the Ames test results of the respective Targets. For RA 12, the large majority of analogues was negative as well, and the target was predicted as negative. The correct predictions from these Read Across analyses were 7 out 8.

As additional information, Table 20 shows analogues also for the cases in which 1:1 RA was applied.

An interesting observation regards three Targets (in blue in the table) whose Ames mutagenicity was erroneously predicted by all QSARs (Objective 2): COM_ID = 15576, 50021, and 2098. As a matter of fact, the first two chemicals above have Structural Alerts (results not shown); however their experimental Ames results are negative. In two out three cases, Read Across generated correct predictions.

Two case studies: a more in depth analysis

Whereas the overall evidence from this exercise is quite self-explanatory, a couple of cases deserve further comments. These are Read Across 4 and 12.

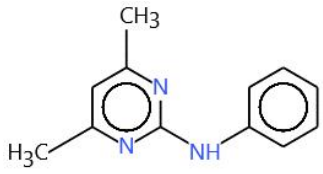
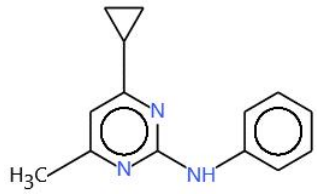
The target of Read Across 4 is very similar to its source chemical, so 1:1 RA can be applied and provides the correct prediction. Because this is an exploratory work, we investigated further by looking for analogues of the target (even though this was not strictly necessary within the scheme we applied). It appears that this prediction is also supported by the fact that in the EFSA database there are several analogues with negative Ames results: 7 negative analogues with at least 70% Structural Similarity, and 11 negative analogues with at least 60% Structural Similarity (Table 20). When analogues are searched

for by SMARTS, only 3 out 13 are actually positive. In terms of majority vote, this is again a support for the (correct) negative prediction of 1:1 RA.

We elaborated further and we ordered the analogues in terms of their overall similarity with the target. Tables 21 and 22 show the target, together with the analogues that fulfill the similarity requirements (as set above) for the three parameters MW, Log KoW, and Structural Similarity.

In the case of Read Across 4 (Table 21) four negative, highly similar analogues were identified, thus supporting again the negative call from Read Across for this chemical. The inspection of the structures also shows that the Nitrogen in the aromatic amine moiety is sterically hindered by the two rings, and thus it is not available for the metabolic transformation to reactive species.

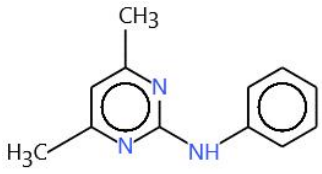
Table 21: Strategy I; Read Across 4: Target and analogues

Target / analogues	SMILES	CAS N°	Similarity	Mol Weight	log Kow	Ames test results
 Pyrimethanil (Target)	Cc1cc(C)nc(Nc2ccccc2)n1	53112-28-0	100%	199	3.19	0
confidential	confidential	NA	84.80 %	241	2.12	0
 Cyprodinil	Cc1cc(nc(Nc2ccccc2)n1)C1CC1	121552-61-2	81.30 %	225	3.99	0
confidential	confidential	confidential	81.30 %	223	3.46	0

confidential	confidential	NA	78.80 %	243	2.63	0

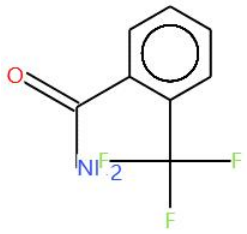
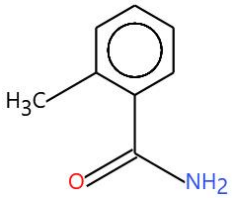
Table 20 shows that the SMARTS search identified also three positive analogues for Read Across 4. They are shown in Table 21bis, together with the target metabolite.

Table 21 bis: Strategy I; Read Across 4: Target and subset of positive analogues

Target / analogues	SMILES	CAS N°	Similarity	Mol Weight	log Kow	Ames test results
 Pyrimethanil (Target)	Cc1cc(C)nc(Nc2ccccc2)n1	53112-28-0	100%	199	3.19	0
confidential	confidential	NA	62.90%	268	2.9	1
confidential	confidential	NA	57.10%	268	3.85	1
confidential	confidential	NA	57.10%	275	0.74	1

It appears that all three chemicals are under the 70% Similarity cut-off, and also differ –to different extents- from the target for the other parameters. In addition, they are structurally different from the target, since they possess a nitro-aromatic Structural Alert, or a moiety that has the potential to generate an iminoquinone. These structural characteristics justify the Ames positivity, and are a further reason not to use these analogues in Read Across.

Table 22: Strategy I; Read Across 12: Target and analogues

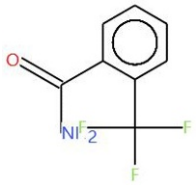
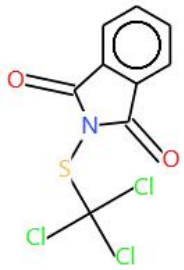
Target / analogues	SMILES	CAS N°	Similarity	Mol Weight	log Kow	Ames test results
 2-(Trifluoromethyl) benzamide) (Target)	C1=CC=C(C(=C1)C(=O)N)C(F)(F)F	360-64-5	100%	189	0.97	1
 2,3-difluoro-6-(trifluoromethyl)benzamide	NC(=O)c1c(F)c(F)ccc1C(F)(F)F	NA	71.40%	225	0.97	0
	Cc1ccccc1C(N)=O	527-85-5	69.60%	135	0.55	0

Regarding Read Across 12, it should be said that the case of the metabolite COM_ID = 2098 is less straightforward. The EFSA genotoxicity database reports experiments in different Salmonella strains: the chemical is positive in strain TA100, so accordingly it was given a positive Overall outcome. The chemical has no Structural Alerts specific for genotoxicity, and was predicted as negative by all QSARs. The analogues retrieved in the EFSA database with Options 1 and 2 are negative in the Ames test. Options 3 and 4 gave, together with few positives, a majority of negative analogues. Thus, to follow the same approach used for the other cases the metabolite is predicted as negative by Read Across.

In line with the previous case, we checked the most similar analogues (Table 22). The first analogue shown fulfills all similarity criteria for the three parameters and is negative. The second one is close to the criteria, and is negative as well.

Table 22bis shows two positive analogues retrieved by the SMARTS search in the EFSA database. It appears that their structural similarity with the target is really very low, both in terms of Similarity score and of composition in functional groups. The target is a benzamide, whereas the core structure of the analogues is quite different chemically, i.e., phthalimide. Thus the analogues should not be used as a basis for Read Across predictions. Similar patterns of results are with analogues from other databases (results not shown).

Table 22 bis: Strategy I; Read Across 12: Target and subset of positive analogues

Target / analogues	SMILES	CAS N°	Similarity	Mol Weight	log Kow	Ames test results
 2-(Trifluoromethyl) benzamide) (Target)	C1=CC=C(C(=C1)C(=O)N)C(F)(F)F	360-64-5	100%	189	0.97	1
	COP(=S)(OC)SCN1C(=O)c2ccccc2C1=O	732-11-6	37.50%	317	2.48	1
	ClC(Cl)(Cl)SN1C(=O)c2ccccc2C1=O	133-07-3	41.40%	297	2.84	1

In conclusion, both the majority votes in Options 1 to 4, and the analysis of the closest analogues point to a final negative prediction for this metabolite in Read Across 12. The reasons for the discrepancy between predictions (QSAR and Read Across) and experimental results are beyond the scope of the present analysis (maybe presence of impurities in the tested sample?).

Overall, the accuracy of Read Across for the Ames test (in the different approaches) was 13 / 14. It should be remarked as well that 1:1 RA predictions (Pairs 1 to 6) were supported by the analysis of analogues.

3.2.1.4. Strategy I: The case studies analyses: prediction of *in vitro* Chromosomal Aberrations

In order to increase the comparative value of this Read Across exercise, we selected chemicals that have also *in vitro* Chromosomal Aberrations data (except Pair 3 Metabolite). Since the approach applied

to the Ames data in the previous section was successful, we used the same approach for the *in vitro* Chromosomal Aberrations. Table 23 presents the Read Across results.

Table 23: Read Across analysis for *in vitro* Chromosomal Aberrations test case studies. The table reports the original experimental data and the predictions for the target chemical

RA	Source CHA	Target CHA	RA 1:1	POS_Anal >70% sim EFSA db	POS_Anal. >60% sim EFSA db	POS_Anal. SMARTS EFSA db	POS_Anal. SMARTS All db	RA many to 1
1	1	0	1	1__1	1__2	1__1	1__1	
2	0	0	0	0__4	0__4	1__12	1__12	
3								
4	1	0	1	1__2	1__3	2__3	2__3	
5	0	1	0	1__3	6__12			
6	0	0	0	0__2	0__4	1__15	2__21	
7	0	0	NA	1__5	1__5			0
8	0	1	NA	0__3	0__7	0__2	0__2	0
9	0	1	NA	0__2	2__9	1__3	1__3	0
10	0	1	NA		0__1	1__10	1__11	0
11		1	NA	0__3	0__4	0__1	0__1	0
12	0	0	NA	0__1	0__6	4__29	8__41	0
13	0	1	NA		1__3			0
14	0	0	NA					NA

RA = code of the chemical pair (see identifiers in Tables 18 and 19);
Source CHA = *in vitro* Chromosomal Aberrations test experimental result for the Source;
Target CHA = *in vitro* Chromosomal Aberrations test experimental result for the Target;
RA 1:1 = direct Read Across prediction from Source to Target (NA = Not Applicable);
POS_Anal. >70% sim EFSA db = Number of Positive Analogues _out of _ Total number of analogues with at least 70% similarity with the target (search in the EFSA Genotoxicity DB);
POS_Anal. >60% sim EFSA db = Number of Positive Analogues _out of _ Total number of analogues with at least 60% similarity with the target (search in the EFSA Genotoxicity DB);
POS anal SMARTS EFSA db = Number of Positive Analogues _out of _ Total number of analogues sharing the same basic skeleton (SMARTS) of the target (search in the EFSA Genotoxicity DB);
POS anal SMARTS All db = Number of Positive Analogues _out of_ Total number of analogues in the Toolbox genotoxicity databases sharing the same basic skeleton (SMARTS) of the target (search in all genotoxicity DBs in the Toolbox);
RA many to 1 = Read Across predictions based on the majority of positive analogues.
NA = Not Applicable
Color codes:
Blue = in vitro Chromosomal Aberrations outcomes incorrectly predicted by all QSARs;
Light Blue = *in vitro* Chromosomal Aberrations outcomes incorrectly predicted by the majority of QSARs;
Yellow = Correct Read Across predictions.

As remarked above, five (six for the Ames test) pairs of chemicals were similar enough as to apply 1:1 RA. The correct predictions were 2 out 5 (Pairs 2 and 6).

For the other cases, the Read Across predictions based on the majority vote from several analogues were correct 2 out 7 times (RA 7 and 12). No analogues were found for Metabolite 14.

Overall, the accuracy of Read Across for the *in vitro* Chromosomal Aberrations was 33% (correct predictions: 4 out 12).

In the Read Across exercise above for the Ames test, we further checked the majority vote from analogues by investigating if some analogues fulfilled the high similarity criteria set for MW, Log KoW, and Structural Similarity. This is particularly interesting in sets of analogues with contrasting results. For the Ames test, it appeared that the majority vote based on all analogues was confirmed in the narrower sets.

We performed the same exercise for the Chromosomal Aberrations Read Across. The overall result was similar, with a few exceptions mainly due the limited number of analogues. For example in RA 5 (Table 23), the Target is positive (Com_id =50576) whereas the Source (Com_id = 1668) is negative. Since RA 1:1 is applicable, an incorrect negative call was attributed to the Target.

In the extended Many-to-1 RA, we found in the EFSA database 3 analogues within 70% Similalrity, and 12 analogues within 60% similarity (with 6 positives out of 12). Considering the 3 most similar analogues (fulfilling all criteria) (Com_id 1668, 1470, and 1604), it appears that the majority (2 out 3) are negative, thus confirming the call of the 1:1 analogue (source). The other positives (not shown) are all below the accepted thresholds for the three parameters. This example supports the reliability of the majority vote based on the extended set of analogues.

We checked also the other cases (results not shown) and we got similar results, with the exception of RA 10, 12, and 13 where "very similar" analogues were not found. In any case, the majority vote from analogues was sistematically in agreement with the call of the parent compound, even if the Chromosomal Aberrations database is less dense (hence potentially more erratic) than that of the Ames test.

Table 23 also highlights Targets: a) erroneously predicted by all QSARs (n = 4, in blue), and b) erroneously predicted by the majority of QSARs (n = 4, light blue) (data from Objective 2). Out of the 8 incorrect Read Across predictions, QSARs predictions were incorrect 6 times, indicating that -in this set of chemicals- it is not possible to use the QSAR results to improve Read Across. This is different from what is apparent in the exercise with the Ames test (Table 20).

3.2.1.5. Strategy I: expanding Read Across with mechanistic reasoning; a case study

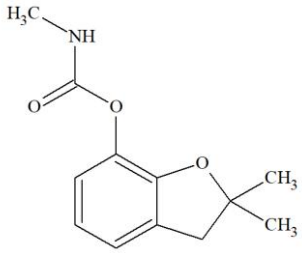
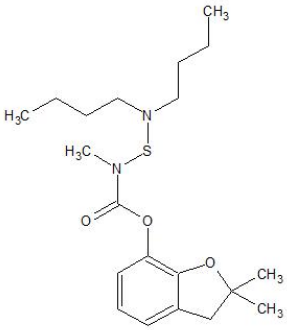
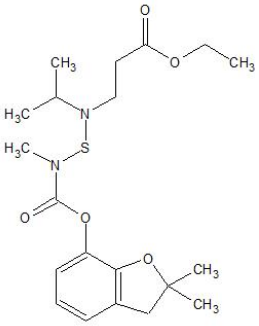
This section presents Read Across applications for the Ames test for the chemical substance Carbofuran (target), which has the peculiarity of being a metabolite of two parent compounds, namely Carbosulfan and Benfuracarb (sources). In addition, the availability of a large range of ADME and toxicological information for the sources permits to apply a more mechanistically-based Read Across analysis. Thus, this case study is developed with an approach different from the previous 1:1 RA examples.

Looking for structural similarity and mechanism of action

As shown by the data reported in Table 24, structure similarity among the target and test substances is around 60%. The description of the molecules by the profilers for organic functional groups highlight the presence of common structural features (Table 24). In particular, the three substances share the same structural core, differing in the N-substitutions in the carbamoyl moiety. These chemicals belong in fact to the class of carbamate pesticides, which exert the toxic action by inactivation of the enzyme

acetylcholinesterase (Fukuto, 1990). The process involves formation of the enzyme-inhibitor complex with subsequent carbamylation of a serine hydroxyl, resulting in inhibition of the enzyme.

Table 24: Data matrix for the metabolite Carbofuran and its parent compounds, Carbosulfan and Benfuracarb. Physicochemical information, similarities and predictions were retrieved from the OECD QSAR Toolbox *(Ames test experimental data: Overall scores generated from EFSA genotoxicity database by genotoxicity experts, Objective 2)*

Name	Carbofuran	Carbosulfan	Benfuracarb
COM_ID	1606	1607	1488
Sub_ID	1141	1141	1139
			
CAS	1563-66-2	55285-14-8	82560-54-1
SMILE	CNC(=O)Oc1cccc2CC(C)(C)Oc12	CCCCN(CCCC)SN(C)C(=O)Oc1cccc2CC(C)(C)Oc12	CCOC(=O)CCN(C(C)C)SN(C)C(=O)OC1=CC=CC2=C1OC(C2)(C)C
Molecular formula	C12H15NO3	C20H32N2O3S	C20H30N2O5S
Molecular weight	221	381	411
log KoW	2.3	5.57	4.06
similarity		62%	59%
OECD QSAR Toolbox Profilers			
Organic functional groups	Aryl|Benzofuran/ Dihydrobenzofuran| Carbamate| Coumaran	Aryl|Benzofuran/ Dihydrobenzofuran| Carbamate| Coumaran	Alkane, branched with secondary carbon| Aryl|Benzofuran/ Dihydrobenzofuran| Carbamate| Carboxylic acid ester| Coumaran| Isopropyl
OASIS Ames	No Alert found	No Alert found	No Alert found
ISS Ames	No Alert found	No Alert found	No Alert found
DNA binding (OASIS)	No Alert found	No Alert found	No Alert found
DNA binding (OECD)	No Alert found	No Alert found	No Alert found
Oncologic	Carbamate Type Compounds	Carbamate Type Compounds	Carbamate Type Compounds
Ames test experimental data	1	0	0

The application of three list of SAs (available in the QSAR Toolbox) relevant for the Ames test, did not point to potential positivity of the three chemicals (see Table 24).

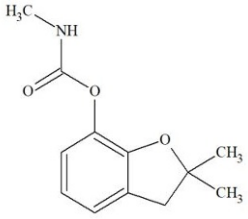
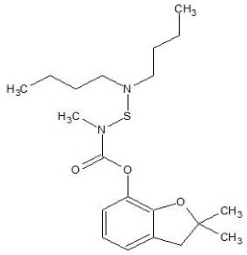
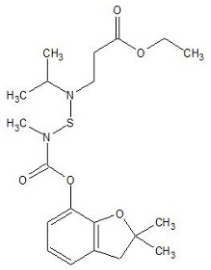
Although no genotoxic reactivity has been detected by the profilers used, it is known that the carbamate moiety can be associated with genotoxic mode of action of carcinogenic substances (Benigni and Bossa, 2011). This is confirmed by the application of the Oncologic profiler for carcinogenicity (as implemented in Toolbox), which highlights this common feature of the three chemicals (Table 24). Carbofuran is classified by Oncologic as with a "low concern". It is not possible to analyze in detail the differences in potential reactivity of the three chemicals, because the software cannot go into the details of the N,N disubstitution of the test substances. Nevertheless, the indication for carcinogenicity is that N,N-disubstitution or substitution with bulky groups generally decreases the concern.

This result points to the possibility that the target carbofuran represents the worst case with respect to the other two parent chemicals, toward a genotoxic reactivity.

Exploring all available experimental results

As recommended for using RA in the framework of REACH legislation *(Grouping of substances and read-across approach – an illustrative example, ECHA-13-R-02-EN Publ.date: April 2013)*, a matrix with available data should be constructed, in order to compare the whole toxicity profile of chemicals of interest. In the present case study, as carbofuran is an active substance *per se*, plenty of data are available for the three chemicals and it was possible to collect a dense data matrix with information on ADME and toxicological properties (Table 25). The data reported show that carbofuran is much more toxic than the parents for range of endpoints (e.g. LD50 and LC50 values for acute toxicity, NOAELs for oral-short term, long term and maternal toxicity). This evidence confirms that the target is the worst case with respect to the other two chemicals and it is not possible to justify a RA of the negative Ames test result.

Table 25: Data matrix for the metabolite Carbofuran and its parent compounds, Carbosulfan and Benfuracarb. Toxicological data taken from EFSA conclusions of the three active substances (EFSA, 2009a; b; c).

Name	Carbofuran	Carbosulfan	Benfuracarb
			
CAS	1563-66-2	55285-14-8	82560-54-1
SMILE	CNC(=O)Oc1cccc2CC(C)(C)Oc12	CCCCN(CCCC)SN(C)C(=O)Oc1cccc2CC(C)(C)Oc12	CCOC(=O)CCN(C(C)C)SN(C)C(=O)OC1=CC=CC2=C1OC(C2)(C)C
Molecular formula	C12H15NO3	C20H32N2O3S	C20H30N2O5S
ADME			
Rate and extent of absorption	83-92 % (urine and air) within 32 hour (0.4 mg/kg bw, rat)	High bioavailability (> 70 %) within 24 h	Relatively rapid, 70-81 % (urine within 144 h)
Distribution	Large, highest residue in liver	Large, highest level in excretory organs and carcass	Large, highest level in excretory organs and carcass
Potential for accumulation	No evidence of accumulation	No evidence of accumulation	No evidence of accumulation
Rate and extent of excretion	92 % of phenyl part within 48 h mainly via urine (89 %) and faeces (2.5 %); carbamate moiety excreted within 32 h in air as CO2	Rapid and extensive (app. 90 %) within 24 h mainly via urine (63 - 78 %)	Extensively excreted, 66-76 % in urines; 10-12 % in faeces within 48 h
Metabolism in animals	Oxidation at C-3, generating 3-OH-metabolites, further oxidation (3-ketocarbofuran) and/or excretion as conjugates Hydrolysis of carbamate bond into CO2	Extensive metabolism (> 80 %): hydrolysis at C-7 to form carbofuran-7-phenol and at N-S to form carbofuran and dibutylamine. Carbofuran-7-phenol and carbofuran are oxidized at C-3 generating 3-OH-metabolites. Dibutylamine is oxidized to CO2 and volatiles.	Extensive; Benfuracarb breaks down to carbofuran, which is further hydroxylated/oxidated into 3-keto-carbofuran-phenol, 3-hydroxy-carbofuran, 3-hydroxy-carbofuran-phenol, carbofuran-phenol
Acute toxicity			
Rat LD50 oral	7 mg/kg bw	Rat: 101 mg/kg bw (♀); 180 mg/kg bw (♂) Rabbit: 42.7 mg/kg bw	205 mg/kg bw
Rat LD50 dermal	1000< LD50<2000 mg/kg bw	3700 mg/kg bw	> 2000 mg/kg bw
Rat LC50 inhalation	0.05 mg/L (13 mg/kg bw)	0.61 mg/L air /1h (whole body, aerosol exposure = 164 mg/kg bw)	0.344 mg/L air/4 h (nose only as liquid droplet aerosol)
Skin irritation	Non-irritant	Non- irritant	Non-irritant

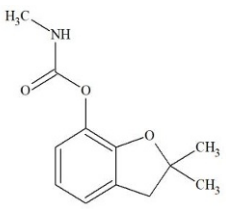
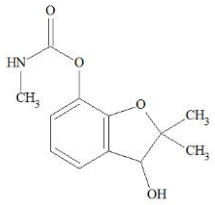
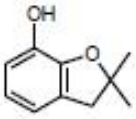
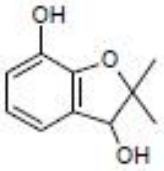
Name	Carbofuran	Carbosulfan	Benfuracarb
Eye irritation	Non-irritant, but mortality reported (rabbits)	Non- irritant	Non-irritant
Skin sensitisation	Non-sensitiser (Bühler and M&K)	Sensitising (guinea pig patch test)	Non-sensitizer (M&K test)
Short term			
Target / critical effect	Testicular degeneration, clinical signs of neurotoxicity related to AChE inhibition (rats and dogs)	Inhibition of acetylcholinesterase (rat) Changes in red blood cells parameters and spleen weight (dog)	Clinical signs of neurotoxicity, inhibition of acetyl cholinesterase, thymus involution (dogs)
Relevant oral NOAEL	0.1 mg/kg bw/day, 1-year dog and 60 day, rat (published study)	90-day, rat: 2 mg/kg bw/day (20 ppm) 6-month, dog: 1.6 mg/kg bw/day (500 ppm)	1 mg/kg bw/day (13-week ; 6-month and 12-24 month dog)
Relevant dermal NOAEL	25 mg/kg bw/day (21 day, rabbit)	21-day, rabbit: 5 mg/kg bw/day	5 mg/kg bw/day (28-day, rat)
Relevant inhalation NOAEL	No data - not required	28-day, rat: 0.15 mg/m3	No data - not relevant
Long term			
Target / critical effect	Body weight and AChE inhibition	Acetylcholinesterase inhibition, focal iris atrophy and degenerative retinopathy (rat) AChE inhibition in the brain, erythrocytes and plasma (mouse)	Clinical signs of neurotoxicity, acetylcholinesterase inhibition (rat)
Relevant NOAEL	0.462 mg/kg bw/day, 2-year rat	2-year, rat: 1 mg/kg bw/day (20 ppm) 2-year, mouse: 2.5 mg/kg bw/day (20 ppm)	5.5 mg/kg bw/day, 104-week, rat
Carcinogenicity	No carcinogenic potential	No carcinogenic potential	ND
Reproductive toxicity			
Reproduction target / critical effect:	Reduced litter parameters in rat multigeneration study; Testicular and sperm toxicity at parental toxic doses.	Reduced number born pups at parental toxic doses (decreased body weight and food consumption)	Reduced pregnancy rate and male fertility indices, reduced pup survival
Relevant parental NOAEL	1.2 mg/kg bw/day	1.2 mg/kg bw/day	1.2 mg/kg bw/day
Relevant reproductive NOAEL	1.2 mg/kg bw/day	1.2 mg/kg bw/day	1.2 mg/kg bw/day
Relevant offspring NOAEL	1.2 mg/kg bw/day	1.2 mg/kg bw/day	1.2 mg/kg bw/day
Developmental toxicity			
Developmental target / critical effect	Foetotoxicity and developmental neurotoxicity at maternal toxic doses (rat).	Rat: incomplete ossification at maternal toxic dose. Rabbit: no developmental effects at maternal toxic doses (decreased body weight and deaths)	Delayed or incomplete ossification and delayed foetal weight (rat). Reduced foetal weight and abortions (rabbit)

Name	Carbofuran	Carbosulfan	Benfuracarb
Relevant maternal NOAEL	Rat: 0.1 mg/kg bw/day Rabbit: 0.5 mg/kg bw/day	Rat: 2 mg/kg bw/day Rabbit: 5 mg/kg bw/day	Rabbit: 15 mg/kg bw/day Rat: 2 mg/kg bw/day
Relevant developmental NOAEL	Rat: 1 mg/kg bw/day Rabbit: 0.5 mg/kg bw/day	Rat: 2 mg/kg bw/day Rabbit: 10 mg/kg bw/day	Rabbit: 10 mg/kg bw/day Rat: 10 mg/kg bw/day
Neurotoxicity			
Short term neurotoxicity			1.81 mg/kg bw/day, 28-day rat
Acute neurotoxicity		NOAEL = 0.5 mg/kg bw, based on AChE inhibition in the brain	
Delayed neurotoxicity	No delayed neuropathy in hens NOAEL neurotoxicity 0.5 mg/kg bw	LD50 hens= 376 mg/kg bw: no delayedneuropathy	No delayed neuropathy in hens LD50 92 mg/kg bw
Subchronic neurotoxicity test:	3.2 mg/kg bw/day, 13-week rat	90-day, rat: NOAEL = 1.2 mg/kg bw/day (20ppm), based on body weight gain, clinicalsigns of neurotoxicity, altered FOB, locomotor activity	
Acute neurotoxicity studies in rats (add Jan 2009):	brain AChE inhibition, LOAEL pups 0.03 mg/kg bw, NOAEL adults 0.03 mg/kg bw		
Acceptable daily intake	0.00015 mg/kg bw/day (acute neurotoxicity study in rat (pups), SF: 200)	0.005 mg/kg bw/day (Rat, acute neurotoxicity study)	0.01 mg/kg bw/day
AOEL		0.005 mg/kg bw/day (Rat, acute neurotoxicity study)	0.01 mg/kg bw/day
ARfD	0.00015 mg/kg bw (acute neurotoxicity study in rat (pups), SF: 200)	0.005 mg/kg bw (Rat, acute neurotoxicity study)	0.02 mg/kg bw

Carbofuran as active substance and its metabolites

The mechanistic arguments discussed can be applied for the RA analysis of carbofuran metabolites. Three carbofuran metabolites are present in EFSA genotoxicity database and reported in Table 26. Analyzing their structural characteristics, it seems plausible to perform 1:1 RA from the parent to the metabolite 3-hydroxy carbofuran. This substance has a similarity of 79% with the parent compound: they differ only for a hydroxyl group. The attribution of the positive genotoxicity outcome from the parent pesticide to the metabolite can be also rationalized by means of mechanistic considerations. The two substances share the same reactive substructure, N-methyl-carbamoyl moiety, associated with genotoxic reactivity (as highlighted by Oncologic profiler, see Table 26). On the same ground of evidence, it is not possible to 1:1 RA the positive Ames test outcome from the parent to the other two metabolites, with similarities less than 70%. In particular, these substances lack the reactive substructure, putative responsible for the genotoxicity. This evidence, not sufficient *per se* to rule out the concern, could be taken into account in a weight of evidence approach for assessing the genotoxicity of the chemicals.

Table 26: Data matrix for the active substance Carbofuran and its metabolites. Physicochemical information, similarities and predictions were retrieved from the OECD QSAR Toolbox. Ames test experimental data, Overall scores generated from EFSA genotoxicity database by the genotoxicity experts (Objective 2)

Name	**AS: Carbofuran**	**3-hydroxy carbofuran**	**carbofuran-phenol (carbofuran-7-phenol)**	**3-OH carbofuran-phenol (3-OH carbofuran-7-phenol)**
				
COM_ID	1606	15270	15274	15272
CAS	1563-66-2	16655-82-6	1563-38-8	17781-15-6
SMILE	CNC(=O)Oc1cccc2CC(C)(C)Oc12	CC1(C(C2=C(O1)C(=CC=C2)OC(=O)NC)O)C	CC1(CC2=C(O1)C(=CC=C2)O)C	CC1(C(C2=C(O1)C(=CC=C2)O)O)C
Molecular formula	C12H15NO3	C12H15NO4	C10H12O2	C10H12O3
Molecular weight	221	237	164	180
similarity		79%	64%	48%
Ames test experimental data	1	1	0	0
OECD QSAR Toolbox Profilers				
Ames by OASIS	No Alert found	No Alert found	No Alert found	No Alert found
Ames by ISS Ames	No Alert found	No Alert found	No Alert found	No Alert found
DNA binding by OASIS	No Alert found	No Alert found	No Alert found	No Alert found
DNA binding by OECD	No Alert found	No Alert found	No Alert found	No Alert found
Oncologic	Carbamate Type Compounds	Carbamate Type Compounds	Phenol Type Compounds	Phenol Type Compounds

3.2.1.6. Conclusions on Strategy I

The main result of this exercise is that a simple Read Across approach gives good predictions for the Ames test (correct 13 out 14 cases), whereas it does not for the *in vitro* Chromosomal Aberrations (correct 4 out 12 cases).

This result parallels and complements those obtained with QSARs in Objective 2. The failure in predicting the *in vitro* Chromosomal Aberrations is impressive. It should be emphasized that exactly the same procedure was applied to the experimental results from the Ames and *in vitro* Chromosomal Aberrations assays. At odds of the Ames test, the *in vitro* Chromosomal Aberrations assay is neither suited for extracting general structure-activity rules (such as those at the basis of QSARs), nor gives comparable and consistent results for chemically similar compounds (as seen through Read Across analysis of analogues).

The selection of chemicals for this exercise included also a number of metabolites erroneously predicted by the QSARs in Objective 2. Out of 14 metabolites, 3 had erroneous Ames predictions and 8 had erroneous Chromosomal Aberrations predictions. It appears that Read Across for Ames test predicted correctly 2 out 3 metabolites missed by QSAR. This suggests that the local analysis of analogues is a powerful support for an overall *in silico* assessment, and can be combined with QSAR analysis. The present discussion is based on the results of the Ames test predictions, since both RA and QSARs for the *in vitro* Chromosomal Aberrations are too unreliable and erratic to be representative of an assessment strategy.

We also provide an example in which additional information (e.g., data on a range of toxicities) is available for the analogues, and permits a more mechanistically based Read Across.

3.2.2. Read Across exercises: Strategy II (Weight of Evidence)

This section presents **thirty Read Across exercises**, performed *ad hoc* for this project for the two endpoints **Ames test** and ***in vitro* Chromosomal Aberrations**.

Recent EFSA guidance on weight of evidence approaches (EFSA, 2017) emphasized the aspect of rationally assembling evidence of similar types, weighing, and integration. The specificity of the work presented in this section is that information is combined by applying a formalized weight-of-evidence method to predict the outcome and estimate uncertainties. Workflows and methods similar to those implemented in the ChemTunes•ToxGPS® platform are employed; however, the results are platform-independent and can be reproduced with public tools.

Often in conventional Read Across (RA) analysis, only experimental data of selected analogues are considered (In this section, the words "source" and "analogue" are used interchangeably.) Analogue prioritization is based on the experimental data reliability as well as analogue quality, evaluated based on chemical similarity (structural fingerprints and properties). RA based only on analogue evidence is referred to as the Tier 1 approach in this report. RA approaches can be further enriched by consideration of *in silico* data. This study is designed to evaluate the inclusion of QSAR data in the RA analysis, referred to herein as the Tier 2 approach. The approach uses the information on the metrics of models performance and reliability, which obviously depend on the quality and treatment / curation of the training sets data.

As explained above, a unique aspect of this Read Across study of pesticide/metabolite substances in EFSA database is that the goal is to estimate the genetic toxicity of metabolites (components) by the information provided in the submission of parents: the RA pair in this case is a parent as the analogue and a metabolite as the target compound. The experimental data from the parents are used as the source data.

Another unique aspect of this particular RA study is that the biology of the parent-metabolite pairs are pre-assigned. In usual RA cases, metabolites are considered as possible related species of the target compound. "Reading" data for metabolites from a parent as an analogue can work if the transformation preserves the essential structural motifs that define the Mode Of Action (MOA) of the parent compound. On the other hand, when transformation products do not resemble the parent closely due to extensive reactions of bond breaking and formation, a new strategy needs to be considered to qualify the analogue and hence the RA feasibility.

Figure 33 depicts the process used in this study to qualify an analogue-metabolite pair and determine the types of similarity metrics to be used. First, pesticidal MOA was applied for biological grouping. Within each MOA, active substance and metabolite pairs are identified. The next step is to recognize the reaction types involved in the metabolic pathways from a parent to a particular metabolite.

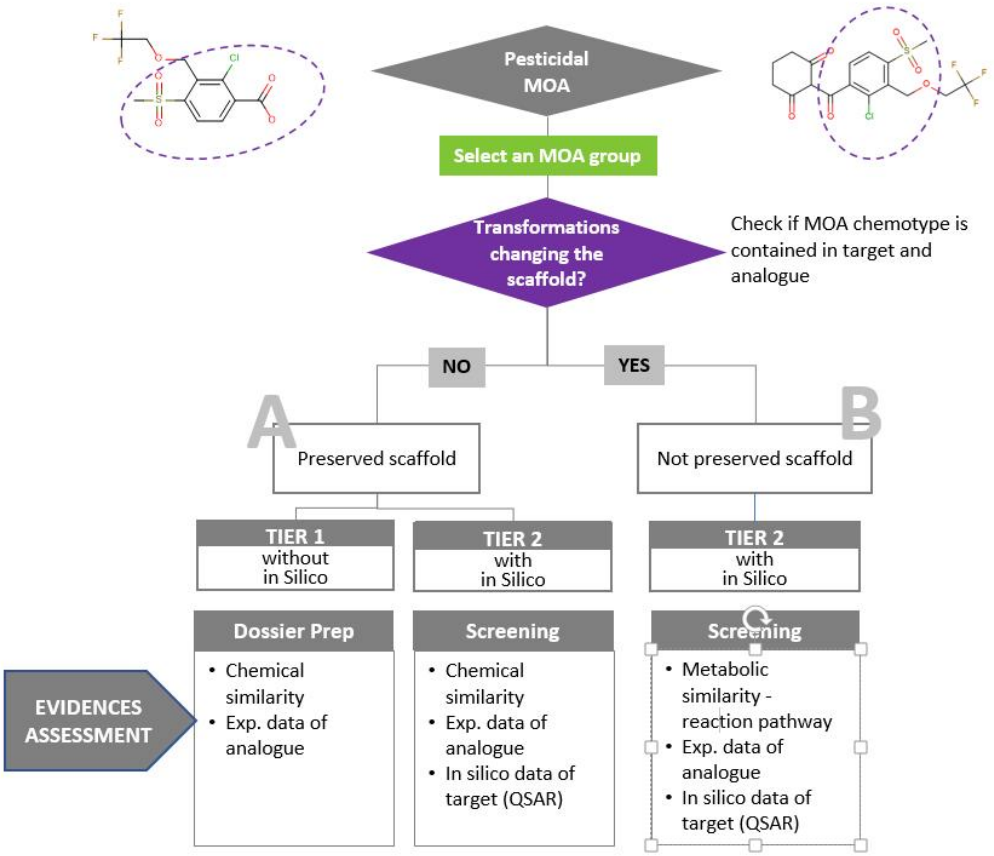


Figure 33: Read Across approach for evaluation of metabolites based on parent

Overall, the following aspects were applied when conducting this Read Across analysis:

- If the metabolic pathways preserve the MOA chemotypes (metabolic group A), chemical similarities based on both structural fingerprints and molecular/physicochemical properties are used to calculate analogue quality.
- If the metabolic pathways involve reactions that break the MOA chemotypes (metabolic group B), analogue quality is addressed by metabolic reactivity using the metabolic fingerprints.

In every case, the contribution of QSAR was used to fine-tune the results of similarity analysis. A decision theory approach based on Dempster-Shafer theory (DST) was used to estimate uncertainty and combine multiple sources of information to obtain the weight-of-evidence (WoE) final outcome (Rathman et al., 2018).

Appendix G extensively presents the method, and details the various similarity measures, as well as the approach used to integrate them into a final WoE measure.

3.2.2.1. Selection of Read Across Cases for Strategy II

As reasoned in previous sections and Appendix G, this RA analysis employed the parent compound as the analogue and its metabolites as the target molecules. These pairs were selected based on a few criteria including the following:

- The effect of experimental data reliability and variation

- Conflicting experimental results are found when multiple studies are available. Multiple experiments from one analogue can be combined to one outcome with the study reliability scores as one of the sources of the uncertainties.

- Metabolites are shared by other parents, hence, increasing chances of experimental variations

- RA and QSAR

- Compounds for which in silico predictions were "erroneous".

- Can RA complement or even improve the accuracy of in silico data?

- Effect of metabolic reactivity

- Metabolites are formed within pathways where the MOA chemotype is not preserved.

The resulting selections for the RA case study are summarized in Table 27.

Table 27: Rationales for Selection of Read Across Cases (1 CAR: Conflicting assay results; 2 IQP: Inaccurate QSAR predictions from many packages)

RA Case ID	SUB ID_COM ID	Pesticide Parent Name	MOA Category	Rationale
RA 1 & 2	15061_15898	tembotrione	HPPD: Triketone	Metabolic Reactivity A (MOA chemotype preserved in metabolic pathway); Baseline for general case for HPPD inhibitors
RA 3 & 4	15061_15899	tembotrione	HPPD: Triketone	Metabolic Reactivity A (MOA chemotype preserved in metabolic pathway); QSAR results for ivtCA vs. the data quality
RA 5 & 6	15061_15900	tembotrione	HPPD: Triketone	Metabolic Reactivity A (MOA chemotype preserved in metabolic pathway); Analogue data quality
RA7 & 8	1172_2061	sulcotrione	HPPD: Triketone	Metabolic Reactivity A (MOA chemotype preserved in metabolic pathway); CAR1
RA9 & 10	35031_50553	mesotrione	HPPD: Triketone	Metabolic Reactivity A (MOA chemotype preserved in metabolic pathway); IQP2
RA11 & 12	35031_50554	mesotrione	HPPD: Triketone	Metabolic Reactivity A (MOA chemotype preserved in metabolic pathway); CAR1
RA13 & 14	1170_50017	imidacloprid	Neonicotinide	Metabolic Reactivity A (MOA chemotype preserved in metabolic pathway); CAR1
RA15 & 16	1170_50019	imidacloprid	Neonicotinide	Metabolic Reactivity A (MOA chemotype preserved in metabolic pathway); Baseline for general case of neonicotinamideinhibitors; CAR1
RA17 & 18	1170_50021	imidacloprid	Neonicotinide	Metabolic Reactivity A (MOA chemotype preserved in metabolic pathway); CAR1, IQP2
RA19 & 20	1166_15576	napropamide	Acetamide	Metabolic Reactivity B (MOA chemotype preserved in metabolic pathway); IQP2
RA21 & 22	1166_50576	napropamide	Acetamide	Metabolic Reactivity B (MOA chemotype preserved in metabolic pathway)
RA23 & 24	1416_2098	cyflumetofen	Beta-keto nitrile	Metabolic Reactivity B (MOA chemotype preserved in metabolic pathway); IQP2
RA 25 & 26	1416_1928	cyflumetofen	Beta-keto nitrile	Metabolic Reactivity B (MOA chemotype not preserved in metabolic pathway)
RA 27 & 28	1174_1877	tetraconazole	Sterol biosynthesis: triazole	Metabolic Reactivity B (MOA chemotype not preserved in metabolic pathway), metabolite common to several parents

RA Case ID	SUB ID_COM ID	Pesticide Parent Name	MOA Category	Rationale
RA 29 & 30	1181_1877	pentaconazole	Sterol biosynthesis: triazole	Metabolic Reactivity B (MOA chemotype not preserved in metabolic pathway), metabolite common to several parents

3.2.2.2. Strategy II: the case studies analysis

The Read Across analysis has been carried out for 15 targets (metabolites) against two endpoints, namely bacterial reverse mutagenesis (Ames assay) and *in vitro* mammalian chromosome aberration (human lymphocytes, CHO or CH V79 cell line assays).

Table 28 summarizes Read Across results and accuracy. The following sections discuss the results in detail.

Table 28: Accuracy of Read Across predictions (TIER 1 & TIER 2)

RA Case ID	SUB ID_COM ID	Name	Ames TIER 1	Ames TIER 2	ivtCA TIER 1	ivtCA TIER 2
RA1 & 2	15061_15898	M6 of tembotrione(A)	True Negative	True Negative; Uncertainty reduced	False Positive; large uncertainty	Equivocal Uncertainty reduced
RA3 & 4	15061_15899	M5 of tembotrione(A)	True Negative	True Negative; Uncertainty reduced	False Positive;	Equivocal Uncertainty reduced
RA5 & 6	15061_15900	M2 of tembotrione(A)	True Negative; large uncertainty	True Negative; Uncertainty reduced	Equivocal; large uncertainty	Equivocal Uncertainty reduced
RA7 & 8	1172_2061	M01 of sulcotrione(A)	False Positive	True Negative	True Negative; large uncertainty	True Negative; Uncertainty reduced
RA9 & 10	35031_50553	M-1 of mesotrione(A)	True Negative	True Negative	Equivocal	False Positive
RA11 & 12	35031_50554	M-2 of mesotrione(A)	True Negative	True Negative	Equivocal; large uncertainty	True Positive
RA13 & 14	1170_50017	M-1 of imidacloprid (A)	True Negative	True Negative; Uncertainty reduced	Positive (No validation data)	Positive (No validation data)
RA15 & 16	1170_50019	M-2 of imidacloprid (A)	True Negative	True Negative: Uncertainty reduced	Positive (No validation data)	Positive (No validation data)
RA17 & 18	1170_50021	M-3 of imidacloprid (A)	True Negative	True Negative	False Positive	False Positive
RA19 & 20	1166_15576	M-1 of napropamide(B)	Equivocal	False Positive	Equivocal	True Positive

RA Case ID	SUB ID_COM ID	Name	Ames TIER 1	Ames TIER 2	ivtCA TIER 1	ivtCA TIER 2
RA21 & 22	1166_50576	M-2 of napropamide(B)	Equivocal	True Negative	Equivocal	True Positive
RA23 & 24	1416_2098	M-1 of cyflumetofen(B)	Equivocal; large uncertainty	False Negative	Equivocal; large uncertainty	True Negative
RA25 & 26	1416_1928	M-2 of cyflumetofen(B)	Equivocal; large uncertainty	True Negative	Equivocal; large uncertainty	True Negative
RA27 & 28	1174_1877	M-5 of tetraconazole (B)	Equivocal; large uncertainty	True negative	Equivocal; large uncertainty	True negative; Uncertainty reduced
RA29 & 30	1181_1877	M of pentaconazole(B)	Equivocal; large uncertainty	True negative	Equivocal1	True Negative1

3.2.2.3. Strategy II: The case studies analyses: prediction of the Ames test

The accuracy of the Read Across was assessed using the experimental data in the database assuming the data truly represented the genetic toxicity of the identified metabolites. Also evaluated was the effect of *in silico* data (QSAR outcome) on the Read Across accuracy.

When considering **only the experimental results** weighted by the analogue and data qualities:

- Nine RA cases were grouped as the metabolic group A (preserving the pesticidal MOA chemotype). In this group of nine cases, only one false positive was observed. All others are correctly predicted.
- Six RA cases were grouped as the metabolic group B (breaking the pesticidal MOA chemotypes during the metabolic reactions). All six cases gave "equivocal" results with large uncertainties. "Reading" from experimental data of only the parent may not be sufficient for robust prediction.

When considering **both QSAR data and the experimental results** weighted by the analogue and data qualities:

- Nine RA cases were grouped as the metabolic group A (preserving the pesticidal MOA chemotype). In this group of nine cases, all 9 were correctly predicted.
- For bacterial mutagenesis endpoint, six RA cases were grouped as the metabolic group B (breaking the pesticidal MOA chemotypes during the metabolic reactions). The results for this group of six cases include four true negatives, one false positive, and one false negative. The false positive case was for the compound Henna (1,4-naphthalenedione, 2-hydroxy-, CAS 83-72-7), for which the QSAR was trained to be positive.

As summarized in Figure 34, the use of QSAR data from Ames mutagenicity models improves the Read Across accuracy and dramatically reduces the number of equivocal outcomes. For the 9 RA cases in Group A, 8 were correctly predicted using only experimental results; using both experimental results and QSAR predictions, all 9 were correctly predicted. For the 6 RA cases in Group B, using only experimental results gives equivocal results for all 6; however, including QSAR predictions results in 4 correct outcomes, 1 false positive and 1 false negative. Overall, for the 15 total cases in Group A and B combined, experimental-only RA gives 8 correct results (53% accuracy) while including QSAR results gives 13 correct results (87% accuracy).

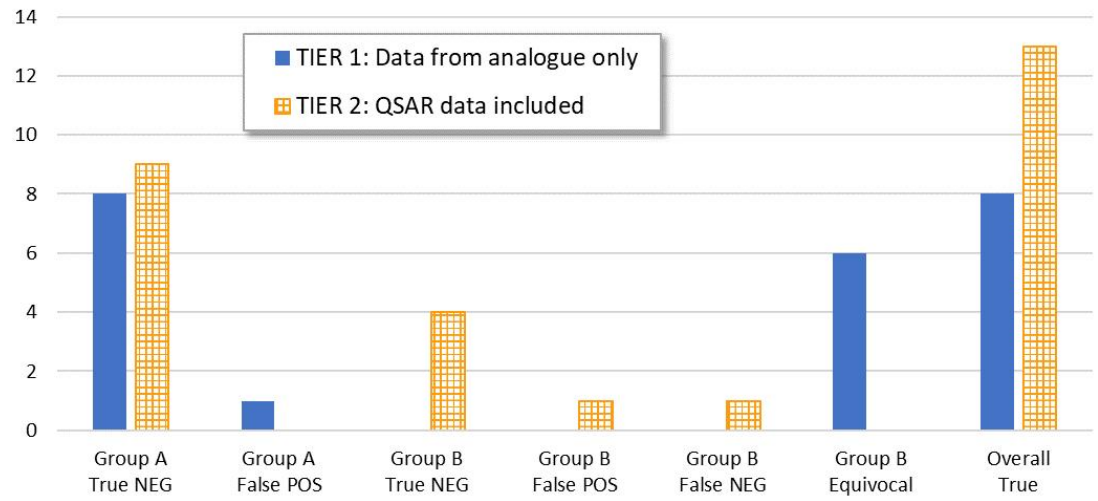


Figure 34: Ames assay assessment by Read-Across

3.2.2.4. Strategy II: The case studies analyses: prediction of *in vitro* Chromosomal Aberrations

When considering **only the experimental results** weighted by the analogue and data qualities:

- Of the nine RA cases of metabolic group A, only one true negative was predicted along with three equivocals and three false positives. (Two had no data to validate.)
- Of the six RA cases of metabolic group B, all results were "equivocal", three with large uncertainties.

When considering **both QSAR data and the experimental results** weighted by the analogue and data qualities:

- Of the nine RA cases (RA scenario group A), one true positive and one true negative were found. There were three equivocals and two false positives. Two cases had no validation data in the database.
- Of the six RA cases of metabolic group B, all outcome were correct: four true negatives and two true positives.

As can be seen in the Figure 35, the addition of QSAR definitely improves the accuracy of the Read-Across, dramatically reducing the false positive and equivocal results.

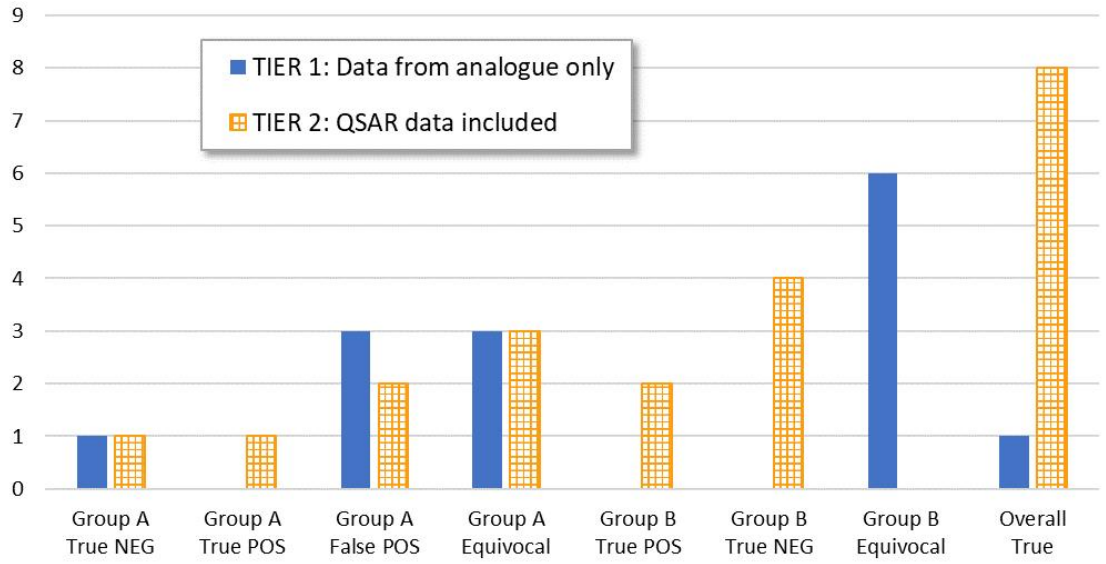


Figure 35: *In vitro* Chromosome Aberration assessment by Read-Across

3.2.2.5. Conclusions on Strategy II

The Read Across approach illustrated here (Strategy II) gives good predictions for the Ames test (out of 15 total cases: 13 correct, 2 incorrect), and fair prediction performance for in vitro Chromosomal Aberrations (out of 13 total cases: 8 correct, 2 incorrect, 3 equivocal).

It's worth noting that one advantage of this approach is that RA in some cases generates an "equivocal" outcome. This occurs when the evidence is weak, conflicting, or of low reliability, so that the uncertainty is sufficiently high to prevent a prediction of either POS or NEG. Although an equivocal outcome may at first seem disappointing, it is much better to an incorrect prediction (a false positive or false negative). An equivocal result indicates that more information is needed before a reliable answer can be given; this itself is obviously a useful and important result.

Finally, we note that all results here involved using evidence from a single analogue to read-across for a target. The approach used here is easily extensible to the case where multiple analogues are available, and we would expect this would be the best way to improve prediction accuracies.

3.2.3. General conclusions on Read Across

Whereas QSARs have undergone during the years many performance evaluations, with special emphasis on comparative prospective exercises, nothing analogous exists for Read Across. The literature is rich in proposals for general workflows and criteria, but the published examples of applications –even though often quite detailed- are limited in number and do not provide sufficient material for assessing the real value of the proposed workflows.

A Read Across analysis usually starts with identifying the best analogues for a given target compound with a specific endpoint in mind. In the present work, analogues are pre-determined by the selection of metabolites, since a primary need of EFSA is that of predicting the metabolites genotoxicity by exploiting all the available information on the parent compound (for which full dossiers are submitted to EFSA). Whenever possible, this is performed with 1:1 Read Across applications.

Within the above constraint, we performed an exploratory work by using two different strategies (Strategy I and II). The test sets selected for the two analyses fulfilled a number of common requirements: a) the chemicals had both Ames and Chromosomal Aberrations data; b) they included both positives and negatives; as well as c) some erroneous QSAR predictions.

In Strategy 1, for each parent / metabolite pair we first calculated a number of basic properties: Molecular Weight, Log KoW, Dice / Atom Centered Structural Similarity. If the chemicals in the pair were similar for the entire profile of the three parameters, the genotoxicity of the parent pesticide was attributed to the metabolite. The metabolites for which the overall similarity with the parent compound was not judged sufficient for directly reading across its genotoxicity, were assessed with a "one-target / many-source chemicals" approach, and analogues both in EFSA and other genotoxicity databases were looked for in different ways.

Central to Strategy II is the systematic use of a decision theory approach (based on Dempster-Shafer theory (DST)) to estimate uncertainty and combine multiple sources of information to obtain the Weight-of-Evidence (WoE) final outcome. Also the biological outcome of the parent is derived by combining the existing pieces of evidence through DST. Overall, the following aspects were applied when conducting this Read Across analysis: a) if the metabolic pathways preserve the basic scaffold, chemical similarities based on both structural fingerprints and molecular/physicochemical properties are used to calculate the overall similarity to the analogue; b) if the metabolic pathways involve reactions that break the scaffold, the overall similarity to the analogue is measured by metabolic reactivity similarity (using metabolic fingerprints). In every case, the contribution of QSAR was used to fine-tune the results of similarity analysis.

Both strategies faced the fact that some Target / Source pairs were considerably similar, whereas in other cases the "metabolism" produced extensive degradations / changes in the Target; the approaches were fine-tuned accordingly.

The results of the two strategies point to interesting similar evidence. A first noteworthy common result is that Read Across appears to be largely successful for predicting the Ames test results: Strategy I generated 1 incorrect prediction out of 14 cases, and Strategy II generated 2 incorrect predictions out of 15 cases.

The performance of the two strategies was partially different with *in vitro* Chromosomal Aberrations. In Strategy II, Read Across generated 8 correct predictions out of 13 cases (with 2 False Positives and 3 Equivocal results), with an overall performance lower than that of the prediction of the Ames test. Strategy I had an even worse performance. Read Across was correct only 4 out of 12 cases. This was paralleled by the fact that QSAR was incorrect for 6 out of 8 incorrect Read Across, thus emphasizing the general difficulties in predicting such an endpoint.

In the two strategies QSAR and Read Across were used differently. Local analysis of analogues (Read Across) helped to better qualify the QSAR predictions in Strategy I (for the Ames test), whereas in Strategy II QSAR provided an important piece of information formally incorporated into the Read Across. In any case, it appears that a synergy between the two approaches, with mutual benefit, is advisable.

For example, when using QSAR for fast priority setting, the Read Across may be used to analyze in depth –under expert supervision- predictions for individual chemicals of special interest.

For EFSA, a central issue in regulatory science is that of estimating uncertainty (EFSA, 2012; 2017). For example, it is stated that: "…All EFSA scientific assessments must include consideration of uncertainties, reporting clearly and unambiguously what sources of uncertainty have been identified and what their impact on the assessment outcome is. Reporting should be consistent with EFSA's general principles regarding transparency and reporting. In a weight of evidence assessment, this should include justifying the choice of methods used, documenting all steps of the procedure in sufficient detail for them to be repeated, and making clear where and how expert judgement has been used. Reporting should also include referencing and, if appropriate, listing or summarising all evidence considered, identifying any evidence that was excluded; detailed reporting of the conclusions; and sufficient information on intermediate results for readers to understand how the conclusions were reached…" (EFSA, 2017).

Whereas this is a general requirement, in the case of QSAR and Read Across the issue has different aspects. The long history of QSAR has led to recognize the factors on which the overall predictive ability depends (Cherkasov et al., 2014; Kubinyi, 2005) (see further details in the Overall Conclusions). The situation with Read Across is fuzzier. First of all, Read Across is much more a case-by-case analysis where different cases may have different and partial pieces of information. In addition, the criteria for rigorous validation required for QSAR are largely not available for Read Across, so that the user has at hand only the information on the uncertainties relative to the composing elements, and not on the overall predictivity of the approach. The uncertainty on the composing elements are only proxies for the key information on the overall predictivity. Strategy II in this report provides an elegant approach to how to use the partial uncertainties. Thus, in the longer period, it is important that more scientific investigations on the predictive ability of Read Across are carried out, and that objective performance measures are established.

3.2.3.1. More on the confidence in Read Across results

The ability of quantifying the performance in Read Across and its associated uncertainties is a key challenge going forward. The issue has different aspects.

One aspect is that the confidence in the final outcomes is linked to the uncertainties on the parameters used in Read Across. For example, in Strategy I we used: Molecular Weight, Log KoW, Structural Similarity, and experimental genotoxicity endpoint.

MW and Structural Similarity are calculated in a deterministic way from the molecular formula, so there is no uncertainty. Log KoW is calculated with the program KOWWIN from Episuite, as implemented in the Toolbox. Its accuracy is reported to be very high: R2=0.95; sd=0.435;me=0.316; n=12805. https://yosemite.epa.gov/sab/sabproduct.nsf/02ad90b136fc21ef85256eba00436459/CCF982BA9F9CFCFA8525735200739805/$File/sab-07-011.pdf.

Thus, the Log KoW uncertainty is extremely low (see also (Hansch and Leo, 1995)).

Regarding the biological endpoint, the reported repeatibility (from laboratory to laboratory) for the Ames test is 80 – 84 % (Piegorsch and Zeiger, 1991). This is unknown for the *in vitro* Chromosomal Aberrations test. The specific biological data used in this work have been first scrutinized and considered acceptable in the preparation of the EFSA database.

On top of this, the procedure followed for Read Across is critical. The key issue in Read Across is the selection of analogues of the Target. This is dictated by the "similarity" of structures. In Strategy I, we have used in a flexible way the whole profile of MW, Log KoW, and Dice Similarity, or Dice Similarity alone, or the shared SMARTS. In a way similar to what happens for the validation of QSARs, one should explore on large databases of results –possibly in a quantitative way- how the predictivity of Read Across depends on "similarity" (however it is defined). One may expect that the Read Across performance is linearly proportional to "similarity", or that it falls abruptly after a certain threshold (more likely). Such an exploration could provide information on the predictive accuracy of Read Across in a way similar to

how predictive accuracy is defined for the QSARs. This information would indicate to the Read Across user the inherent reliability / uncertainty of the procedure. Unfortunately this type of validation is not available (at least publicly), and the user can only rely on information on the uncertainties relative to the composing elements (here MW, experimental data, etc…) that are proxies for the key information on predictivity.

3.3. Impact of structural changes in the molecule in result of metabolic or degradation processes

Objective 5 of this project focuses on: "Evaluation of the impact of the structural changes in the molecule in result of metabolic or degradation processes to the genotoxic potential of the substances". The goal is to provide EFSA with additional means to predict the genotoxicity of metabolites by relating the structures of the parent compounds and of the metabolites.

Relevant results of the previous objectives of this research have shown that: a) the reliability of the (Q)SAR models for assay systems / endpoints different from *in vitro* bacterial mutagenicity (Ames test) is still far from optimality, and the systems need further development; and that b) Read Across was largely successful for predicting the Ames test results, whereas it is less successful in predicting the *in vitro* chromosomal aberrations. As a consequence, the analysis of Objective 5 was limited to structural changes that impact the Ames test genotoxicity results, since only these data seem to be a sound basis for this research.

In this part of the work, we first better qualify the EFSA genotoxicity database, then we illustrate two lines of research followed. In the research, we consider the distinction between Structural Alerts that condense the extensive knowledge on the mechanisms of genotoxicity (Benigni and Bossa, 2011), and other structural changes that may have impact on the genotoxicity of the metabolites. As a matter of fact, the major, more drastic structural changes that impact on genotoxicity are when the transformation from parent to metabolites involves a different pattern of Structural Alerts: substructural motifs responsible of toxicity may disappear, or may be generated *ex novo.* Within this perspective, Analysis 1 focuses on all the subgroups of substances (identified by SUB_ID) in which some metabolites have Ames results different from that of the parent, and examines to what extent this difference can be explained by the expert analysis of Structural Alerts and of other structural motifs. This includes also – when necessary- consideration of analogues retrieved in the same EFSA database.

The second line of research (Analyses 2 and 3) uses chemoinformatics tools in order to identify structural differences –different from the Structural Alerts- between parent and metabolites that may influence changes in Ames mutagenicity, and attempts to identify those that are neutral in respect to genotoxicity and those that may contribute to its enhancement.

Analysis 2 presents a global statistical analysis of the chemical composition of the whole EFSA database, rather than focusing on the identification and mechanistic analysis of selected cases. A complex analysis of the experimental results and their correlation with the presence of the different structural moieties both in active substances and metabolites has been performed: for each couple active substance / metabolite, a comparative study of structural groups/fragments has been done, with special attention to the couples where the metabolite(s) has a genotoxicity status different from the parent compound.

In Analysis 3, we complemented and refined Analysis 2 using a different perspective. To magnify the results, we selected Pesticide / Metabolite pairs in which the structural changes are not too drastic, setting a similarity threshold of 70% between the Active Substance (AS) and its metabolites. We then analysed all this "similar" pairs, and identified the chemical substructures that are more often involved in the chemical transformation from each AS to its metabolites.

As a result of Analyses 2 and 3, a classification of structural fragments / motifs and their effects is provided.

Detailed results for Analyses 2 and 3 are shown in Annex 3.

3.3.1. Preliminary considerations: The EFSA genotoxicity database, and the requirements for the analysis of structural factors

The general experience on structure-activity relationships indicates that the optimal condition for identifying structural factors that influence a biological activity is when different structural changes are observed in a set of congeneric chemicals sharing the same basic skeleton, with different substituents attached to it. This applies to QSARs, that are most informative and sound when are derived for congeneric series of chemicals. A requirement that applies specifically to dichotomic results (e.g., negative and positive), is that the two biological classes are sufficiently balanced in number, and that they are representative of a large range of chemical motifs.

Thus to put into a proper context the results of Objective 5, we developed a further characterization of the EFSA genotoxicity database (namely, regarding the Ames test results), in the light of the above general principles.

3.3.1.1. Positives and negatives in the EFSA genotoxicity database

A first consideration regards the proportion of positives in the EFSA genotoxicity database for the Ames test. This is around 4%, which corresponds to 42 chemicals with defined structures. On the other hand, the Salmonella results in the historical database in the public domain, such as that organized into the ISSSTY database (implemented also in the OECD QSAR Toolbox), contains 7367 chemicals, with a proportion of 48% positives (n=3576) (Benigni et al., 2008). More recently, the National Institute of Health Sciences of Japan has established a new unique proprietary Ames mutagenicity database containing 12,140 new chemicals that have not been previously used for developing QSAR models. Here the positives were 1757 (around 15%); out of them, 672 were strong positives (Honma et al., 2018).

The large proportion of positives in the historical, general database (ISSSTY) derives from the fact that investigators have purposely focused on chemical classes of special interest, aiming at investigating hypotheses on mechanisms of action and elucidating structure-activity relationships. In this sense, the publicly available database was largely the result of a scientific pursuit. On the other hand, the new Japan database reflects more the present situation of chemicals in the market.

In any case, both in the ISSSTY and the Japan database, there is a very large number of mutagens from which to learn structural patterns that may induce mutagenicity. On the contrary, the possibility of contrasting positives and negatives in the EFSA database is quite limited.

3.3.1.2. The congenericity of the subgroups parent / metabolites in the EFSA database

A first characterization of the congenericity of the EFSA genotoxicity database has been already provided in Section 2.2. Given the importance of the subject, we present here further analyses.

In the EFSA database the chemicals are organized into subgroups, each consisting of a parent compound and of a (variable) number of metabolites. In principle, this resembles to the classical congeneric classes of chemicals where the structural differences between members can be compared, thus helping to unveil the modulating factors of activity. However, in the EFSA database the changes undergone by the parent compounds are often so dramatic that the "congenericity" of the series is broken, and it is difficult to make meaningful comparisons between the different chemicals in the subgroup.

Table 29 reports all the pairs parent / metabolite in which there is a change of mutagenicity status (in either way). In the table, Diff_sty = 1 means that the metabolite is positive, whereas the parent is negative; Diff_sty = -1 indicates that the parent is positive and the metabolite is not. Selected pairs with no change of activity are shown as well for a comparison. The table also reports the chemical similarity in each pair (Dice similarity calculated with the OECD QSAR Toolbox).

It appears that changes of activity (and lack of) are independent from the similarity between parent and metabolites, and that the similarity spans the entire range of values (from limited, to vey drastic structural changes).

Table 29: Mutagenicity of selected parent / metabolite pairs, and their similarity

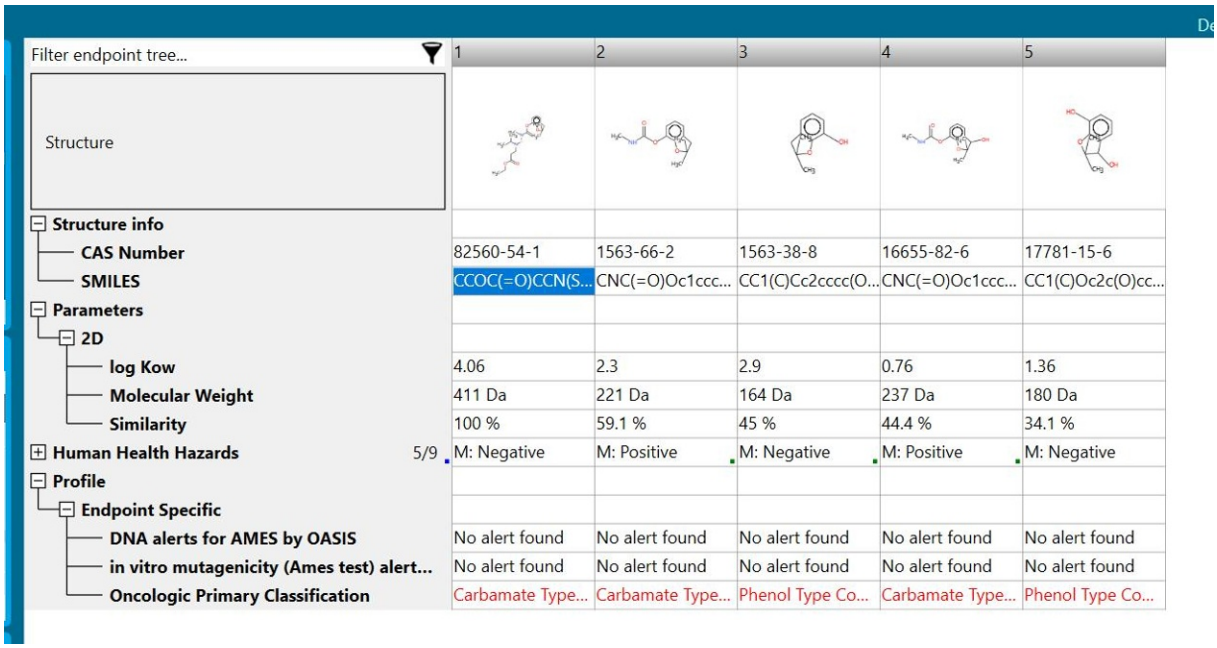
SUB_ID	com_id_Metab	STY_Metab	com_id_Parent	STY_Parent	DIFF_STY	Similarity
1203	1872	0	1844	1	-1	NA*
1203	15785	0	1844	1	-1	NA*
1191	50208	1	1559	0	1	90.3
35060	50632	0	50630	0	0	87.2
1259	15544	1	1502	0	1	83.3
1239	15734	0	1605	1	-1	82.8
1232	75404	1	1573	0	1	81.3
35058	50622	1	50616	0	1	81.1
15035	15487	0	15486	0	0	76.9
1196	75067	0	1582	0	0	76.6
4162	75049	0	1580	0	0	76.4
35058	50621	1	50616	0	1	75.7
35031	50553	0	50309	1	-1	71.8
1180	75002	0	1635	1	-1	71.4
85015	75268	0	75265	0	0	69.4
1172	2061	0	1579	1	-1	68.6
1368	1902	1	1645	0	1	67.9
1239	15925	0	1605	1	-1	66.7
1140	15274	0	1606	1	-1	64.3
15013	1605	1	15061	0	1	62.9
1141	1606	1	1607	0	1	61.9
1139	1606	1	1488	0	1	59.1
1209	15440	0	1848	0	0	59.1
4187	50397	1	6363	0	1	56
35031	50554	0	50309	1	-1	54.1

SUB_ID	com_id_Metab	STY_Metab	com_id_Parent	STY_Parent	DIFF_STY	Similarity
35058	50620	1	50616	0	1	54.1
1217	15105	1	1853	0	1	51.3
1416	2098	1	1509	0	1	48.9
1139	15270	1	1488	0	1	44.4
1106	50474	1	1676	0	1	42.1
1244	15693	0	1508	0	0	41
1100	75595	0	1701	0	0	38.5
1187	75158	0	1703	1	-1	38.5
15013	15926	1	15061	0	1	37.8
1137	15889	1	1549	0	1	33.3
85018	75367	0	75368	1	-1	28.6
1321	1883	1	1882	0	1	26.3
85018	75369	0	75368	1	-1	26.1
3688	1883	1	1688	0	1	20
15040	15652	0	15649	0	0	20
3688	50296	1	1688	0	1	19.5

*It was not possible to calculate the similarity for the first two metabolite (Sub_ID: 1203) with the Toolbox, due to the presence of ionized forms. A calculation with ChemFolder gave 50.9 and 40.0, respectively.

Com_id_Metab: com_id of the metabolite
Com_id_Parent: com_id of the parent
STY_Metab: Ames test result of the metabolite (0 = negative; 1 = positive)
STY_Parent: Ames test result of the parent (0 = negative; 1 = positive)
Diff_STY: -1: parent is positive, and metabolite is negative
1: parent is negative, and metabolite is positive
Similarity: Dice structural similarity (OECD QSAR Toolbox v4.2)

As an example, Figure 36 displays the structures of chemicals in the subgroup identified by SUB_ID = 1139 (snapshot from the Toolbox).This subgroup is composed of a parent and four metabolites; two of the metabolites are Ames positive, whereas the parent and the other metabolites are negative. The figure also displays the structural similarity of the metabolites with the parent (Compound 1 in the figure), and a number of other descriptors. The inspection of the figure shows how drastic can the structural changes be in the same subgroup of EFSA chemicals.



Filter endpoint tree...	1	2	3	4	5
Structure					
Structure info					
CAS Number	82560-54-1	1563-66-2	1563-38-8	16655-82-6	17781-15-6
SMILES	CCOC(=O)CCN(S...	CNC(=O)Oc1ccc...	CC1(C)Cc2cccc(O...	CNC(=O)Oc1ccc...	CC1(C)Oc2c(O)cc...
Parameters					
2D					
log Kow	4.06	2.3	2.9	0.76	1.36
Molecular Weight	411 Da	221 Da	164 Da	237 Da	180 Da
Similarity	100 %	59.1 %	45 %	44.4 %	34.1 %
Human Health Hazards 5/9	M: Negative	M: Positive	M: Negative	M: Positive	M: Negative
Profile					
Endpoint Specific					
DNA alerts for AMES by OASIS	No alert found	No alert found	No alert found	No alert found	No alert found
in vitro mutagenicity (Ames test) alert...	No alert found	No alert found	No alert found	No alert found	No alert found
Oncologic Primary Classification	Carbamate Type...	Carbamate Type...	Phenol Type Co...	Carbamate Type...	Phenol Type Co...

Figure 36: Snapshot from the OCDE QSAR Toolbox v4.2 identified by SUB_ID 1139. The structures of the substances are better visualized in Appendix H.

3.3.2. Analysis 1, transformations involving Structural Alerts

In this attempt to rationalize the cases in which parent compounds and metabolites have different genotoxicity (Ames) outcomes, we apply human expert reasoning to the knowledge provided by Structural Alerts (in combination with other information such as other structural motifs, analogues from the EFSA database).

In practice, we identified all the subgroups (SUB_ID) in which one or more metabolites have genotoxicity outcomes different from that of the parent compound. The total number of subgroups with changes of the Ames outcomes from parent to metabolites is 23. As pointed out above, only Ames test results are considered, and the Overall Ames outcomes generated by us for Objective 2 were used.

We characterized the compounds with two sets of SAs present in the OECD QSAR Toolbox: ISS rule-base and Oncologic. We found (results not shown) that other sets of SAs were overlapping or confounding. The SAs from the two sets were considered together through expert reasoning.

It should be noted that the ISS rule-base and the Oncologic implementation in the Toolbox have different characteristics: the ISS rule-base (present in Toolbox and Toxtree) is a classical list of SAs, whereas Oncologic in the Toolbox implementation only points to large classes of potentially harmful chemicals, with no fine-tuned assessment of the individual chemicals. The complete Oncologic system is going to be implemented in a future release of the Toolbox. We found that the Oncologic classes remarkably contributed to our expert reasoning.

In some cases, our reasoning on the Structural Alerts was supported also by the analysis of close analogues. These were all retrieved from the EFSA genotoxicity database.

Each individual subgroup (Sub_id) is discussed in detail in Appendix H, that provides also tables with structures, identifiers, Ames overall outcomes and SAs found. In the figures, the parent is in the first position and the metabolites are ordered in descending similarity order from the parent. Note that a number of substances are repeated in more than one table: this happens when they belong to more

than one subgroup (e.g., a chemical may be a metabolite in one subgroup, but also a parent in another subgroup, or it can be a metabolite common to different subgroups).

3.3.2.1. Conclusions on Analysis 1

The results of the analysis are summarized in Table 30, where each row represent an individual subgroup.

The headings are self-explanatory. The symbol "?1?" indicates that there are doubts about the mutagenicity call (see discussion in the appropriate paragraphs in Appendix H). Cells in yellow indicate cases in which the knowledge on SAs and mutagenicity calls diverge.

Table 30: Results of the analysis of subgroups of substances (active substance plus metabolites, identified by SUB_ID) in which some metabolites have Ames results different from those of the parent

SUB_ID	Positives	Positives correctly predicted	Negatives	Negatives correctly predicted	Main SA
3688	2	2	1	1	ab unsaturated carbonyls
1321	1	1	1	1	ab unsaturated carbonyls
1137	1	1	3	2	ab unsaturated carbonyls
1259	1	1	6	3	ab unsaturated carbonyls
1180	1	1	1	1	halogenated aliphatic
1172	1	0	1	1	halogenated aromatic
1368	1	1	2	2	aromatic amine
1416	?1?	0	2	2	
1106	1	1	1	1	carbamate
1139	2	2	3	3	carbamate
1140	2	2	1	1	carbamate
1141	1	1	1	1	carbamate
1187	1	1	1	1	carbamate
1203	1	1	2	2	carbamate
1217	?1?	0	4	4	carbamate
1239	1	1	4	3	carbamate
15013	1	1	3	2	carbamate

SUB_ID	Positives	Positives correctly predicted	Negatives	Negatives correctly predicted	Main SA
85018	1	1	2	2	carbamate
1191	1	1	1	0	hydrazine
4187	1	1	2	2	hydrazine
35031	1	1	2	2	nitro aromatic
35058	3	2	9	8	nitro aromatic
1232	?1?	0	2	2	

This analysis shows that the expert (not automatic) use of SAs permits the rationalization of the large majority of the patterns of genotoxicity in the subgroups of substances in which parent and (some) metabolites have different Ames outcomes. In terms of individual chemicals, this translates into a sensitivity of 0.82 (23 / 28 correctly predicted mutagens) and a specificity of 0.85 (47 / 55 correctly predicted non-mutagens). These figures are considerably higher than those attained in Objective 2 with the automatic application of expert rules software (e.g., DEREK, Toxtree, etc..), when sensitivities were below 0.65 and specificities in the range 0.80 – 0.90. This improvement should be attributed to the expert reasoning on each individual case, including the combination of two types of SAs and the inspection of close analogues when appropriate.

Examples of how the inspection of close analogues helped to put in a better perspective the evidence from SAs are in the discussion relative to Sub_id 1259, 1172, 1416.

In a number of cases, we were not able to rationalize the discrepancy between SAs and mutagenicity outcomes. For example, in Sub_id 1137, there are two chemicals almost identical (Com_id 15889 and 75575). Both have the same SA (alpha beta unsaturated carbonyl), but with different Ames results. It is suggested that more sophisticated calculations may help to solve the dilemma.

Other cases of discrepancy between SA and Ames results seem to point to weakness of the experimental data, for example for Sub_id 1217, 1232, 1416. Regarding Sub_id 1416, the metabolite Com_id 2098 is a special case. It is reported to be positive in TA100 with and without S9: however, the chemical has no SA, and all QSARs applied in Objective 2, as well as the Read Across of Objective 4 gave negative predictions. In this case, evidence from QSAR and Read Across would suggest re-testing (if necessary).

While considering the above results, it is important to recall that both the experimental results and the knowledge of SAs have a probabilistic character. As already noted, the repeatability of Ames outcomes from laboratory to laboratory is 80 – 84 % (Piegorsch and Zeiger, 1991). On the other hand, the positive predictivity of SAs is (almost always) lower of 100% in large databases (Benigni and Bossa, 2011), thus the intersection of the two lines of evidence has intrinsic uncertainty. In this perspective, the predictive performance attained in this analysis is encouraging.

Finally, it should be emphasized that some rules (e.g., hindered versus non hindered carbamates; presence of nitro-aromatic moiety) are confirmed as highly predictive in this analysis, and can be considered as take-home lessons of this work.

3.3.3. Analysis 2: Global analysis of structural changes in the transformation from parent to metabolite

The general goal of this analysis is the identification of correlations between presence of different structural moieties and genotoxic potential of the molecules, in the optics of structural changes resulting from metabolic or degradation processes. To complement the analyses presented in the previous sections, we used a different perspective by performing a global statistical analysis of the whole database.

For each couple of an active substance and metabolite a comparative study of structural groups/fragments has been done. A particular attention has been paid to the Parent / Metabolite pairs where the metabolite(s) was more toxic/less toxic than the parent compound.

As the result of this analysis, we provide here a distribution of different functional groups/structural fragments present in the database with indication of the total number of cases, the total number of positive cases and the estimation of the toxifying/detoxifying potential associated with each group/fragment.

These results are summarised in a series of tables and figures, that can be used by EFSA in the assessment of the genotoxicity potential of pesticides metabolites and/or degradation products.

3.3.3.1. Methods

The first phase of the functional groups/fragments analysis included:

- a query of the original database and a collection of all information in the Excel matrix where each row corresponds to a couple of AS and MET;
- a collection and analysis of functional groups/structural fragments for each substance in a couple.

Description of the first phase is reported in Figure 37.

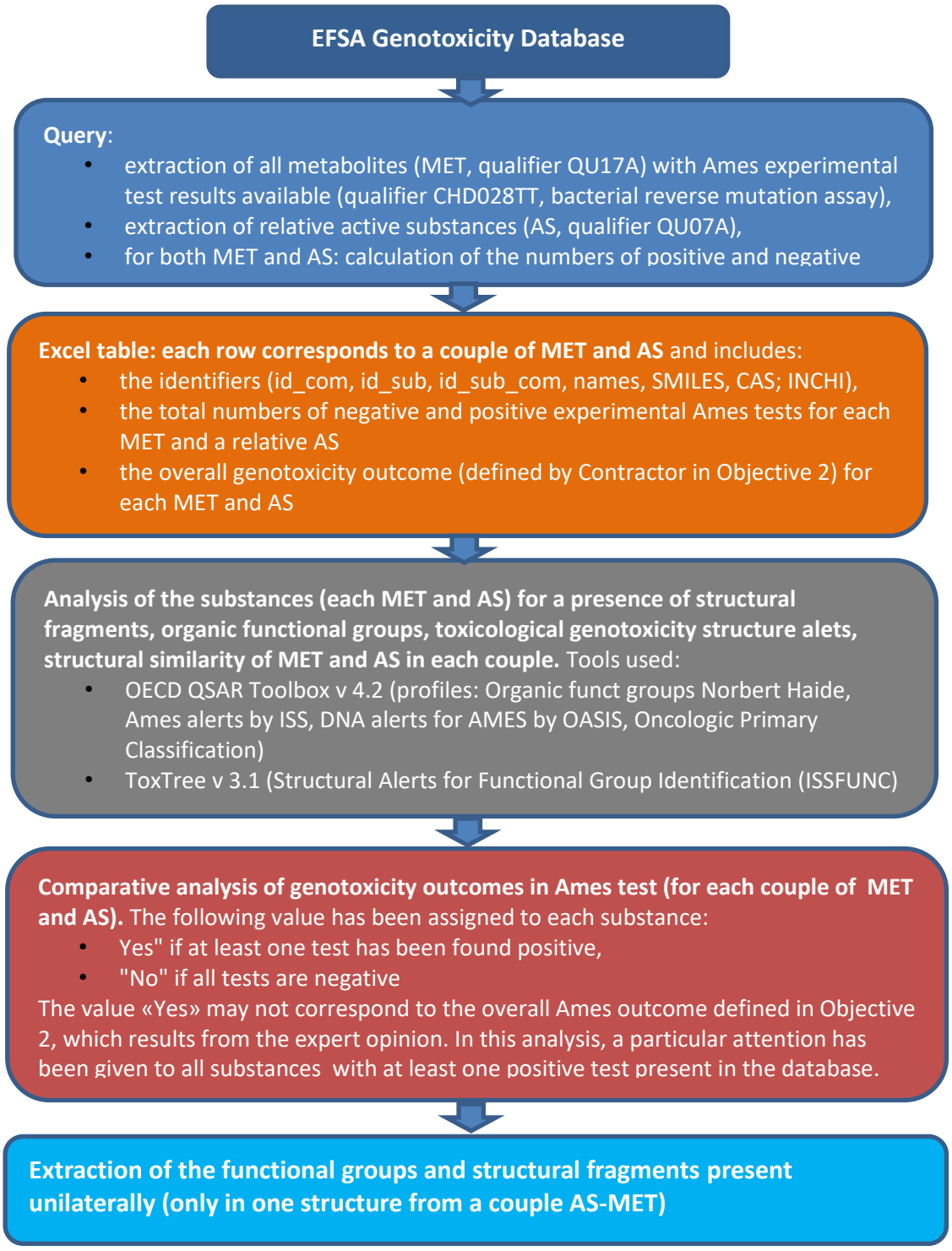


Figure 37: Description of the first phase of the functional groups/fragments analysis

The second phase of the functional groups/fragments analysis included the following steps:

- A detailed analysis of all possible cases based on the experimental test results
- Calculation of number of cases in which a functional group (present unilaterally in a couple AS-MET):
- probably has a TOXIFYING EFFECT;
- probably has a DETOXIFYING EFFECT OR DOES NOT ALTER GENOTOXICITY

Figure 38 reports the description of the second phase of the functional groups/fragments analysis.

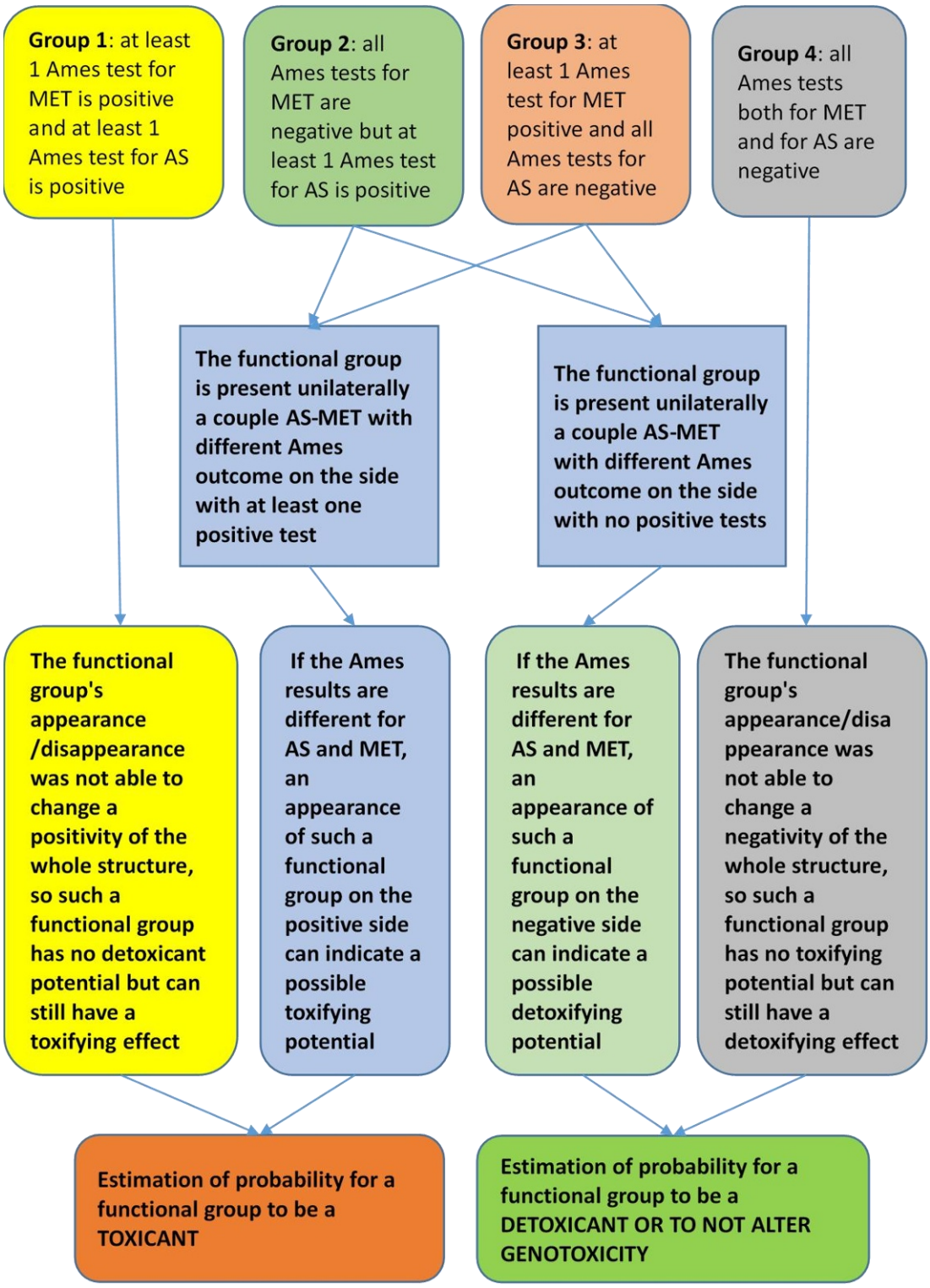


Figure 38: Description of the second phase of the functional groups/fragments analysis

The results obtained during the functional groups/structural fragments analysis are summarised in the tables:

- Table 31. Estimated classification of toxifying potential of functional groups.
- Table 32. Estimated classification of functional groups based on their potential to have a detoxifying role or at least do not alter the genotoxicity.
- Table 33. Evaluation of toxifying potential of structural fragments identified using the Ames Structural Alerts by ISS
- Table 34. Evaluation of toxic potential of structural fragments identified using the DNA alerts for AMES by OASIS
- Table 35. Evaluation of toxic potential of structural fragments identified using the Oncologic Primary Classification

The figures provide a graphical distribution of different functional groups/structural fragments with indication of the total number of cases, the total number of positive cases and the estimation of the toxifying/detoxifying potential associated with each group/fragment:

- Figure 39: Functional groups with possible toxifying effect identified using the Norbert Haider functional groups profile
- Figure 40: Structural fragments with possible toxifying effect identified using Ames alerts by ISS profile
- Figure 41: Structural fragments with possible toxifying effect identified using the DNA alerts for AMES by OASIS profile
- Figure 42: Structural fragments with possible toxifying effect identified using the Oncologic Primary Classification

Table 31: Estimated classification of toxifying potential of functional groups. The colour gradient changes from dark orange to light indicating the decrease of the toxifying potential

Functional group name	Number of cases in which the functional group is present unilaterally in a couple AS-MET	Estimated probablilty for a functional group to be a TOXICANT, %
Nitro compound	3	100.0
Carbamic acid derivative	13	46.2
Carbamic acid ester (uretane)	15	40.0
Thioether	17	35.3
Oxime ether	12	33.3
Oxohetarene	12	33.3
Thiocarbonic acid derivative	15	26.7
Carboxylic acid sec. amide	53	15.1
Anion	76	14.5
Cation	77	14.3
Ketone	29	13.8
Carboxylic acid amide	59	13.6
Carbonyl compound	33	12.1
1.2-diol	9	11.1
Secondary alcohol	38	10.5
Sulfone	10	10.0
Secondary mixed amine (aryl. alkyl)	10	10.0
Heterocyclic compound	95	8.4
Primary aromatic amine	36	8.3
Alkyl halide	62	8.1
Aryl chloride	80	7.5
Alkyl chloride	40	7.5
Halogen derivative	122	7.4
Amine	61	6.6

Functional group name	Number of cases in which the functional group is present unilaterally in a couple AS-MET	Estimated probablilty for a functional group to be a TOXICANT, %
Phenol	63	6.3
Aromatic compound	79	6.3
Aryl halide	95	6.3
Primary amine	54	5.6
Alkyl fluoride	36	5.6
Carboxylic acid tert. amide	20	5.0
Carboxylic acid prim. amide	20	5.0
Hydroxy compound	144	4.9
Alcohol	83	4.8
Secondary amine	22	4.5
CO2 derivative (general)	113	4.4
Carbonic acid derivative	99	4.0
Primary alcohol	30	3.3
Carboxylic acid derivative	167	2.4
Carboxylic acid	152	0.7

Table 32: Estimated classification of functional groups based on their potential to have a detoxifying role or at least do not alter the genotoxicity. The colour gradient changes from dark green to light indicating the decrease of the detoxifying potential

Functional group name	Number of cases in which the functional group is present unilaterally in a couple AS-MET	Estimated probablilty for a functional group to be a DETOXIFYING OR TO NOT ALTER GENOTOXICITY. %
Carboxylic acid	44	99.3
Carboxylic acid derivative	167	97.6
Primary alcohol	30	96.7
Carbonic acid derivative	99	96
CO2 derivative (general)	113	95.6
Secondary amine	22	95.6
Alcohol	83	95.2
Hydroxy compound	144	95.1
Carboxylic acid tert. amide	20	95.0
Carboxylic acid prim. amide	20	95.0
Primary amine	54	94.4
Alkyl fluoride	36	94.4
Aryl halide	95	93.7
Aromatic compound	79	93.7
Phenol	63	93.7
Amine	61	93.4
Halogen derivative	122	92.6
Aryl chloride	80	92.5
Alkyl chloride	40	92.5
Alkyl halide	62	91.9
Primary aromatic amine	36	91.7
Heterocyclic compound	95	91.6

Sulfone	10	90.0
Functional group name	**Number of cases in which the functional group is present unilaterally in a couple AS-MET**	**Estimated probablilty for a functional group to be a DETOXIFYING OR TO NOT ALTER GENOTOXICITY. %**
Secondary mixed amine (aryl. alkyl)	10	90.0
Secondary alcohol	38	89.5
1.2-diol	9	88.9
Carbonyl compound	33	87.9
Carboxylic acid amide	59	86.4
Ketone	29	86.2
Cation	77	85.7
Anion	76	85.5
Carboxylic acid sec. amide	53	84.9
Thiocarbonic acid derivative	15	73.3
Oxime ether	12	66.7
Oxohetarene	12	66.7
Thioether	17	64.7
Carbamic acid ester (uretane)	15	60.0
Carbamic acid derivative	13	53.8
Nitro compound	3	0.0

Table 33: Evaluation of toxifying potential of structural fragments identified using the Ames Structural Alerts by ISS. The colour gradient changes from dark blue to light indicating the decrease of the toxifying potential

Structural fragments identified using the Ames Structural Alerts by ISS	Number of cases in which the alerts has been found	Positive predittivity in the genotoxicity pesticides database, %
Anthrones	4	100.0
Heterocyclic Polycyclic Aromatic Hydrocarbons	2	100.0
Alkyl (C<5) or benzyl ester of sulphonic or phosphonic acid	1	100.0
Nitro-aromatic	5	80.0
Quinones	6	66.7
Aliphatic N-nitro group	10	60.0
Monohaloalkene	18	38.9
Alkyl carbamate and thiocarbamate	8	37.5
alpha.beta-unsaturated carbonyls	50	30.0
Aliphatic halogens	35	8.6
Primary aromatic amine.hydroxyl amine and its derived esters	46	6.5
Hydrazine	44	4.5
Aromatic mono-and dialkylamine	23	0
Isocyanate and isothiocyanate groups	4	0
Simple aldehyde	4	0
Aromatic ring N-oxide	2	0
Alkenylbenzenes	2	0
Epoxides and aziridines	1	0
Aromatic N-acyl amine	1	0

Table 34: Evaluation of toxic potential of structural fragments identified using the DNA alerts for AMES by OASIS. The colour gradient changes from dark blue to light indicating the decrease of the toxifying potential

DNA alerts for AMES by OASIS	Number of cases in which the alerts has been found	Positive predittivity in the genotoxicity pesticides database, %
Alkyl Sulfate Type Compounds	1	100.0
Thiocarbonyl Type Compounds	5	60.0
Carbamate Type Compounds	80	17.5
Phenol Type Compounds	36	16.7
Thiocarbamate Type Compounds	33	15.2
Organophosphorus Type Compounds	20	15.0
ortho-Haloganated Heterocyclic Type Compounds	32	12.5
Aromatic Amine Type Compounds	126	11.1
Halogenated Aromatic Hydrocarbon Type Compounds	241	2.5
Alpha-Halothioether Reactive Functional Groups	2	0
Alpha- and beta-Haloether Reactive Functional Groups	76	0
Aldehyde Type Compounds	9	0
Sultone Reactive Functional Groups	1	0
Hydrazo Type Compounds	13	0
Lactone Type Reactive Functional Groups	2	0
Acrylamide Reactive Functional Groups	3	0
Dicarbonyl Type Compounds	1	0
Urea Type Compounds	1	0
Halogenated Nitroaromatic Type Compounds	3	0

Table 35: Evaluation of toxifying potential of structural fragments identified using the Oncologic Primary Classification. The colour gradient changes from dark blue to light indicating the decrease of the toxifying potential

Oncologic Primary Classification	Number of cases in which the Oncologic class has been found	Percentage of cases with at least one positive test pesticides database, %
Non-covalent interaction_DNA intercalation_Fused-Ring Primary Aromatic Amines	2	100.00
Radical mechanism via ROS formation (indirect)_N.N-Dialkyldithiocarbamate derivatives	2	100.00
Alkylation_Alkylphosphates. Alkylthiophosphates and Alkylphosphonates	3	66.67
Michael-type addition. quinoid structures_Quinones and Trihydroxybenzenes	6	66.67
Nucleophilic addition reaction with cycloisomerization_Hydrazine Derivatives	4	50.00
Direct acting epoxides formed after metabolic activation_ Quinoline Derivatives	3	33.33
Schiff base formation by aldehyde formed after metabolic activation_Geminal Polyhaloalkane Derivatives	30	0
Incorporation into DNA/RNA. due to structural analogy with nucleoside bases_Specific Imine and Thione Derivatives	6	0
Radical mechanism via ROS formation (indirect)_Single-Ring Substituted Primary Aromatic Amines	1	0

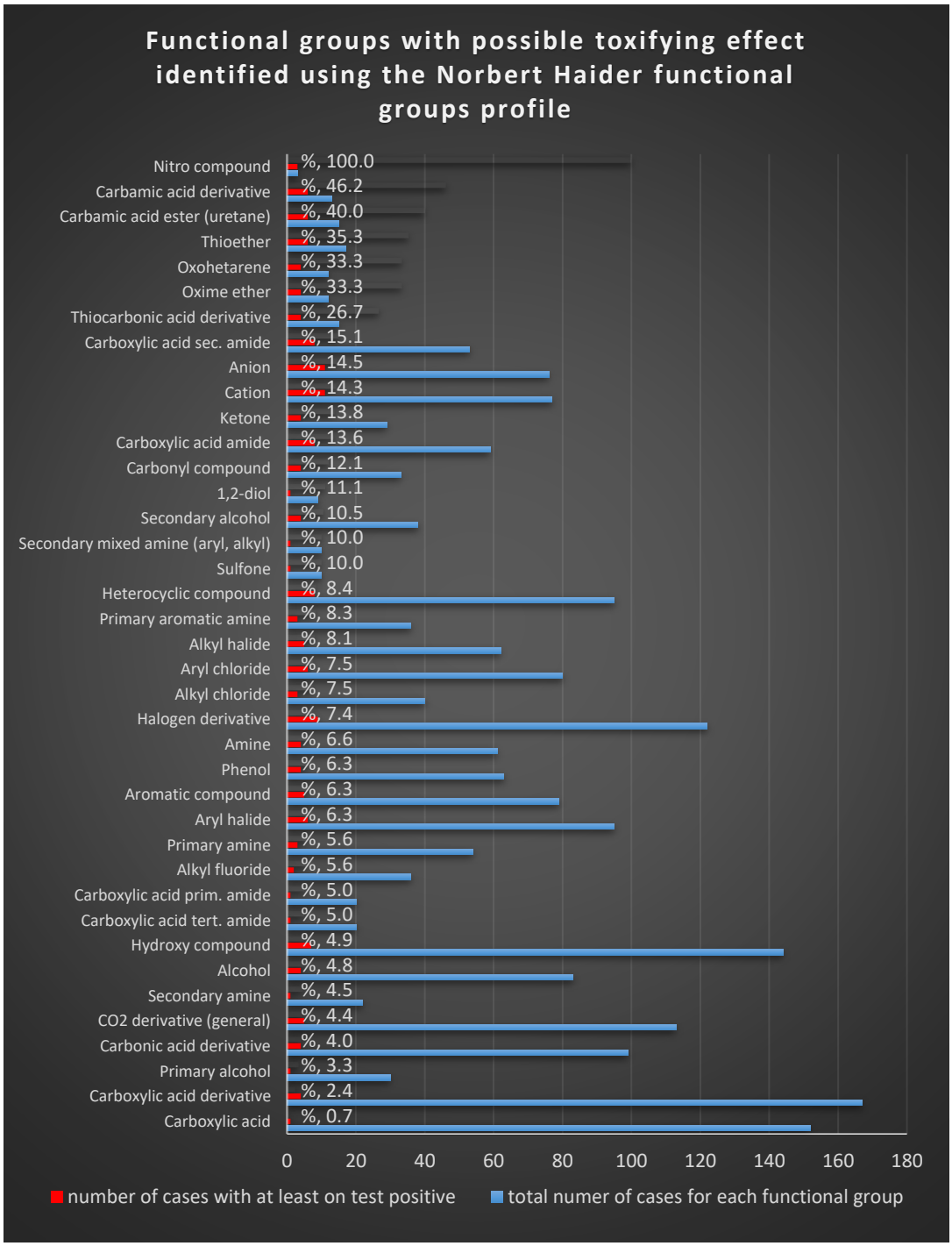


Figure 39: Functional groups with possible toxifying effect identified using the Norbert Haider functional groups profile

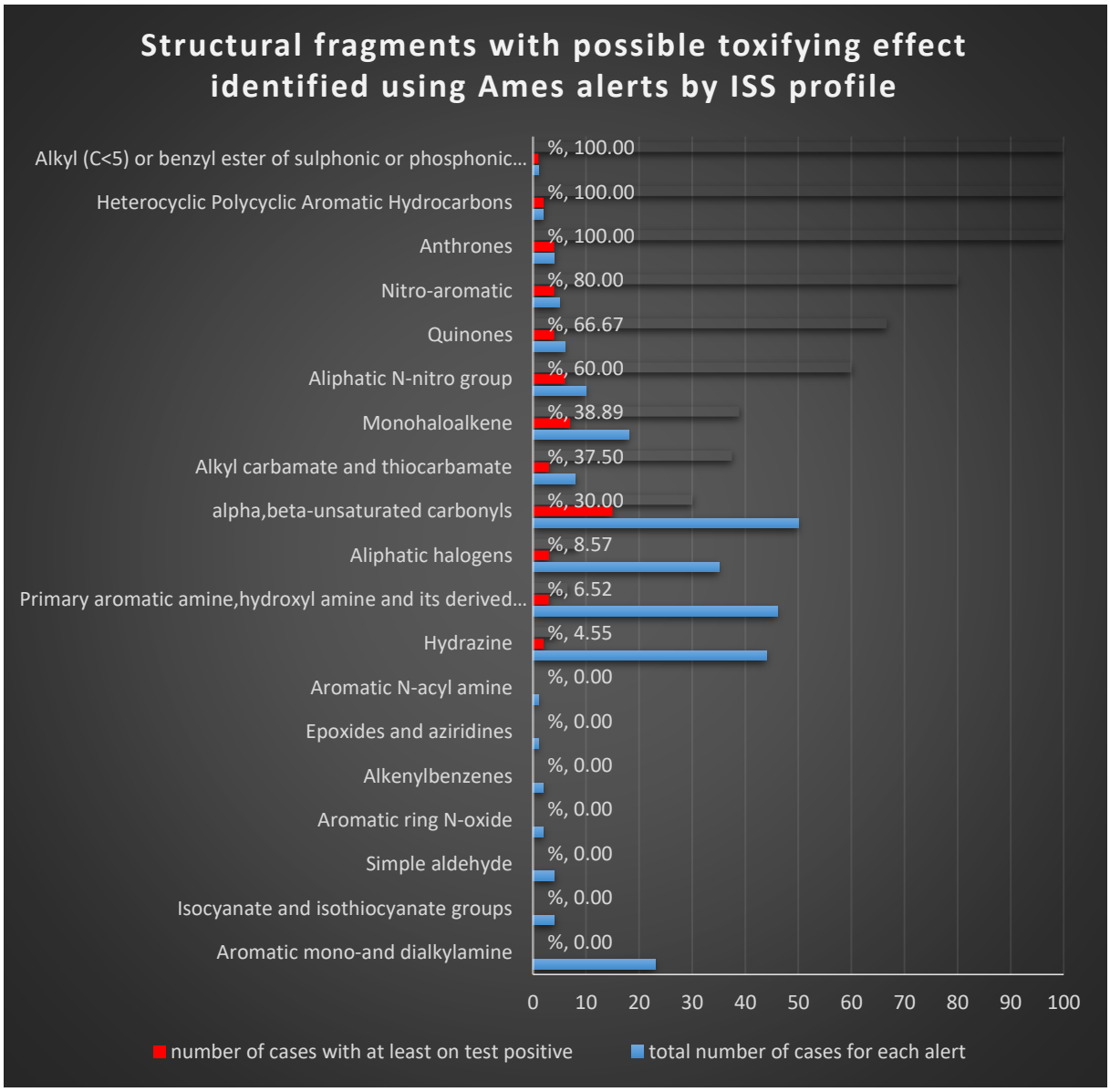


Figure 40: Structural fragments with possible toxifying effect identified using Ames alerts by ISS profile

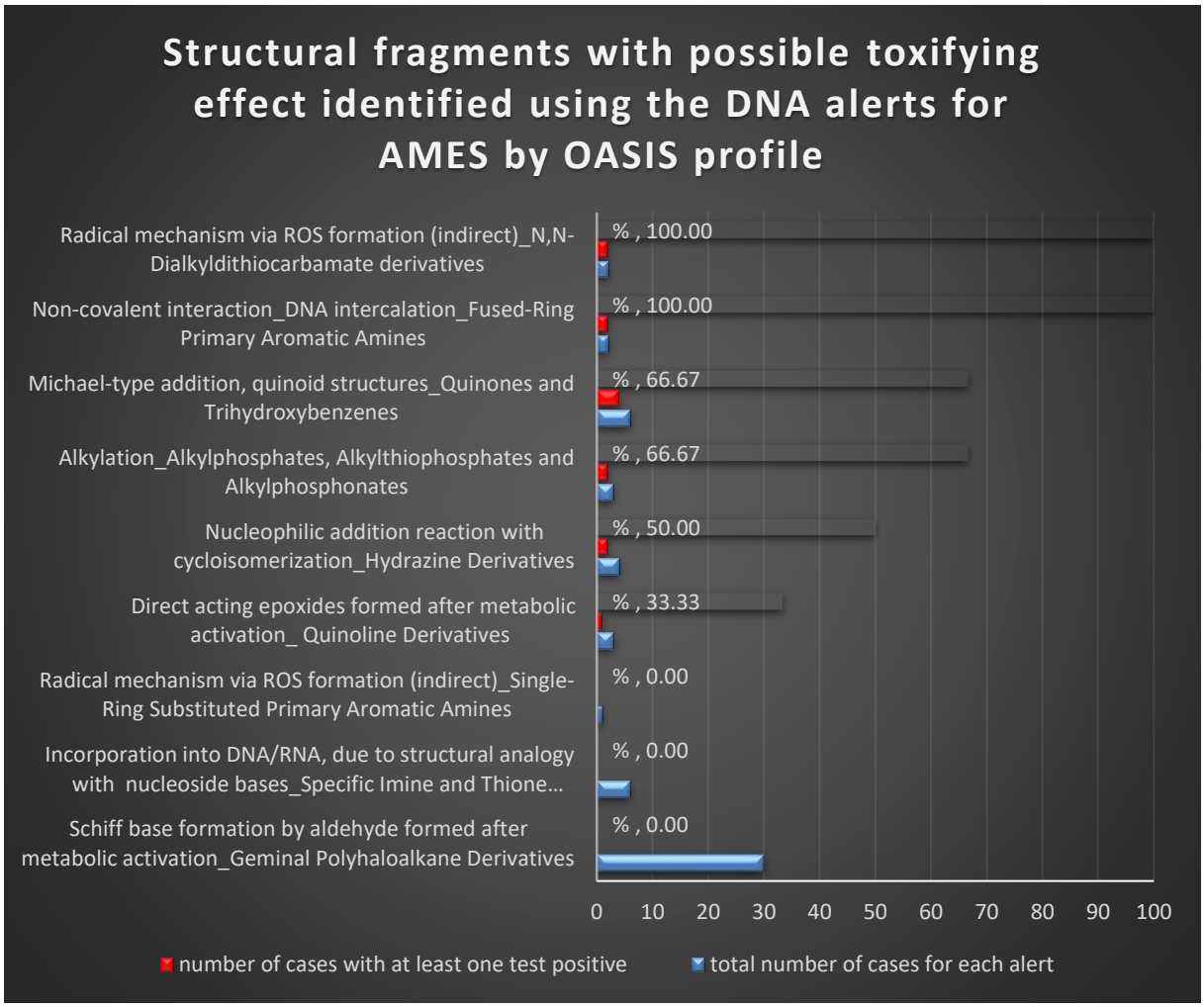


Figure 41: Structural fragments with possible toxifying effect identified using the DNA alerts for AMES by OASIS profile

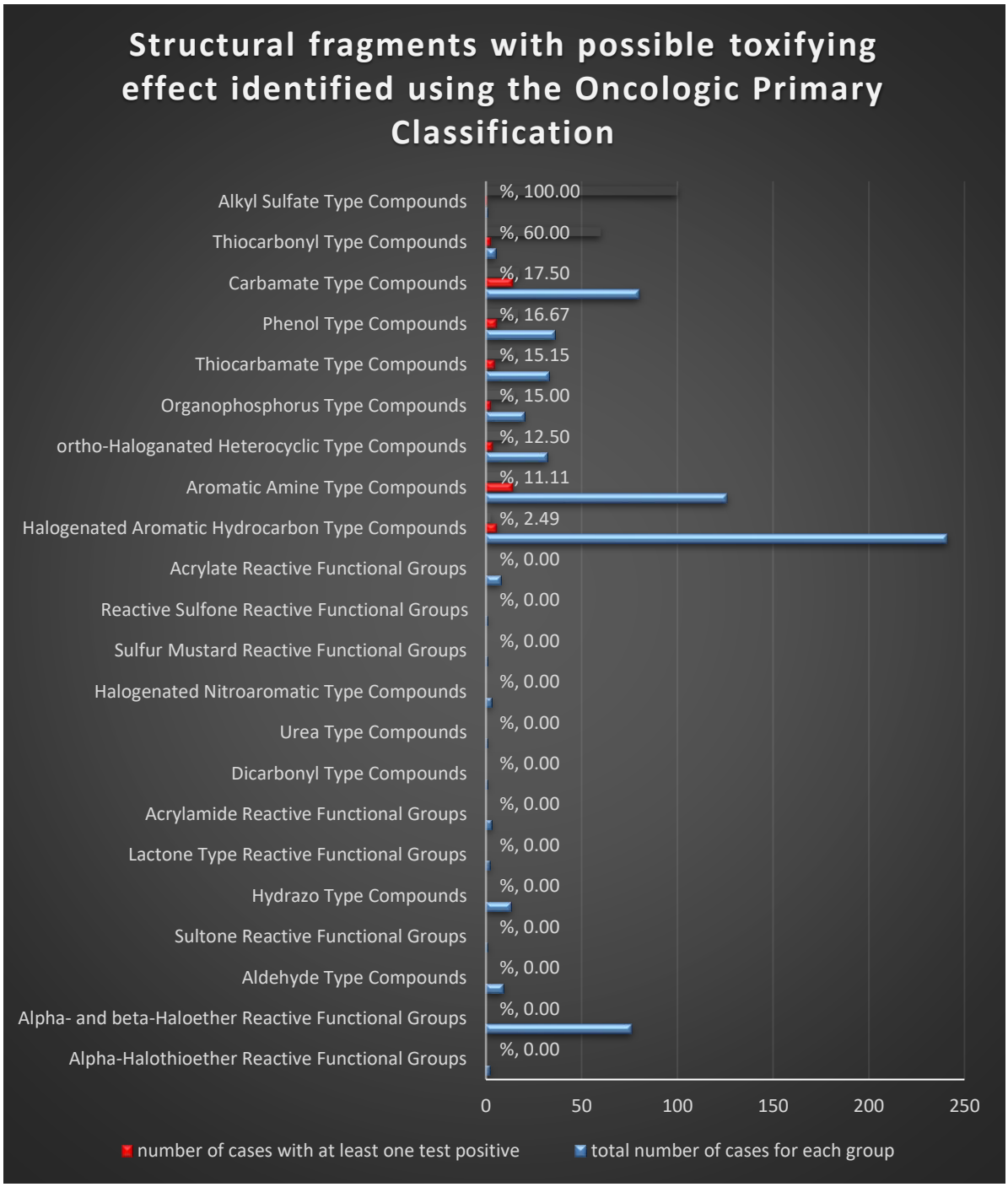


Figure 42: Structural fragments with possible toxifying effect identified using the Oncologic Primary Classification

3.3.3.2. Results of Analysis 2

After the analysis of the 622 couples of pesticides active substance (AS) and their metabolites (MET) was performed within 4 different groups:

- **Group 1:** at least one Ames test strain for MET is positive and at least one Ames test strain for AS is positive (9 cases)
- **Group 2**: all Ames tests for MET are negative but at least one Ames test strain for AS is positive (46 cases)
- **Group 3**: at least one Ames test strain for MET positive and all Ames tests for AS are negative (24 cases)
- **Group 4**: all Ames tests both for MET and for AS are negative (543 cases)

The filtering of the structure fragments present only unilaterally in each couple allowed us to identify the possible role of the structural groups (the phase 2 of the analysis described in Methods Section) in the genotoxic outcomes. All 4 cases including the couples with a negative genotoxicity on both sides (which is the most common situation in the EFSA genotoxicity database) were analysed.

We would like to underline, that the assignment to the structure fragment of a score as:

- no detoxifying potential but can still have a toxifying effect
- a possible toxifying potential
- indicates a possible detoxifying potential
- no toxifying potential but can still have a detoxifying effect

was based on the probabilistic approach, since other factors rather than one group change may be involved in the determination of the genotoxic potential. As a matter of fact, the use of the functional group profiling may not capture the entire complexity of structural changes, including their influence on physical chemical properties and conformational changes. The functional group decision tree identifies only the presence (yes/no) of a certain functional group. In some cases the same group may be present in 2-3 different sites of a molecule, but such cases cannot be detected with the methodology used by us. Therefore an uncertainty is always present. However we believe that the global analysis performed by us depicts well a general trend of the genotoxic potentials of the functional group in the chemical domain of pesticides.

According to the results obtained in this study, we used different colors in the tables in Result section in order to indicate the possible toxifying potential.

Since the structure alerts (SA) were developed for some known mechanisms only, it was important to perform a complementary analysis of the functional groups (different from SAs) covering the complete chemical space of the database. This analysis is performed from scratch, and it is independent from the knowledge on SAs.

Figures 43 displays the functional groups that are likely to have a toxifying potential (with the potentially most "dangerous" groups on the top), and the functional groups that have a very low probability to be a toxicant, and even -in some cases- could have a detoxifying potential (on top the less "dangerous" groups, possible detoxifying). It should be noticed that some of the functional groups (e.g., nitro) are coincident with known SAs.

Figure 43: Results of Analysis 2: the figure classifies substructures / functional groups according to their tendency to: a) increase; or b) leave unaltered or decrease the genotoxic potential (see details in the text)

3.3.4. Analysis 3: non-toxifying structural changes in Parent / Metabolite pairs with high similarity

The knowledge of on structural changes that do not alter, or even downgrade the toxicity status of the parent may be very useful in the assessment of the toxicity of metabolites. This is the subject of the research presented in this section. The possibility of identifying such changes is maximal when the core structure of the parent pesticide is maintained and only minor changes happen, and thus can be related to toxicity.

Preliminarily to this analysis, we performed a similarity characterization of the pairs Active Substance (AS) / Metabolite (with Ames results). This preliminary work used Dice coefficients and the profiler "Structural Alerts for Functional Groups Identification", as implemented in Toxtree software. Other options (different metrics and coefficients) gave equivalent results (results not shown).

In Figure 44, the distribution of the similarity values for the pairs AS-metabolites is shown.

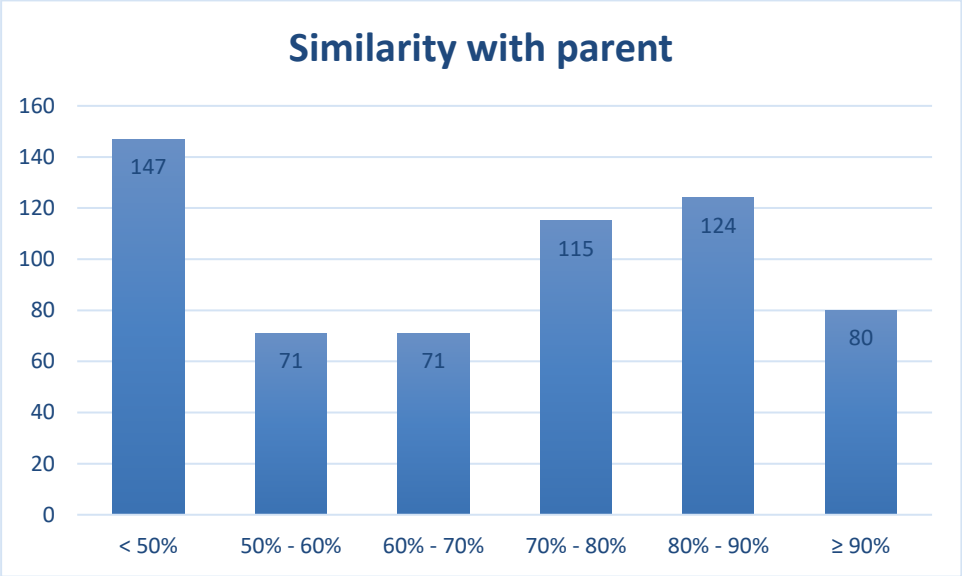


Figure 44: Distribution of the similarity values of metabolites to parents

In what follows, we report the results of the analysis performed for the characterization of chemical features/substructures that are likely not to alter the reactivity of the parent compound, when transformed to the metabolites. For the analysis, we selected the pairs Pesticide-Metabolite with a similarity value greater than or equal to 70%, because no simple conclusions can be drawn for chemicals whose chemical structure differs to a large extent.

Chemicals pairs with similarity 70% or more, were 319. Among these pairs, in ten cases the experimental value for the Ames test (overall) changed in the transformation from the parent to the metabolite. In particular, only in seven cases out 319 (2 %) the metabolite resulted mutagenic as opposed to the Pesticide. In all the other cases, the metabolites showed similar reactivity, if not lower (three cases), than the parents.

This finding implies that –at least with confidence for the present dataset- the chemical modifications encountered in the transformations relative to pairs with similarity 70% or more, may be considered in the majority of cases as neutral changes with respect to the Ames test mutagenicity (i.e., they may not give rise to changes of mutagenicity).

This has important consequences.

First, it supports the results of the Read Across for the Ames test according to Strategy I, which showed that - for very similar chemical pairs - 1:1 Read Across is highly predictive.

Second, it circumscribes the sample of data where to look for structural changes with low probability of generating genotoxic metabolites.

In practice, having selected the pairs with similar parent / metabolite, each chemical was described in terms of chemical groups. The chemical groups composition for each chemical was obtained with the chemical characterization profilers from the software Toxtree and OECD QSAR Toolbox. The results from the two software were adapted and merged after preliminary check of overlaps. Subsequently, we analyzed the couples of Pesticide-metabolites and we selected the chemical features that are present only in one of the members of the pair.

In this way, we focused on the chemical groups that were gained or lost in the chemical transformation from the parent substance to the metabolite. The chemical features were characterized by a measure called "Delta": features that appear in the metabolites will have a positive Delta, whereas features that disappear in the metabolite will have a negative Delta. A zero Delta corresponds to the situation in which the structural feature remains unchanged in the transformation, or is absent in both the parent and the metabolite.

In Table 36 the principal chemical modifications, i.e. the functional groups which differ more frequently in the transformation from parent to metabolites, are listed. After analysis of a great number of functional groups, we arbitrarily selected features above 10% of frequency in occurrence. We also considered a number of features, below this threshold, which appear to be relevant for this analysis.

Table 36: Description of the functional groups selected.

Pairs involved (%)	number of occurrences	Main chemical functionality		gained in transformation (positive Delta)	lost in transformation (negative Delta)
35.94	115	carbonyl		78 (1)	37
17.5	56		alkyl	34	22
15	48		aryl	33 (1)	15
28.44	91	carboxylic acid		91 (1)	0
11.25	36		alkyl	36	0
15.94	51		aryl	50 (1)	1
12.19	39	carboxamide		10	29
14.69	47	carboxylate ester		7	40
31.25	100	Aromatic*		15	85 (1)
15	48	alkyl benzene		1	47
6.56	21	amino benzene		10	11
6.88	22	halo benzene		1	21
4.69	15	hydroxy benzene		15	0
4.69	15	hydroxymethyl benzene		15	0
26.25	84	alcohol		84	0
15.31	49		alkyl	49	0
10.31	33		aryl	33	0
19.38	62	halide		4	58
11.25	36		alkyl	2	34
9.06	29		aryl	3	26
22.19	71	ether		10	61
21.56	69		alkyl	9	60
16.88	54		methyl	6	48
16.56	53		aryl	5	48
14.69	47	alkyl amine		11	36
6.88	22	alkene		11	11
3.75	12	carbamate		4 (2)	8
9.69	31	sulfonyl group		18 (1)	13

In parentheses the number of occurrences of the feature in mutagenic metabolites (with non-mutagenic parent).
Details on frequency of alkyl and aryl functional groups are reported when relevant. These two types are the most abundant but not necessarily the only ones possible.
*This feature refers to any non-aromatic atom attached to an aromatic one.

Figure 45 displays the distribution of the number of chemical modifications - Delta - in each pair, with respect of the principal chemical functionalities selected. The histogram highlights the fact that a limited number of chemical modifications appear to be involved in each transformation in the selected sample of pairs (with similarity at least 70%).

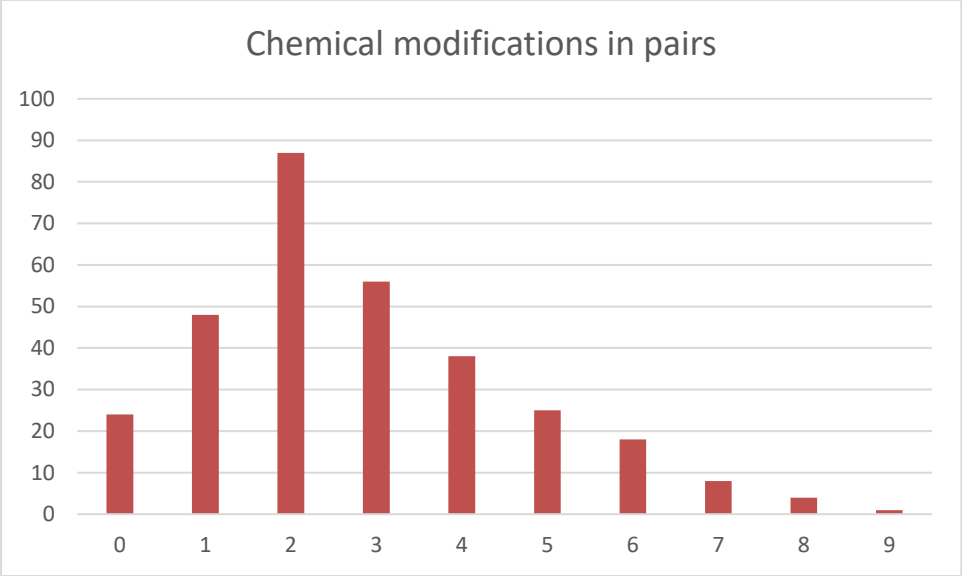


Figure 45: Distribution of the number of chemical modifications – called Delta - in each pair, with respect to the principal chemical functionalities selected

On the other hand, in each pair more than one of the same chemical feature can be modified during the chemical transformation. This can be observed in Figures 46 and 47, where the distribution of the Delta values (both positive or negatives), for each chemical feature, is depicted. The figure provides two different perspectives of the results.

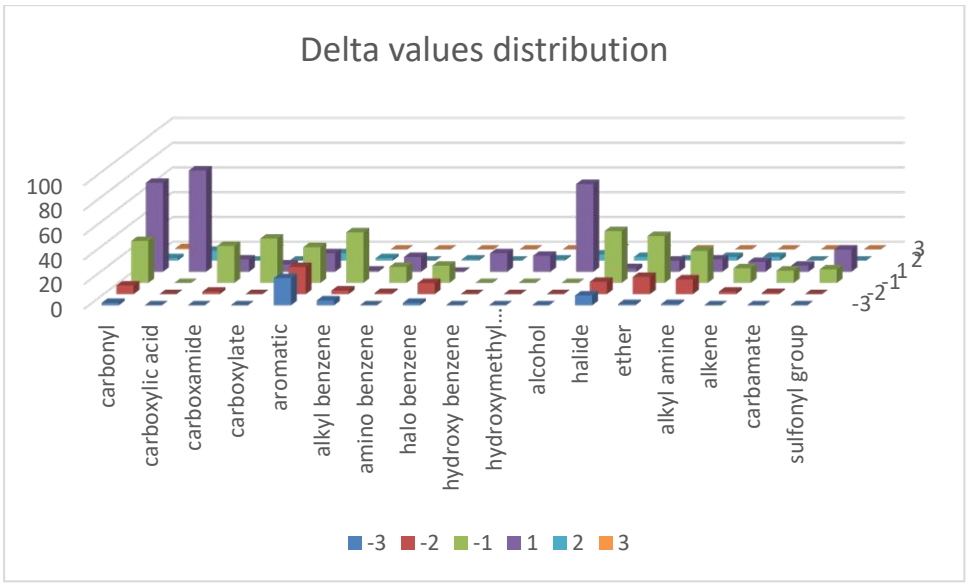


Figure 46: Distribution of the Delta values (both positive or negatives), for each chemical feature

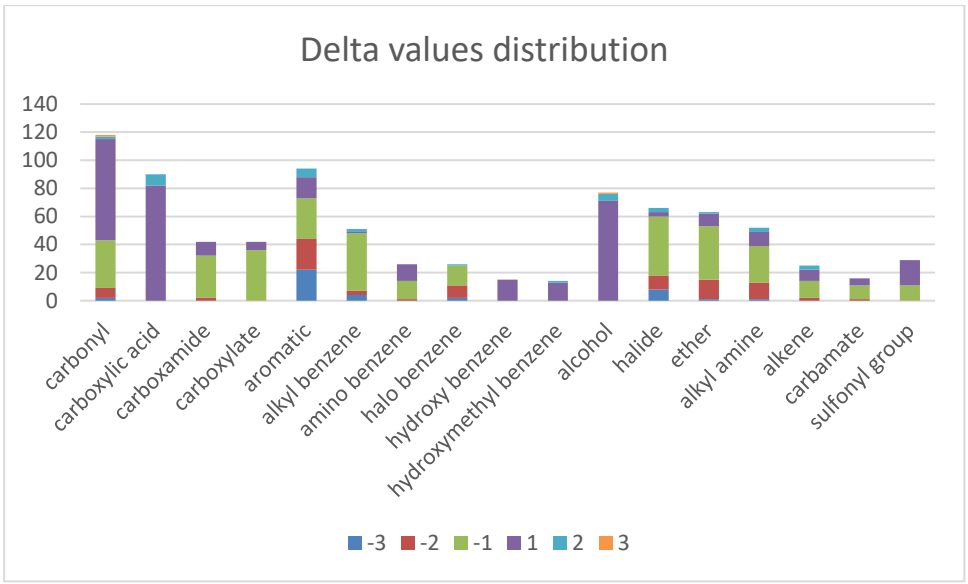


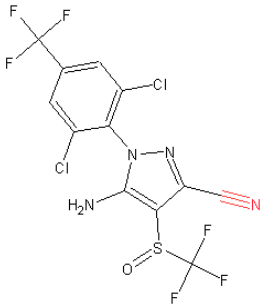
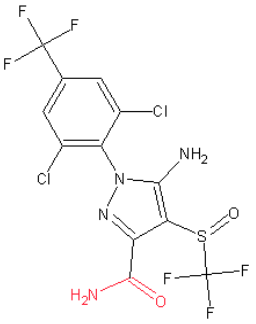
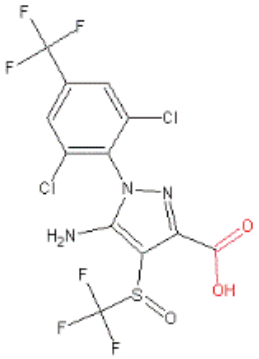
Figure 47: Distribution of the Delta values (both positive or negatives), for each chemical feature

3.3.4.1. More on the results of Analysis 3

The most frequent chemical modification in the transformation from parent to metabolite (within 70% similarity) is represented by the **carbonyl group**, which has a frequency of 38%. The modification can arise as a gain in this functional group (in 78 couples, positive Delta) as well as a loss (in 43 cases, negative Delta). These chemical modifications are to be considered neutral in respect to changes in genotoxicity; in fact, changes from non-mutagenic pesticides to mutagenic metabolites happen only in 1.3 % of the cases (1 / 78, see Table 36).

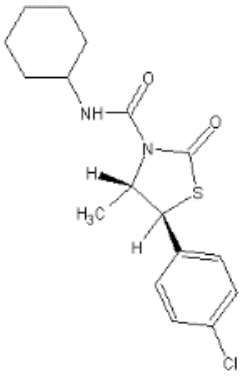
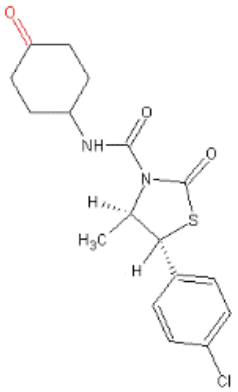
Going into the detail of this chemical modification, we can observe that the carbonyl group can be in partial overlapping with other categories, which contain the carbonyl functionality, i.e. carboxylic acid, carboxylate or carboxamide. This is the case of the substances with sub_id 1127, depicted below (Table 37). The Nitrile function present in the AS is substituted, after chemical transformation, into carbonyl features, in one case a carboxylic acid, in the other a carboxamide.

Table 37: Example of the **carbonyl group** definition, in the case of sub_id 1127.

			
	AS		
sub_id	1127	1127	1127
com_id	1821	2041	2042
carbonyl		1	1
carboxylic acid		0	1
carboxamide		1	0

Nevertheless, this is not always the case and the carbonyl functional group may appear also not in conjunction with the other features, as for example in the case below (sub_id 1257, Table 38).

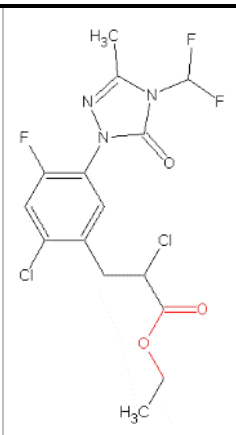
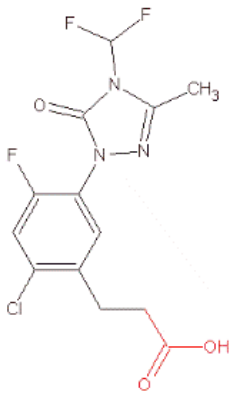
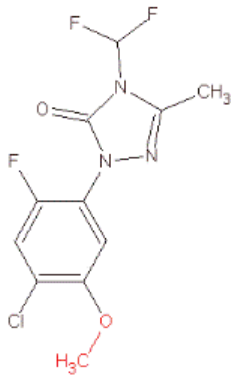
Table 38: Example of the **carbonyl group** definition, in the case of sub_id 1257.

		
	AS	
sub_id	1257	1257
com_id	1651	75467
carbonyl		1
carboxylic acid		0
carboxamide		0
carboxylate		0

The **carboxylic acid group** is another chemical modification frequently encountered in the transformation process. It appears equally distributed among alkyl and aryl carboxylic acids. It is worth noting that the Delta for this feature is positive in almost all the pairs, thus the carboxylic acid group often appears in the metabolites as a result of the chemical transformation, without affecting in general the mutagenicity of the AS.

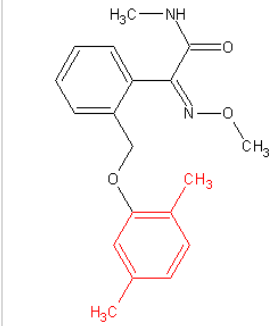
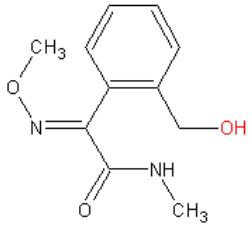
The **carboxylate** ester feature is often transformed going from AS to metabolite, as highlighted by the prevalence of negative Deltas in table 36. In the following example (sub_id 15043, Table 39), the carboxylate is hydrolysed in a carboxylic acid in one metabolite and transformed in an ether feature in the other one.

Table 39: Example of the **carboxylate** ester definition, in the case of sub_id 15043.

			
	AS		
sub_id	15043	15043	15043
com_id	15674	15676	15683
carbonyl		0	-1
carboxylic acid		1	0
halide		-1	-1
ether		0	1
carboxylate		-1	-1

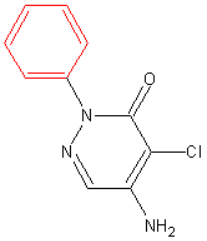
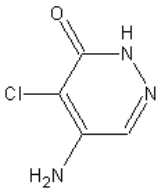
The **aromatic** delta counts the non-aromatic atoms connected to aromatic ones, thus accounting both to the changes in content of aromatic rings and to the changes in the groups directly connected to them. Understanding of this chemical feature may be complemented by the others related features, detailing benzene substituents variations. In the example below (sub_id 1078, Table 40), 3 aromatic atoms and 2 alkyl aromatic substituents are lost in the metabolite, which in fact lacks the para-xylene subgroup.

Table 40: Example of the **aromatic** definition, in the case of sub_id 1078.

		
	AS	
sub_id	1078	1078
com_id	1625	2022
alcohol		1
aromatic		-3
alkyl benzene		-2
amino benzene		0
halo benzene		0
hydroxy benzene		0
hydroxymethyl benzene		1

Anyhow, the aromatic feature may arise independently from the related ones. As illustrated in the example below (sub_id 1144, Table 41), the aromatic negative delta is accounting of the benzyl substituent loss.

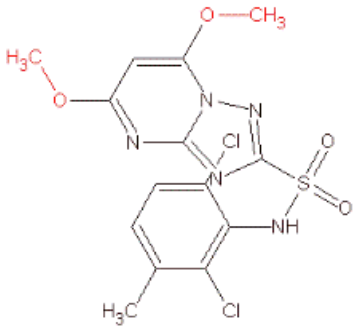
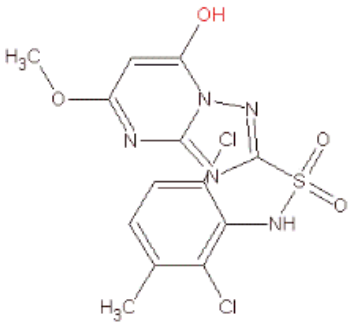
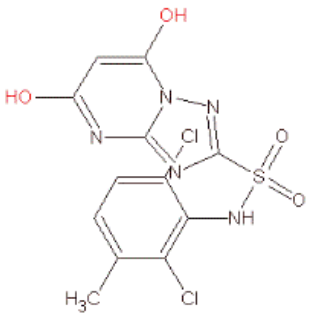
Table 41: Example of the **aromatic** definition, in the case of sub_id 1144

		
	AS	
sub_id	1144	1144
com_id	1611	15627
aromatic		-1
alkyl benzene		0
amino benzene		0
halo benzene		0

hydroxy benzene		
hydroxymethyl benzene		

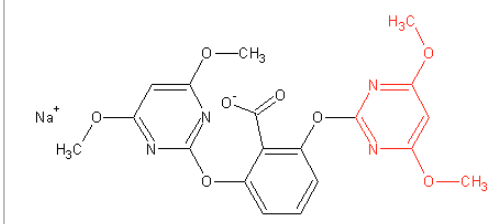
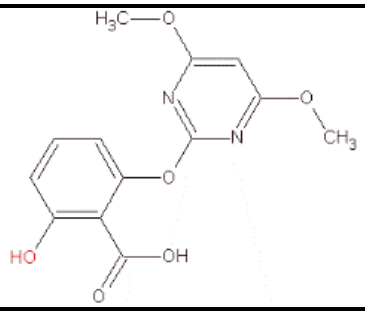
The **Alcohol** functionality is always gained in metabolites, being either aromatic, which often implies a ring hydroxylation reaction, or alkyl alcohol, i.e. a side-chain oxidation reaction. The presence of the alcoholic feature may arise also from ethers O-dealkylation, thus coupled with a negative Delta of the **ether** functionality, as in the example below (sub_id 1238, Table 42).

Table 42: Example of the **alcohol** functional group definition, in the case of sub_id 1238

			
	AS		
sub_id	1238	1238	1238
com_id	1666	2088	2089
alcohol		1	2
ether		-1	-2

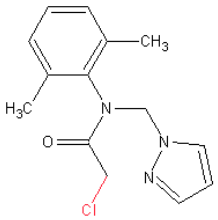
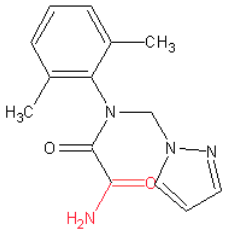
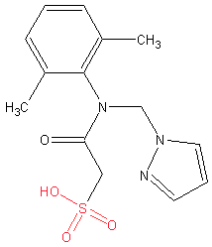
Ether feature is not always connected with the alcoholic one, as exemplified in the case below (sub_id 1246. Table 43). It appears that two of the three ether functionalities lost in the metabolite are related with the **aromatic** negative Delta, i.e. the lack of the aromatic ring and attached methoxy groups in the metabolite. Whereas one alcohol group is formed by O-dealkylation of the other ether group.

Table 43: Example of the **ether** functional group definition, in the case of sub_id 1246

		
	AS	
sub_id	1246	1246
com_id	1860	15525
aromatic		-3
alcohol		1
ether		-3

Halides, chlorine and fluorine above all, are predominantly lost in the transformation (see Table 36). It can be due to a reductive dehalogenation, as well as transformation in other features (e.g. carboxamideor sulfonyl), as in the following example (sub_id 1168, Table 44).

Table 44: Example of the **halides** functionality definition, in the case of sub_id 1168.

			
	AS		
sub_id	1168	1168	1168
com_id	1664	2056	2057
carbonyl		1	0
carboxamide		1	0
halide		-1	-1
sulfonyl group		0	1

Alkyl amine feature is related with the amine transformation, which for example could generate a primary amine, as illustrated below (sub_id 1347, Table 45).

Table 45: Example of the **Alkyl amine** feature definition, in the case of sub_id 1347.

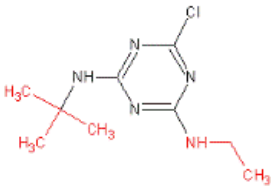
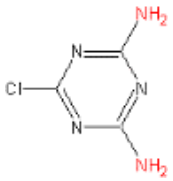
		
	AS	
sub_id	1347	1347
com_id	1698	1896
alkyl amine		-2
amino benzene		0

Figure 48 summarizes the main groups involved in metabolic transformations that are neutral in respect to genotoxicity.

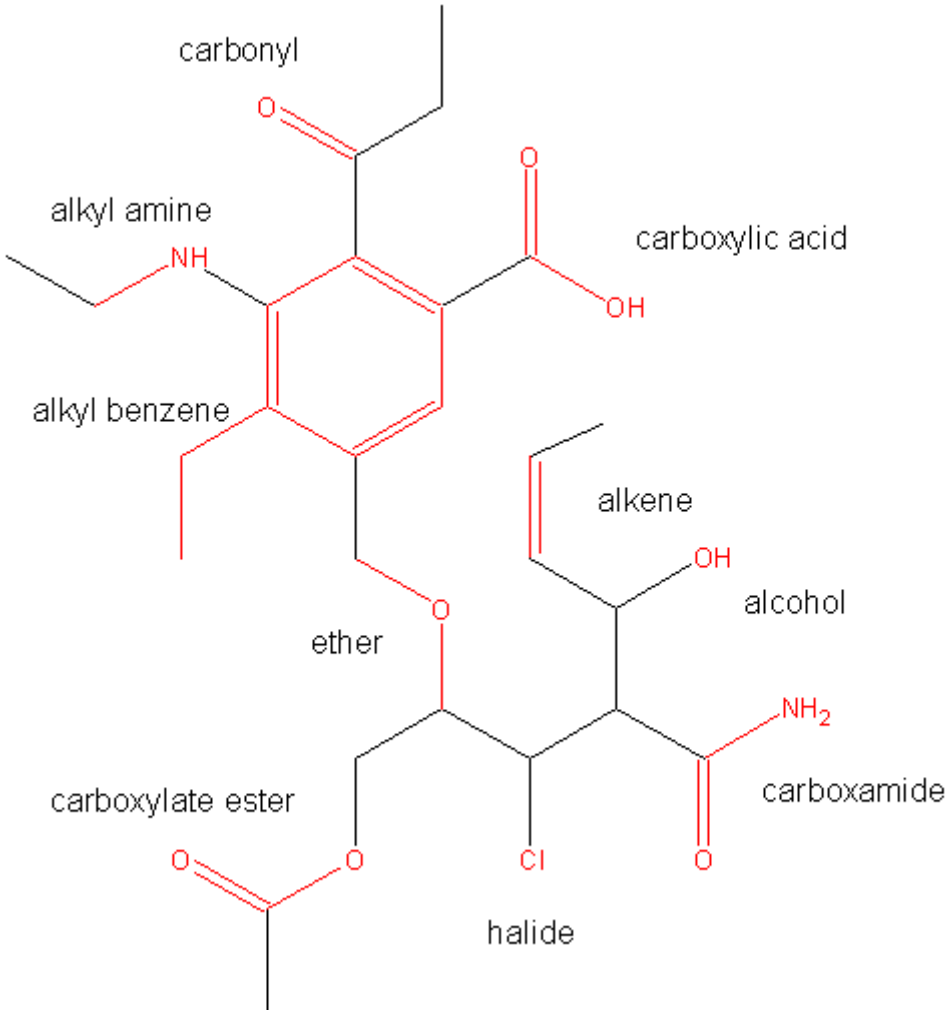


Figure 48: Structural changes that do not enhance toxicity

The functional groups displayed in Figure 48 are chemical modifications that -in the transformation from parents to metabolites- can be considered neutral with regards to the Ames mutagenicity concern. In combination with the knowledge on the toxicity of the parent pesticides, this information can be used as supporting evidence within Read Across or Weight-of-Evidence assessments of the genotoxicity of metabolites.

3.3.4.2. Structural transformations in the (minority) of negative to positive Ames results changes

In Table 46, for the seven couples in which the metabolites are mutagenic in the Ames test while the AS is negative, the characterization in terms of the functional groups is reported. For each metabolite, the change in functional groups with respect to its parent substance is shown. A negative value (negative Delta) represents a loss of the structural feature in the metabolite, whilst a positive value (positive Delta) represents a gain in the functionality. When there is no change in the functional group (or the functional group is absent in both the parent and the metabolite), a null value is reported.

Two metabolites in Table 46 contain a positive Delta for the **carbamate** group. The **carbamate** group represents a Structural alert for the Ames mutagenicity of the pesticide and, if present, it cannot be considered as a factor not altering the reactivity of the parent compound.

In the other pairs (see above), the Ames test positivity is questionable. Anyhow, it cannot be explained in terms of the pattern of the chemical features selected, showing only a positive Delta in carboxylic or sulfonyl groups.

Table 46: Pesticide / Metabolite pairs with positive Ames metabolites.

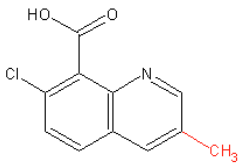
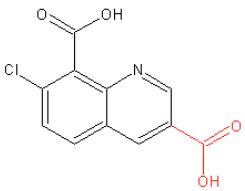
sub_id	**1139**	**1139**	**1191**	**1232**	**1259**	**4187**	**15013**
metab. com_id	**1606**	**15270**	**50208**	**75404**	**15544**	**50397**	**1605**
AS com_id	**1488**	**1488**	**1559**	**1573**	**1502**	**6363**	**15061**
deltaSTY	**1**	**1**	**1**	**1**	**1**	**1**	**1**
carbonyl	-1	-1	0	1	0	0	-1
carboxylic acid	0	0	0	1	0	0	0
carboxamide	0	0	0	0	0	0	0
carboxylate ester	-1	-1	0	0	0	0	0
aromatic	0	0	0	0	0	-1	-1
alkyl benzene	0	0	0	0	0	0	0
amino benzene	0	0	0	0	0	0	0
halo benzene	0	0	0	0	0	0	0
hydroxy benzene	0	0	0	0	0	0	0
hydroxymethyl benzene	0	1	0	0	0	0	0
alcohol	0	1	0	0	0	0	0
halide	0	0	0	0	0	0	0
ether	0	0	0	0	0	0	0
alkyl amine	1	1	0	0	0	0	-1
alkene	0	0	0	0	0	0	0

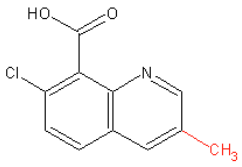
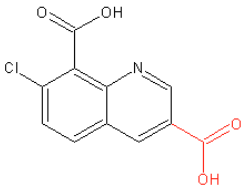
sub_id	**1139**	**1139**	**1191**	**1232**	**1259**	**4187**	**15013**
metab. com_id	**1606**	**15270**	**50208**	**75404**	**15544**	**50397**	**1605**
AS com_id	**1488**	**1488**	**1559**	**1573**	**1502**	**6363**	**15061**
deltaSTY	**1**	**1**	**1**	**1**	**1**	**1**	**1**
carbamate	1	1	0	0	0	0	0
sulfonyl group	0	0	0	0	1	0	0

Cells in red indicate pairs in which parent and/or metabolite contain a carbamate moiety.

As an illustrative example of the calculations at the basis of Table 46, results relative to sub_id 1232 are displayed in Table 47.

Table 47: Changes in relevant functional groups in the chemical transformation from the parent substance to the metabolite of the sub_id: 1232

		
	AS	
sub_id	1232	1232
com_id	1573	75404
carbonyl		1
carboxylic acid		1
carboxamide		0
carboxylate ester		0
aromatic		0
alkyl benzene		0
amino benzene		0
halo benzene		0
hydroxy benzene		0

		
hydroxymethyl benzene		0
alcohol		0
halide		0
ether		0
alkyl amine		0
alkene		0
carbamate		0
sulfonyl group		0

3.3.5. Integration of the knowledge on structural changes

In this project, we have evaluated several types of evidence (QSAR, Structural Alerts, Read Across, Structural Factors) that can be combined to assess the potential toxicity of metabolites.

There can be several ways of integrating such evidence. One possibility is of e.g., implementing sequential, tiered approaches that may enable more efficient evaluation of large numbers of chemicals, for example for prioritization purposes. Another possibility is that of using it as supporting evidence in a read across from the parent chemical to its (one or more) metabolites. In the following sections, we demonstrate some examples of these applications.

3.3.5.1. Integration in tiered approaches

In the following two exercises, we show how the evidence analysed and reported in this document may support the evaluation of the Ames mutagenicity potential of a large inventory of pesticides metabolites. To this aim, the findings from this work are exploited in the form of tiered approaches.

The exercises below use the EFSA genotoxicity data.

Tiered approach 1

The initial situation is represented by 566 metabolites available in the EFSA-QSAR DB, which are screened for the presence of structural alerts for the Ames mutagenicity. The numerosity of the initial sample depends on the applicability domain of the structural alerts used. In this first exercise, we applied the ISS Ames mutagenicity module of Toxtree. A similar strategy can be based on other (Q)SAR models (or combinations of), as we show in the next exercise.

In the first tier, the initial screening generates two different sub-samples: Sample 1, containing substances (metabolites) without SAs; and Sample 2 in which one or more SAs were detected in the chemicals. Sample 1 could be considered of low mutagenicity concern. This assumption results in 1% error (5 out 431 substances are erroneously predicted as negatives).

The second tier of the strategy involves the use of a structural similarity filter on Sample 2, which contains substances with at least one SA. Structural similarity is calculated between each metabolite in Sample 2, and its own parent. This Similarity Index calculation (based on the Toxtree profiler "Structural Alerts for Functional Groups Identification") further subdivides the chemicals into a set of substances with low similarity (< 70%) (Sample 3), and another sample (Sample 4) containing metabolites more similar to their parents (≥ 70%).

Sample 3 includes metabolites containing one or more SAs (irrespectively from the presence of SAs in the parent chemical), which are structurally very dissimilar from their parents. In this situation, the experimentally known mutagenicity value of the parent chemical cannot help in the evaluation of the metabolites, like in a read across approach. The positive prediction for these chemicals, suggested by the alerts, should be evaluated on an individual basis, through expert judgement. Analysis of specific SAs (especially those with high positive predictivity in this dataset, see **Analyses 1 and 2**), can support the evaluation. It should be emphasized that Sample 3 is greatly enriched in positive substances (14 %), with respect to the initial one (4 %).

Regarding the metabolites with chemical structures very similar to their parents (Sample 4), a further step may be envisaged. The differences in structure between parents and metabolites are analysed in terms of changes in functional groups.

If the changes in metabolites structure consist of functional groups which are likely to not alter the mutagenicity of the parent (i.e. the "neutral" ones shown in figure 48), a 1:1 Read Across could be performed from the parents to their metabolites (Sample 5). On the other hand, when structural changes are not included among the "neutral" ones (Sample 6), a change in the mutagenicity potential of the metabolite cannot be ruled out. As a matter of fact, the proportion of positives in Sample 6 is 7 / 28 (25%).

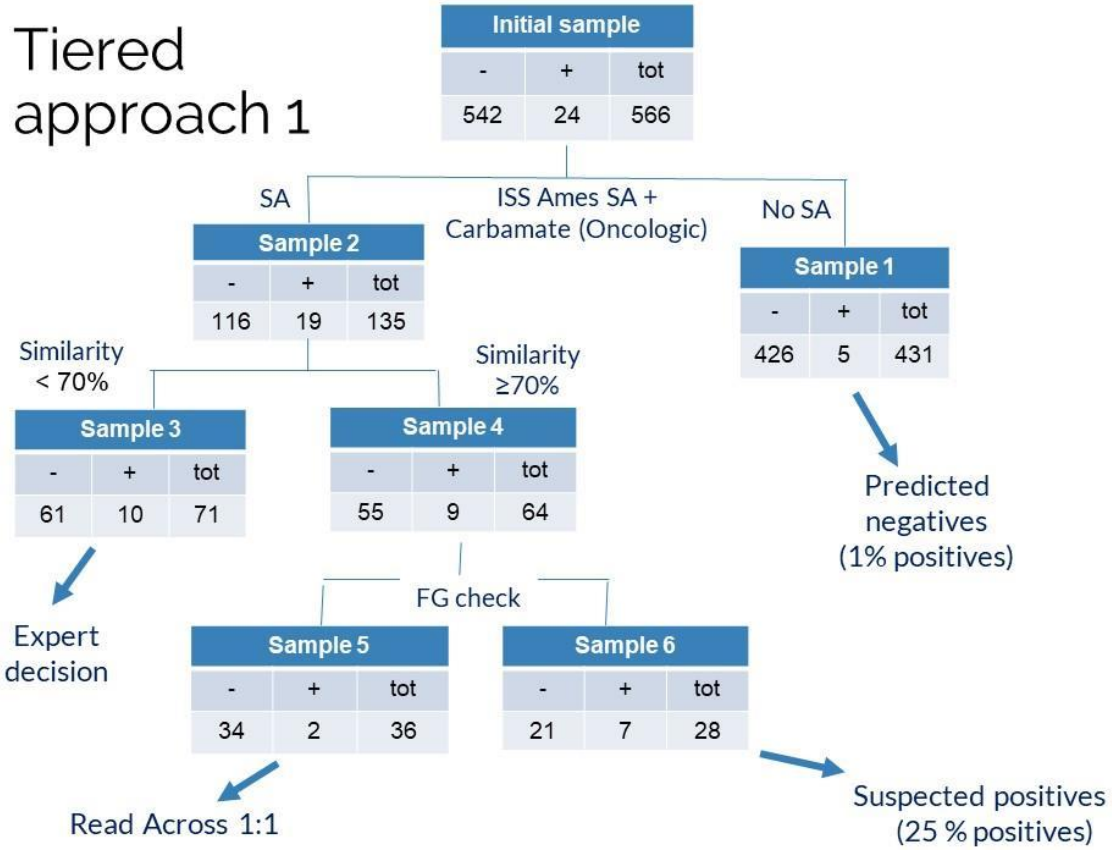


Figure 49: Tiered approach 1 for the analysis of metabolites. The case study uses the components studied in this work.

Along this scheme, different implementations may be designed. An example is provided in the following section.

Tiered approach 2

This second exercise follows the lines of the first one, but a different predictive system is applied for the first screening of the initial sample. Iinstead of the ISS Ames mutagenicity module of Toxtree, two QSAR models, namely Sarah (statistical based model) and Leadscope Model Applier (rule based model), are applied in combination (Figure 50).

Similarly to tiered approach 1, the sample containing metabolites with negative QSAR predictions (Sample 1) has a low percentage of positives (1 %). Also in this case, the sample of metabolites with positive predictions and with low similarity (< 70%) with their parents (Sample 3) is greatly enriched in positive chemicals (27 %). Metabolites similar to their parents (similarity ≥ 70%, sample 4) can be evaluated by means of read across, using the supporting information of the functional groups. This results in only one error in the overall prediction of the sample (details not shown).

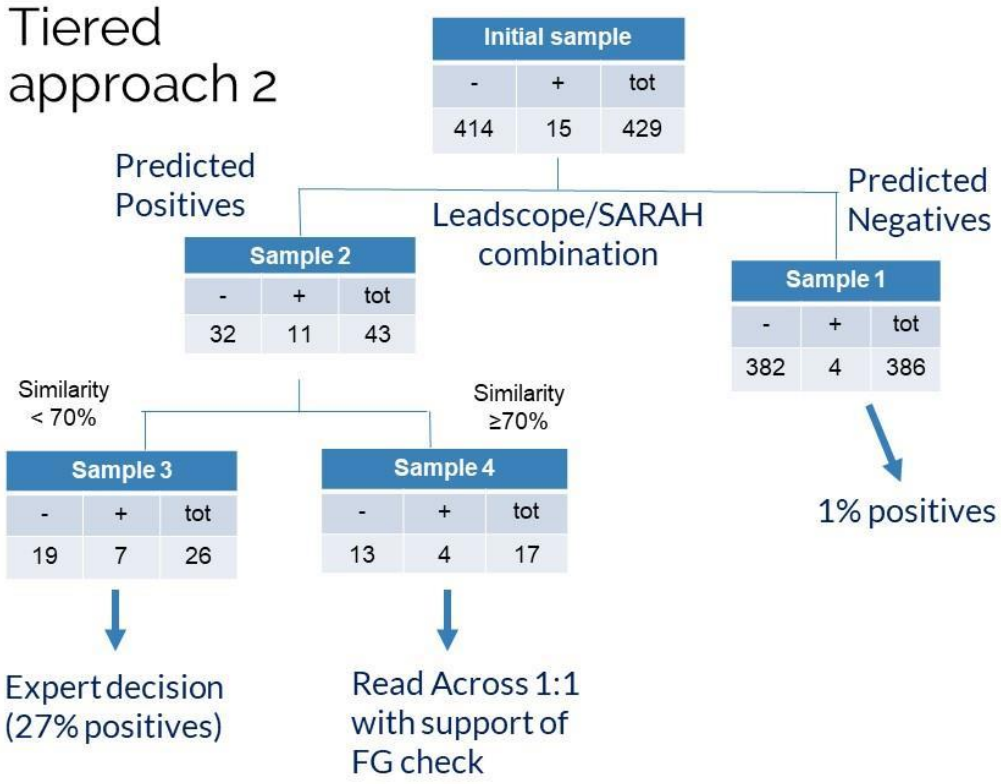


Figure 50: Tiered approach 2 for the analysis of metabolites. The case study uses the components studied in this work.

Overall, the above exercises show that the integration evidence from this work permits the discrimination between sub-sets of metabolites with markedly different probabilities of positivity, ranging from 1% (Samples 1), to 25 – 27 % (Sample 6 in exercise 1, and Sample 3 in exercise 2). At the same time, this provides a description of the landscape of the genotoxicity / structural factors in the EFSA database.

In addition, the integration of evidence can provide the basis for WoE assessments of individual chemicals, as shown in the following section.

3.3.5.2. Weight of evidence approach in mutagenicity evaluation, a case study

In this exercise, the possibility to read across the Negative mutagenicity value of the parent compound (com_id 1554) to one of its metabolites (com_id 50063) is evaluated, taking into account different evidence collected in this study.

In this case, depicted in Figure 51, the following information supports the Negative outcome of the read across:

- High similarity with the parent chemical (≥ 70%). This evidence is associated with a good performance in the 1:1 Read Across approach (93 % of predictivity), as illustrated in Section 3.2.1.3.

- Absence of Structural alerts for mutagenicity. This evidence is associated with a correct negative prediction in the 99% of cases, when applied to the EFSA genotoxicity database (see tiered approaches in the previous section).
- In the transformation from parent to metabolite, the chemical modifications that have occurred can be considered neutral with regards to the Ames mutagenicity concern. In particular, transformation of carboxylate and carboxylic acid functional groups are associated with no concern for mutagenicity in 100 % and 99 % of cases respectively (calculated the basis of Table 36).

Collectively, the weight of this evidence strengthens the Negative outcome of the Read Across leading to a correct negative prediction.

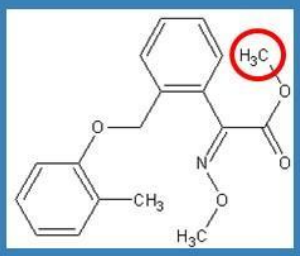
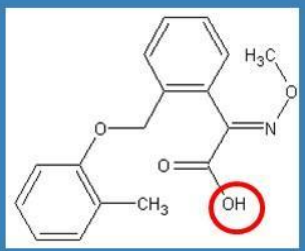
		
	Active Substance	metabolite
sub_id	1314	1314
com_id	1554	50063
Ames test	**Negative**	
Ames test prediction		**Negative**
Similarity with AS		75 %
		Delta
alcohol		0
carbonyl		0
carboxylic acid		1
carboxylate		-1
aromatic		0

Figure 51: Summary of evidence for the mutagenicity assessment of metabolite com_id 50063

3.3.6. Conclusions on the influence of metabolic structural changes

The goal of Objective 5 is: "Evaluation of the impact of the structural changes in the molecule in result of metabolic or degradation processes to the genotoxic potential of the substances". This goal was achieved by performing a number of complementary analyses, some more mechanistically oriented (Analysis 1) and others more based on chemoinformatics and statistics (Analyses 2 and 3).

Analysis 1 showed that the expert (not automatic) use of SAs permits the rationalization of the large majority of the patterns of genotoxicity in the subgroups of substances in which parent and (some) metabolites have different Ames outcomes. Expert reasoning on each individual case, including the combination of different types of SAs and the inspection of close analogues when appropriate, attained

better predictive performance in respect to the automatic application of expert rules (e.g., DEREK, Toxtree, etc..).

The second series of analyses used chemoinformatics / statistical tools in order to identify structural differences -between parent and metabolite- that go beyond the known SAs, and may, or may not produce changes in the Ames mutagenicity. The analyses were performed from different perspectives (i.e., in closely congeneric Parent / Metabolite pairs; in all pairs irrespective of similarity). The potential effects of the appearance / disappearance of functional groups / SAs on the genotoxicity of the metabolites is described in a series of tables.

These results should be seen in connection with those from Objective 2 (evaluation of QSARs) and Objective 4 (evaluation of Read Across). QSARs, Read Across, expert analysis of Structural Alerts, consideration of structural changes, Chemical Similarity are complementary tools to be used in the assessment of metabolites. Since both experiments and *in silico* methods are probabilistic in nature, the combined use of a wide array of tools can surely increase the reliability and confidence in the assessment. Examples of how this information can be combined in sequential (tiered) and Weight-of Evidence approaches is given in the previous section.

4. Overall Conclusions

To facilitate the practical implementation of the guidance on the residue definition for dietary risk assessment, a large scale evaluation of applicability of existing *in silico* models for prediction of genotoxicity of pesticides and their metabolites, together with analyses of the impact of structural factors related to metabolic changes, has been organized by EFSA.

This will be beneficial for the work of the risk assessors when applying the guidance for residue definition as well as in other areas of risk assessment of pesticides. This endeavour has also a remarkable scientific dimension, since it is the first study -at this large scale- on pesticides.

4.1. The landscape of the results of this work

A first investigation in this project is the **evaluation of (Q)SARs.** This was based both on: a) an extensive literature search; and on: b) an *ad hoc* exercise of prediction of the genotoxicity of the Pesticides and Metabolites contained in the EFSA Genotoxicity Database.

Point a) revealed the existence of a very rich literature, including comparative prediction exercises, as well as more specific topics like combination of QSAR models, effect of Applicability Domain on the predictions, integration between models and expert knowledge. Almost all studies retrieved focused on modelling the Ames test, and not the other assays or endpoints. The results from literature agreed with what was found in the *ad hoc* exercise in Point b).

Point b) was performed by applying a large range of commercial and publicly available (Q)SAR models to the EFSA genotoxicity database. Five experimental assays were selected: Bacterial Reverse Mutation Assay (Ames test), Mammalian Bone Marrow Chromosome Aberration Test, Mammalian Erythrocyte Micronucleus Test, *In vitro* Mammalian Chromosome Aberration Test, *In vitro* Mammalian Cell Gene Mutation Test.

Overall, the results of this investigation point to a substantial difference between the performance in the prediction of the Ames test on one hand, and that of the other experimental assays on the other hand.

For the Ames test, all (Q)SAR models generated statistically significant predictions. Sensitivity ranges between 46% (Toxtree) and 71% (a model from Leadscope), Specificity between 66% (Lazar) and 98% (Percepta). Overall, this result confirms the statistically significant predictions reported in previous exercises available in the literature.

On the opposite, the reliability of the (Q)SAR models for assays / endpoints different from *in vitro* bacterial mutagenicity (Ames) appears to be still quite far from optimality. There is no possibility of comparing these results with previous studies in the literature, since an extensive literature search did not retrieve similar prediction exercises. Thus, this EFSA projects contributes with original information to the research on the predictivity of QSARs for genotoxicity endpoints different from bacterial mutagenicity.

Combinations of QSAR predictions of the EFSA genotoxicity results were explored, confirming evidence from literature: as a general trend, the combination of QSARs increases Sensitivity, but at the expense of Specificity. On the other hand, predictions within and outside the Applicability Domain of the models do not seem to be drastically different. The general advice is not to dismiss as insignificant the predictions outside the Applicability Domain.

Whereas QSARs have undergone during the years many performance evaluations, with special emphasis on comparative prospective exercises, nothing analogous can be found in the literature for **Read Across.** The literature is rich in proposals for general workflows and criteria, but the published examples of applications –even though often quite detailed- are limited in number and do not provide sufficient material for assessing the real value of the proposed workflows.

In this work, we present around sixty Read Across case studies in which we try to predict the genotoxicity of metabolites from the information on the parent pesticide. It should be recalled that the pesticides properties are systematically documented in the dossiers provided to the Regulatory Authorities, thus are the primary source of information on which to base the Read Across for their metabolites. When necessary, further information from a wider range of analogues was used in our exercises.

Read Across was applied to both Ames and *in vitro* Chromosomal Aberrations assays, with two different strategies that consider different sets of information. A common result is that Read Across appears to be largely successful for predicting the Ames test results. The performance of the two strategies was partially different with *in vitro* Chromosomal Aberrations, but overall it was lower than that obtained with the Ames test.

Considering the different degree of success of Read Across and QSAR for the various genotoxicity assays, a crucial issue is why the Ames test can be predicted with reasonable accuracy, whereas other genotoxicity assays cannot. This is discussed in detail in Section 3.1.6, in connection with the evaluation of QSARs. Here we can only repeat that it may be attributed to the lower quality level of the databases for tests different from the Ames test.

A third dimension of this research was the evaluation of the **impact of the structural changes** -in result of metabolic or degradation processes- to the genotoxic potential of the substances. One line of research of this project studied the potentially most dramatic structural changes, represented by changes in the pattern of Structural Alerts. The information on Structural Alerts was not applied in an automatic way, but was filtered through human expert knowledge: this permitted the rationalization of the large majority of the patterns of genotoxicity in the subgroups of substances in which parent and (some) metabolites have different Ames outcomes. As a matter of fact, the supervision by the human expert permitted a better predictive performance in respect to automatic applications of rules.

In addition, an extensive analysis of Parent / Metabolite structural differences -beyond the known Structural Alerts- was performed with chemoinformatics tools. This resulted in a list of structural changes that were catalogued into those related to changes in the Ames mutagenicity, and others which are neutral in this respect. The knowledge on these structural factors complements the knowledge on Structural Alerts, and –as demonstrated with a number of exercises in this work- may be used in combination in the assessment of the genotoxicity of metabolites.

4.2. Suggestions for future improvements of *in silico* models and approaches

The evidence collected in this study points to areas to be explored to improve the predictive ability of the models.

For the (Q)SARs, the coverage of, and generalizability to the whole chemical space should be improved. This requires that training sets represent better the diversity of chemical structures. This involves not only search and curation of data by the developers, but also new experimentation that –unfortunately- can only come from initiatives of large scientific bodies. At present, the only initiative of this type is that of National Institute of Health Sciences of Japan, that has made public new Ames data on around 12,000 chemicals (Honma et al., 2018).

The need for better curation of data is of utmost importance for assays different from the Ames test: the changes in the history of protocols and criteria for these assays require a special attention to the critical revision and curation of the experimental databases used as training sets. The present state-of-art of genotoxicity databases does not permit to base on firm ground the (Q)SARs for genotoxicity tests different from the Ames test.

The coverage of, and generalizability to the whole chemical space is also related to the critical issue of the Applicability Domain rules. Direct, even though limited evidence from this work indicates that at present there is no dramatic difference in the reliability of predictions within and outside the Applicability Domains of QSARs. This is confirmed by the fact that the same QSAR, when applied to different data sets, may have largely different performance, implicitly pointing to failures in recognizing the proper domain for the QSAR. Until more progress is made in this area, a practical advice it not to disregard as insignificant the predictions outside the Applicability Domains.

Other evidence from this work indicates that the integration of the (Q)SAR (including Structural Alerts) predictions with expert knowledge is a way to generate a more equilibrated increase of both Sensitivity and Specificity in respect to simple combinations of (Q)SARs. This points to the existence of a large area of context-dependent expert knowledge, which has not been formalized yet in the QSAR models (even in the expert systems), and has the potential to substantially improve the prediction systems.

Several issues related to Read Across are different from those of QSAR. The long history of QSAR has led to recognize the criteria and conditions to develop a valid QSAR. The construction of a QSAR model is a process in which the weights to be given to different factors (e.g., physical chemical properties, substructures) are assigned by a data-based statistical analysis of training sets in a rigorous and reproducible process, and the predictive ability can be objectively measured in well-designed exercises. Among other factors, it is recognized that: a) collinearity or multi-collinearity (inter-correlations) among variables has to be avoided; b) a number of chemical descriptors too high in respect to the number of chemicals may give rise to chance correlations; c) the fitting of the model has to be characterized by high F-statistics values and low standard deviations; d) the goodness of fitting has to be validated by cross-validation and external validation. In addition, the model has to make sense from a physical chemical / structural scientific point of view (Benigni et al., 2007; Cherkasov et al., 2014; Dearden et al., 2009; Franke and Gruska, 2003; Hansch et al., 2002; Hansch and Leo, 1995; Kubinyi, 2005).

On the contrary, the situation with Read Across is fuzzier. First of all, Read Across is much more a case-by-case analysis, where different cases may have different and partial pieces of information. In addition, the development of validation procedures -that has been central to the QSAR research for many years- has not been performed in the same systematic and rational way for Read Across. Thus, the Read Across practitioner only has an initial list of pieces of information –maybe characterized by their uncertainties-, but this list is not incorporated into an overall model calibrated on the endpoint to be predicted. In the longer period, it is important that further scientific investigations on the predictive ability of Read Across are carried out, and that objective performance measures are established. This EFSA project contributes to such an investigation with one of the largest exercises available today, numbering around sixty case studies.

The importance of such research can be emphasized by recalling some previous, unpublished work obtained by the ISS team (R. Benigni; C. Bossa; C.L. Battistelli; O. Tcherememskaia) in connection with a Scientific Review of the QSAR Toolbox and usability improvements (Project ECHA/2013/167). Read Across were performed for the Ames and aquatic toxicity tests. Several characteristics of the information used (quality and size of databases, chemical diversity of the databases, Tanimoto similarity of analogues, etc…) were contrasted with the predictivity of Read Across: it appeared that only Tanimoto similarity had a strong correlation with the predictivity. In other terms, the quality (chemical similarity) of analogues was the bottle neck of the entire process, independently from how and from where they were retrieved.

Obviously, both QSAR and Read Across require better data curation for the assays different from the Ames test.

4.3. On the use of *in silico* models for predicting genotoxicity

A recurrent question is whether the present performance of (Q)SARs is sufficient for ensuring a reliable practical use. The work in this project indicates that the QSAR prediction accuracy for the *in vitro* bacterial mutagenicity of EFSA chemicals resulted in an average 0.86 Accuracy, with most of the models in the range 0.70 – 0.90. This compares fairly well with the experimental variability of the Ames test itself, whose repeatability from laboratory to laboratory has been shown to be 0.80 – 0.85 (Piegorsch and Zeiger, 1991). This suggests that (Q)SARs for the Ames test have sufficient reliability for use in prioritization processes, as well as support for regulatory decisions in combination with other types of evidence. On the contrary, QSARs for other assays endpoints still need improvements.

This work has also shown that Read Across of the Ames test may give reliable predictions as well, especially when supported by careful analysis of analogues and of Structural Alerts. Read Across for *in vitro* Chromosomal Aberrations seems to be not as reliable as for the Ames test. A recommendation is that Read Across for non-Ames tests is applied with extreme caution, for example by accepting its results only when a wide range of analogues is available and all have convincingly similar patterns of activities and properties.

The need for expert supervision of all applications merits special emphasis.

4.4. More on the integration of evidence

Since both experiments and *in silico* methods are probabilistic in nature, the combined use of a wide array of tools within integrated testing strategies can surely increase the reliability and confidence in the assessment.

In this project, we have evaluated several types of evidence (QSAR, Structural Alerts, Read Across, Structural Factors, Structural Similarity) that can be used to assess the potential toxicity of metabolites. Integration of evidence –at the best of professional judgement -must take place in order to reach conclusions. Examples of integration for evidence for assessing individual chemicals have been provided throughout this work. The usefulness of sequential, or tiered approaches deserves an additional discussion: these may enable more efficient evaluation of large numbers of chemicals. On this respect, it can be useful to refer to a thorough analysis performed at the Organization for Economic Co-operation and Development (OECD) taking as an example the prediction of carcinogenicity(OECD, 2017a).

Two opposite types of skepticism inhibit a wider use of alternative approaches (e.g., QSARs). Regulatory hazard assessors are suspicious of negative model predictions for the non-carcinogens, because of concerns about possible false negatives from the prediction models. Industrial safety assessors are suspicious of positive model predictions for the carcinogens, because of possible false positives from the prediction models. Both of the above concerns tend to encourage continued default use of the rodent carcinogen assay. It was suggested to solve the dilemma of false positives *versus* false negatives,

by separating the process into two arms: i) use a sequence of conservative (very sensitive) (Q)SARs or *in vitro* models to arrive to the conclusion that no further testing is necessary for negatives; ii) use a mechanistically based, Weight-Of-Evidence approach to evaluate the chemicals showing positive results for (Q)SARs or *in vitro* tests (OECD, 2017a). Within this scheme, intelligent use of (Q)SARs with different Sensitivity / Specificity can give an answer to different types of issues that may arise during the assessments.

Along these lines, we have sketched a preliminary scheme in the form of a tiered approach using different factors in succession. At the end, the decision tree places the metabolites into different categories, i.e., low probability (1%) of being mutagens, high probability (25%) of being mutagens, uncertain. This classification can be used to prioritize the assessment of large numbers of chemicals, as well as to inform and guide evaluations of individual chemicals, by relying on a data-based characterization of the case under study.

4.5. A final quotation

Finally, we would like to quote a superb description of the role of QSAR by Rainer Franke (even though applied to drug design, it has of general value): " … As the drug design process is of a very complex nature, effective drug design requires an entire spectrum of techniques in which QSAR methods still play an important role… The real power of drug design methods is to extract and synthesize information from data to obtain hypotheses that can be put to experimental test. No dramatic overnight discoveries of wonder drug will result, but an increase in the chance of success due to indications of promising directions is a realistic expectations…"(Franke and Gruska, 2003) . Thus, QSAR in toxicology contributes to understand experimental data and to defend human health in combination with other tools.

Glossary/Abbreviations

Term	Explanation
AD	Applicability Domain
ADI	Applicability Domain Index
ADME	Absorption, Distribution, Metabolism and Excretion
Ames test	Mutagenicity in Salmonella typhimurium (Bacterial Reverse Mutation Assay)
Aneugenicity	Numerical chromosomal aberrations
AOEL	Acceptable operator exposure level
ARfD	Acute Reference Dose
AS	Active Substances
BCF	Bioconcentration factor
CAS number	Unique numerical identifier assigned by Chemical Abstracts Service (CAS) to every chemical substance described in the open scientific literature
CBI	Categorical Bayesian Integration Approach
CCRIS	Chemical Carcinogenesis Research Information System
CE	Cloning Efficiency
CFSAN	Center for Food Safety and Applied Nutrition
CHA	Chromosomal Aberrations
ChemACE	Chemical Assessment Clustering Engine by US EPA
ChemID Plus	Toxnet Database from US National Library of Medicine
CHL	Chinese Hamster Lung
CHO	Chinese Hamster Ovary
Clastogenicity	Breaks or rearrangements of chromosomes
COM_ID	ID number, identifying individual chemicals (both Parents and Metabolites) in the EFSA genotoxicity database
Comet	Comet assay (single cell gel electrophoresis SCRE) on eukaryotic cell
COREPA	Pattern recognition approach
CPDB	Carcinogenic Potency Database
Critical Effect	Adverse effect seen at the lowest dose when a vulnerable population is exposed to a substance such as an environmental or food toxin.
CV	Cross validation
DAR	Draft Assessment Report
DB	Database
DDSR	Division of Drug Safety Research Staff
DfW	Derek for Windows
DGM/NIHS	Division of Genetics and Mutagenesis, National Institute of Health Sciences
Dice chemical similarity coefficient	Coefficient used for comparing the similarity of two chemicals. It is twice the number of elements common to both chemicals, divided by the sum of the number of elements in each chemical
DL	Dominant lethal
DNA	Deoxyribose nucleic acid
DS	Datasets
DST	Dempster-Shafer theory
E.coli	Escherichia coli
ECHA	European Chemicals Agency
eChemportal	The Global Portal to Information on Chemical Substances
EFSA DB	Genotoxicity EFSA database
EFSA	European Food Safety Authority
EU	European Union
EURL ECVAM	European Union Laboratory for Alternatives to Animal Testing

Term	Explanation
Genotoxicity OASIS Database	OASIS genotoxicity database implemented in the OECD QSAR Toolbox, by the Laboratory of Mathematical Chemistry (LMC)
GLP	Good laboratory practice
GSH	Glutathione
HOMO	Energy of the Highest Occupied Molecular Orbital
ICH M7	Guideline for Assessment and control of DNA reactive (mutagenic) impurities in pharmaceuticals to limit potential carcinogenic risk
ICH	International Conference on Harmonisation of Technical Requirements for the Registration of Pharmaceuticals for Human Use
ISS	Istituto Superiore di Sanità (Italian National Institute of Health)
ISSCAN database	Long-term carcinogenicity bioassay on rodents (rat, mouse) database by ISS
ISSCTA database	Cell transformation database by ISS
ISSMIC database	*in vivo* mutagenicity (micronucleus test) database by ISS
ISSSTY database	*in vitro* mutagenicity in Salmonella typhimurium (Ames test) database by ISS
ITC	Interagency Testing Committee
JRC	Joint Research Center
kNN	k-Nearest Neighbours (pattern recognition approach)
Kow	Octanol/water partition coefficient
Lazar	Lazy Structure-Activity relationships
LC50	Lethal Concentration, on half of the sample population
LD50	Lethal Dose, on half of the sample population
LMO	Leave-many-out
LogP, Log Kow	Logarithm (base 10) of the octanol/water partition coefficient
LOO	Leave-one-out
LSMA	Leadscope Model Applier
LUMO	Energy of the Lowest Unoccupied Molecular Orbital
MET	Metabolites
MF	Mutant frequency
MLA	Mouse lymphoma assay
MN	Micronucleus
MOA	Mode of Action
MultiCase	Multiple Computer Automated Structure Evaluation
MW	Molecular Weight
NB	Naive Bayesian model
NOAEL	No Observed Adverse Effect Level
OECD	Organisation for Economic Co-operation and Development
OFAS	Office of Food Additive Safety
Oncologic (software)	Computer-based expert system by US EPA
OPS	Optimum Predictive Space
PCA	Principal Component Analysis
PDR	Physician's Desk Reference
Pearson coefficients	Correlation coefficient between variables; it is the covariance of the two variables divided by the product of their standard deviations
PEG	Polyethylene glycol
Pesticide	Substance used to kill or control pests, including disease-carrying organisms and undesirable insects, animals and plants.
PLS	Partial Least Squares
PLS-DA	Partial Least Squares Discriminant Analysis
PPR	EFSA Pesticides Unit, and Panel on Plant Protection Products and their Residues
QPRF	Prediction reporting format

Term	Explanation
(Q)SAR	(Quantitative) Structure Activity Relationship
QSAR Toolbox	OECD QSAR Toolbox, version 4.2 (by the Laboratory of Mathematical Chemistry (LMC), Bourgas, Bulgaria)
RA	Read Across
RAR	Renewal Assessment Report
REACH	Registration, Evaluation, Authorisation and Restriction of Chemicals
RIFM	Research Institute for Fragrance Materials
RMS	Rapporteur Member State
ROC	Receiver Operating Characteristics
SA	Structure Alerts
SCE	Sister Chromatid Exchange assay
Sensitivity	Correctly predicted positive/total number of positive
SMARTS	Language for Describing Molecular Patterns, allowing to specify substructures using rules that are straightforward extensions of SMILES
SMILES	Simplified Molecular Input Line Entry System, a specification in form of a line notation for describing the structure of molecules using short ASCII strings.
Specificity	Correctly predicted Negative/total number of negative
STY	Mutagenicity in Salmonella typhimurium (Ames test)
SUB_ID	ID number, identifying the group of Parent and Metabolites belong to, in the EFSA genotoxicity DB
SVM	Machine learning algorithm
Tanimoto (or Jacard) chemical similarity coefficient	Coefficient used for comparing the similarity of two chemicals. It corresponds to the number of elements in common, divided by total number of the elements in each chemical
T.E.S.T.	Toxicity Estimation Software Tool by US EPA
TGR mutation	Transgenic rodent mutation
Toxicity Japan MHLW	Databases by the Japan, Ministry of Health, Labour and Welfare toxicity, under Japanese Existing Chemical Programme
Toxread	Software by Istituto di Ricerche Mario Negri (Milano) and KODE
ToxTree	Toxic Hazard Estimation by decision tree approach, version v. 2.6.13 by Ideaconsult Ltd (Sofia, Bulgaria)
TTC	Threshold of Toxicological Concern
UDS	Unscheduled DNA synthesis
US EPA	Environmental Protection Agency, United States
US FDA	Food and Drug Administration, United States
VEGA	Virtual Models for evaluating the properties of chemicals
WoE	Weight-of-Evidence approach

Appendix A – Survey on QSAR systems: literature search strategy

Study question

Given the central role of the assessment of genotoxicity of pesticides and their metabolites for regulatory purposes, an evaluation of applicability of existing (Q)SAR models for its prediction is particularly important. The compilation of a list of recommended (Q)SAR models with the best performance, and with the most reliable predictions of genotoxicity of pesticides active substances and their metabolites, will be beneficial for the work of the risk assessors when applying the guidance for residue definition as well as in other areas of risk assessment of pesticides.

General eligibility criteria for inclusion or exclusion of studies

This review is complementary to the external scientific report done by the Joint Research Center (JRC) in 2010 (Worth et al., 2010). It covers free available and commercial models as well as literature models for the period from 1 January 2009 to 31 December 2016. The final list of studies included only papers providing new evidence, giving a priority to those results amenable to quantitative treatment and presentation (e.g., new comparative evaluations of QSAR systems). General presentations, even well written but without new information or data, were not treated. A number of research papers focusing on specific aspects (e.g., methodological improvements, potential contribution of sophisticated modeling approaches such as docking) were described as well.

Description of the review method and literature search

Search strings were developed by combining pairs of Key Words from two sets of conceptual components; A) QSAR; Structure-Activity Relationships; *in silico* on one side; and B) mutation; mutagenicity; genotoxicity; chromosomal aberrations; DNA damage; and more specific terms (e.g., Ames, Mouse Lymphoma, aneugenicity, etc…) on the other side.

Examples of strings are: "QSAR AND mutation"; "QSAR AND mutagenicity"; "QSAR AND genotoxicity"; "QSAR AND chromosomal aberrations"; "QSAR AND DNA damage"; "QSAR AND Ames"; "QSAR AND Mouse Lymphoma"; "QSAR AND aneugenicity", and so on.

The search strings were applied to the databases Pubmed, Google Schoolar, and several websites (OECD, JRC, ECHA, US EPA, Canada, Danish EPA, ECETOC). The search dates ran from 2009 – 2016.

Iterations of search term combinations were tested to refine the search, until no more new results were found.

The initial electronic search generated thousands of results that were reviewed for relevance by the Consortium scientists. This initial screening was performed based on titles and abstracts. Some terms and combination of terms (i.e., mutation and structure-activity relationships) pointed mostly to papers were the reference to chemical structure was irrelevant in respect to models and predictions. Search in Google Scholar generally confirmed the Pubmed findings, with a minor number of additional papers.

Upon refinement, around 60 papers were shortlisted. The text of these papers was screened more carefully, and only papers of high relevance where accepted.

In agreement with the principles of evidence-based reviews, data were extracted from the selected papers, and –whenever possible- plotted as ROC graphs for an easier visualization.

Methodological quality assessment of included Studies

The quality of the papers was carefully checked by the experts of the Consortium, based on their extensive experience with QSAR and mutagenicity research. Main criteria for inclusion of the papers in the review were the presence of factual data, reporting of quantitative predictions, figures of merit (e.g., sensitivity, specificity, etc…).

The experienced people involved in this phase of the project were six, namely Romualdo Benigni, Paola Leopardi, Alessandro Giuliani, Cecilia Bossa, Olga Tcheremenskaia, and Chiara Laura Battistelli.

General presentations, even well written but without new information, were discarded. The final list included only papers with new information (e.g., new comparative evaluations of QSAR systems, new approaches).

Reporting of study results

The study results consist of literature on: a) description of existing models; b) performance evaluations reported in the literature; and c) methodological improvements. Whenever possible, the presentation of data as ROC graphs provided an easy visualization of performance and comparisons. The data from the selected papers were organized into a number of different topics (e.g., performance in training and test sets, applicability domain, combination of systems).

Synthesis

Results were presented and discussed in Section 2.1, and were used to interpret and put in a larger perspective the outcomes of the QSAR prediction exercise with the EFSA database (Objective 2).

Appendix B – Software tools available for genotoxicity assessment

Genotoxicity prediction is featured in a wide range of commercial and freely available software tools. The most commonly used systems were already described in JRC, 2010 (Worth et al. 2010). Below we update the list and description. Main sources of description are JRC, 2010, (Worth et al. 2010) Fioravanzo et al., (Fioravanzo et al. 2012) and the websites of software developers. Summary descriptive details of predictive (Q)SAR models are reported in Table B.1.

Models in the public domain

T.E.S.T. https://www.epa.gov/chemical-research/toxicity-estimation-software-tool-test

T.E.S.T. is able to estimate various properties using a variety of approaches, and can be applied to different types of data to develop *ad hoc* QSAR models.

The following are the available QSAR methods:

- Hierarchical method – The toxicity for a given query compound is estimated using the weighted average of the predictions from several different models. The different models are obtained by using Ward's method to divide the training set into a series of structurally similar clusters. A genetic algorithm-based technique is used to generate models for each cluster. The models are generated prior to runtime.
- FDA method – The prediction for each test chemical is made using a newly developed model that is fit to the chemicals that are most similar to the test compound. Each model is generated at runtime.
- Single-model method – Predictions are made using a multilinear regression model that is fit to the training set (using molecular descriptors as independent variables) using a genetic algorithm-based approach. The regression model is generated prior to runtime.
- Group contribution method – Predictions are made using a multilinear regression model that is fit to the training set (using molecular fragment counts as independent variables). The regression model is generated prior to runtime.
- Nearest neighbor method – The predicted toxicity is estimated by taking an average of the three chemicals in the training set that are most similar to the test chemical.
- Consensus method – The predicted toxicity is estimated by taking an average of the predicted toxicities from each of the above QSAR methodologies.

The required molecular descriptors are calculated within the program. The software is based on the Chemistry Development Kit, an open-source Java library for computational chemistry.

Before any T.E.S.T. model is used to make a prediction for a chemical, the software determines whether the test chemical falls within the domain of applicability of the model. In case of the consensus method, the predicted toxicity is estimated by taking an average of the predicted toxicities from the previously mentioned QSAR methods, provided the predictions are within the respective applicability domains. The requirements for the AD are not satisfied for any of the models used in this software, then that model is not used by the consensus model. Thus, T.E.S.T. provides results only if the compound falls within the AD of at least one model.

In the present version, Ames test mutagenicity is predicted using the consensus method, where the predicted toxicity is the average of the predicted toxicities from the various T.E.S.T.'s QSAR methodologies.

CAESAR (VEGA) https://www.vegahub.eu/

CAESAR model for genotoxicity is implemented in the VEGA platform (Virtual Models for evaluating the properties of chemicals). It is an integrated model made of two complementary techniques: a machine learning algorithm (SVM), to build an early model with the best statistical accuracy, equipped with an expert facility for false negatives removal based on known structural alerts, to refine its predictions.

CAESAR is based on the use of Structural Alerts (SA) for Ames test mutagenicity, derived by studies done by human experts or from computer programs, like SARpy. All SAs lists include both active and inactive fragments, with the exception of those from Toxtree. The final Ames test mutagenicity outcome is derived with a CONSENSUS model that combines the different models.

The applicability domain of predictions is assessed using an Applicability Domain Index (ADI) that has values from 0 (worst case) to 1 (best case). The ADI is calculated by combining several indices (e.g., calculation of the most similar compounds found in the training and test set of the model, etc…).

In addition to the applicability domain, the system provides also a number of other parameters useful to evaluate the prediction. These include the identification of structurally similar compounds and the calculation of the similarity index, the check of unusual structural fragments, the check of the descriptor range, and the analysis of the sensitivity of the descriptors. Structural analogues from the training and test sets, along with experimental and predicted results, are provided as well.

Lazar https://www.in-silico.ch/

Lazar (Lazy Structure-Activity relationships) takes a chemical structure as input, and finds the most similar chemicals in a database by comparing the composition in substructures. The comparison is performed locally, with an automated and reproducible Read-Across procedure. Rationales for predictions, applicability domain estimations and validation results are presented in a clear graphical interface for the critical examination by toxicological experts. Lazar is built on top of the OpenTox framework.

Lazar provides predictions for Ames test bacterial mutagenicity (together with a range of endpoints outside of genotoxicity).

Toxtreehttps://eurl-ecvam.jrc.ec.europa.eu/laboratories-research/predictive_toxicology/qsar_tools/toxtree

Toxtree, by the European Union Laboratory for Alternatives to Animal Testing (EURL ECVAM), developed by IdeaConsult Ltd., is a flexible and user-friendly open-source application that places chemicals into categories and predicts various kinds of toxic effect by applying decision tree approaches. It includes two modules with Structural Alerts for mutagenicity and carcinogenicity prediction – the ISS Benigni-Bossa rulebase (which expands on the Ashby supermutagen model) for Ames test mutagenicity, and the ToxMic rulebase for the *in vivo* micronucleus assay.

Commercial QSAR models

Derek and Sarah, Lhasa ltd. https://www.lhasalimited.org/products/

Derek (originally, Derek for Windows (DfW)) is a commercial system developed and marketed by Lhasa Ltd. The development of knowledge-based rules in Derek is overseen by collaborative group which consists of representatives from commercial, educational and non-profit organisations. The program applies structure–activity relationship and expert knowledge rules (Structural Alerts, SA) to derive a reasoned conclusion about the potential toxicity of the query chemical. The program provides supporting

evidence for its predictions in the form of comments, literature references, and toxicity data, allowing its predictions to be evaluated. It provides predictions for the different genotoxicity endpoints.

Lhasa has developed a new software, Sarah Nexus, which is statistically-based software aimed at giving fast, automatic predictions for Ames mutagenicity. The structures are fragmented, and these fragments are reviewed for activity versus inactivity. The model then arranges those 'interesting' fragments into a self-organizing network of hypotheses (or nodes), and relevant hypotheses are used to inform an overall prediction of toxicity. The Sarah Nexus prediction includes an overall conclusion about the toxicity in a structure, confidence rating in that prediction, as well as supporting examples.

For bacterial *in vitro* mutagenicity only among the genotoxicity endpoints, Derek Nexus contains expert-derived functionality to provide negative predictions. It has been designed to support post-processing workflows, for example by highlighting structural features for further expert assessment.

HazardExpert http://www.compudrug.com/hazardexpertpro

The HazardExpert models are proprietary, the software now being marketed by CompuDrug Ltd. The program works by searching the query structure for known toxicophores that are derived from the literature in the field of QSAR or from the US EPA and Interagency Testing Committee (ITC) monographs. Predictions are made taking into account the effects of bioavailability and bioaccumulation. Results are given for seven different toxicity classes, including aspecific mutagenicity.

MultiCASE http://www.multicase.com/

In MultiCase (Multiple Computer Automated Structure Evaluation) (MultiCASE Inc., Beachwood, OH, USA) software, complementary QSAR methodologies (statistics based and expert rule based) are built in for the evaluation of the chemical's potential to cause a large range of effects, including bacterial mutagenicity and genotoxicity models.

MultiCASE is a machine-learning tool. It investigates the SMILES structures of organic chemicals and splits them to all possible 2– 10 consecutive (non-hydrogen) atom structural fragments. Subsequently, the fragments of active and inactive molecules are compared in order to identify all the fragments associated with the active compounds. The molecular descriptors, correlating with enhanced or diminished activity of chemicals sharing a common structural alert, are selected. These data are combined and utilized for the development of a quantitative estimate of the potential toxicity of test chemicals.

The applicability domain assessment is based on two methodologies. The first approach compares all of the two- and three-atom fragments of test chemicals with those of all control database chemicals on the basis of the Office of Pharmaceutical Science, Informatics and Computational Safety Analysis expert rules applied in the program. Warnings are provided for all the test chemicals that contain unknown fragments. The second approach describes the certainty of the prediction on the basis of the coverage.

The performance / reliability of the models is automatically calculated by the software with internal validation (leave-many-out (LMO) and leave-one-out (LOO) crossvalidation (CV)) statistics.

MultiCase predicts: Bacterial mutagenicity(Ames test), *in vitro* mammalian cells gene mutation, *in vitro* chromosome aberrations, yeast mutagenicity, Drosophila mutagenicity, UDS, *in vitro* SCE, *in vivo* mutagenicity.

OASIS TIMES, by LMC http://oasis-lmc.org/products/software.aspx

OASIS TIMES is based on the use of Structural Alerts (SA). Each SA is accompanied by modulating factors, to account for the influence of the rest of the molecule, as well as with defined and documented mechanism of interaction with DNA (for the mutagenicity model) and/or nuclear proteins and enzymes (for the chromosomal aberration model). Expert knowledge was used to define the SAs and the mechanistic basis for prediction (interaction with biological macromolecules) is well documented. A pattern recognition approach (COREPA) was used to derive modulating factors for each SA.

A unique characteristics of TIMES is that the system combines tissue metabolism simulator (toxicokinetics) and reactivity models (toxicodynamics) in the same modeling platform. The metabolism simulator is based on a heuristic algorithm to generate metabolic maps from a comprehensive library of biotransformations and abiotic reactions. Reactivity models for binding of chemicals with protein, DNA, lipids etc. predict toxicity of generated metabolites and parent chemicals. Thus, TIMES allows prioritization of chemicals according to toxicity of their metabolites. Toxic metabolites are highlighted in generated metabolic maps and the associated mechanisms of interactions with macromolecules are provided. Using TIMES platform one could predict toxic endpoints without metabolic activation; metabolism only (without predicting toxic outcome) and toxicity as a result of metabolic activation.

The following genotoxicity endpoints are predicted without and with metabolic activation: a) *In vitro* genotoxicity: activation by rat liver S9 metabolism is simulated, b) AMES mutagenicity; c) Chromosomal aberrations. For *in viv*o genotoxicity, activation by rat *in vivo* metabolism is simulated, and the following endpoints are predicted: a) Comet Genotoxicity; b) Liver TGR; c) Liver Clastogenicity; d) Micronucleus Test (MNT) in bone marrow.

TOPKAT http://accelrys.com/products/collaborative-science/biovia-discovery-studio/qsar-admet-and-predictive-toxicology.html

TOPKAT is an *in silico* method developed by Accelrys, Inc. for assessing toxicity predictions of organic compounds. The proprietary TOPKAT models include Ames mutagenicity model developed on large training sets. The system uses 2D Kier and Hall chemical descriptors, electrotopological E-state and other descriptors to generate discriminant model, which provides a discriminant score, in the form of probability of mutagenicity ranging from 0 to 1 (100%).

TOPKAT developed a special system to evaluate the applicability domain, through the so-called Optimum Predictive Space (OPS). A chemical is identified if it is within or outside the OPS or in a borderline position. Predictions are expected to be reliable if the chemical is within the OPS. Moreover, the presence of fragments not represented in the training set is also considered as a condition for an unreliable estimation.

BioEpisteme (Prous Institute for Biomedical Research) (http://www.prousresearch.com/Home.aspx)

BioEpisteme is a modelling system that utilizes a methodology based on a wide range of molecular descriptors and binding profiles.

The tool is organized into two integrated modules, namely a data prediction and a model building module, interoperating and exchanging data. The prediction module provides the probability of a given molecule to display a certain combination of mechanisms of action on the basis of various experimental results. The model building module includes a large set of molecular descriptors (topological descriptors, connectivity indices, partial charge descriptors, surface area, volume and shape descriptors, physicochemical properties), calculated to derive the best representation of the molecule. In the model building module, a modified version of the k-nearest neighbours algorithm is used as a prediction algorithm. In order to develop the model, the system discards the descriptors which are poorly correlated with the endpoint, or which contain little information. The set of relevant descriptors is selected on the basis of a 10-fold crossvalidation. The report summary of the whole modelling procedure as well as the lists of discarded and selected descriptors is provided by the software. BioEpisteme has

a fully automated procedure for calculating the predictive performance of the derived QSAR models according to the LMO and LOO cross-validation statistics implemented.

Bioepisteme is a general methodology that can be applied to develop *ad hoc* models. It includes a model to predict Ames test mutagenicity.

Leadscope Model Applier (Leadscope, Inc.) http://www.leadscope.com/

Leadscope Model Applier is a chemoinformatic platform that provides QSAR for the prediction of potential toxicity and adverse human clinical effects. The Models are constructed at the Food and Drug Administration (FDA) by the Division of Drug Safety Research Staff (DDSR) based on both proprietary and non-proprietary data and are intended to support regulatory decision making processes.

The Leadscope suite includes both statistical QSAR and Expert Alerts models, namely:

Leadscope/Genotoxicity statistical models: Leadscope models are developed with molecular descriptors that include structural features and seven calculated properties, which are molecular weight, LogP, polar surface area, hydrogen bond acceptors, hydrogen bond donors, number of rotational bonds and, Lipinski score (rule violation). The prediction results for each model are provided as the qualitative "prediction call" and the "positive prediction probability". The robustness of the prediction can be evaluated through Model Features Count (i.e., a parameter used to verify that the target compound contains a significant number of features that are present in the prediction model), and the "30% Similarity Training Neighbours" Count (i.e., number of training compounds structurally similar to the target (with at least 30 % similarity)). The QSAR models are based on Partial logistic regression. Models are available for the prediction of mutagenicity based on microbial *in vitro* assays (i.e., Ames); mammalian *in vivo* assays (i.e., *in vivo* mammalian gene mutation and mammalian dominant lethal (DL)), mammalian *in vitro* assays (i.e., combined CHO/CHL hgprt gene mutation tests and mouse lymphoma gene mutation assay); *in vitro* and *in vivo* clastogenicity; *in vitro* sister chromatide exchange (SCE); *in vivo* micronucleus tests in mouse.

Leadscope/Genetox Expert Alert Model. To develop this system, an initial library of mutagenicity structural alerts was identified from the literature. Information on plausible mechanisms was collected alongside the structural definitions. Factors that deactivate the alerts were also identified from the literature and through an analysis of the corresponding data using the Leadscope data mining software. Over 200 distinct alerts are encoded in the system. A confidence score based upon information collected for each alert is provided alongside the positive or negative call. Structural analogues from the alert reference set, along with experimental results and any identified alert, are provided. Leadscope predicts Ames test, *in vivo* mammalian gene mutation and mammalian dominant lethal (DL) tests, mammalian *in vitro* hgprt gene mutation test, mouse lymphoma gene mutation, *in vitro* and *in vivo* clastogenicity, *in vitro* sister chromatide exchange (SCE), *in vivo* micronucleus tests in mouse.

ACD/Percepta (Advanced Chemistry Development, Inc., Pharma Algorithms, Inc.) http://www.acdlabs.com/home/

ACD/Percepta (release 2016) is a suite of comprehensive tools for the prediction of different toxicity endpoints. Predictions are made from chemical structure and based upon large validated databases and QSAR models, in combination with expert knowledge of organic chemistry and toxicology. The majority of ACD/Percepta models were developed using the GALAS modelling methodology (Global, Adjusted Locally According to Similarity), which consists in two parts: 1) a global (baseline) statistical model based on PLS with multiple bootstrapping, using a predefined set of fragmental descriptors; 2) local correction to baseline prediction based on the analysis of model performance for similar compounds from the training set. ACD/Percepta allows to evaluate the robustness of the prediction by examining compounds similar to the target from the training set, together with literature data and reference. Some models also provide an estimation of the reliability of the prediction, by a reliability index (RI). This

index provides values in a range from 0 to 1 and gives an evaluation of whether a submitted compound falls within the Model Applicability Domain. In particular: RI < 0.3 (Not Reliable), RI in range 0.3-0.5 (Borderline Reliability), RI in range 0.5-0.75 (Moderate Reliability), RI ≥ 0.75 (High Reliability). Estimation of the RI takes into account the following two aspects: similarity of the tested compound to the training set and the consistency of experimental values for similar compounds.

It predicts Ames test, Mouse Lymphoma Assay, *in vivo* Micronucleus test, Chromosomal Aberrations, Unscheduled DNA Synthesis.

ChemTunes Studio (ChemTunes Studio, Toxicity Knowledge Base, Altamira LLC and Molecular Networks GmbH) https://www.mn-am.com/products/chemtunes

ChemTunes is a knowledgebase software consisting of experimental *in vitro* and *in vivo* toxicity information, and *in silico* models for a series of human health toxicity endpoints, comprising key genetic toxicity endpoints. The software is made of multiple components, including genotoxic chemotypes (structural alerts), mechanistically-informed (mode-of-action driven) QSAR models, and comparison of the prediction results to nearest neighbours. A quantitative weight-of-evidence (WoE) decision theory approach is used to obtain the final overall assessment and to provide a quantitative estimation of the uncertainty associated with the prediction.

All ChemTunes Studio QSAR models consist of chemical mode-of-action category models as well as a general global model. The computational modelling approach is a hybrid of partial least squares (PLS)/ordinal logistic regression methods. For model building, global molecular and shape descriptors (from CORINA Symphony) and quantum-mechanic parameters are used. Applicability domain analysis reports whether the target compound is out-of-domain.

The reliability of each alert is determined by exploring the ability of the alert to hit positive compounds in a large training set. It predicts Ames test, *in vivo* micronucleus, *in vitro* chromosome aberrations.

B.1. Summary descriptive table of predictive (Q)SAR models

QSAR model name	Endpoint	Type	Algorithm	Applicability Domain	Public or Commercial
Caesar (VEGA)	Salmonella typhimurium (Ames test)	Hybrid	Combination (cascading) of a) statistical data mining; and b) expert knowledge (Structural Alerts) models	Combined index based on: a) chemical similarity; b) concordance for similar molecules; c) accuracy for prediction of similar molecules; d) descriptors range	Public
Lazar	Bacterial Reverse Mutation Test (Ames test)	Statistical	modified k-nearest neighbour classification; MP2D fingerprints	Number of, and similarity with training set compounds	Public

QSAR model name	Endpoint	Type	Algorithm	Applicability Domain	Public or Commercial
T.E.S.T.	Bacterial Reverse Mutation Test (Ames test)	Hybrid (combination of statistical and expert knowledge)	Suite of methods (statistical and expert). The final call is consensus from all approaches	Specific Applicability Domain rules for each method	Public
Toxtree	Bacterial Reverse Mutation Test (Ames test); *in vivo* micronuclei in rodents	Expert (Structural Alerts)	Decision tree	No	Public
Bioepisteme	Bacterial Reverse Mutation Test (Ames test)	Statistical	modified k-nearest neighbour classification	Similarity with training set compounds	Commercial
HazardExpert	Aspecific Genotoxicity	Expert (Structural Alerts)	Decision tree (including ADME consideration)	No	Commercial
Leadscope	Ames test, *in vivo* mammalian gene mutation and mammalian dominant lethal (DL) tests, mammalian *in vitro* hgprt gene mutation test, mouse lymphoma gene mutation, *in vitro* and *in vivo* clastogenicity, *in vitro* sister chromatide exchange (SCE), *in vivo* micronucleus tests in mouse	Statistical and expert	Both statistical and expert knowledge based models	Structural analogues and confidence scores	Commercial
MultiCase	Bacterial mutagenicity(Ames test), *in vitro* mammalian cells gene mutation, *in vitro* chromosome aberrations, yeast mutagenicity, Drosophila mutagenicity, UDS, *in vitro* SCE, *in vivo* mutagenicity	Statistical	Molecular fragments-based machine learning	Similarity with training set chemicals, and fragments coverage	Commercial

QSAR model name	Endpoint	Type	Algorithm	Applicability Domain	Public or Commercial
OASIS Times	Ames test, *in vitro* chromosomal aberrations, *in vivo* micronucleus, *in vivo* liver genotoxicity, *in vivo* liver TGR mutation, *in vivo* liver clastogenicity	Hybrid	Decision tree	Applicability domains consisting of three sub-domain layers: general parametric requirements, structural features and alert(s) reliability	Commercial
TopKat	Ames test mutagenicity	Statistical	Discriminant model	Optimum Predictive Space	Commercial
ChemTunes	Ames test, *in vivo* micronucleus, *in vitro* chromosome aberrations	Hybrid (Statistical and expert)	Combination of PLS and ordinal logistic regression	Similarity with training set chemicals	Commercial
ACD Percepta	Ames test, Mouse Lymphoma Assay, Micronucleus test, Chromosomal Aberrations, Unscheduled DNA Synthesis	Statistical	Statistical model based on PLS with multiple bootstrapping	Chemical and biological similarity with training set compounds	Commercial
DEREK Nexus, Lhasa Ltd.	Ames test, *in vitro* chromosomal aberrations and mammalian cells gene mutation, *in vivo* micronucleus	Expert	Decision tree	No	Commercial
Sarah Nexus, Lhasa Ltd.	Ames test	Statistical	Structural fragments-based pattern recognition	Coverage of substructures	Commercial

Appendix C – Read Across: Literature search strategy

Study question

Given the central role of the assessment of genotoxicity of pesticides and their metabolites for regulatory purposes, an evaluation of applicability of existing (Q)SAR models as well as of read Across approaches for its prediction is particularly important. In particular, the review of the literature on Read Across, the compilation of a list of proposed workflows and software tools, will be beneficial for the work of the risk assessors when applying the guidance for residue definition as well as in other areas of risk assessment of pesticides.

General eligibility criteria for inclusion or exclusion of studies

This review covers the literature for the period from 1 January 2006 to 31 December 2016. The final list of studies includes papers on general principles of Read Across (even though not directly related to genotoxicity assessment), available tools, and genotoxicity cases studies.

Description of the review method and literature search

Search strings were developed by combining pairs of Key Words from two sets of conceptual components; A) Read Across; and B) mutation; mutagenicity; genotoxicity; chromosomal aberrations; DNA damage; and more specific terms (e.g., Ames, Mouse Lymphoma, aneugenicity, etc…) on the other side.

Examples of strings are: "Read Across AND mutation"; "Read Across AND mutagenicity"; "Read Across AND genotoxicity"; "Read Across AND chromosomal aberrations"; "Read Across AND DNA damage"; "Read Across AND Ames"; "Read Across AND Mouse Lymphoma"; "Read Across AND aneugenicity", and so on.

The search strings were applied to the databases Pubmed, Google Schoolar, and several websites (OECD, JRC, ECHA, US EPA, Canada, Danish EPA, ECETOC). The search dates ran from 2006 – 2016.

Iterations of search term combinations were tested to refine the search, until no more new results were found.

The initial electronic search generated results that were reviewed for relevance by the Consortium scientists. This initial screening was performed based on titles and abstracts. Some terms and combination of terms (i.e., mutation and Read Across) pointed mostly to papers were the reference to genotoxicity / mutagenicity was secondary and irrelevant. Search in Google Scholar generally confirmed the Pubmed findings, with some additional papers.

It should be noticed that the number of retrieved papers connected with genotoxicity / mutagenicity Read Across was by far much lower than the thousands of papers connected in some way to genotoxicity / mutagenicity QSAR. For example, on the same period of time the numbers were: a) Read Across / genotoxicity n= 36; b) Read Across / mutagenicity n = 22.

Upon refinement, fifteen of high relevance where accepted.

Appendix D – Software tools available for Read Across

This section focuses on currently available tools that can be used for grouping, including tools (like the OECD QSAR Toolbox, ToxRead, and REACH-across) that are environments for performing the full process of Read Across, and other tools designed for broader applications that can potentially be used to identify similar compounds. Summary of the details of software for Read Across applications are reported in Table D.1.

Software environments for full development of Read Across analyses

OECD QSAR Toolbox

One of the most important tools for grouping and Read Aacross is the OECD QSAR Toolbox. It is in the public domain, and can be freely downloaded. The development of the Toolbox was co-ordinated by the Organisation for Economic Co-operation and Development (OECD) and the work undertaken by the Laboratory of Mathematical Chemistry at the University "Prof. Assen Zlatarov", Bourgas, Bulgaria under the leadership of Professor Ovanes Mekenyan. The software was developed in collaboration with the European Chemicals Agency (ECHA).The Toolbox was designed specifically for the purpose of category formation and Read Across to fill gaps in data needed for safety/hazard assessment of chemicals. It was specifically developed to be used by the chemical industry and other stakeholders to help with regulatory submissions where in silico tools were employed for predicting toxicity. The first version of the Toolbox was launched in March 2008; the most recent version (at time of writing) is version 4.2 which was released in February 2018. The Toolbox is freely downloadable, along with detailed user guides and supporting information (http://www. qsartoolbox.org).

The OECD Toolbox incorporates information and tools from various sources into a workflow. The workflow mimics that described in the OECD grouping guidance (OECD, 2017b), aimed at grouping compounds into rational and chemically/mechanistically justifiable categories. These categories can be built using structural or mechanistic features of chemicals that are relevant to the toxicological endpoint being investigated. Further sub-categorization can then be performed as required using profilers, already defined within the Toolbox, or user-defined profilers to ensure that the members of the category fall within a clearly defined structural domain, representative of the target chemical. Read Across can then performed from those members of the category with known experimental data to those where data are lacking. A full report on the process can be generated in-keeping with regulatory requirements.

The software supports a suggested workflow, which is designed to mimic the manner in which an assessor would make a judgement on a chemical. The workflow consists of six steps: Input, Profiling, Data gathering, Category formation, Data gap filling, Report.

It should be emphasized that the many searching and calculation options in the Toolbox can also be used separately, in no previously defined order, thus leaving to the user the maximum flexibility and freedom in the use of the Toolbox resources.

In the suggested workflow, a target chemical is introduced into the Toolbox using a chemical identifier such as a name, a CAS registry number or by drawing a chemical structure using the inbuilt drawing tool.

The target can then be "profiled", i.e., characterized in terms of different properties. There are several types of 'profilers' in the Toolbox namely: predefined, general mechanistic, endpoint specific, empiric and toxicological. This step is used to identify relevant structural features, or potential mechanisms of action of the chemical. Profilers of interest in genotoxicity / mutagenicity assessments are: Carcinogenicity (genotox and non genotox) alerts by ISS; DNA alerts for AMES, MN and CA by OASIS; in vitro mutagenicity (Ames test) alerts by ISS; in vivo mutagenicity (micronucleus) alerts by ISS; Oncologic primary classification, DNA binding by OECD; DNA binding by OASIS, Protein binding. (see also: *QSAR Toolbox User Manual Strategies for grouping chemicals to fill data gaps to assess genetic toxicity and genotoxic carcinogenicity:*

http://www.oecd.org/chemicalsafety/risk-assessment/46985336.pdf). The profilers Carcinogenicity by ISS, and Oncologic contain primarily alerts for genotoxic carcinogenicity, together with a (lower) number of alerts for non-genotoxic carcinogenicity.

It is also possible to obtain a metabolic profile for chemicals of interest, by identifying observed or simulated metabolites. Once metabolites have been identified for a given chemical the metabolites can then be profiled using any of the profilers.

The next step is used to gather data on the endpoint(s) of interest. Data gathering can be performed on an individual, specific endpoint or on a range of endpoints selected by the user. This step involves gathering available endpoint data from multiple sources that have been provided to the Toolbox. Some datasets are focused on a specific endpoint whereas other sources are more encompassing. Datasets of interest for mutagenicity assessment are: Bacterial mutagenicity ISSSTY; Carcinogenicity Potency Database CPDB; Carcinogenicity & Mutagenicity ISSCAN; Cell transformation assay ISSCTA; Genotoxicity OASIS; Micronucleus ISSMIC; Micronucleus OASIS; Toxicity Japan MHLW, EFSA genotoxicity database.

The category definition step directs a user to select one or more of the profilers to identify source analogues, and sub-categorize the analogues retrieved so that the set of final analogues identified are similar with respect to all the profiling outcomes chosen. Chemicals can be grouped based on structural (e.g., presence of specific functional groups) and/or mechanistic similarity (e.g., presence of alerts related to toxic effects). If the mechanism of action is known then chemicals should be grouped based on descriptors or structures related to that mechanism; chemicals identified as being structurally dissimilar (either by visual inspection by the user or by subsequent refinement of the category using additional sub-categorization criteria within the software) can later be excluded from the category. If the mechanism of action is unknown then chemicals can be grouped according to their structural features. Using profilers based on the presence of specific organic functional groups ensures a defined structural domain for the category.

Once a category has been defined and populated with sufficient structures and their associated (toxicological) data, a Read Across prediction of activity for the unknown chemical can be made. For a Read Across prediction, data that have been gathered on analogues in the category are used to make a prediction of the activity of the target chemical. Read Across is useful where quantitative predictions are required that can be based on a small number of analogues in the category or where a qualitative (or semi-quantitative) prediction is required (active/ inactive/weakly active *etc*). Usually, a majority rule is used to predict the unknown data from those of e.g., five analogues.

The last step in the workflow is documenting the prediction made. Prediction templates that follow a similar structure to the QSAR Prediction reporting format (QPRF) as reported in the ECHA guidance (ECHA, 2017a) can be created which document the logic and steps a user has made in deriving the prediction. Export files in IUCLID can also be generated which is particularly pertinent for Industry users submitting registration dossiers to ECHA.

ToxRead

ToxRead was originally developed by the Istituto Mario Negri, Milan, as a standalone Java tool to help in the assessment of Ames mutagenicity. It can be accessed freely. The research was funded by two EU projects CALIEDOS (http://www.caleidos-life.eu/) and PROSIL (http://www.life-prosil. eu/). The current version of the tool is v0.11 (http://www.tox- read.eu/) and includes modules to make Read Across predictions of both Ames mutagenicity and Bioconcentration factor (BCF).

An end-user inputs a structural identifier in the form of a SMILES string and chooses the number of nearest neighbours (source analogues) and endpoint in order to run a Read Across prediction. By default, three analogues are typically presented. The application relies on the VEGA core library (http:// www.vega-qsar.eu/index.php) that implements a similarity index. The VEGA library also provides other features, such as parsing of the SMILES string, SMARTS matching and molecule depiction.

In the output, a Read Across plot is constructed and the target chemical of interest is represented as the centroid. Surrounded by it, are 3 red circles labelled by CAS numbers. These are source analogues identified on the basis of the VEGA similarity algorithm in the underlying database of ToxRead, which contains 6065 chemicals with experimental data for mutagenicity from several well-known sources . The size of the circles of these source analogue is proportional to their similarity index. The smaller the circle, the less similar the analogue. Double clicking on each of the 3 analogues will reveal the Tanimoto similarity index of the source analogue relative to the target, its identity by CAS and SMILES string, as well and its experimental mutagenicity outcome and any other experimental information that might be available e.g. Log Kow, BCF.

ToxRead includes four main libraries for mutagenicity comprising some 759 rules in total. The structural alerts are taken from the SARpy rulebase developed within the VEGA program, the Benigni-Bossa rulebase that is implemented in both the OECD Toolbox and Toxtree platforms as well as 281 alerts that were manually extracted by experts and rules empirically extracted.

Clicking on the output reveals a large array of information for the user to be inspected. QSAR model predictions from ISS, Caesar, SARPy and KNN models are also provided with reliability scores and a consensus score of all 4 models. An overall assessment for the target chemical is also provided that integrates both the QSAR and Read Across predictions. In this way, the automatically derived Read Across prediction can be controlled and put in a wider perspective by the user.

REACH-across

REACH-across is a commercial, web-based tool, which aims to support and automate structure-based Read Across (https://www.ulreachacross.com/index.html) (Hartung, 2016).

The basic idea is to develop a tool to guide ReadAcross directly responding to the requirements of the European Chemicals Agency (ECHA) RAAF scheme (ECHA, 2017b). It exploits the existence of the large public database of ECHA with all the animal test data currently submitted to the agency for REACH registrations. Due to the requirements of the REACH Directive, about 800,000 studies on 10,000 chemicals were already been registered in the ECHA database by December 2014.

Fundamentally, the REACH-across™ tool calculates a similarity measure (usually, Tanimoto) of the input chemical with all the chemicals in the database, and predicts the toxicity endpoint of interest based on the average toxicity of the closest analogues. At present, 8 toxicity endpoints can be predicted, including mutagenicity. A Tanimoto similarity of 0.7 is often taken as a cut-off for Read Across Alternative similarity measures (other than the Tanimoto index) and prediction models as well as filters (e.g., considering only certain Klimisch quality scores) are foreseen. The tool will benefit from the inclusion of additional databases and an optional inclusion of proprietary data by the user.

CBRA

Chemical Biological ReadAcross (CBRA) was a term coined by (Low et al., 2013) in a research aimed at extending the chemical similarity principle to predict toxicity by incorporating biological activity data in an effort to account for biological similarity. This hybrid approach of using both chemical and biological activity data was expected to be more predictive of *in vivo* toxicity. The actual toxicity prediction was a similarity weighted average of the activities of nearest neighbours visualised as a radial plot. A software implementation of the approach was developed and is freely available from http://www.fourches-laboratory.com/software.

Users need to introduce three different input files: 1) a file of chemical structural descriptors, 2) biological activity information which are structured as descriptors (e.g., gene expression data); and 3) a file of toxicity information – namely the activity to be predicted. In the presentation of the software in Low et al.(Liew et al., 2012), the chemical descriptors used are from Dragon (v.5.5, Talete SRL, Milan, Italy). Based on the above information, similarity between chemicals are computed. Options are available to modify the number of neighbours or the Tanimoto similarity threshold. The default values

are 5 and 0.4 respectively. A radial plot with predictions is automatically generated. The radial plot is structured to reflect 2 sets of neighbours – those on the basis of biological similarity and those on the basis of chemical similarity. The neighbours identified are not necessarily the same but may overlap. Chemical neighbours are reflected on the right hand side of the plot whereas biological neighbours are shown on the left hand side of the plot. The toxicity activity and the prediction is reflected by colour – red for active and green for inactive. The target chemical is represented as the centroid in the radial plot and its colour will indicate its Read Across prediction outcome.

CIIPro

CIIPro is a cheminformatics web portal freely available at http://ciipro.rutgers.edu/. It is intended to facilitate Read Across predictions of a target on the basis of chemical and/or biological similarity. The prediction result can be visualised by a similarity chart along with associated similarity and confidence values. The novelty of the approach lies in taking advantage of the wide array of bioassay data publicly available from PubChem (https://pubchem.ncbi.nlm.nih.gov/).

The starting point in using the portal is to upload a training and test set of chemicals on the basis of their chemical identifiers (PubChem compound identifier, CID). The training set of chemicals is used to create a biological profile using the CIIProfiler tab. This is created by extracting relevant bioassay data from Pubchem. The biological profile derived is represented as a heatmap: the density of the colour will dictate whether a response is active (dark blue = 1), inconclusive (grey = 0) or inactive (light blue = -1). CIIP Predictor is then used to calculate a Weighted Estimated Bio- logical Similarity (WEBS) between the chemicals in the test set and the chemicals in the training set. The WEBS tool calculates two values for each chemical pair, the biological similarity (from 0 to 1) and its confidence score. The confidence score is an estimate of the reliability of the calculated biological similarity, the higher the score, the more reliable the biological similarity value. The confidence score represents the number of assays that have results for both chemicals in a given pair but gives less weight to the assays that only have inactive results for both chemicals. Biological nearest neighbours are then calculated by the WEBS tool by setting suitable parameter cutoffs for both the biological similarity and the confidence scores. The biological similarity cutoff is the minimum biological similarity score for a chemical to be considered as a nearest neighbour to the target chemical. The confidence score cutoff is the percentage of assays in the biological profile that both chemicals need to have responses in for a biological similarity calculation to be meaningful. The number of biological nearest neighbours (from 1 to 5) to be used for predictions is also selected by the end-user. The activities of each test chemical's biological nearest neighbours' are averaged together to predict the target chemical's activity. Biological nearest neighbours are presented on the right hand side of the plot whereas chemical nearest neighbours are on the left of the target chemicals' predicted activity.

Software tools supporting Read Across

The following are systems that can support the Read Across process, mainly for the process of finding analogues.

AMBIT

AMBIT, was developed with funding from industry via a European Chemical Industry Council Long Range Initiative (CEFIC-LRI) (http://cefic-lri.org/toolbox/ambit/) and is freely downloadable via the Sourceforge website (http://ambit.sourceforge.net/). The software was developed originally by Procter and Gamble and IdeaConsult; recently it has been rewamped by Clariant Corporate Product Stewardship – Global Toxicology & Ecotoxicology in collaboration with Idea Consult. It comprises a relational database of compounds (including over 450,000 chemical structures and their identifiers), associated properties, QSAR models, references and tables of data including pre-calculated fingerprints (that allow substructure and similarity searching to be performed more rapidly). Similarity searching of molecules based on Tanimoto coefficient values can be performed and substructure searches for similar molecules are also possible.

The use of a workflow utilising AMBIT to identify analogues within a dataset can be summarized as follows:

a) The starting set of structures are defined using CAS/EINECS number, names, SMILES, MOL or SDF files or using a structure editor.

b) An analogue search is performed (by default hashed fingerprints are compared by Tanimoto distance)

c) The results are displayed and the user can decide to restrict further queries within the set of selected structures

Substructure search can be performed by user-defined fragments. Results can be further filtered by compound profiles (e.g. by experimental or calculated data e.g. octanol:water partition coefficient), Selected structures can be grouped into typical chemical classes or clustered to identify groups of analogues.

The Read Across prediction and justification are user derived, based on subjective expert judgement.

The Read Across predictions are recorded by launching the working matrix and adding new records within the relevant endpoint cells. Information on the Read Across approach, the rationale, the source analogue used for the prediction, the toxicity value being used as the Read Across value can be annotated by the end user as a record. Any outlier data can be deleted within the working matrix or missing records not reflected in the database can be added in the relevant endpoint cell. The last step is to finalize the matrix by saving any edits made. This will enable creation of an assessment report which mimics the category and analogue reporting formats that are described in the ECHA and OECD grouping guidance documents.

ToxMatch

The Toxmatch program was commissioned by the European Commission's Joint Research Centre (JRC) and developed by Ideaconsult, Sofia. Version 1.07 was released in 2009 and can be freely downloaded, along with the User Manual (from which much of the information below was obtained) and other supporting information, from the JRC website (http://ihcp.jrc.ec.europa.eu/our_labs/computational_toxicology/qsar_tools/toxmatch). Toxmatch is a flexible, open-source software program that can be used to group chemicals together and predict activity or classify new chemicals into an appropriate group based on similarity measures.

Toxmatch can be used to generate a small number of descriptors but a user can also upload descriptors, obtained using alternative software, into the Toxmatch environment. The software uses knowledge of the endpoint data (i.e. a supervised training technique) to classify new chemicals and predict activity.

Some training sets of toxicity are pre-loaded (including carcinogenicity / mutagenicity), but can also be uploaded by the user.

Toxmatch can be used to quantify the level of similarity between two chemicals based on similarity in descriptor space (Euclidean distance, Hodgkin-Richards index, Tanimoto index or cosine-like (Carbo) index) or structural similarity (Tanimoto index, Hellinger distance, or Maximum Common Substructure similarity (MCSS)). Once the similarity index between two compounds is established, a similarity index can be determined between an individual chemical and a set of chemicals. This can be performed between a representative chemical within the dataset and the query chemical. This approach uses the Tanimoto distance (Fingerprints, kNN) method or the Hellinger distance (atom environments, summary atom environment method) method. Another approach is to determine the average similarity between the query chemical and its k nearest neighbours (by default 10 nearest neighbours are selected, but the user can specify an alternative number). Toxmatch can use this similarity information to predict (read-across) the activity of a query chemical, based on activity of other chemicals in the category. The category comprises a number of nearest neighbours (as defined by the user) and activity is based on a weighted average of the activity of these nearest neighbours (where the most similar chemical has the

highest weighting). Alternatively, Toxmatch can be used as a tool to group chemicals together based on their calculated degree of similarity. A query chemical can then be assigned to one of the groups based on its similarity to other members of the group i.e. it will be assigned to the group that contains most of its k nearest neighbours.

AIM

The AIM software was developed by the US Environmental Protection Agency as part of the Sustainable Futures Initiative and is freely downloadable from the EPA website (http://www.epa.gov/opptintr/sf/tools/aim.htm). The tool was designed such that analogues of a chemical of interest could be identified and the software would indicate the publically available sources from where toxicity data for the analogues could be obtained (the data are not contained within the software itself). Structure searching is performed using CAS registry numbers, SMILES strings or (sub)structure searching. Analogues are searched using over 700 structural features (atoms, groups and super fragments) as characteristics and matched against an inventory of source analogues with available experimental data (in total the inventory comprises over 86,000 analogues pre-indexed with publicly experimental data and links to data sources). The software provides hyperlinks to the experimental data sources available but does not actually provide the underlying data themselves. AIM uses a two-tiered system for identifying analogues. The default approach is for analogues to be selected if all fragments/atoms and super fragments in the target chemical are contained in the source analogues proposed. This type of query assumes a one to one match and is the most stringent means of identifying analogues. If no analogues are identified that satisfy these criteria, a second tier is performed. Many of the large super fragments specify orientation of atoms and these types of criteria are relaxed in the subsequent search.

The toxicity data for these analogues can then be used for read-across. The end-user must apply subjective judgement to determine the validity of any of the suggested analogues for the decision context of interest.

In the context of AIM, the Chemical Assessment Clustering Engine (ChemACE) developed by US EPA should be mentioned as well: https://www.epa.gov/tsca-screening-tools/chemical-assessment-clustering-engine-chemace . ChemACE is designed cluster chemicals that are analogs of one another using the same fragment generation system found in AIM, but uses a more complex method for identifying analogs for the clustering exercises that is based on advanced queries in multidimensional space and by allowing users to design rules to modify the approach. The software is freely downloadable from the site above.

D.1. Summary descriptive table of software for Read Across applications

Software name	Read Across specific	Analogue search method	Other functions	Genotox database	Public or Commercial
OECD QSAR Toolbox	Yes	Multifunctional, flexible	Databases interrogation, calculation of phys chem paramters, Structural Alerts, QSARs	Yes	Public
ToxRead	Yes	Chemical similarity		Yes	Public
REACH-across	Yes	Chemical similarity		Yes (REACH)	Commercial
AMBIT	No	Chemical similarity, substructures			Public
ToxMatch	No	Similarity based on different descriptors	Descriptors calculation		Public
AIM	No	Substructure searching	Links to biological databases		Public
ChemACE	No	Substructure searching	Clustering using AIM descriptors		Public
CBRA	Yes	Similarity of biological / chemical patterns			Commercial
CIIPro	Yes	Similarity of biological / chemical patterns			Public

Appendix E – Chemical space analysis

The analysis was performed to evaluate the representativeness of the EFSA genotoxicity dataset in respect to a broader "pesticides" chemical space, and to assess the chemical space described by the pesticide active compounds and their metabolites.

In the chemical space analysis, a selection of physical chemical properties was predicted for the selected compounds extracted from the EFSA DB active substances 349 (DS1) and 607 metabolites (DS2) and for the reference pesticides dataset 1667 (DS3). The characterization of the chemical space of the datasets was performed with the aim of assessing structural similarities and dissimilarities in the datasets. The following descriptors were used for the analysis and were calculated with the (Q)SAR tool Percepta from ACD Labs (2016): HDonors, HAcceptors, FRB, MW, PSA, LogP_Classic, LogD_Classic_7.40, LogP_Galas, LogP_RI_Galas, LogD_Galas_7.40, and the following with a node in Knime (molecular properties): Atomic Polarizabilities, Aromatic Atoms Count, Aromatic Bonds Count, Element Count, Bond Polarizabilities, Bond Count, Eccentric Connectivity Index, Fragment Complexity, VABC Volume Descriptor, Largest Chain, Largest Pi Chain, Petitjean Number, Vertex adjacency, information magnitude, Zagreb Index, Formal Charge, Formal Charge (pos), Formal Charge (neg), Heavy Atoms Count, Molar Mass, SP3 Character.

A few basic descriptors were selected and visualised to ease the comparison of the datasets and are presented in **Error! Reference source not found.**

The DS1, DS2 and DS3 datasets were investigated also by means of Principal Component Analysis (PCA). PCA gives an overview of the information in data tables. The results of the PCA are presented for DS1 and DS2 in score plots for phisyco-chemical descriptors and also as the Hotelling T^2 for phisyco-chemical descriptors. To analyse the physico-chemical dataset by means of PCA, variables were log-transformed, in order to approximately conform to normality. Then the data were UV-centered in order to compare observations with descriptors having different measurement scales. A PCA model with 4 principal components was built to represent approximately the 86% of the total variability (R^2X=0.858). The score plot for the first two dimensions is depicted in Figure E.8a, and for the second and the third dimension in Figure E.8b thus enabling the visualisation of the chemical space described by some of the components of the model. Figure E.8c displays the Hotelling's T^2, as a summary plot for the observations (substances). The T^2 Range plot shows the distance from the origin in the model plane (score space) for each selected observation. The plot shows the T^2 calculated for the range of selected components (in this case all the 4 components of the model). Observations (substances) over the dotted line (99% confidence) were considered as outliers.

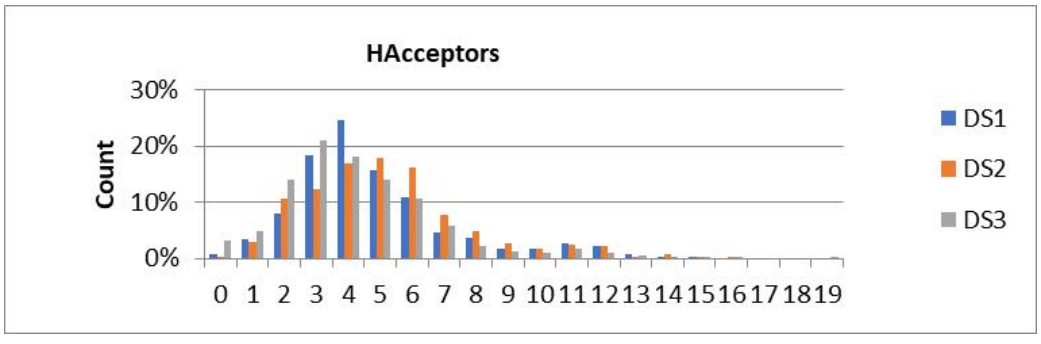


Figure E.1: Physico-chemical properties of pesticides in the three datasets.

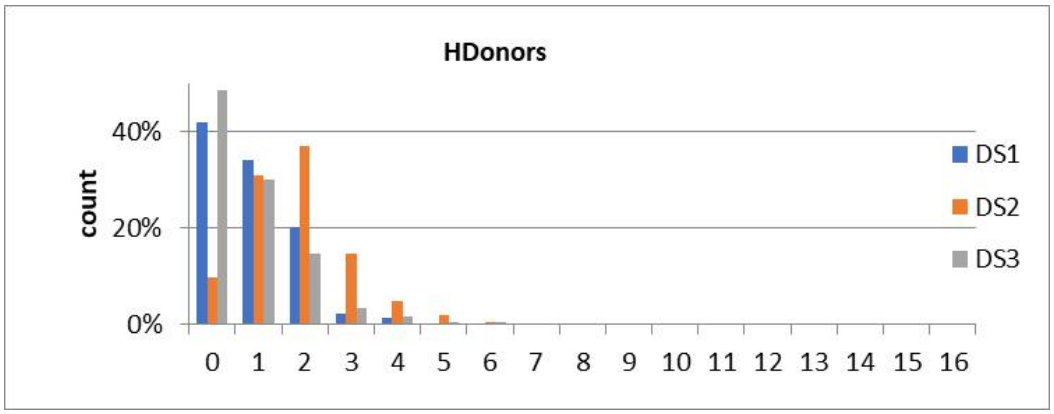


Figure E.2. Physico-chemical properties of pesticides in the three datasets.

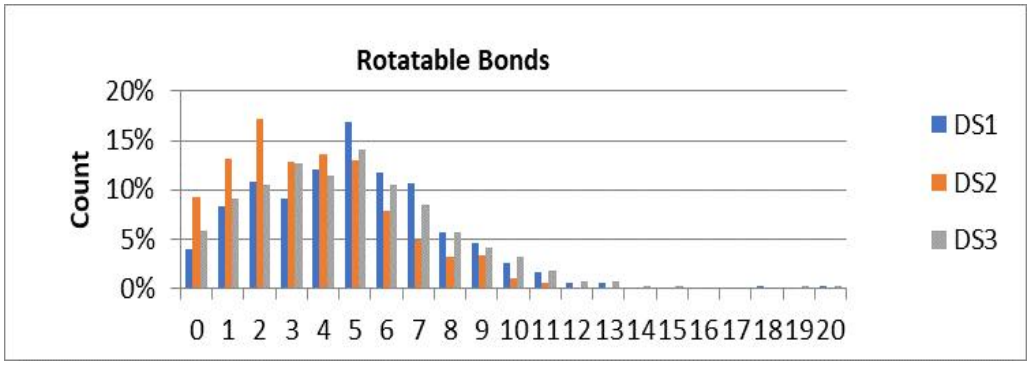


Figure E.3. Physico-chemical properties of pesticides in the three datasets

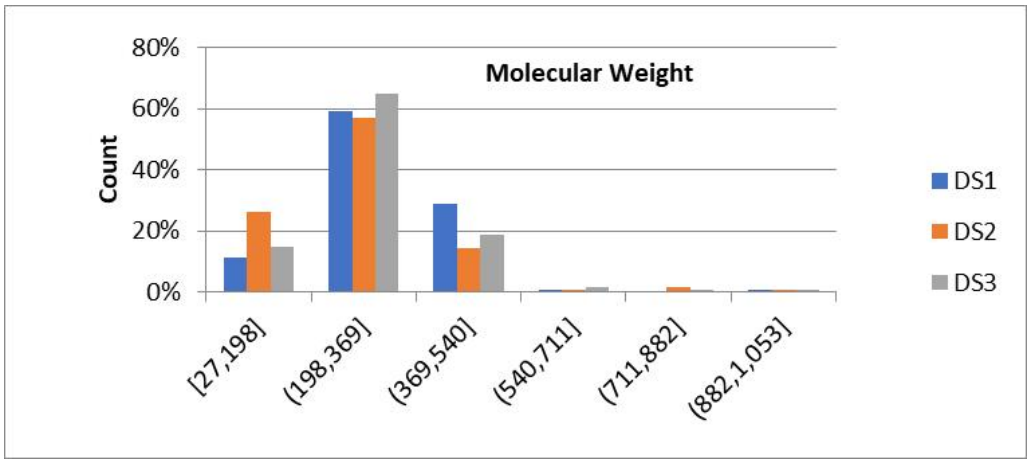


Figure E.4. Physico-chemical properties of pesticides in the three datasets

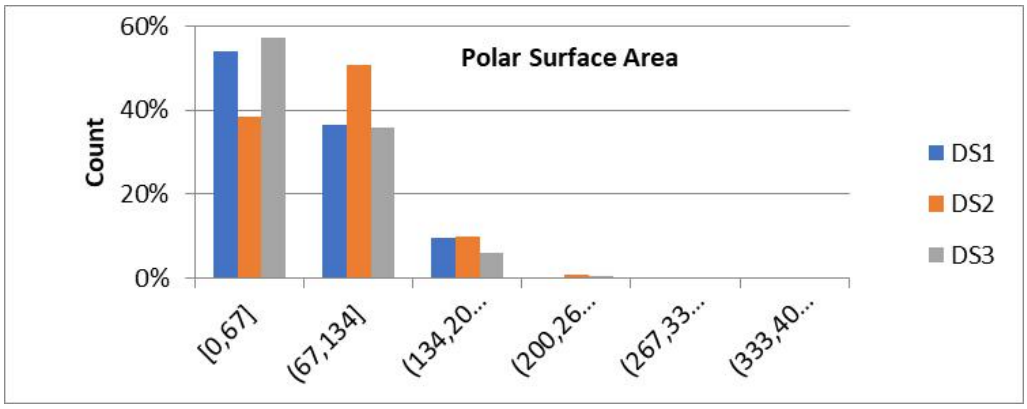


Figure E.5. Physico-chemical properties of pesticides in the three datasets

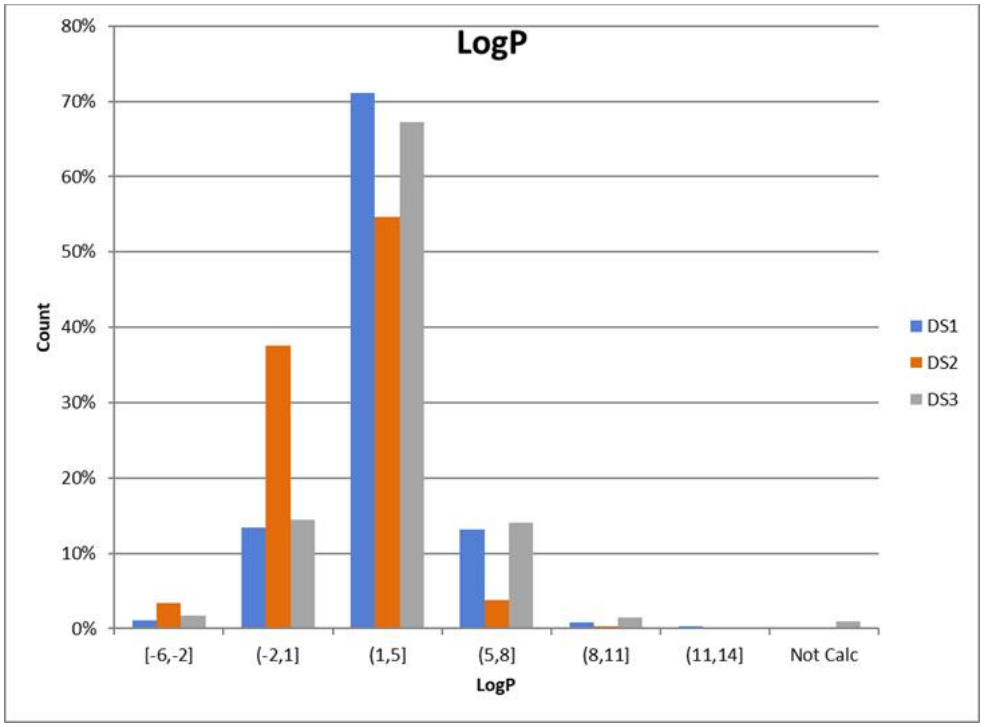


Figure E.6. Physico-chemical properties of pesticides in the three datasets

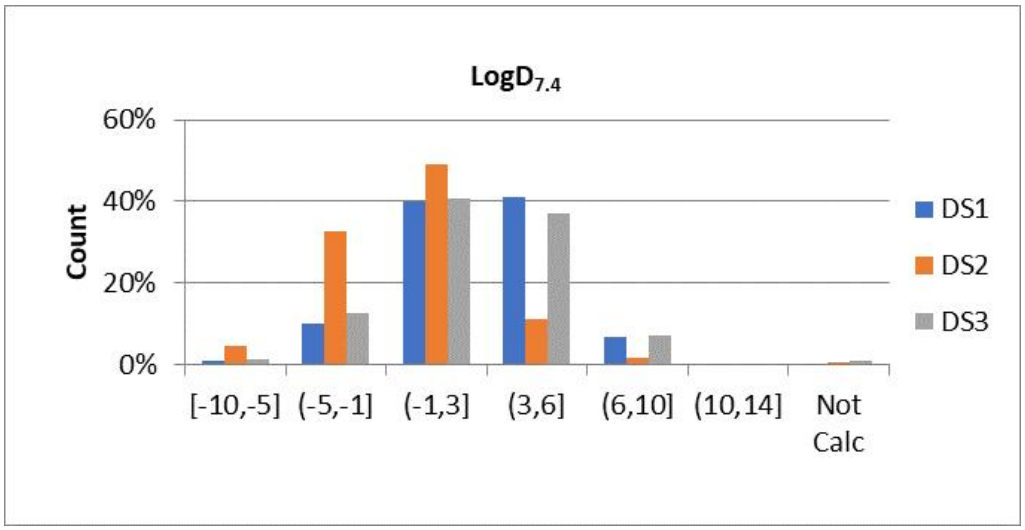


Figure E.7. Physico-chemical properties of pesticides in the three datasets

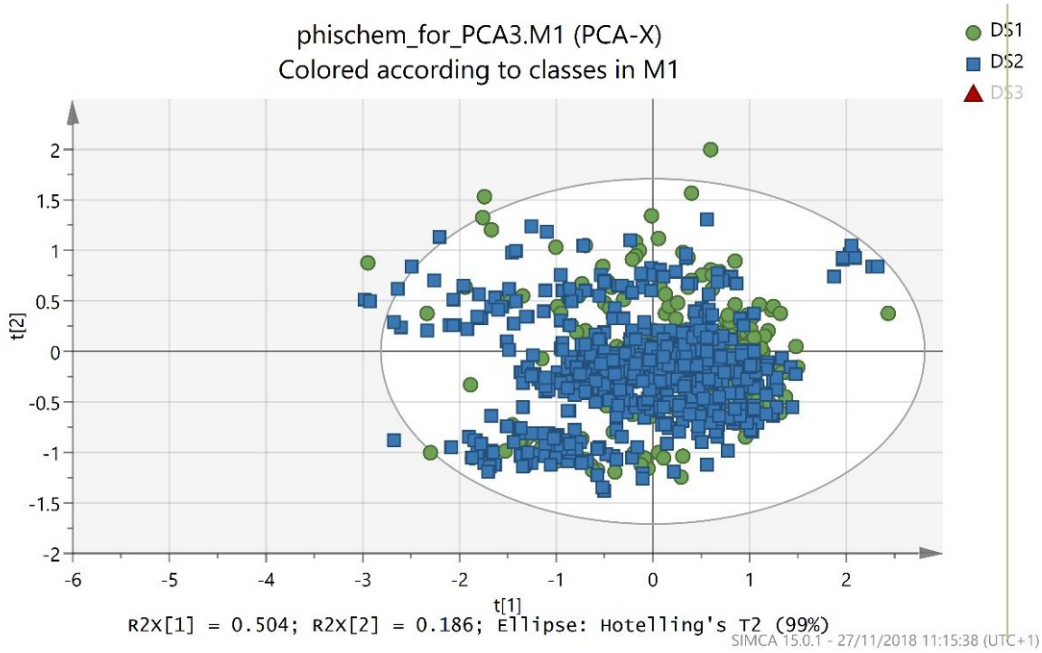


Figure E.8a. Chemical space of pesticides in the two groups of Physico-chemical dataset for the first two PC (observations of DS3 have been hidden for a better visualization of DS1 and DS2 ones).

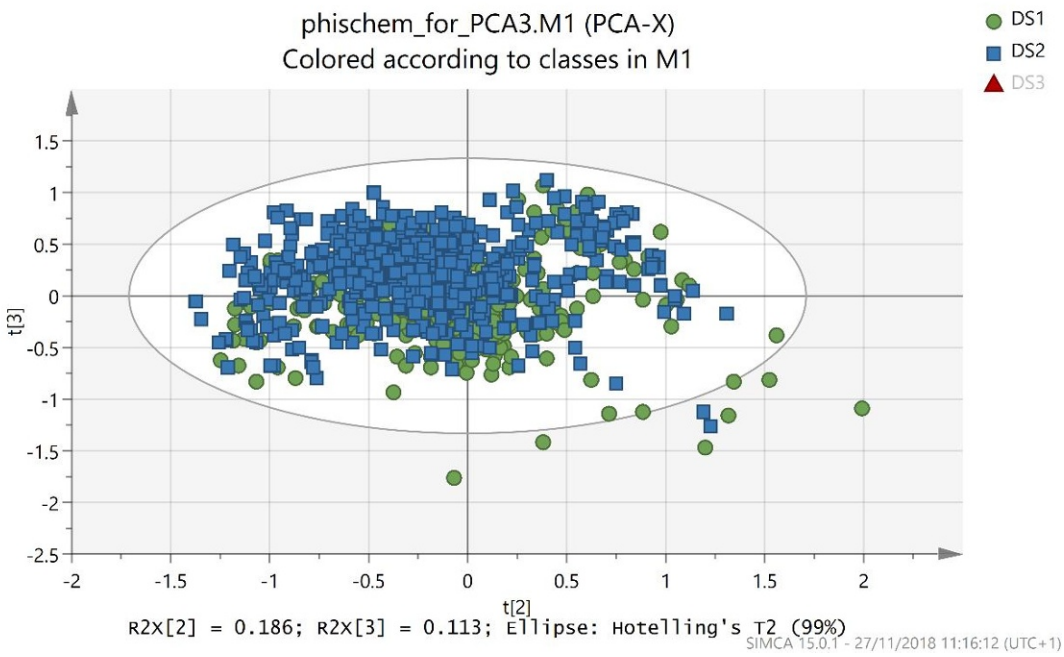


Figure E.8b. Chemical space of pesticides in the two datasets of Physico-chemical dataset for the second and the third PC (observations of DS3 have been hidden for a better visualization of DS1 and DS2 ones).

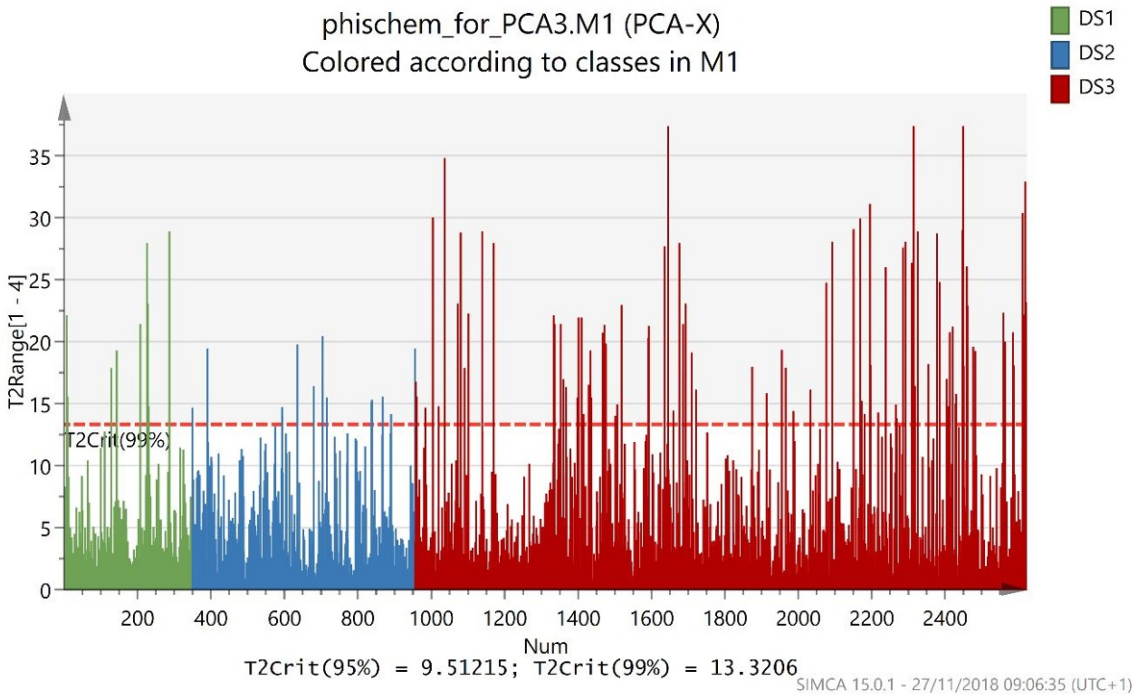


Figure E.8c. Scores from the first to the last component summarized by Hotelling's T^2 for the Physico-chemical dataset. Observations over T^2Crit are considered as outliers.

The Principal Component Analysis (PCA) model, summarised in Hotelling's T2 plot showed the presence of a rather homogeneous group of structures, with few structures laying above the 99% confidence interval. Those were further identified and refer mainly to components that are metabolites of the substance e.g. COM_ID 15180 - [8,9-Z]-isomer of Avermectin B1a and COM_ID 5924 - N-formyl-175-L. A full list of identified outliers is provided in Table E.1.

Table E.1: Structures from DS1 and DS2 identified as outliers in the PCA analysis of the Physico-chemical descriptors

COM_ID	COM_NAME	DATSET
1469	1,4-Dimethylnaphthalene	DS1
1471	2,4,6,8-Tetramethyl-1,3,5,7-tetraoxacyclooctane	DS1
1601	Fenbutatin oxide	DS1
1623	Didecyldimethylammonium chloride	DS1
1811	1-Methylcyclopropene	DS1
1833	Mepiquat chloride	DS1
1836	Dodemorph acetate	DS1
1838	Chlormequat chloride	DS1
50389	Lindane	DS1
15959; 15960; 15961	Azadirachtin extract (Mitsui AgriScience International S.A/B.V source)	DS2
15180	[8,9-Z]-isomer of avermectin B1a	DS2
16037	Endosulfan-ether	DS2
16038	Endosulfan-hydroxyether	DS2
2004	Avermectin B1a	DS2
3711	d-limonene	DS2
50139	4-hydroxy-1,1-dimethylpiperidinium chloride	DS2
50174	3-hydroxy-oxetane	DS2
5922	N-demethyl-175-J	DS2
5923	N-formyl-175-J	DS2
5924	N-formyl-175-L	DS2
75155	XDE-105 factor B	DS2
75156	XDE-105 factor K	DS2
75310	(5s,8s)-3-(2,5-dimethylphenyl)-3,4,8-trihydroxy-1-azaspiro[4.5]decan-2-one	DS2
75611	Emamectin benzoate	DS2

Appendix F – Read Across exercise, Strategy I. Molecules selected

Molecules investigated in Strategy I

In the following tables (F.1-F.14), details on molecules selected in the Read Across strategy I exercise are reported.

Table F.1: RA1, SUB_ID 1170 (AS: Imidacloprid)

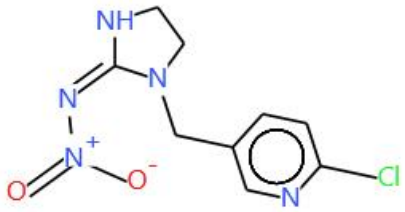
RA	Source	Target
1		
SMILES	[O-][N+](=O)N=C1NCCN1Cc1ccc(Cl)nc1	Clc1ccc(CN2CCNC2=NN=O)cn1
SUB_ID	1170	
COM_ID	1653	50021
Name	Imidacloprid	1-[(6-chloro-3-pyridinyl)methyl]-,oxohydrazone
Ames test results	0	0
In vitro CHA test results	1	0

Table F.2: RA 2, SUB_ID 1347 (AS: Terbuthylazine)

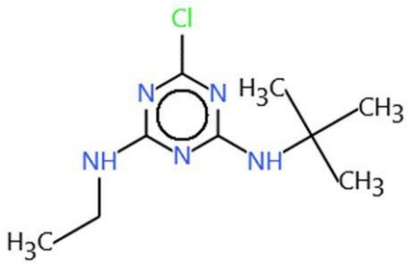
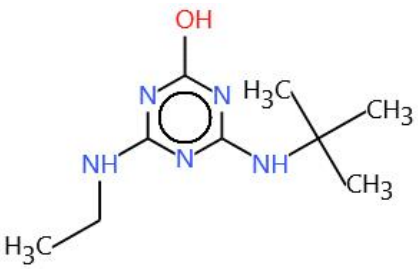
RA	Source	Target
2		
SMILES	CCNc1nc(Cl)nc(NC(C)(C)C)n1	CCNc1nc(O)nc(NC(C)(C)C)n1
Name	Terbuthylazine	Hydroxy-terbuthylazine
Ames test results	0	0
In vitro CHA test results	0	0
SUB_ID	1347	
COM_ID	1698	1895

Table F.3: RA 3, SUB_ID 1133 (AS: Dimethoate)

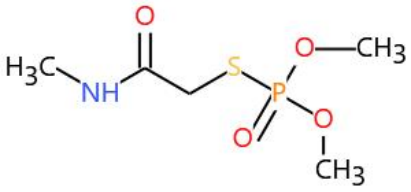
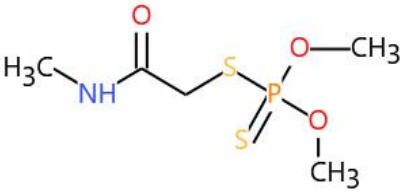
RA	Source	Target
3		
SMILES	CNC(=O)CSP(=O)(OC)OC	CNC(=O)CSP(=S)(OC)OC
SUB_ID	1133	
COM_ID	1624	6028
Name	Dimethoate	Omethoate
Ames test results	1	1
In vitro CHA test results	NA	NA

Table F.4: RA 4, SUB_ID 35058 (AS: confidential)

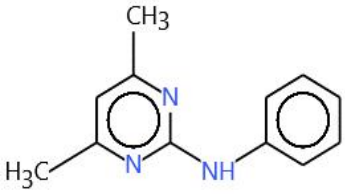
RA	Source	Target
4		
SMILES	confidential	Cc1cc(C)nc(Nc2ccccc2)n1
SUB_ID	35058	
COM_ID	50616	1689
Name	confidential	Pyrimethanil
Ames test results	0	0
In vitro CHA test results	1	0

Table F.5: RA 5, SUB_ID 1166 (AS: Napropamide)

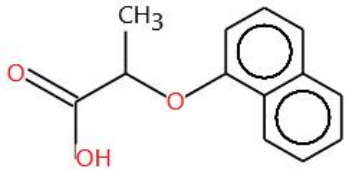
RA	Source	Target
5		
SMILES	CC(Oc1cccc2ccccc12)C(O)=O	CCN(CC)C(=O)C(C)Oc1cccc2ccccc12
SUB_ID	1166	
COM_ID	1668	50576
Name	Napropamide	2-(naphthalen-1-yloxy)propanoic acid
Ames test results	0	0
In vitro CHA test results	0	1

Table F.6: RA 6, SUB_ID 1347 (AS: Terbuthylazine)

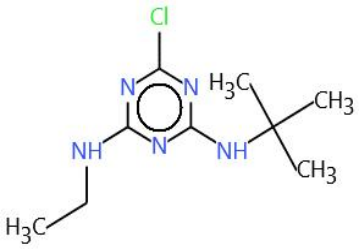
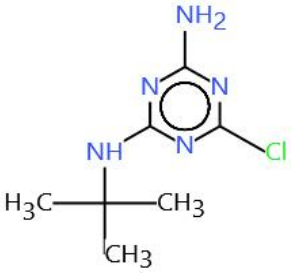
RA	Source	Target
6		
SMILES	CCNc1nc(Cl)nc(NC(C)(C)C)n1	CC(C)(C)Nc1nc(N)nc(Cl)n1
SUB_ID	1347	
COM_ID	1698	1897
Name	Terbuthylazine	Desethyl-terbuthylazine
Ames test results	0	0
In vitro CHA test results	0	0

Table F.7: RA 7, SUB_ID 15043 (AS: Carfentrazone-ethyl)

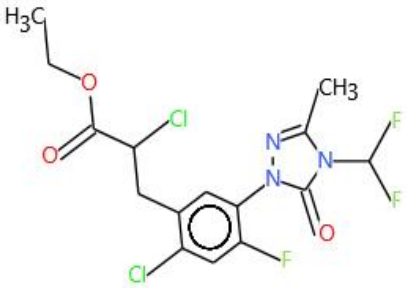
RA	Source	Target
7		
SMILES	CCOC(=O)C(Cl)Cc1cc(N2N=C(C)N(C(F)F)C2=O)c(F)cc1Cl	CC1=NN(C(=O)N1C(F)F)c1cc(C=CC(O)=O)c(Cl)cc1F
SUB_ID	15043	
COM_ID	15674	15678
Name	Carfentrazone-ethyl	(2E)-3-{2-Chloro-5-[4-(difluoromethyl)-3-methyl-5-oxo-4,5-dihydro-1H-1,2,4-triazol-1-yl]-4-fluorophenyl}acrylic acid
Ames test results	0	0
In vitro CHA test results	0	0

Table F.8: RA 8, SUB_ID 85027 (AS: confidential)

RA	Source	Target
8	confidential	confidential
SMILES	confidential	confidential
SUB_ID	85027	
COM_ID	75507	75509
Name	confidential	confidential
Ames test results	0	0
In vitro CHA test results	0	1

Table F.9: RA 9, SUB_ID 85027 (AS: confidential)

RA	Source	Target
9	confidential	confidential
SMILES	confidential	confidential
SUB_ID	85027	
COM_ID	75507	75511
Name	confidential	confidential
Ames test results	0	0
In vitro CHA test results	0	1

Table F.10: RA 10 SUB_ID 3842 (AS: Lambda-cyhalothrin)

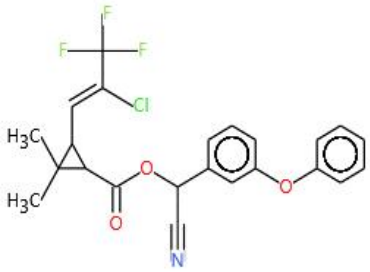
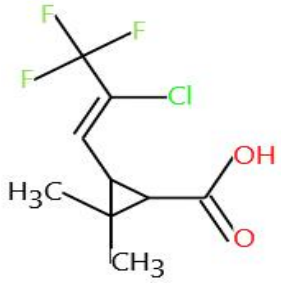
RA	Source	Target
10		
SMILES	CC1(C)C(C=C(Cl)C(F)(F)F)C1C(=O)OC(C#N)c1cccc(Oc2ccccc2)c1	CC1(C)C(C=C(Cl)C(F)(F)F)C1C(O)=O
SUB_ID	3842	
COM_ID	6185	15493
Name	Lambda-cyhalothrin	(1R,3R)-3-[(1Z)-2-chloro-3,3,3-trifluoro-1-propen-1-yl]-2,2-dimethylcyclopropanecarboxylic acid
Ames test results	0	0
In vitro CHA test results	0	1

Table F.11: RA 11, SUB_ID 35031 (AS: Mesotrione)

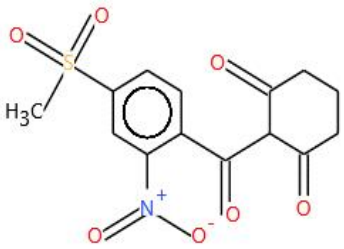
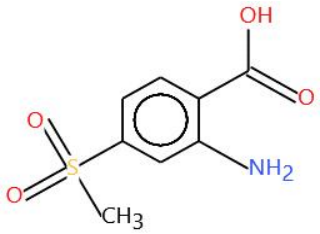
RA	Source	Target
11		
SMILES	CS(=O)(=O)c1ccc(C(=O)C2C(=O)CCCC2=O)c(c1)[N+]([O-])=O	CS(=O)(=O)c1ccc(C(O)=O)c(N)c1
SUB_ID	35031	
COM_ID	50309	50554
Name	Mesotrione	2-amino-4-methylsulfonyl benzoic acid
Ames test results	1	0
In vitro CHA test results	NA	1

Table F.12: RA 12, SUB_ID 1416 (AS: Cyflumetofen)

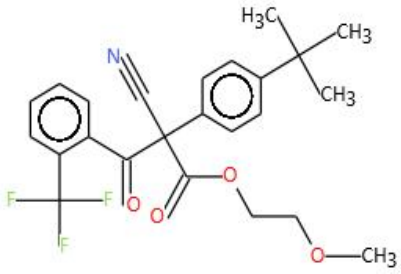
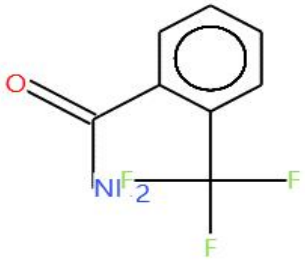
RA	Source	Target
12		
SMILES	COCCOC(=O)C(C#N)(C(=O)c1ccccc1C(F)(F)F)c1ccc(cc1)C(C)(C)C	NC(=O)c1ccccc1C(F)(F)F
SUB_ID	1416	
COM_ID	1509	2098
Name	Cyflumetofen	2-(Trifluoromethyl) benzamide
Ames test results	0	1
In vitro CHA test results	0	0

Table F.13: RA 13, SUB_ID 1166 (AS: Napropamide)

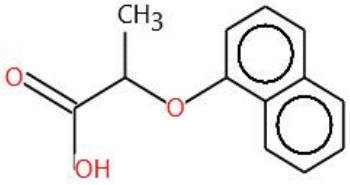
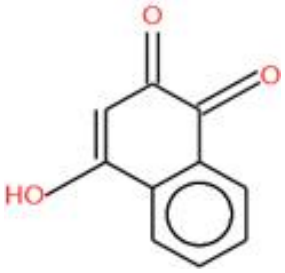
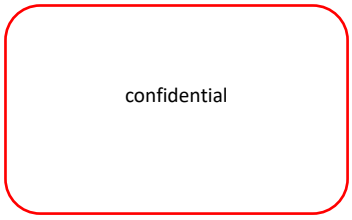
RA	Source	Target
13		
SMILES	CC(Oc1cccc2ccccc12)C(O)=O	OC1=CC(=O)C(=O)c2ccccc12
SUB_ID	1166	
COM_ID	1668	15576
Name	Napropamide	2-hydroxynaphthalene-1,4-dione
Ames test results	0	0
In vitro CHA test results	0	1

Table F.14: RA 14, SUB_ID 85018 (AS: confidential)

RA	Source	Target
14		
SMILES	confidential	confidential
SUB_ID	85018	
COM_ID	75368	75367
Name	confidential	confidential
Ames test results	1	0
In vitro CHA test results	0	0

Appendix G – Read Across exercise: Strategy II. Materials and Methods

Biological Measures of Similarity

Pesticidal Mode of Action. Biological pathways, targets, or primary sites of action associated with the active substances of this database were first considered to address the biological similarity. The MOA classifications are represented with chemical classes for herbicides (AG Canada 2018)[1], insecticides (IRAC 2018)[2], and fungicides (FRAC 2018)[3]. These chemical groups were identified using the public software ChemoTyper (ChemoTyper software. https://chemotyper.org) with the public ToxPrint chemotypes (ToxPrint chemotypes https://toxprint.org/) (Yang et al., 2015).

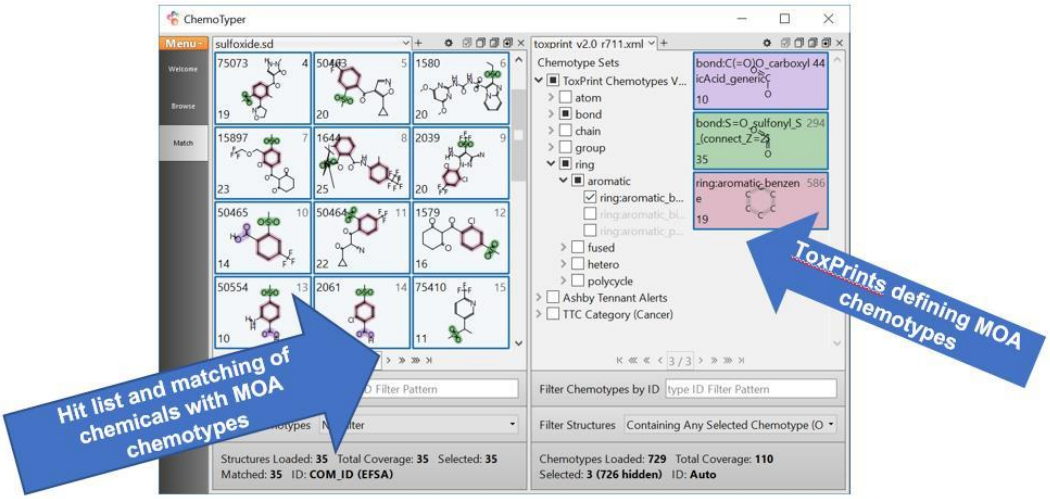


Figure G.1: Identification of MOA chemotypes by ToxPrints

Figure G.1 shows an example of chemicals in the pesticide database being grouped by the pesticidal MOAs within the ChemoTyper using ToxPrints. For each pesticidal MOA, one or more ToxPrint chemotypes are identified that can then be used to programmatically assign a given molecule into a pesticidal MOA group. While the ToxPrint chemotypes are similar to traditional structure scaffolds, chemotype representation offer the potential for more complex types of features. Chemotypes can encode features that cannot be captured using conventional substructure representations (e.g., SMARTS). For example, a chemotype representation can include not only molecular graph information (atoms and connectivity), but also specifications based on atom-and molecular-based properties. We can, for example, create a chemotype that defines a particular fragment in which the partial charge on one or more atoms is specified to fall within a particular range, and the partition coefficient logP for the molecule in which the fragment is observed also meets some specified criterion.

[1] AG Canada. Herbicide mode of action classification. https://www1.agric.gov.ab.ca/$department/deptdocs.nsf/all/prm6487
Last accessed: May 30, 2018

[2] IRAC Insecticide Resistance Action Committee Mode of Action Classification Scheme http://www.irac-online.org/documents/moa-classification/
Last accessed: May 30, 2018.

[3] FRAC Fungicide Resistance Action Committee (FRAC) Code List 2018 http://www.frac.info/docs/default-source/publications/frac-code-list/frac_code_list_2018-final.pdf?sfvrsn=6144b9a_2
Last accessed: May 30, 2018.

Figure G.2 depicts the MOA hit list and frequency of chemicals found in the EFSA database. The pesticidal mode of action is used to group compounds that are biologically similar. The role of biologically similarity in our ReadAcross approach is described in more detail in the next section. Briefly, in this particular study we consider a parent and metabolite to be biologically similar if the structural fragment responsible for pesticidal mode of action is retained in the metabolite. We used this to decide whether analogue quality (defined in the next section) is based on structure/property similarity or metabolic reactivity similarity. It is important to note that we not assuming the pesticidal model of action to be the same as the toxicity mode of action.

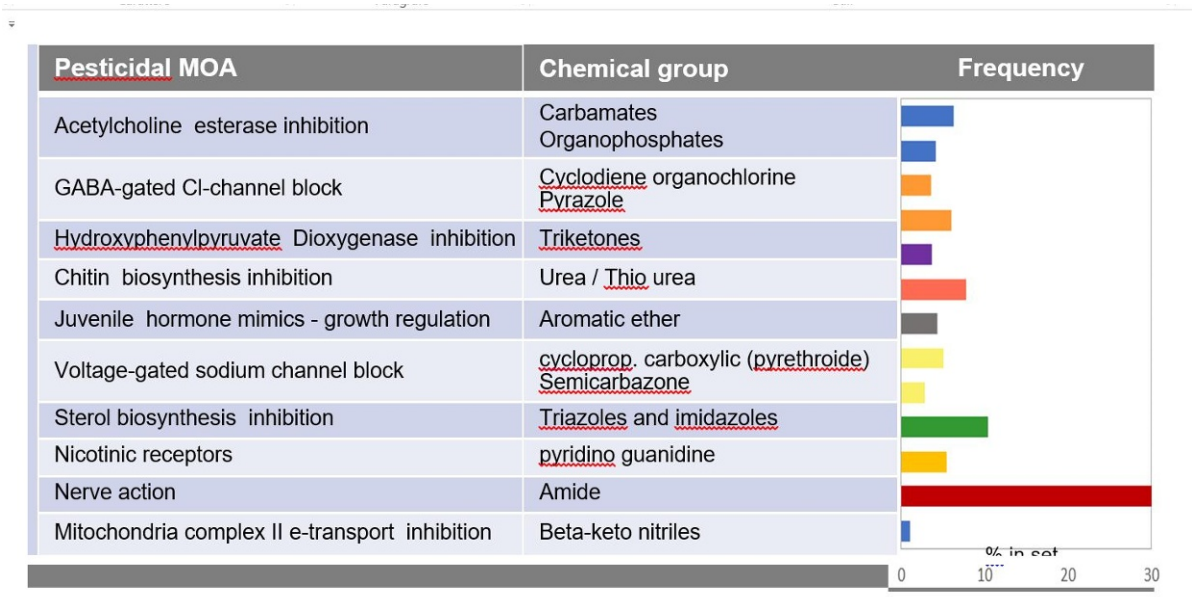
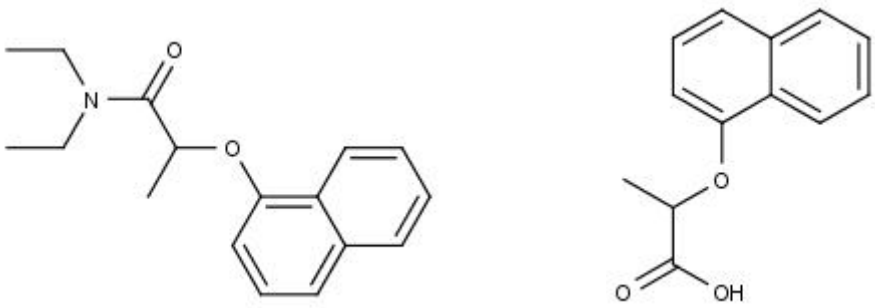


Figure G.2: Biological and chemical grouping by pesticide MOAs

Chemical Measures of Similarity

Structure-based Similarity

ToxPrint Chemotypes: While ToxPrint chemotypes were originally developed to capture the structural motifs of chemicals representing MoAs, it also provides generic atoms, bonds, rings, and chains to describe molecules (Yang et al., 2015). The ToxPrints are used as the fragment fingerprints with the ultimate purpose of database characterization as well as descriptors for QSAR models (Richard et al., 2016). Since ToxPrint chemotypes were designed with knowledge of chemical functions, it often differentiates fragments based on mechanistic details. For example, the similarity of napropamide and one of its metabolites



napropamide metabolite of napropamide

is calculated to be 0.40 by ToxPrints, while it is 0.70 according to the RDKit circular fingerprints (RDKit is an open source cheminformatics software http://www.rdkit.org/) (Rogers and Hahn, 2010). ToxPrint distinguishes the mode of action of acetamide group while RDKit does not. As said above, ToxPrint chemotypes are freely downloadable from https://chemotyper.org / and https://toxprint.org/ .

Pair-Wise Similarity Indices: Tanimoto coefficients were calculated for all pairs of structurable organic chemicals (yielding a 967 x 967 matrix of similarity coefficients) in the EFSA pesticide database. Using this similarity metric, the nearest neighbors of each chemical were identified. The Tanimoto similarity for a pair of molecules is a number between 0 and 1, with 1 indicating the molecules are identical, or at least highly similar since they have the same ToxPrint fingerprint.

Metabolic Reactivity Chemotypes: For a given target-analogue pair, if one compound is a possible metabolite of the other, then it's possible the analogue may have a relatively low similarity to the target based on structure or properties. In such cases, similarity based on metabolic reactivity is a more relevant measure of analogue quality. A publicly available set of metabolic rules (SyGMa rules, (systematic generation of potential metabolites)) have been published in the literature (Ridder and Wagener, 2008). There is also a recent reimplementation of SyGMa in Python library (https://pypi.org/project/SyGMa/). A commercial implementation is in the ChemTunes•ToxGPS® software (https://www.mn- am.com/products/toxgps). For the use of metabolic similarity in this report, the published implementation is sufficient.

The presence of a given chemical transformation chemotype indicates the presence in the molecule of a particular type of metabolic reaction site. The metabolic reactivity similarity for a given target-analogue pair is then

metabolic reactivity similarity = M/P

where P is the total number of metabolic reaction sites in the parent, and M is the number of sites common to both parent and metabolite.

For example, applying the rules to Tetraconazole (parent) and Triazole acetic acid (metabolite), a total of 8 reaction sites are matched for this pair. Figure G.3 demonstrates the match and mismatch of the reaction sites between the two structures. The metabolic reactivity similarity is then calculated to be 3/5=0.6. This metric was in turn used to qualify the similarity between the parent and the metabolite in RA cases where the metabolic similarity needed to be considered.

Figure G.3: Identification of Metabolic Sites

Molecular / Physicochemical Properties

CORINA Symphony Properties: From the public CORINA Symphony Community edition, (Richard et al., 2016) molecular properties representing characteristics of absorption and penetration are included to

compare the similarity. In addition, for pesticides, properties differentiating mode of action and metabolic reactivity are also important; hence, a few commercial parameters to capture shape and quantum mechanical properties are calculated using CORINA Symphony v1.1.

A total of 11 molecular / physicochemical properties called "CORINA Symphony Descriptors" were used to evaluate property-based similarity: number of hydrogen bond acceptors (HAcc), molecular weight (Weight), molecular complexity (Complex), approximate surface area (ASA), McGowan volume (McGowan), topological polar surface area (TPSA), dipole moment (dipole), polarizability (Polariz), water solubility (logS), octanol/water partition coefficient (XlogP), and heat of formation (HoF). The publicly free CORINA Symphony Community edition was mainly used to calculate the properties in this study from the web services (https://www.mn-am.com/services/corinasymphonydescriptors). To compare the reactivity, a few other properties were calculated from the commercial CORINA Symphony: water solubility, molar volume (McGowan), polarizability, and heats of formation. However, similar properties can also be calculated by other free software programs, e.g., EPISUITE (https://www.epa.gov/tsca-screening-tools/epi-suitetm-estimation-program-interface); the quantum mechanical parameters such as heats of formation can be calculated by MOPAC (http://openmopac.net/).

Skyline Plot, a Pattern Recognition Plot for Chemical Similarity Comparisons

The similarity between molecules can be defined based on structural features (e.g., ToxPrint chemotypes) and/or molecular properties.

Skyline plot. A column graph of standardized property values in a fixed order for a given compound can be compared with that of other compounds. These plots are a convenient visual tool for exploring property-based similarity. In addition, Pearson correlation coefficients between two chemicals can be calculated with the values in the plots. An example of Skyline plot is presented in Figure G.4.

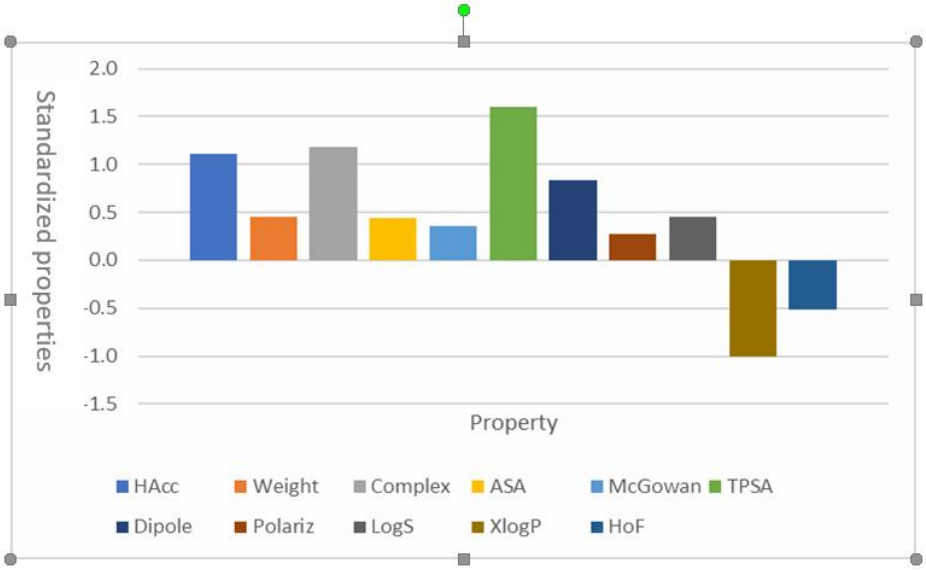


Figure G.4: Example of a Skyline plot prepared for a single compound using CORINA Symphony descriptors

Analog quality based on both structure- and property-based similarities. Once the Tanimoto (structural similarity) and Pearson (properties-based) coefficients have been calculated for a given target-analogue pair of molecules, a quantitative measure of analogue quality is calculated taking both structure- and property-based similarities:

$$\text{analogue quality} = \sqrt{(Tanimoto)\left(\frac{1+r}{2}\right)}$$

Note that the operation (1 + r)/2 simply rescales the Pearson correlation coefficient to the range 0 to 1.

Experimental Study Reliability

Accounting for study reliability is one of the most important aspects in weighing evidence sources for Read-Across. Estimating the reliability of a study reliability considers both the information completeness and the confidence one has in "believing" the results. For example, this database reports that about 65% of the Ames-negative studies used E-coli strain of WP2s or TA 102 to cover for the oxidizing or cross-linking mutagens. For the remaining 35% of the Ames-negative results, study reliability is being challenged since the results from these particular strains are necessary to judge whether a study is negative. The "believability" of these "negative" results reported in these studies should be lower than those studies that properly used the WP2s or TA102 test strains.

The following aspects of genetic toxicity data were included in the database describing the regulatory guidelines, protocols, study design, test system, and remarks. Only parameters for *in vitro* studies are listed here.

- ***OECD guideline and deviation.*** Meeting the guidelines or well-documented explanation of deviations (if any).
- ***Testing system.*** Species, strains or cell lines / cells, metabolic activation indicators (on/off) are well noted.
- ***Negative/Positive controls.*** No information was given about what negative or positive controls were used, although we assumed that appropriate controls were used for each strain with and without the metabolic activation system according to the OECD guideline.
- ***Metabolic activation system.*** No detailed descriptions in the database on the source of S-9 species (rat or hamster), concentration (10% or 30%) or activation chemicals although these conditions are known to affect outcomes.
- ***Assay techniques.*** No information on the culture techniques (preincubation, standard incorporation, etc.) were reported in the EFSA database.
- ***Acceptability.*** All data in the database received in Oct 2017 were deemed acceptable by EFSA database.
- ***Remarks.*** Important information source to give insights for the reliability of the calls. It provides the dose/concentration levels and ranges, duplicate experiments, strains or cells if out of deviations.

Five aspects of data elements were considered when rating a study:

1. OECD guideline and deviation
2. GLP compliance
3. Study design (species, strains, cell lines, metabolic activation)
4. Study design (concentration / dose levels and ranges, number duplicates, repeats)
5. Control information.

Experimental study reliability scores were assigned as follows (Table G.1).

Table G.1: Study reliability scores

Reliability Score	Description	Example
1.0	Meets all five listed in Section 2.4.2. Also the number of revertant counts at a given conc. level are available along with the precipitation and cytotoxicity information.	If we reviewed the conc. / dose level data from study records that satisfy all , then the study reliability would be 1.0.
0.95	Meets all five listed in Section 2.4.2., but no conc. level detailed reading.	
0.85	Studies either missing records or not conducted and at least one deficiency in the five aspects.	OECD equivalent guideline. The deviation included the highest concentration did not cover the full range recommended.
0.70	Studies either missing records or not conducted and at least two deficiencies in the five aspects.	If the OECD guideline deviation was the test system with strains lacking WP2 or TA102 and the outcome was negative.
0.50	Studies either missing records or not conducted and at least two deficiencies in the five aspects.	If the OECD guidelines had deviation of the test system, and there were only one test done with control data not providing details.

The study reliability scores are further adjusted for control information. If relatively detailed information of control is given (concurrent vehicle vs. no treatment), then the full score was given. A 5% penalty (-0.05 decrement in reliability score) is given if no detail is described, but the control was used; if there were no data whether the control was used, a 10% penalty was applied.

In Silico Evidence (QSAR Models)

QSAR model predictions for Ames and chromosome aberration endpoints were obtained using the ChemTunes.ToxGPS® software platform (model version 1.0, 2017-09-22). The logistic PLS modelling approach fits the categorical dependent variable to the PLS factors, linear combinations of the ToxPrint chemotypes, physicochemical properties, and quantum mechanical descriptors. Each model is optimized with respect to the number of pre-selected ToxPrints chemotypes and number of PLS factors. The overall QSAR model for a given endpoint includes global models as well as "local" mode-of-action models that take mechanistic knowledge into account. A quantitative weight-of-evidence (WoE) method was applied to combine predictions from individual models to arrive at a final prediction. The WoE method is based on Dempster-Shafer decision theory and takes into account the reliability of each model, which is obtained from model validation metrics [Rathman, 2018].

A knowledge base was built from large datasets of bacterial reverse mutagenicity and *in vitro* chromosome aberration data from public sources. The chemical curation process followed a general two step process: (1) Confirmation of the chemical compound's identity was performed by comparision of chemical structures, names, and CAS numbers. (2) Chemical structure curation was split into several sub steps. Validation of atom/bond features, valence, radical states, charges, and multiple fragments was done by an in-house set of CSRML rules. Next, confirmation of stereochemistry and double bond geometry was performed. Finally, identification and removal organometallic compounds, metal complexes and metals, natural products, mixtures, polymers, etc. was performed. Preparation of the structures for the computational analysis was performed by applying CORINA CLEAN workflow in CORINA Symphony. All toxicity data was manually evaluated according to the study inclusion criteria. When necessary, experts in the field have been consulted.

Combination of Outcomes by Weight of Evidence

A decision theory approach based on Dempster-Shafer theory (DST) was used to estimate uncertainty and combine multiple sources of information to obtain the weight-of-evidence final outcome. A detailed description of this approach, including numerous illustrative examples, is available in (Rathman et al., 2018). An overview is provided herewith.

The first step is to construct the DST structure for each individual evidence, represented graphically throughout this report by the probability bar.

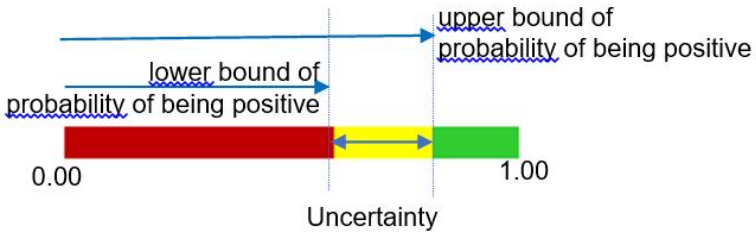


Figure G.5: Probability bar to represent positive, negative, and uncertainty

The red and green bars (Figure G.5) represent the probabilities of a positive or negative outcome, respectively, while the yellow bar represents the uncertainty. One source of uncertainty is that the method used to obtain the evidence (e.g., the experimental study or the QSAR model) is not 100% reliable. As described previously, reliability scores for experimental studies were assigned based on careful analysis of the study details. When considering evidence from an analogue in a read-across, a second source of uncertainty arises due to the fact that the analogue is a different molecule than the target. Reliability scores derived from structural similarity, property similarity, and metabolic similarity metrics (described previously), are used to capture this type of uncertainty.

After obtaining the probability bar (the DST structure) for each individual source of evidence, the second step in the DST approach is to combine the individual structures. The Inagaki combination rule was used (for details see (Rathman et al., 2018).

Example calculations. The evidence used in Read Across case study RA9 includes three independent experimental study results (Ames tests) for the analogue and a QSAR prediction for the target. The structure-based similarity (Tanimoto) between target and analogue is 0.60 and the property-based similarity (Pearson correlation coefficient) between target and analogue is 0.72. As described previously, the analogue quality metric is calculated by combining these two similarities:

analogue quality $\sqrt{(0.60)\left(\frac{1+0.72}{2}\right)}$ = 0.72

The first experimental study result (Ames-1 test) for the analogue was POSITIVE and the reliability score for this study was 0.50. The DST method interprets the degree to which this result for the analogue can be used as evidence that the target will also be POSITIVE; specifically, DST assigns a probability of (0.50)(0.72) = 0.36 to the belief that the target will be POSITIVE, and the remaining (1 - 0.36) = 0.64 is assigned UNCERTAIN. The probability bar is then:

The same approach is repeated for the second experimental study result available for the analogue. The Ames-2 experimental study result was NEGATIVE and reliability score 0.90. The DST method interprets the degree to which this result for the analogue can be used as evidence that the target will also be NEGATIVE; specifically, DST assigns a probability of (0.50)(0.90) = 0.45 to the belief that the target will be NEGATIVE, and the remaining (1 - 0.45) = 0.55 is assigned as UNCERTAIN. The probability bar for this piece of evidence is thus:

The Ames-3 experimental study result was NEGATIVE and reliability score 0.80.

The next step is to combine the analogue evidence. Combination using the Inagaki rule gives:

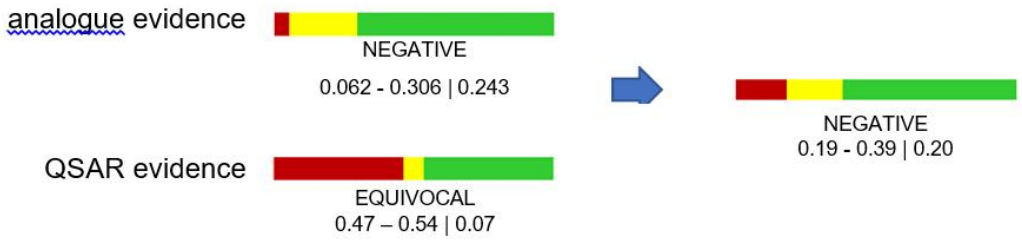
Next, the *in silico* evidence for the target is considered. The QSAR Ames model is a partial least squares logistic regression model that provides probabilistic predictions (Honma et al., 2018). For the RA9 target molecule, the QSAR prediction is based on predictions from two independent QSAR models, a global model and a model specific to aromatic nitro compounds. (In general, one or more other QSAR models could be used in addition to or in place of the QSAR models used here. The only requirement for using more than one model is that the models are independent.) For a given model, the reliability associated with a POSITIVE prediction is the positive predictivity value (PPV), the probability that a positive prediction is correct. Similarly, the reliability associated with a NEGATIVE prediction is the negative predictivity value (NPV), the probability that a negative prediction is correct. PPV and NPV are determined by appropriate validation methods when the model is developed. If these values are not provided by the model provider, they can be independently obtained by generating predictions for a test set of compounds with known outcomes and approximately (ideally) equal numbers of POSITIVEs and NEGATIVEs.

The global model predicts a probability of 0.20 that the target is Ames POSITIVE, and therefore a corresponding probability of 0.80 that the molecule is Ames NEGATIVE. The model reliability metrics, PPV and NPV, for this model are 0.72 and 0.78, respectively. The DST approach re-assigns the prediction probabilities, taking into account that the models are less than 100% reliable. The DST-based probability of a POSITIVE outcome based on the global model prediction is thus (0.20)(0.72) = 0.14, while the

probability of a NEGATIVE outcome is (0.80)(0.78) = 0.63. The remaining portion of the probability is the UNCERTAINTY, 1 – 0.14 – 0.63 = 0.23. The aromatic nitro model predicts a probability of 0.74 that the target is Ames POSITIVE, and therefore a corresponding probability of 0.26 that the molecule is Ames NEGATIVE. The model reliability metrics, PPV and NPV, for this model are 0.88 and 0.69, respectively. The DST-based probability of a POSITIVE outcome is (0.74)(0.88) = 0.65, while the probability of a NEGATIVE outcome is (0.26)(0.69) = 0.18. The remaining portion of the probability is the UNCERTAINTY, 1 – 0.65 – 0.18 = 0.17.

DST is then used to combine *in silico* data, two QSAR models in this case, to obtain the overall QSAR prediction:

Disagreement between the models results in an equivocal outcome in this case. The final step is to combine the analogue evidence with the QSAR evidence. Using the Inagaki combination rule gives the result shown below.



Selection of Read Across Cases for Strategy II

This RA analysis employed the parent compound as the analogue and its metabolites as the target molecules. The pairs selected, with the rationale for selection, are in table G.2.

Table G.2. Rationales for Selection of Read Across Cases *(1 CAR: Conflicting assay results; 2 IQP: Inaccurate QSAR predictions from many packages)*

RA Case ID	SUB ID_COM ID	Pesticide Parent Name	MOA Category	Rationale
RA 1 & 2	15061_15898	tembotrione	HPPD: Triketone	Metabolic Reactivity A (MOA chemotype preserved in metabolic pathway); Baseline for general case for HPPD inhibitors
RA 3 & 4	15061_15899	tembotrione	HPPD: Triketone	Metabolic Reactivity A (MOA chemotype preserved in metabolic pathway); QSAR results for ivtCA vs. the data quality

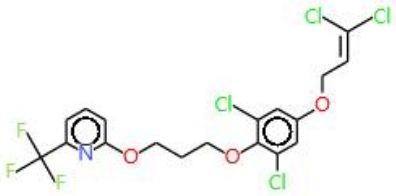
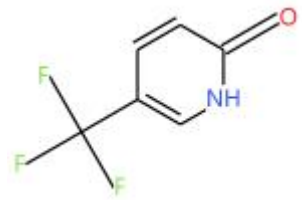
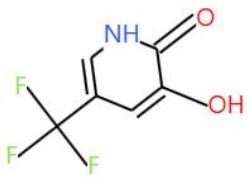
RA Case ID	SUB ID_COM ID	Pesticide Parent Name	MOA Category	Rationale
RA 5 & 6	15061_15900	tembotrione	HPPD: Triketone	Metabolic Reactivity A (MOA chemotype preserved in metabolic pathway); Analogue data quality
RA7 & 8	1172_2061	sulcotrione	HPPD: Triketone	Metabolic Reactivity A (MOA chemotype preserved in metabolic pathway); CAR1
RA9 & 10	35031_50553	mesotrione	HPPD: Triketone	Metabolic Reactivity A (MOA chemotype preserved in metabolic pathway); IQP2
RA11 & 12	35031_50554	mesotrione	HPPD: Triketone	Metabolic Reactivity A (MOA chemotype preserved in metabolic pathway); CAR1
RA13 & 14	1170_50017	imidacloprid	Neonicotinide	Metabolic Reactivity A (MOA chemotype preserved in metabolic pathway); CAR1
RA15 & 16	1170_50019	imidacloprid	Neonicotinide	Metabolic Reactivity A (MOA chemotype preserved in metabolic pathway); Baseline for general case of neonicotinamideinhibitors; CAR1
RA17 & 18	1170_50021	imidacloprid	Neonicotinide	Metabolic Reactivity A (MOA chemotype preserved in metabolic pathway); CAR1, IQP2
RA19 & 20	1166_15576	napropamide	Acetamide	Metabolic Reactivity B (MOA chemotype preserved in metabolic pathway); IQP2
RA21 & 22	1166_50576	napropamide	Acetamide	Metabolic Reactivity B (MOA chemotype preserved in metabolic pathway)
RA23 & 24	1416_2098	cyflumetofen	Beta-keto nitrile	Metabolic Reactivity B (MOA chemotype preserved in metabolic pathway); IQP2
RA 25 & 26	1416_1928	cyflumetofen	Beta-keto nitrile	Metabolic Reactivity B (MOA chemotype not preserved in metabolic pathway)
RA 27 & 28	1174_1877	tetraconazole	Sterol biosynthesis: triazole	Metabolic Reactivity B (MOA chemotype not preserved in metabolic pathway), metabolite common to several parents
RA 29 & 30	1181_1877	pentaconazole	Sterol biosynthesis: triazole	Metabolic Reactivity B (MOA chemotype not preserved in metabolic pathway), metabolite common to several parents

Appendix H – Structural factors in metabolic transformations: the Structural Alerts

SUB_ID 3688 Pyridalyl

The parent compound is negative and has no SAs. On the opposite, the two metabolites are positive with the alpha,beta-unsaturated carbonyls SA (Table H.1).

Table H.1: SUB_ID 3688 (AS: Pyridalyl)

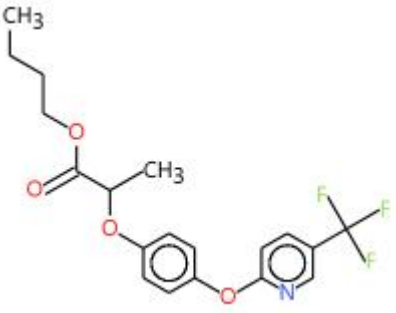
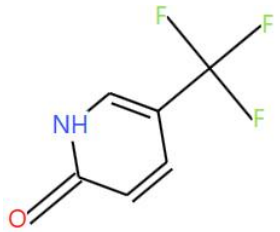
#	CAS N	CHEMICAL STRUCTURE/SMILES	Similarity	Oncologic Primary Classif	Ames test alerts by ISS	COM_ID	SUB_ID	Ames test
1	179101-81-6	Parent  FC(F)(F)c1cccc(OCCCOc2c(Cl)cc(OCC=C(Cl)Cl)cc2Cl)n1	100%	Not classified	No alert found	1688	3688	Neg
2	332526-63-0	 FC(F)(F)C1C=CC(=O)NC=1	20%	Not classified	alpha,beta-unsaturated carbonyls	1883	3688	Pos
3	No CAS	 OC1=CC(=CNC1=O)C(F)(F)F	19.5%	Not classified	alpha,beta-unsaturated carbonyls	50296	3688	Pos

SUB_ID 1321 Fluazifop-P-butyl

The parent compound is negative and has no SAs. On the opposite, the metabolites is positive with the alpha,beta-unsaturated carbonyls SA (Table H.2).

It should be noticed that the metabolite is included also in the previous subgroup, indicating that it may derive from two different parents. This occurrence is not rare in this database.

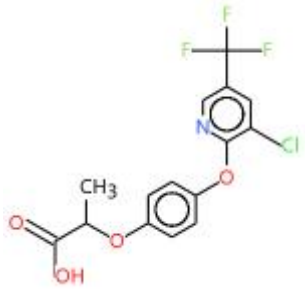
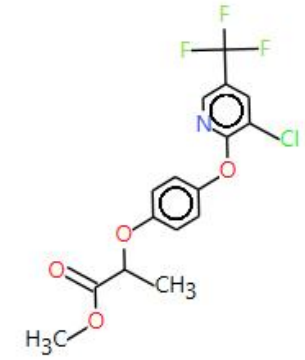
Table H.2: SUB_ID: 1321 (AS: Fluazifop-P-butyl)

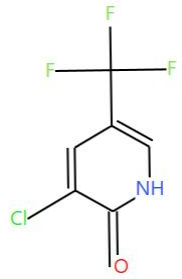
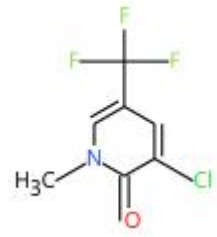
#	CAS	CHEMICAL STRUCTURE / SMILES	Similarity	Oncologic Primary Classif	Ames test alerts by ISS	COM_ID	SUB_ID	Ames test
1	79241-46-6	Parent  CCCCOC(=O)C(C)Oc1ccc(Oc2ccc(cn2)C(F)(F)F)cc1	100%	Not classified	No alert found	1882	1321	Neg
2	33252-63-0	 FC(F)(F)C1C=CC(=O)NC=1	26.30%	Not classified	alpha,beta-unsaturated carbonyls	1883	1321	Pos

SUB_ID 1137 Haloxyfop-P

The parent compound and a metabolite are negative, and have no SAs (Table H.3). Two metabolites (Compounds 3 and 4) are very similar, and both have the same SA. Whereas Compound 3 is positive as expected, there is not simple explanation for the negativity of Compound 4 (Inductive effect of the Methyl on the Nitrogen involved ? May the Methyl stabilize the Amide tautomer compared to the enol, thus inhibiting the reactivity of the SA?).

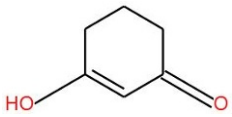
Table H.3: SUB_ID: 1137 (AS: Haloxyfop-P)

#	CAS	CHEMICAL STRUCTURE / SMILES	Similarity	Oncologic Primary Classif	Ames test alerts by ISS	COM_ID	SUB_ID	Ames test
1	95977-29-0	Parent  CC(Oc1ccc(Oc2ncc(cc2Cl)C(F)(F)F)cc1)C(O)=O	100%	Not classified	No alert found	1549	1137	Neg
2	72619-32-0	 COC(=O)C(C)Oc1ccc(Oc2ncc(cc2Cl)C(F)(F)F)cc1	89.80%	Not classified	No alert found	1825	1137	Neg

#	CAS	CHEMICAL STRUCTURE / SMILES	Similarity	Oncologic Primary Classif	Ames test alerts by ISS	COM_ID	SUB_ID	Ames test
3	76041-71-9	 FC(F)(F)C1=CNC(=O)C(Cl)=C1	33.30%	Not classified	alpha,beta-unsaturated carbonyls	15889	1137	Pos
4	No CAS	 CN1C=C(C=C(Cl)C1=O)C(F)(F)F	32.40%	Not classified	alpha,beta-unsaturated carbonyls	75575	1137	Neg

Sub_ID 1259 Clethodim

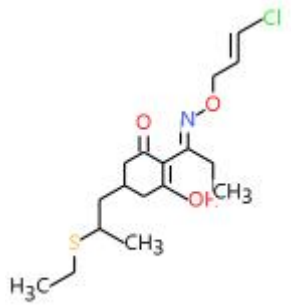
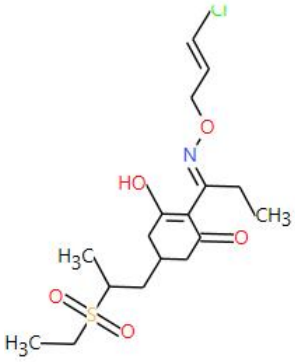
Three metabolites (Cps 5, 6, 7) have no SAs and are negative, whereas the parent and three other metabolites have SAs (Table H.4). However, only Compound 2 is positive.

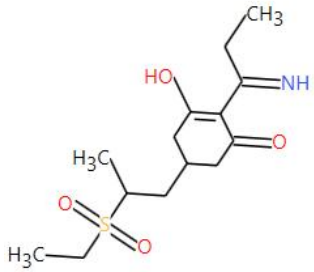
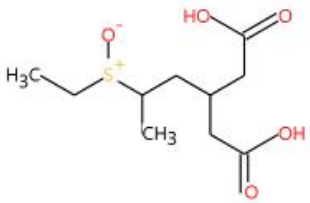
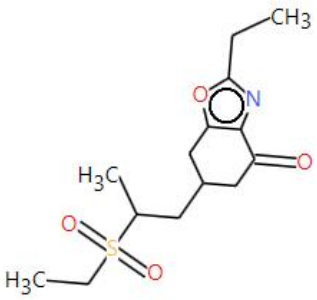
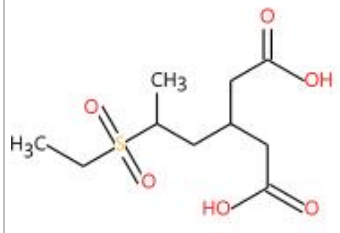
Inspection of the entire EFSA database shows that the specific motif  is negative in 9 / 10 occurrences, and the monohaloalkene CC=CCl is positive in only 3 / 38 occurrences. According to (Modi et al., 2012), the alkylating activity of alpha-beta unsaturated ketones may be substantially inhibited by substitution at the double bond, particularly by bulky or hydrophilic groups.

The statistics on the above SAs provided by the EFSA database may stimulate in the future fine-tuning of the SAs themselves.

In addition, it should considered that Compound 2 is positive in two Ames strains, but is negative in two other *in vitro* assays (chromosomal aberrations and mammalian cells gene mutation). This may raise doubts on its positive Ames call.

Table H.4: SUB_ID: 1259 (AS: Clethodim)

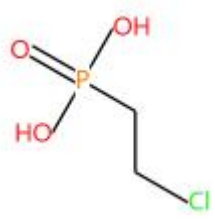
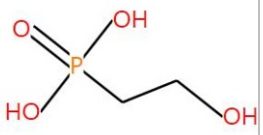
#	CAS	CHEMICAL STRUCTURE/SMILES	Similarity	Oncologic Primary Classif	Ames test alerts by ISS	COM_ID	SUB_ID	Ames test
1	99129-21-2	Parent  CCSC(C)CC1CC(=O)C(=C(O)C1)C(CC)=NOCC=CCl	100%	Not classified	alpha,beta-unsaturated carbonyls|Monohaloalkene	1502	1259	Neg
2	No CAS	 CCC(=NOCC=CCl)C1C(=O)CC(CC(C)S(=O)(=O)CC)CC=1O	83.30 %	Not classified	alpha,beta-unsaturated carbonyls|Monohaloalkene	15544	1259	Pos
3	No CAS	 CCC(=NOCC=CCl)C1C(=O)CC(O)(CC(C)S(=O)(=O)CC)CC=1O	65.30 %	Not classified	alpha,beta-unsaturated carbonyls Monohaloalkene	15556	1259	Neg

#	CAS	CHEMICAL STRUCTURE/SMILES	Similarity	Oncologic Primary Classif	Ames test alerts by ISS	COM_ID	SUB_ID	Ames test
4	No CAS	 CCC(=N)C1C(=O)CC(CC(C)S(=O)(=O)CC)CC=1O	60.50 %	Not classified	alpha,beta-unsaturated carbonyls	15547	1259	Neg
5	No CAS	 CC[S+]([O-])C(C)CC(CC(O)=O)CC(O)=O	51.30 %	Not classified	No alert found	15551	1259	Neg
6	No CAS	 CCc1nc2C(=O)CC(CC(C)S(=O)(=O)CC)Cc2o1	46.50 %	Not classified	No alert found	15549	1259	Neg
7	No CAS	 CCS(=O)(=O)C(C)CC(CC(O)=O)CC(O)=O	40%	Not classified	No alert found	15552	1259	Neg

SUB_ID 1180 Ethephon

The positive Ames of the parent, in contrast to the negativity of the metabolite, can be explained by the genotoxicity "aliphatic halogens" SA, which is lost in the metabolite, giving rise to detoxification (Table H.5).

Table H.5: SUB_ID: 1180 (AS: Ethephon)

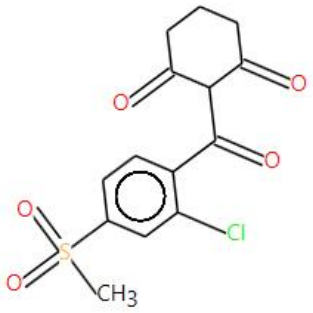
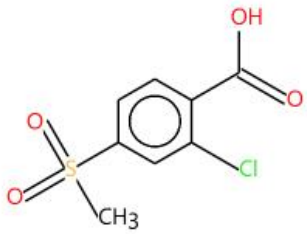
#	CAS	CHEMICAL STRUCTURE/SMILES	Similarity	Oncologic Primary Classif	Ames test alerts by ISS	COM_ID	SUB_ID	Ames test
1	16672-87-0	Parent  OP(O)(=O)CCCl	100%	Organophosphorus Type Compounds	Aliphatic halogens	1635	1180	Pos
2	22987-21-9	 OCCP(O)(O)=O	71.40%	Organophosphorus Type Compounds	No alert found	75002	1180	Neg

SUB_ID 1172 Sulcotrione

The mechanisms of action of the class of the halogenated aromatics are particularly complex. The ISS rule base includes SAs for non-genotoxic carcinogenicity only (Table H.6).

The inspection of analogues in the entire EFSA database pointed to 17 similar compounds with a carboxylic moiety ortho to the halogen, with only 1 / 17 positives. The positivity of the parent is more difficult to rationalize.

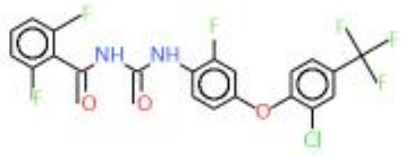
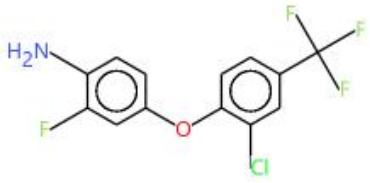
Table H.6: SUB_ID: 1172 (AS: Sulcotrione)

#	CAS	CHEMICAL STRUCTURE/SMILES	Similarity	Oncologic Primary Classif	Ames test alerts by ISS	COM_ID	SUB_ID	Ames test
1	99105-77-8	Parent  CS(=O)(=O)c1ccc(C(=O)C2C(=O)CCCC2=O)c(Cl)c1	100%	Halogenated Aromatic Hydrocarbon Type Compounds	No alert found	1579	1172	Pos
2	53250-83-2	 CS(=O)(=O)c1ccc(C(O)=O)c(Cl)c1	68.60%	Halogenated Aromatic Hydrocarbon Type Compounds	No alert found	2061	1172	Neg

SUB_ID 1368 Flufenoxuron

The parent is negative in the Ames test. Out of the two metabolites, the first one is negative, the second is positive (Table H.7). For this second metabolite the Primary Aromatic Amine Structural SA, due to the hydrolysis of the amide, explains the experimental data.

Table H.7: SUB_ID: 1368 (AS: Flufenoxuron)

#	CAS	CHEMICAL STRUCTURE/SMILES	Similarity	Oncologic Primary Classif	Ames test alerts by ISS	COM_ID	SUB_ID	Ames test
1	101463-69-8	Parent  Fc1cc(Oc2ccc(cc2Cl)C(F)(F)F)ccc1NC(=O)NC(=O)c1c(F)cccc1F	100%	Halogenated Aromatic Hydrocarbon Type Compounds	No alert found	1645	1368	Neg
2	No CAS	 NC(=O)Nc1ccc(Oc2ccc(cc2Cl)C(F)(F)F)cc1F	75%	Halogenated Aromatic Hydrocarbon Type Compounds	No alert found	1903	1368	Neg
3	No CAS	 Nc1ccc(Oc2ccc(cc2Cl)C(F)(F)F)cc1F	67.90 %	Aromatic Amine Type Compounds| Halogenated Aromatic Hydrocarbon Type Compounds	Primary aromatic amine,hydroxyl amine and its derived esters	1902	1368	Pos

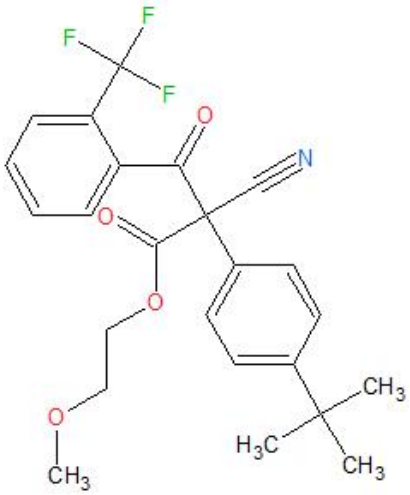
SUB_ID 1416 Cyflumetofen

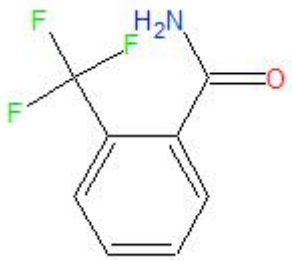
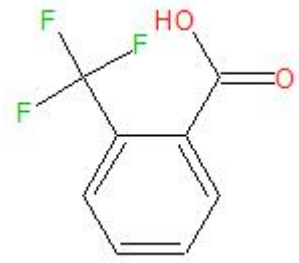
Parent and metabolites have no SAs, but Com_id 2098 is reported as positive (Table H.8).

This is a special case, since the above compound was predicted as Negative by all the QSARs (Objective 2) and by Read Across (Objective 4). In the EFSA database, there are 69 benzamides (analogues) with only 3 positives. In other terms, there are no structural elements that permit the rationalization of the Ames test positivity.

In addition, Com_id 2098 is negative in two other *in vitro* tests. Thus seems to be a typical case in which (Q)SAR analyses point to the need of re-testing (if necessary).

Table H.8: SUB_ID: 1416 (AS: Cyflumetofen)

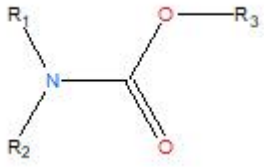
#	CAS	CHEMICAL STRUCTURE/SMILES	Similarity	Oncologic Primary Classif	Ames test alerts by ISS	COM_ID	SUB_ID	Ames test
1	400882-07-7	Parent  COCCOC(=O)C(C#N)(C(=O)c1ccccc1C(F)(F)F)c1ccc(cc1)C(C)(C)C	100%	Not classified	No alert found	1509	1416	Neg

#	CAS	CHEMICAL STRUCTURE/SMILES	Similarity	Oncologic Primary Classif	Ames test alerts by ISS	COM_ID	SUB_ID	Ames test
2	360-64-5	 NC(=O)c1ccccc1C(F)(F)F	48.90%	Not classified	No alert found	2098	1416	Pos
3	433-97-6	 OC(=O)c1ccccc1C(F)(F)F	48.90%	Not classified	No alert found	1928	1416	Neg

SUB_ID 1106 Phosalone

The parent is negative and belongs to three Oncologic classes: Carbamate, Halogenated aromatic and Organophosphorus (Table H.9).

According to (Modi et al., 2012), simple alkyl carbamates have specific structural requirements for optimal carcinogenicity:



a two-carbon moiety (e.g., ethyl and vinyl) at the carboxyl end (R3), and

(ii) a relatively free amino end available for N-hydroxylation and N-acyloxylation (R1 and/or R2).

The parent is a sterically hindered carbamate.

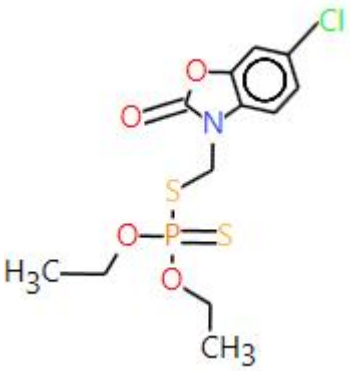
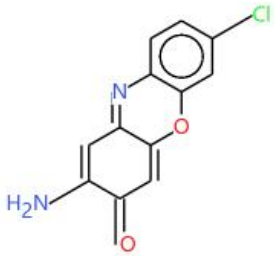
The type of Halogenated aromatic (SMILES: N1C(Oc2c1ccc(c2)Cl)=O) is quite peculiar, and no other examples are found in the EFSA database.

The organophosphorus moiety is present in 8 chemicals in the EFSA database, out of them only 2 are positive.

Thus the negativity of the parent is supported by the analysis of SAs and substructures.

On the other hand, the positivity of the metabolite is supported by the presence of the alpha,beta-unsaturated carbonyls SA.

Table H.9: SUB_ID: 1106 (AS: Phosalone)

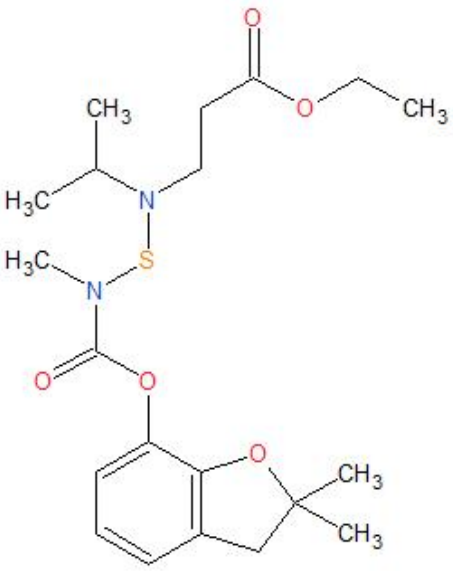
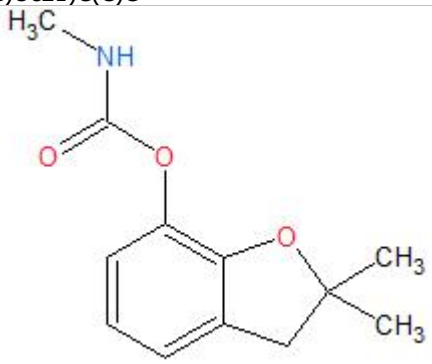
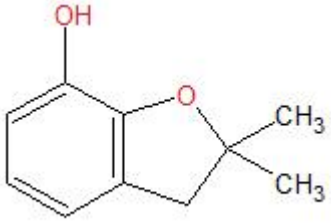
#	CAS	CHEMICAL STRUCTURE/SMILES	Similarity	Oncologic Primary Classif	Ames test alerts by ISS	COM_ID	SUB_ID	Ames test
1	2310-17-0	Parent  CCOP(=S)(OCC)SCN1C(=O)Oc2cc(Cl)ccc12	100%	Carbamate Type Compounds| Halogenated Aromatic Hydrocarbon Type Compounds| Organophosphorus Type Compounds	No alert found	1676	1106	Neg
2	No CAS		42.10%	Halogenated Aromatic Hydrocarbon Type Compounds	alpha,beta-unsaturated carbonyls	50474	1106	Pos

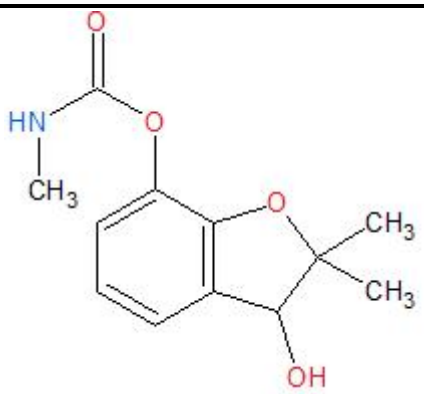
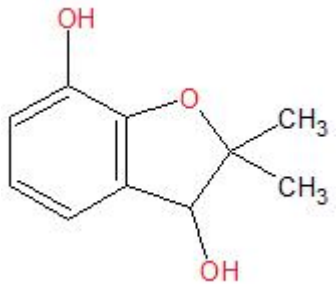
		NC1=CC2=Nc3ccc(Cl)cc3OC2=CC1=O						

SUB_ID 1139 Benfuracarb

The parent is a hindered carbamate, and is negative (Table H.10). Two metabolites are non-hindered (N accessible) carbamates, and are positive (as explained in SUB_ID 1106 Phosalone). Two more metabolites are phenols with no SAs (and are negative).

Table H.10: SUB_ID: 1139 (AS: Benfuracarb)

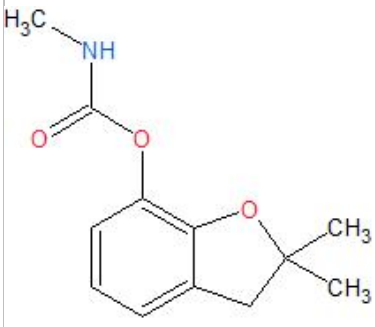
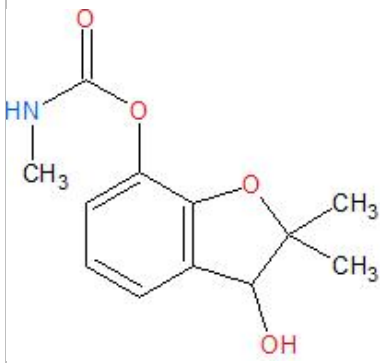
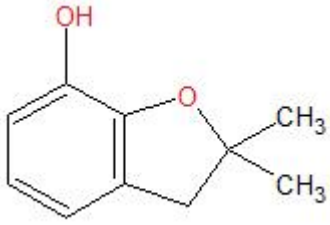
#	CAS	CHEMICAL STRUCTURE/SMILES	Similarity	Oncologic Primary Classif	Ames test alerts by ISS	COM_ID	SUB_ID	Ames test
1	82560-54-1	Parent  CCOC(=O)CCN(SN(C)C(=O)Oc1cccc2CC(C)(C)Oc21)C(C)C	100%	Carbamate Type Compounds	No alert found	1488	1139	Neg
2	1563-66-2	 CNC(=O)Oc1cccc2CC(C)(C)Oc21	59.10%	Carbamate Type Compounds	No alert found	1606	1139	Pos
3	1563-38-8	 CC1(C)Cc2cccc(O)c2O1	45%	Phenol Type Compounds	No alert found	15274	1139	Neg

#	CAS	CHEMICAL STRUCTURE/SMILES	Similarity	Oncologic Primary Classif	Ames test alerts by ISS	COM_ID	SUB_ID	Ames test
4	16655-82-6	 CNC(=O)Oc1cccc2C(O)C(C)(C)Oc12	44.40%	Carbamate Type Compounds	No alert found	15270	1139	Pos
5	17781-15-6	 CC1(C)Oc2c(O)cccc2C1O	34.10%	Phenol Type Compounds	No alert found	15272	1139	Neg

SUB_ID 1140 Carbofuran

The parent and one metabolite are non-hindered (N accessible) carbamates, and are positive (as explained in SUB_ID 1106 Phosalone). Another metabolite is a simple phenol, and is negative (Table H.11).

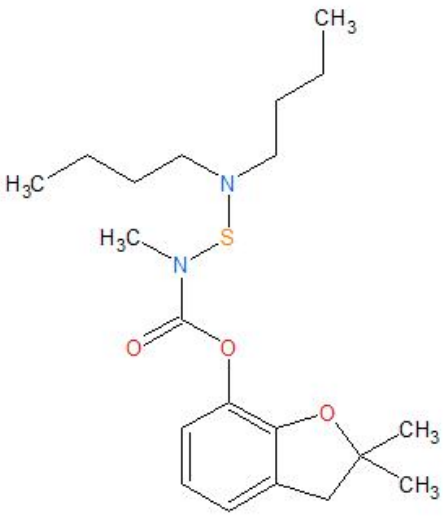
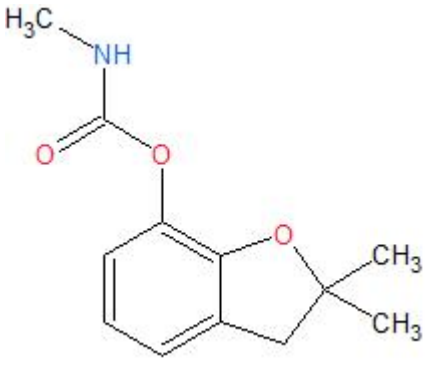
Table H.11: SUB_ID: 1140 (AS: Carbofuran)

#	CAS	CHEMICAL STRUCTURE/SMILES	Similarity	Oncologic Primary Classif	Ames test alerts by ISS	COM_ID	SUB_ID	Ames test
1	1563-66-2	 Parent CNC(=O)Oc1cccc2CC(C)(C)Oc21	100%	Carbamate Type Compounds	No alert found	1606	1140	Pos
2	16655-82-6	 CNC(=O)Oc1cccc2C(O)C(C)(C)Oc12	78.80%	Carbamate Type Compounds	No alert found	15270	1140	Pos
3	1563-38-8	 CC1(C)Cc2cccc(O)c2O1	64.30%	Phenol Type Compounds	No alert found	15274	1140	Neg

SUB_ID 1141 Carbosulfan

The hindered carbamate parent is negative (Table H.12), whereas the non hindered carbamate metabolite (N accessible) is positive (as explained in SUB_ID 1106 Phosalone).

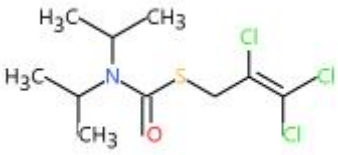
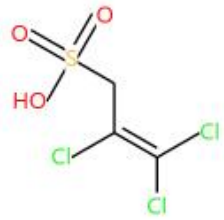
Table H.12: SUB_ID: 1141 (AS: Carbosulfan)

#	CAS	CHEMICAL STRUCTURE/SMILES	Similarity	Oncologic Primary Classif	Ames test alerts by ISS	COM_ID	SUB_ID	Ames test
1	55285-14-8	Parent  CCCCN(CCCC)SN(C)C(=O)Oc1cccc2CC(C)(C)Oc21	100%	Carbamate Type Compounds	No alert found	1607	1141	Neg
2	1563-66-2	 CNC(=O)Oc1cccc2CC(C)(C)Oc21	61.90 %	Carbamate Type Compounds	No alert found	1606	1141	Pos

SUB_ID 1187 Tri-allate

The non hindered, alkyl carbamate parent is positive (N accessible, explained in SUB_ID 1106 Phosalone), whereas the metabolite, devoid of SAs, is negative (Table H.13).

Table H.13: SUB_ID: 1187 (AS: Tri-allate)

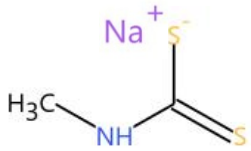
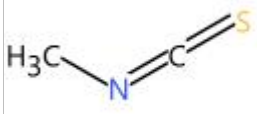
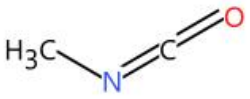
#	CAS	CHEMICAL STRUCTURE/SMILES	Similarity	Oncologic Primary Classif	Ames test alerts by ISS	COM_ID	SUB_ID	Ames test
1	2303-17-5	Parent  CC(C)N(C(C)C)C(=O)SCC(Cl)=C(Cl)Cl	100%	Carbamate Type Compound| Thiocarbamate Type Compounds	Alkyl carbamate and thiocarbamate	1703	1187	Pos
2	No CAS	 OS(=O)(=O)CC(Cl)=C(Cl)Cl	38.50%	Not classified	No alert found	75158	1187	Neg

SUB_ID 1203 Metam-sodium

The non hindered carbamate parent is positive (N accessible), whereas the two metabolites are negative (as explained in SUB_ID 1106 Phosalone).

The two metabolites have the Isocyanate and isothiocyanate SA (ISS rule base). However, this SA has a quite low positive predictivity (7 positives / 19 counts in the ISSSTY database): so its main role is more for alerting in very conservative analyses, and not for predictions (Table H.14).

Table H.14: SUB_ID: 1203 (AS: Metam-sodium)

#	CAS	CHEMICAL STRUCTURE/SMILES	Similarity	Oncologic Primary Classif	Ames test alerts by ISS	COM_ID	SUB_ID	Ames test
1	137-42-8	Parent  [Na+].CNC([S-])=S	100%	Carbamate Type Compounds| Thiocarbamate Type Compounds	No alert found	1844	1203	Pos
2	556-61-6	 CN=C=S	NA (50.9%)*	Not classified	Isocyanate and isothiocyanate groups	1872	1203	Neg
3	624-83-9	 CN=C=O	NA (40%)*	Not classified	Isocyanate and isothiocyanate groups| Primary aromatic amine,hydroxyl amine and its derived esters	15785	1203	Neg

*The presence of ions does not permit the calculation of similarity of the metabolites by the Toolbox. In bracket, the similarity values calculated with ChemFolder are reported.

SUB_ID 1217 Benthiavalicarb-isopropyl

The parent and one metabolite are hindered carbamates (N not accessible; for an effect, the substituent on the carboxyl end should be ethyl or vinyl), and are negative as expected (as explained in SUB_ID 1106 Phosalone).

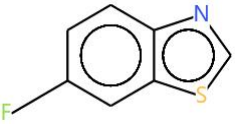
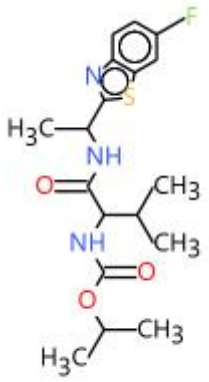
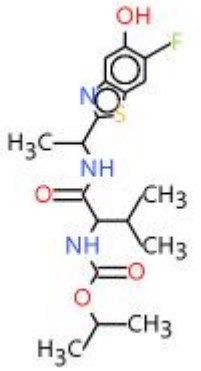
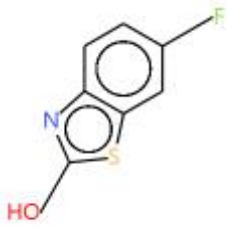
Regarding the halogenated aromatics class in Oncologic, they are mainly effective for non-genotoxic carcinogenicity. The substructure  is present in four other chemicals in the EFSA database and they are all negative. In addition, the positive Com_id 15105 was tested only in Ames and is positive only in TA98 +S9, (no other tests). Thus the positivity evidence is quite weak (Table H.15).

Table H.15: SUB_ID: 1217 (AS: Benthiavalicarb-isopropyl)

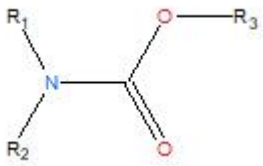
#	CAS	CHEMICAL STRUCTURE/SMILES	Similarity	Oncologic Primary Classif	Ames test alerts by ISS	COM_ID	SUB_ID	Ames test
1	177406-68-7	Parent  CC(C)OC(=O)NC(C(C)C)C(=O)NC(C)c1nc2ccc(F)cc2s1	100%	Carbamate Type Compounds|Halogenated Aromatic Hydrocarbon Type Compounds	Alkyl carbamate and thiocarbamate	1853	1217	Neg
2	No CAS	 CC(C)OC(=O)NC(C(C)C)C(=O)NC(C)c1nc2cc(O)c(F)cc2s1	86.80%	Carbamate Type Compounds|Halogenated Aromatic Hydrocarbon Type Compounds|Phenol Type Compounds	Alkyl carbamate and thiocarbamate	15318	1217	Neg
3	No CAS	 CC(O)c1nc2ccc(F)cc2s1	56.40%	Halogenated Aromatic Hydrocarbon Type Compounds	No alert found	15104	1217	Neg

4	No CAS	 CC(=O)c1nc2ccc(F)cc2s1	51.30%	Halogenated Aromatic Hydrocarbon Type Compounds	No alert found	15105	1217	Pos
5	No CAS	 Oc1nc2ccc(F)cc2s1	48.60%	Halogenated Aromatic Hydrocarbon Type Compounds	No alert found	15103	1217	Neg

SUB_ID 1239 Carbendazim

The parent (Carbendazim) of this SUB_Id is a metabolite in the next SUB_ID.

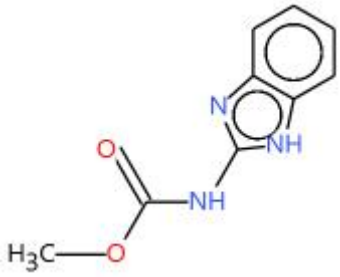
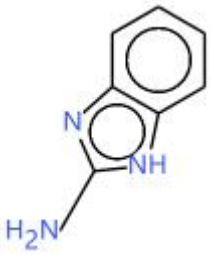
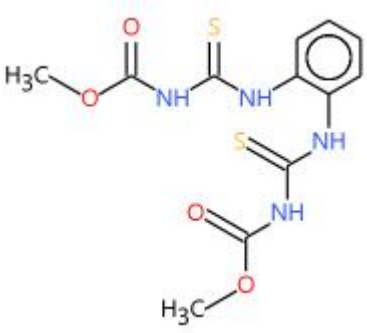
The parent, and metabolites 2 and 4 are carbamate with similar structure, and are negative (Table H.16).



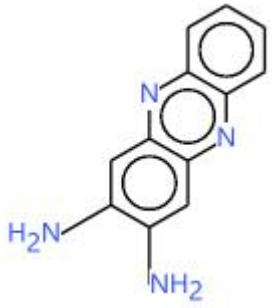
As a matter of fact, in the structure of the carbamate the substituent in R3 is methyl, whereas for optimal activity it should be ethyl or vinyl. The same substitution is negative in metabolites 2 and 4. On the other hand, the parent is reported to be positive for chromosomal aberrations *in vitro*.

The positivity of metabolite 5 can be attributed to the classical primary aromatic amine SA, whereas the negativity of metabolite 3 is not clearly understandable. It can be hypothesized that the basic character of the indole nucleus may inhibit the formation of the nitrenium ion, an intermediate in the primary aromatic amines mechanism of action (Benigni and Bossa, 2011).

Table H.16: SUB_ID: 1239 (AS: Carbendazim)

#	CAS	CHEMICAL STRUCTURE/SMILES	Similarity	Oncologic Primary Classif	Ames test alerts by ISS	COM_ID	SUB_ID	Ames test
1	10605-21-7	Parent  COC(=O)Nc1nc2ccccc2[nH]1	100%	Carbamate Type Compounds	No alert found	1605	1239	Neg [4]
2	22769-68-2	 COC(=O)Nc1nc2ccc(O)cc2[nH]1	82.80%	Carbamate Type Compounds| Phenol Type Compounds	No alert found	15734	1239	Neg
3	934-32-7	 Nc1nc2ccccc2[nH]1	66.70%	Aromatic Amine Type Compounds	Primary aromatic amine,hydroxyl amine and its derived esters	15925	1239	Neg
4	23564-05-8	 COC(=O)NC(=S)Nc1ccccc1NC(=S)NC(=O)OC	50%	Carbamate Type Compounds	No alert found	15738	1239	Neg

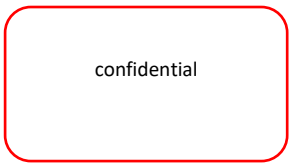
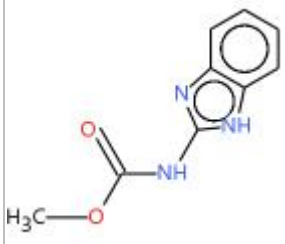
[4] The positive results in the Ames test reported in the EFSA database for Carbendazim is to be attributed to mutagenic impurities EFSA, 2010a.

#	CAS	CHEMICAL STRUCTURE/SMILES	Similarity	Oncologic Primary Classif	Ames test alerts by ISS	COM_ID	SUB_ID	Ames test
5	655-86-7	 Nc1cc2nc3ccccc3nc2cc1N	40%	Aromatic Amine Type Compounds	Heterocyclic Polycyclic Aromatic Hydrocarbons| Primary aromatic amine,hydroxyl amine and its derived esters	15926	1239	Pos

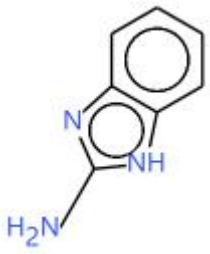
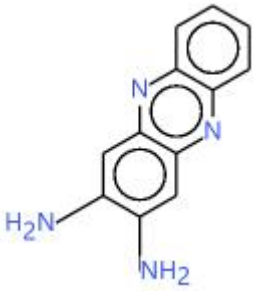
SUB_ID 15013

Chemicals with COM_ID = 1605, 15925, 15926 are also part of the previous Sub_ID 1239 (Table H.17). See considerations above.

Table H.17: SUB_ID: 15013 (AS: confidential)

#	CAS	CHEMICAL STRUCTURE/SMILES	Similarity	Oncologic Primary Classif	Ames test alerts by ISS	COM_ID	SUB_ID	Ames test
1	No CAS	Parent 	100%	Carbamate Type Compounds	No alert found	15061	15013	Neg
2	10605-21-7	 COC(=O)Nc1nc2ccccc2[nH]1	62.90%	Carbamate Type Compounds	No alert found	1605	15013	Neg[5]

[5] The positive results in the Ames test reported in the EFSA database for Carbendazim is to be attributed to mutagenic impurities EFSA, 2010a.

#	CAS	CHEMICAL STRUCTURE/SMILES	Similarity	Oncologic Primary Classif	Ames test alerts by ISS	COM_ID	SUB_ID	Ames test
3	934-32-7	 Nc1nc2ccccc2[nH]1	38.70%	Aromatic Amine Type Compounds	Primary aromatic amine,hydroxyl amine and its derived esters	15925	15013	Neg
4	655-86-7	 Nc1cc2nc3ccccc3nc2cc1N	37.80%	Aromatic Amine Type Compounds	Heterocyclic Polycyclic Aromatic Hydrocarbons| Primary aromatic amine,hydroxyl amine and its derived esters	15926	15013	Pos

SUB_ID 85018 (confidential)

The parent is a carbamate with N accessible, and is positive (as explained in SUB_ID 1106 Phosalone). Two metabolites are devoid of SAs, and are negative (Table H.18).

Table H.18: SUB_ID: 85018 (AS: confidential)

#	CAS	CHEMICAL STRUCTURE/SMILES	Similarity	Oncologic Primary Classif	Ames test alerts by ISS	COM_ID	SUB_ID	Ames test
1	confidential	Parent confidential	100%	Carbamate Type Compounds| Thiocarbamate Type Compounds	No alert found	75368	85018	Pos
2	No CAS number	confidential	28.60%	Not classified	No alert found	75367	85018	Neg

#	CAS	CHEMICAL STRUCTURE/SMILES	Similarity	Oncologic Primary Classif	Ames test alerts by ISS	COM_ID	SUB_ID	Ames test
3	No CAS number	confidential	26.10%	Not classified	No alert found	75369	85018	Neg

SUB_ID 1191 Metamitron

Both parent (negative) and metabolite (positive) have SA hydrazine (Table H.19). The rationalization of the difference in Ames mutagenicity is not clear.

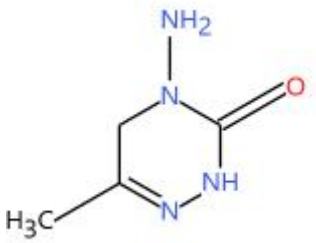
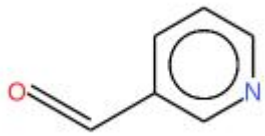
Table H.19: SUB_ID: 1191 (AS: Metamitron)

#	CAS	CHEMICAL STRUCTURE/SMILES	Similarity	Oncologic Primary Classif	Ames test alerts by ISS	COM_ID	SUB_ID	Ames test
1	41394-05-2	Parent  CC1=NN=C(C(=O)N1N)c1ccccc1	100%	Not classified	Hydrazine	1559	1191	Neg
2	No CAS	 CC1=N[N+]([O-])=C(C(=O)N1N)c1ccccc1	90.30%	Not classified	Hydrazine	50208	1191	Pos

SUB_ID 4187 Pymetrozine

Parent with hindered SA Hydrazine is negative. In the metabolite, the SA is not hindered so the metabolite is Positive. The other metabolite is negative since the group linked to the aldehyde is quite large and hinders its reactivity (Table H.20).

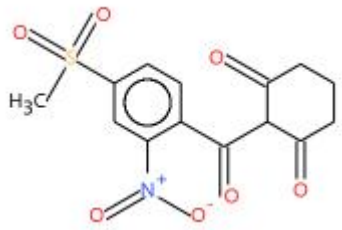
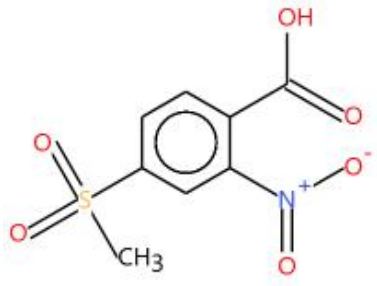
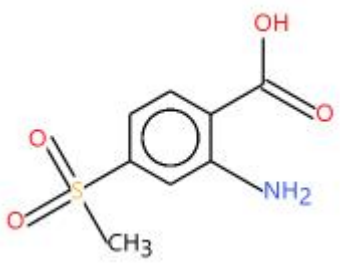
Table H.20: SUB_ID: 4187 (AS: Pymetrozine)

#	CAS	CHEMICAL STRUCTURE/SMILES	Similarity	Oncologic Primary Classif	Ames test alerts by ISS	COM_ID	SUB_ID	Ames test
1	123312-89-0	Parent  CC1CN(N=Cc2cccnc2)C(=O)NN=1	100%	Not classified	Hydrazine	6363	4187	Neg
2	No CAS number	 CC1CN(N)C(=O)NN=1	56%	Not classified	Hydrazine	50397	4187	Pos
3	No CAS number	 O=Cc1cccnc1	50%	Aldehyde Type Compounds	Simple aldehyde	50398	4187	Neg

SUB_ID 35031 Mesotrione

The parent with the Nitro aromatic SA is positive. In the two metabolites, the ISS rule base does not fire the presence of SAs, since the carboxylic moiety in ortho detoxifies both aromatic nitro and amine functionalities (Table H.21).

Table H.21: SUB_ID: 35031 (AS: Mesotrione)

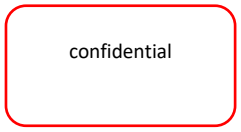
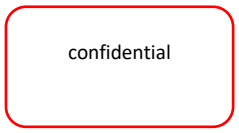
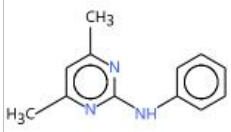
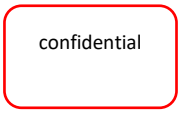
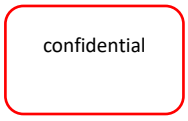
#	CAS	CHEMICAL STRUCTURE/SMILES	Similarity	Oncologic Primary Classif	Ames test alerts by ISS	COM_ID	SUB_ID	Ames test
1	104206-82-8	Parent  CS(=O)(=O)c1ccc(C(=O)C2C(=O)CCCC2=O)c(c1)[N+]([O-])=O	100%	Aromatic Amine Type Compounds	Nitro-aromatic	50309	35031	Pos
2	No CAS	 CS(=O)(=O)c1ccc(C(O)=O)c(c1)[N+]([O-])=O	71.80 %	Aromatic Amine Type Compounds	No alert found	50553	35031	Neg
3	No CAS	 CS(=O)(=O)c1ccc(C(O)=O)c(N)c1	54.10 %	Aromatic Amine Type Compounds	No alert found	50554	35031	Neg

SUB_ID 35058 confidential

The parent is a hindered aromatic amine, and is negative. Two metabolites with Nitroaromatics SA are positive (Table H.22).

The negativity of Compound 11 (with SA aromatic amine) is not easy to explain, as well as the Positivity of Compound 12, which is a hindered aromatic amine. However, the evidence on the positivity of the latter compound is relatively weak based on the EFSA genotoxicity data.

Table H.22: SUB_ID: 35058 (AS: confidential)

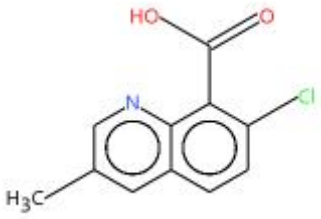
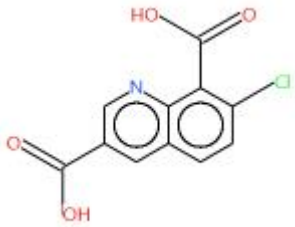
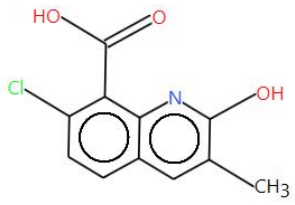
#	CAS	CHEMICAL STRUCTURE/SMILES	Similarity	Oncologic Primary Classif	Ames test alerts by ISS	COM_ID	SUB_ID	Ames test
1	conf iden	Parent 	100%	Aromatic Amine Type Compounds	No alert found	50616	35058	Neg
2	No CAS		85.7 %	Aromatic Amine Type Compounds	No alert found	50618	35058	Neg
3	53112-28-0	 Cc1cc(C)nc(Nc2ccccc2)n1	81.3 %	Aromatic Amine Type Compounds	No alert found	1689	35058	Neg
4	No CAS		81.1 %	Aromatic Amine Type Compounds	Nitro-aromatic	50622	35058	Pos
5	No CAS		80%	Aromatic Amine Type Compounds|Phenol Type Compounds	No alert found	50617	35058	Neg

#	CAS	CHEMICAL STRUCTURE/SMILES	Similarity	Oncologic Primary Classif	Ames test alerts by ISS	COM_ID	SUB_ID	Ames test
6	No CAS	confidential	75.7 %	Aromatic Amine Type Compounds	Nitro-aromatic	50621	35058	Pos
7	No CAS	confidential	74.3 %	Aromatic Amine Type Compounds	No alert found	50619	35058	Neg
8	No CAS	confidential	74.3 %	Aromatic Amine Type Compounds	No alert found	50623	35058	Neg
9	No CAS	confidential	72.2 %	Aromatic Amine Type Compounds	No alert found	50626	35058	Neg
10	No CAS	confidential	72.2 %	Aromatic Amine Type Compounds	No alert found	50624	35058	Neg
11	No CAS	confidential	64.3 %	Aromatic Amine Type Compounds	Primary aromatic amine,hydroxyl amine and its derived esters	50627	35058	Neg
12	No CAS	confidential	54.1 %	Aromatic Amine Type Compounds| Phenol Type Compounds	No alert found	50620	35058	Pos

SUB_ID 1232 Quinmerac

The parent and the two metabolites have no SAs, but metabolite 2 is reported to be positive (Table H.23). This has no straightforward explanation: however, the relative EFSA conclusion points to weak evidence of positivity (Worth et al., 2010).

Table H.23: SUB_ID: 1232 (AS: Quinmerac)

#	CAS	CHEMICAL STRUCTURE/SMILES	Similarity	Oncologic Primary Classif	Ames test alerts by ISS	COM_ID	SUB_ID	Ames test
1	90717-03-6	Parent  Cc1cnc2c(ccc(Cl)c2C(O)=O)c1	100%	Not classified	No alert found	1573	1232	Neg
2	No CAS	 OC(=O)c1cnc2c(ccc(Cl)c2C(O)=O)c1	81.30%	Not classified	No alert found	75404	1232	Pos
3	No CAS	 Cc1cc2ccc(Cl)c(C(O)=O)c2nc1O	77.40%	Not classified	No alert found	75406	1232	Neg

List of Supplementary Material

Annex 1 Curated dataset for QSAR analyses

Annex 2 Performance statistics of QSAR evaluations

Annex 3 Metabolic degradation structural changes analysis

References

Ahlberg E, Carlsson L, Boyer S, 2014. Computational derivation of structural alerts from large toxicology data sets. Journal of Chemical Information and Modeling, 54, 2945-52.

Aiba nee Kaneko M, Hirota M, Kouzuki H, Mori M, 2015. Prediction of genotoxic potential of cosmetic ingredients by an in silico battery system consisting of a combination of an expert rule-based system and a statistics-based system. Journal of Toxicological Sciences, 40, 77-98.

Amberg A, Beilke L, Bercu J, Bower D, Brigo A, Cross KP, Custer L, Dobo K, Dowdy E, Ford KA, Glowienke S, Van Gompel J, Harvey J, Hasselgren C, Honma M, Jolly R, Kemper R, Kenyon M, Kruhlak N, Leavitt P, Miller S, Muster W, Nicolette J, Plaper A, Powley M, Quigley DP, Reddy MV, Spirkl HP, Stavitskaya L, Teasdale A, Weiner S, Welch DS, White A, Wichard J, Myatt GJ, 2016. Principles and procedures for implementation of ICH M7 recommended (Q)SAR analyses. Regulatory Toxicology and Pharmacology, 77, 13-24.

Api AM, Belsito D, Bhatia S, Bruze M, Calow P, Dagli ML, Dekant W, Fryer AD, Kromidas L, La Cava S, Lalko JF, Lapczynski A, Liebler DC, Miyachi Y, Politano VT, Ritacco G, Salvito D, Schultz TW, Shen J, Sipes IG, Wall B, Wilcox DK, 2016a. RIFM fragrance ingredient safety assessment, 1-(1,2,3,4-tetrahydro-4,4-dimethyl-1-naphthyl)propan-1-one, CAS Registry Number 74499-60-8. Food and Chemical Toxicology, 97s, S101-s109.

Api AM, Belsito D, Bhatia S, Bruze M, Calow P, Dagli ML, Dekant W, Fryer AD, Kromidas L, La Cava S, Lalko JF, Lapczynski A, Liebler DC, Miyachi Y, Politano VT, Ritacco G, Salvito D, Schultz TW, Shen J, Sipes IG, Wall B, Wilcox DK, 2016b. RIFM fragrance ingredient safety assessment, 2-ethyl-1-butanol, CAS Registry Number 97-95-0. Food and Chemical Toxicology, 97s, S157-s167.

Api AM, Belsito D, Bhatia S, Bruze M, Calow P, Dagli ML, Dekant W, Fryer AD, Kromidas L, La Cava S, Lalko JF, Lapczynski A, Liebler DC, Miyachi Y, Politano VT, Ritacco G, Salvito D, Schultz TW, Shen J, Sipes IG, Wall B, Wilcox DK, 2016c. RIFM fragrance ingredient safety assessment, 3,7-dimethyl-1,6-nonadien-3-ol, CAS Registry Number 10339-55-6. Food and Chemical Toxicology, 97s, S168-s179.

Api AM, Belsito D, Bhatia S, Bruze M, Calow P, Dagli ML, Dekant W, Fryer AD, Kromidas L, La Cava S, Lalko JF, Lapczynski A, Liebler DC, Penning TM, Politano VT, Ritacco G, Salvito D, Schultz TW, Shen J, Sipes IG, Wall B, Wilcox DK, 2016d. RIFM fragrance ingredient safety assessment, 2-methylundecanol, CAS Registry Number 10522-26-6. Food and Chemical Toxicology, 97s, S119-s128.

Api AM, Belsito D, Bhatia S, Bruze M, Calow P, Dagli ML, Dekant W, Fryer AD, Kromidas L, La Cava S, Lalko JF, Lapczynski A, Liebler DC, Politano VT, Ritacco G, Salvito D, Schultz TW, Shen J, Sipes IG, Wall B, Wilcox DK, 2016e. RIFM fragrance ingredient safety assessment, p-Isopropylbenzyl acetate, CAS Registry Number 59230-57-8. Food and Chemical Toxicology, 97s, S69-s79.

Api AM, Belsito D, Bhatia S, Bruze M, Calow P, Dagli ML, Dekant W, Fryer AD, Kromidas L, La Cava S, Lapczynski A, Liebler DC, O'Brien D, Parakhia R, Penning TM, Politano VT, Ritacco G, Salvito D, Schultz TW, Shen J, Sipes IG, Wall B, Wilcox DK, 2017. RIFM fragrance ingredient safety assessment, isobornyl isovalerate, CAS registry number 7779-73-9. Food and Chemical Toxicology, 110 Suppl 1, S1-s8.

Araya S, Lovsin-Barle E, Glowienke S, 2015. Mutagenicity assessment strategy for pharmaceutical intermediates to aid limit setting for occupational exposure. Regulatory Toxicology and Pharmacology, 73, 515-20.

Bakhtyari NG, Raitano G, Benfenati E, Martin T, Young D, 2013. Comparison of in silico models for prediction of mutagenicity. Journal of Environmental Science and Health. Part C: Environmental Carcinogenesis & Ecotoxicology Reviews, 31, 45-66.

Barber C, Amberg A, Custer L, Dobo KL, Glowienke S, Van Gompel J, Gutsell S, Harvey J, Honma M, Kenyon MO, Kruhlak N, Muster W, Stavitskaya L, Teasdale A, Vessey J, Wichard J, 2015. Establishing best practise in the application of expert review of mutagenicity under ICH M7. Regulatory Toxicology and Pharmacology, 73, 367-77.

Barber C, Cayley A, Hanser T, Harding A, Heghes C, Vessey JD, Werner S, Weiner SK, Wichard J, Giddings A, Glowienke S, Parenty A, Brigo A, Spirkl HP, Amberg A, Kemper R, Greene N, 2016. Evaluation of a statistics-based Ames mutagenicity QSAR model and interpretation of the results obtained. Regulatory Toxicology and Pharmacology, 76, 7-20.

Benfenati E, Belli M, Borges T, Casimiro E, Cester J, Fernandez A, Gini G, Honma M, Kinzl M, Knauf R, Manganaro A, Mombelli E, Petoumenou MI, Paparella M, Paris P, Raitano G, 2016. Results of a round-robin exercise on read-across. SAR and QSAR in Environmental Research, 27, 371-84.

Benigni R, Bossa C, 2011. Mechanisms of chemical carcinogenicity and mutagenicity: a review with implications for predictive toxicology. Chemical Reviews, 111, 2507-36.

Benigni R, Bossa C, Richard AM, Yang C, 2008. A novel approach: chemical relational databases, and the role of the ISSCAN database on assessing chemical carcinogenicity. Annali dell'Istituto Superiore di Sanità, 44, 48-56.

Benigni R, Netzeva TI, Benfenati E, Bossa C, Franke R, Helma C, Hulzebos E, Marchant C, Richard A, Woo YT, Yang C, 2007. The expanding role of predictive toxicology: an update on the (Q)SAR models for mutagens and carcinogens. Journal of Environmental Science and Health. Part C: Environmental Carcinogenesis & Ecotoxicology Reviews, 25, 53-97.

Benigni R, Worth A, Netzeva T, Jeliazkova N, Bossa C, Gruska A, Franke R, 2009. Structural motifs modulating the carcinogenic risk of aromatic amines. Environmental and Molecular Mutagenesis, 50, 152-61.

Beynon M, Curry B, Morgan P, 2000. The Dempster–Shafer theory of evidence: an alternative approach to multicriteria decision modelling. Omega, 28, 37 - 50.

Blackburn K, Stuard SB, 2014. A framework to facilitate consistent characterization of read across uncertainty. Regulatory Toxicology and Pharmacology, 68, 353-62.

Cherkasov A, Muratov EN, Fourches D, Varnek A, Baskin II, Cronin MT, Dearden JC, Gramatica P, Martin YC, Todeschini R, Consonni V, Kuzmin VE, Cramer R, Benigni R, Yang C, Rathman JF, Terfloth L, Gasteiger J, Richard AM, Tropsha A, 2014. QSAR modeling: Where have you been? Where are you going to? Journal of Medicinal Chemistry, 57, 4977-5010.

Contrera JF, 2013. Validation of Toxtree and SciQSAR in silico predictive software using a publicly available benchmark mutagenicity database and their applicability for the qualification of impurities in pharmaceuticals. Regulatory Toxicology and Pharmacology, 67, 285-93.

Dearden JC, Cronin MT, Kaiser KL, 2009. How not to develop a quantitative structure-activity or structure-property relationship (QSAR/QSPR). SAR and QSAR in Environmental Research, 20, 241-66.

ECHA, 2017a. Guidance on information requirements and chemical safety assessment. ECHA-17-G-11-EN https://echa.europa.eu/documents/10162/13632/information_requirements_r7c_en.pdf

ECHA, 2017b. Read-across Assessment Framework (RAAF). ECHA-17-R-01-EN https://echa.europa.eu/documents/10162/13628/raaf_en.pdf.

EFSA (European Food Safety Authority), 2009a. Conclusion on the peer review of the pesticide risk assessment of the active substance Carbosulfan. EFSA Journal 2009; 7(10):1354. 10.2903/j.efsa.2009.1354

EFSA (European Food Safety Authority), 2009b. Conclusion regarding the peer review of the pesticide risk assessment of the active substance carbofuran. EFSA Scientific Report. EFSA Journal 2009, 310, pp. 1 -132. 10.2903/j.efsa.2009.310r

EFSA (European Food Safety Authority), 2009c. Peer review of the pesticide risk assessment of the active substance Benfuracarb. EFSA Scientific Report. EFSA Journal 2009, 239, pp. 1 - 107. 10.2903/j.efsa.2009.239r

EFSA (European Food Safety Authority), 2010. Conclusion on the peer review of the pesticide risk assessment of the active substance quinmerac. EFSA Journal 2010; 8(3):1523, pp. 1 - 68. 10.2903/j.efsa.2010.1523

EFSA (European Food Safety Authority), 2012. Scientific Opinion on Evaluation of the Toxicological Relevance of Pesticide Metabolites for Dietary Risk Assessment. EFSA Journal 2012;10(7):2799. 10.2903/j.efsa.2012.2799

EFSA (European Food Safety Authority), 2016. Guidance on the establishment of the residue definition for dietary risk assessment. EFSA Journal 2016;14(12):4549. 10.2903/j.efsa.2016.4549

EFSA (European Food Safety Authority), 2017. Guidance on the use of the weight of evidence approach in scientific assessments. EFSA Journal 2017;15(8):4971. 10.2903/j.efsa.2017.4971

Feng J, Lurati L, Ouyang H, Robinson T, Wang Y, Yuan S, Young SS, 2003. Predictive toxicology: benchmarking molecular descriptors and statistical methods. Journal of Chemical Information and Computer Sciences, 43, 1463-70.

Ferrari T, Cattaneo D, Gini G, Golbamaki Bakhtyari N, Manganaro A, Benfenati E, 2013. Automatic knowledge extraction from chemical structures: the case of mutagenicity prediction. SAR and QSAR in Environmental Research, 24, 365-83.

Fioravanzo E, Bassan A, Pavan M, Mostrag-Szlichtyng A, Worth AP, 2012. Role of in silico genotoxicity tools in the regulatory assessment of pharmaceutical impurities. SAR and QSAR in Environmental Research, 23, 257-77.

Fitzpatrick RB, 2008. CPDB: Carcinogenic Potency Database, Medical Reference Services Quarterly, 27:3, 303-311, DOI: 10.1080/02763860802198895.

Ford KA, Ryslik G, Chan BK, Lewin-Koh SC, Almeida D, Stokes M, Gomez SR, 2017. Comparative evaluation of 11 in silico models for the prediction of small molecule mutagenicity: role of steric hindrance and electron-withdrawing groups. Toxicology Mechanisms and Methods, 27, 24-35.

Franke R, Gruska A, 2003. General introduction to QSAR., Vol., CRC Press, Boca Raton.

Fukuto TR, 1990. Mechanism of action of organophosphorus and carbamate insecticides. Environmental Health Perspectives, 87, 245-54.

Greene N, Dobo KL, Kenyon MO, Cheung J, Munzner J, Sobol Z, Sluggett G, Zelesky T, Sutter A, Wichard J, 2015. A practical application of two in silico systems for identification of potentially mutagenic impurities. Regulatory Toxicology and Pharmacology, 72, 335-49.

Hansch C, Hoekman D, Leo A, Weininger D, Selassie CD, 2002. Chem-bioinformatics: Comparative QSAR at the interface between chemistry and biology. Chemical Reviews, 102, 783-812.

Hansch C, Leo A, 1995. Exploring QSAR. 1. Fundamentals and applications in chemistry and biology., Vol. 1, American Chemical Society, Washington, D.C.

Hansen K, Mika S, Schroeter T, Sutter A, ter Laak A, Steger-Hartmann T, Heinrich N, Muller KR, 2009. Benchmark data set for in silico prediction of Ames mutagenicity. Journal of Chemical Information and Modeling, 49, 2077-81.

Hartung T, 2016. Making big sense from big data in toxicology by read-across. Altex, 33, 83-93.

Helma C, Cramer T, Kramer S, De Raedt L, 2004. Data mining and machine learning techniques for the identification of mutagenicity inducing substructures and structure activity relationships of noncongeneric compounds. Journal of Chemical Information and Computer Sciences, 44, 1402-11.

Hillebrecht A, Muster W, Brigo A, Kansy M, Weiser T, Singer T, 2011. Comparative evaluation of in silico systems for ames test mutagenicity prediction: scope and limitations. Chemical Research in Toxicology, 24, 843-54.

Honma M, Kitazawa A, Cayley A, Williams RV, Barber C, Hanser T, Saiakhov R, Chakravarti S, Myatt GJ, Cross KP, Benfenati E, Raitano G, Mekenyan O, Petkov P, Bossa C, Benigni R, Battistelli CL, Giuliani A, Tcheremenskaia O, DeMeo C, Norinder U, Koga H, Jose C, Jeliazkova N, Kochev N, Paskaleva V, Yang C, Daga PR, Clark RD, Rathman J, 2018. Improvement of quantitative structure-activity relationship (QSAR) tools for predicting Ames mutagenicity: outcomes of the Ames/QSAR International Challenge Project. Mutagenesis.

Jolly R, Ahmed KB, Zwickl C, Watson I, Gombar V, 2015. An evaluation of in-house and off-the-shelf in silico models: implications on guidance for mutagenicity assessment. Regulatory Toxicology and Pharmacology, 71, 388-97.

Judson PN, Cooke PA, Doerrer NG, Greene N, Hanzlik RP, Hardy C, Hartmann A, Hinchliffe D, Holder J, Muller L, Steger-Hartmann T, Rothfuss A, Smith M, Thomas K, Vessey JD, Zeiger E, 2005. Towards the creation of an international toxicology information centre. Toxicology, 213, 117-28.

Kamath P, Raitano G, Fernandez A, Rallo R, Benfenati E, 2015. In silico exploratory study using structure-activity relationship models and metabolic information for prediction of mutagenicity based on the Ames test and rodent micronucleus assay. SAR and QSAR in Environmental Research, 26, 1017-1031.

Kazius J, McGuire R, Bursi R, 2005. Derivation and Validation of Toxicophores for Mutagenicity Prediction. Journal of Medical Chemistry, 48, 312-320.

Kubinyi H, 2005. Validation and predictivity of QSAR models., Vol., CADDS, Ankara.

Lhasa, 2014. Vitic Nexus http://www.lhasalimited.org/products/vitic-nexus.htm.

Liew CY, Pan C, Tan A, Ang KX, Yap CW, 2012. QSAR classification of metabolic activation of chemicals into covalently reactive species. Molecular Diversity, 16, 389-400.

Low Y, Sedykh A, Fourches D, Golbraikh A, Whelan M, Rusyn I, Tropsha A, 2013. Integrative chemical-biological read-across approach for chemical hazard classification. Chemical Research in Toxicology, 26, 1199-208.

Madden JC, 2013. Tools for Grouping Chemicals and Forming Categories, Vol. 17, Royal Society of Chemistry, London.

Metivier JP, Lepailleur A, Buzmakov A, Poezevara G, Cremilleux B, Kuznetsov SO, Le Goff J, Napoli A, Bureau R, Cuissart B, 2015. Discovering structural alerts for mutagenicity using stable emerging molecular patterns. Journal of Chemical Information and Modeling, 55, 925-40.

Miller EC, Miller JA, 1981a. Mechanisms of chemical carcinogenesis. Cancer, 47, 1055-1064.

Miller EC, Miller JA, 1981b. Searches for ultimate chemical carcinogens and their reactions with cellular macromolecules. Cancer, 47, 2327-2345.

Modi S, Li J, Malcomber S, Moore C, Scott A, White A, Carmichael P, 2012. Integrated in silico approaches for the prediction of Ames test mutagenicity. Journal of Computer-Aided Molecular Design, 26, 1017-33.

Mombelli E, Raitano G, Benfenati E, 2016. In Silico Prediction of Chemically Induced Mutagenicity: How to Use QSAR Models and Interpret Their Results. Methods in Molecular Biology, 1425, 87-105.

NAFTA, 2012. Integrated Approaches to Testing and Assessment (IATA and (Q)SAR-(Quantitative) Structure Activity Relationship ((Q)SAR) Guidance Document https://archive.epa.gov/pesticides/news/web/html/qsar-guidance.html North American Free Trade Agreement Technical Working Group.

OECD, 2017a. Fundamental And Guiding Principles For (Q)SAR Analysis Of Chemical Carcinogens with Mechanistic Considerations, OECD Series on Testing and Assessment, No. 229, OECD Publishing, Paris, https://doi.org/10.1787/9789264274792-en. Series on testing and Assessment.

OECD, 2017b. Guidance on Grouping of Chemicals, Second Edition, OECD Series on Testing and Assessment, No. 194, OECD Publishing, Paris, https://doi.org/10.1787/9789264274679-en. OECD Series on Testing and Assessment.

Ono A, Takahashi M, Hirose A, Kamata E, Kawamura T, Yamazaki T, Sato K, Yamada M, Fukumoto T, Okamura H, Mirokuji Y, Honma M, 2012. Validation of the (Q)SAR combination approach for mutagenicity prediction of flavor chemicals. Food and Chemical Toxicology, 50, 1538-46.

Patlewicz G, Ball N, Booth ED, Hulzebos E, Zvinavashe E, Hennes C, 2013. Use of category approaches, read-across and (Q)SAR: general considerations. Regulatory Toxicology and Pharmacology, 67, 1-12.

Patlewicz G, Fitzpatrick JM, 2016. Current and Future Perspectives on the Development, Evaluation, and Application of in Silico Approaches for Predicting Toxicity. Chemical Research in Toxicology, 29, 438-51.

Patlewicz G, Helman G, Pradeep P, Shah I, 2017. Navigating through the minefield of read-across tools: A review of in silico tools for grouping. Computational Toxicology, 3, 1 -18.

Piegorsch WW, Zeiger E, 1991. Measuring intra-assay agreement for the Ames Salmonella assay., Vol., Springer-Verlag, Heidelberg.

Rathman J, Yang C, Zhou H, 2018. Dempster-Shafer theory for combining in silico evidence and estimating uncertainty in chemical risk assessment. Computational Toxicology, 6, 16 - 31.

Richard AM, Judson RS, Houck KA, Grulke CM, Volarath P, Thillainadarajah I, Yang C, Rathman J, Martin MT, Wambaugh JF, Knudsen TB, Kancherla J, Mansouri K, Patlewicz G, Williams AJ, Little SB, Crofton KM, Thomas RS, 2016. ToxCast Chemical Landscape: Paving the Road to 21st Century Toxicology. Chemical Research in Toxicology, 29, 1225-51.

Ridder L, Wagener M, 2008. SyGMa: combining expert knowledge and empirical scoring in the prediction of metabolites. ChemMedChem, 3, 821-32.

Rogers D, Hahn M, 2010. Extended-connectivity fingerprints. Journal of Chemical Information and Modeling, 50, 742-54.

Rybacka A, Ruden C, Andersson PL, 2014. On the use of in silico tools for prioritising toxicity testing of the low-volume industrial chemicals in REACH. Basic & Clinical Pharmacology & Toxicology, 115, 77-87.

Schultz TW, Cronin MTD, 2017. Lessons learned from read-across case studies for repeated-dose toxicity. Regulatory Toxicology and Pharmacology, 88, 185-191.

Skare JA, Blackburn K, Wu S, Re TA, Duche D, Ringeissen S, Bjerke DL, Srinivasan V, Eisenmann C, 2015. Use of read-across and computer-based predictive analysis for the safety assessment of PEG cocamines. Regulatory Toxicology and Pharmacology, 71, 515-28.

Snyder RD, 2010. Possible structural and functional determinants contributing to the clastogenicity of pharmaceuticals. Environmental and Molecular Mutagenesis, 51, 800-14.

Snyder RD, Holt PA, Maguire JM, Trent JO, 2013. Prediction of noncovalent Drug/DNA interaction using computational docking models: studies with over 1350 launched drugs. Environmental and Molecular Mutagenesis, 54, 668-81.

Swets JA, 1988. Measuring the accuracy of diagnostic systems. Science, 240, 1285-1293.

Valencia A, Prous J, Mora O, Sadrieh N, Valerio LG, Jr., 2013. A novel QSAR model of Salmonella mutagenicity and its application in the safety assessment of drug impurities. Toxicology and Applied Pharmacology, 273, 427-34.

Wichard JD, 2017. In silico prediction of genotoxicity. Food and Chemical Toxicology, 106, 595-599.

Worth A, Fuart-Gatnik M, Lapenna S, Lo Piparo E, Mostrag-SZzlichtying A, Serafimova R, 2010. Applicability of QSAR analysis to the evaluation of the toxicological relevance of metabolites and degradates of pesticide active substances for dietary risk assessment. EFSA External Scientific Report. European Commission Joint Research Centre, Institute for Health & Consumer Protection, Ispra, Italy https://https://doi.org/10.2903/sp.efsa.2010.EN-50.

Wu S, Blackburn K, Amburgey J, Jaworska J, Federle T, 2010. A framework for using structural, reactivity, metabolic and physicochemical similarity to evaluate the suitability of analogs for SAR-based toxicological assessments. Regulatory Toxicology and Pharmacology, 56, 67-81.

Xu C, Cheng F, Chen L, Du Z, Li W, Liu G, Lee PW, Tang Y, 2012. In silico prediction of chemical Ames mutagenicity. Journal of Chemical Information and Modeling, 52, 2840-7.

Yang C, Tarkhov A, Marusczyk J, Bienfait B, Gasteiger J, Kleinoeder T, Magdziarz T, Sacher O, Schwab CH, Schwoebel J, Terfloth L, Arvidson K, Richard A, Worth A, Rathman J, 2015. New publicly available chemical query language, CSRML, to support chemotype representations for application to data mining and modeling. Journal of Chemical Information and Modeling, 55, 510-28.