

Research Article

Integrated Strategy for Mutagenicity Prediction Applied to Food Contact Chemicals

Serena Manganelli¹, Benoît Schilter², Emilio Benfenati¹, Alberto Manganaro^{1,3} and Elena Lo Piparo²

¹Laboratory of Environmental Chemistry and Toxicology, IRCCS-Istituto di Ricerche Farmacologiche Mario Negri, Milano, Italy; ²Chemical Food Safety Group, Nestlé Research Center, Lausanne, Switzerland; ³Kode s.r.l., Pisa, Italy

Summary

Food contamination due to unintentional leakage of chemicals from food contact materials (FCM) is a source of increasing concern. Since for many of these substances only limited or no toxicological data are available, the development of alternative methodologies to rapidly and cost-efficiently establish the level of safety concern is critical to ensure adequate consumer protection. Computational toxicology methods are considered the most promising solutions to cope with this data gap. In particular, mutagenicity assessment has a particular relevance and is a mandatory requirement for all substances released from plastic FCM, regardless how low migration and exposure are. In the present work, a strategy integrating a number of (Quantitative) Structure Activity Relationship ((Q)SAR) models for Ames mutagenicity predictions is proposed. A list of chemicals representing moieties likely migrating from FCM was selected to test the value of the newly defined strategy and the possibility to combine predictions given by the different algorithms was evaluated. In particular, a scheme to integrate mutagenicity estimations into a single final assessment was developed, resulting in an increased domain of applicability. In most cases, a deeper analysis of experimental data, where available, allowed fixing misclassification errors, highlighting the importance of data curation in the development, validation and application of *in silico* methods. The high accuracy of the strategy provided the rationale for its application for toxicologically uncharacterized chemicals. Finally, the overall strategy of integration will be automated through its implementation into a freely available software application.

Keywords: FCM migrating substances, integration, mutagenicity, *in silico* models, (Q)SAR

1 Introduction

Food contact materials (FCM) are materials and articles that are intended to come into contact with food during its production, processing, storage, preparation and serving, before its consumption (EFSA, 2015). Amongst others, these include plastics, paper and board, glass, metal coatings, printing inks and adhesives (Van Bossuyt et al., 2016).

The Framework Regulation (EC) No 1935/2004 includes general requirements for all FCM (EC, 2004), but only few harmonized legislations exist for specific types of FCM, such as the EU Regulation 1282/2011 on plastic materials (EC, 2011). The Framework Regulation states that FCM should be sufficiently chemically inert so that they do not release their

constituents into food in amounts that could endanger consumer health and food quality. Regarding migrating substances, a distinction is made between those having a technological function in the manufacturing of the FCM (the intentionally added substances, IAS) and those originating from impurities in raw materials and from the reaction and degradation of substances in the intended use (the so-called non-intentionally added substances, NIAS). IAS are, in general, toxicologically well studied and can be assessed using standard risk assessment. In contrast, for most NIAS no toxicological data are available and risk assessment is therefore not straightforward. Consequently, there are significant uncertainties regarding the safety of NIAS, triggering increasing public, scientific and regulatory concern (Van Bossuyt et al., 2017).

Received July 17, 2017;
Accepted September 11, 2017;
Epub September 18, 2017;
doi:10.14573/altex.1707171



This is an Open Access article distributed under the terms of the Creative Commons Attribution 4.0 International license (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution and reproduction in any medium, provided the original work is appropriately cited.



Taking into account the time constraint and the pressure to avoid the use of laboratory animals, the development of alternative methodologies to establish a rapid and cost-efficient level of safety concern of identified NIAS appears critical to ensure adequate consumer protection without undue over-conservatism. In this context, computational toxicology methods are recognized as the most promising solutions and are increasingly applied by academic and regulatory scientists (Benfenati et al., 2009; JRC, 2010). *In silico* methods have been most prominently promoted by the European Registration, Evaluation, Authorization and Restriction of Chemicals (REACH, EC 1907/2006) regulation on chemicals (Cassano et al., 2014). They are successfully employed for early identification of toxicological hazard in other regulatory frameworks, such as in the qualification of potentially genotoxic impurities in drug substances (ICH, 2017; Sutter et al., 2013) to limit potential carcinogenic risk.

In the food context, the most commonly applied method to establish the level of safety concern of chemicals in the absence of experimental data has been the threshold of toxicological concern (TTC) (Kroes et al., 2004). More recently, the use of computational models has been highlighted (Van Bossuyt et al., 2017; Schilter et al., 2014). The first and common step of these approaches is the identification of a possible structural alert for genotoxicity. Definitely, the genotoxicity and, more specifically, the mutagenicity endpoint has a particular relevance due to the theoretical lack of threshold of effect that this category of chemicals exhibits. For IAS, genotoxicity data are always requested, regardless of the estimated migration level and resulting exposure (EFSA, 2012). In the frame of risk assessment, the hazard identification step considers genotoxicity and mutagenicity via direct DNA reactivity as the default assumption in the absence of sufficient data to the contrary (Schilter et al., 2014; Jacobs et al., 2015). It is commonly accepted that DNA-reactive mutagenic agents do not exhibit a dose below which no effect is anticipated. Even if it is very well known that all mutagens are genotoxic, however, not all genotoxic substances are mutagenic. The bacterial reverse mutation assay (Ames test) is considered a reliable predictor of genotoxic potential (Schilter et al., 2014), and is the most common *in vitro* test to detect gene mutations (OECD, 1997). Also, mutagenicity is one of the most modelled endpoints due to the quantity and quality of experimental data available. This is the reason why we addressed this specific end-point in our study. The main categories of *in silico* methods for the prediction of mutagenic potential of chemicals are (Q) SAR models, based on numerical descriptors, rule-based expert systems, making use of structural alerts associated with adverse outcomes, and hybrid models combining both approaches (Benfenati et al., 2016; Mombelli et al., 2016).

The integration of models based on complementary algorithms, i.e., statistical and structure activity relationship (SAR) based, is often a mandatory requirement, e.g., for the genotoxicity qualification of pharmaceutical impurities (ICH, 2017). Up to now, different strategies combining a number of (Q) SAR models for predicting Ames mutagenicity have been proposed, moving towards a weight of evidence (WoE) approach. In general, such combinations resulted in improved predictions

over single models (Amaury et al., 2007; Cassano et al., 2014; Kulkarni et al., 2016; Manganelli et al., 2016; Mazzatorta et al., 2007). Recently, EFSA's Scientific Committee has developed a guidance document promoting the use of the WoE in toxicological assessments combining both qualitative and quantitative approaches (EFSA, 2017). The EFSA guidance proposes a strategy for assembling, weighing and integrating different lines of evidence from testing and non-testing methods (NTM) and defines reliability, relevance and consistency in terms of their contribution to the overall assessment. It also provides an example of the use of NTM within a WoE framework, which proposes the integration of a number of (Q)SAR models and read-across for mutagenicity estimations. This example highlights strengths and weaknesses of both methods, which may vary on a case-by-case basis.

Aside from integration, automation allows quick and efficient analysis of chemicals for activity across a battery of *in silico* methods. This is useful for rapidly evaluating large numbers of chemicals. It is commonly realized with the support of pipeline tools, e.g., KNIME and Pipeline Pilot (Warr et al., 2012). Besides saving time, the automation process also offers the advantage of reducing the inconsistencies and errors due to the manual building, validation and application of *in silico* methods (Cox et al., 2013; Dixon et al., 2016; Romano, 2008; Zhang et al., 2006). In the case of read-across, the automation of the key steps (e.g., data search methods that use similarity measures and fragment search) allows overcoming one of the main drawbacks of this method, i.e., the lack of reproducibility (Benfenati, 2016; Gini et al., 2014).

In this context, the present work was aimed at drawing up an automated strategy for integrating a number of (Q)SAR models for Ames mutagenicity predictions applicable to large sets of compounds. The dataset compiled by Price et al. (2014), containing a list of substances migrating from plastic FCM isolated from the FACET dataset (Hearty et al., 2011) plus their mutagenic analogues, was selected as a good candidate to develop, validate and test the approach. In our study, these chemicals were assembled and processed by using three *in silico* (Q)SAR consensus models for mutagenicity. Hence, a scheme to integrate mutagenicity estimations into a single final assessment was defined and applied to toxicologically uncharacterized FCM chemicals. Finally, the overall strategy of integration will be automated through its implementation into a freely available software application.

2 Material and methods

2.1 Chemical structures

For the present work, a list of 183 chemicals obtained from the database identified by Price et al. (2014) was used. It includes substances migrating from plastic FCMs from the FACET dataset (Hearty et al., 2011) and mutagenic structural analogues. Indeed, all compounds used in plastic food packaging go through a rigorous assessment by expert panels, so experimental data, where available, were mainly non-mutagenic. Thus, in order

to validate our consensus approach, we also considered mutagenic analogues of the FCM chemicals as a positive control in accordance with Price et al. (2014). Overall, experimental data referring to Ames mutagenicity were available for 97 (29 mutagenic and 68 non-mutagenic) of the 183 chemicals. The remaining 86 substances were toxicologically uncharacterized for this endpoint and the developed approach was applied to the screening of potential mutagenic compounds.

In detail, chemicals' curation was performed as follows:

- Name to structure conversion was executed using Marvin View¹ / JChem² for parent FCM (if available as single substances), migrants and structural analogues. Parent compounds existing as mixtures, oligomers and polymers were also identified and converted into structures (i.e., single constituents, monomers), partially with the help of chemical databases such as ChemSpider³, ChemIDplus⁴ and PubChem⁵.
- SMILES (Weininger et al., 1988) generated by these tools were compared against the original ones from Price et al. (2014).
- An in-house software (Floris et al., 2014) was used to identify and then remove duplicates within parents, migrants and analogues.
- Canonical SMILES were obtained using the istMolBase software⁶ (Kode, 2013), based on the VEGA core libraries and Chemistry Development Kit libraries (CDK) (Benfenati et al., 2015).

The final list of 183 chemicals with names and structures, and the related SMILES is reported in Table S1⁷.

2.2 Data curation

In the curation process, some of the experimental values were modified and new ones were introduced. Indeed, some gaps emerged from the analysis of further experimental sources and/or from database and literature updates. The single models in part refer to experimental data which often do not include the complete set of the strains used according to official protocols. However, we believe that combining so many substances (thousands) at the basis of each model, and also the different models together, will cover the data gaps for certain strains on individual substances. Most important when combining different predictions is consistency in the endpoint selection (e.g., not mixing the genotoxicity with mutagenicity endpoints). Since the databases of the models used in this study were built using experimental data from the Ames test, we only considered data for mutagenicity obtained with this assay, even if other *in vitro* or *in vivo* tests gave different results for the substances under examination. This is necessary to allow a fair comparison between

experimental values and estimations. For example, 2-isopropyl thioxanthone (ITX) was classified *in vitro* as experimentally equivocal/mutagenic. Based on *in vivo* experimental data, ITX is generally considered non-genotoxic. However, non-genotoxic experimental data did not refer to the Ames results, but to *in vivo* tests. Indeed, EFSA states that ITX induced a borderline increase of revertant colonies in a bacterial reversion test and was inactive in adequate genotoxicity tests in liver and bone marrow (EFSA, 2005).

A further search for experimental data on the other chemical compounds led to assignment of new activity labels. For example, 2-ethylhexyl-4-dimethylaminobenzoate (EHDAB) was classified as non-mutagenic based on expert assessment (EFSA, 2005) reporting no evidence of genotoxicity in the standard Ames test for this substance. 2,2,3-trifluoro-3-(trifluoromethyl) oxirane was reclassified as non-mutagenic, as reported in the ECHA CHEM database⁸ with reliability 1 according to the Klimisch score (Klimisch et al., 1997). In ECHA CHEM database⁷ we also recovered non-mutagenic results to bacteria for bis(2,6-diisopropylphenyl)carbodiimide and 4-(4-methylphenylthio) benzophenone. Moreover, we labeled 2,2'-bisphenol F as non-mutagenic based on the opinion drawn up by EFSA (2009) on the substance "bis(hydroxyphenyl)methane" (bisphenol F), which is a mixture of the 2,2'-, 2,4'- and 4,4'-isomers.

Assessing borderline substances in more detail, we reclassified two experimentally equivocal compounds, methyl methacrylate (MMA) (EFSA, 2008) and propionic acid (EFSA, 2014), as negative based on registration data reported in the ECHA CHEM database⁷ with reliability 1 according to the Klimisch score (Klimisch et al., 1997).

Moreover, analogues were examined to have a positive control and to increase the number of experimental values. ChemIDPlus/Toolbox (OECD, 2017) database matches provided positive mutagenic values for most of the analogues and negative data for two of them; the two non-mutagenic compounds for which all of the Ames assays were conducted according to OECD TG 471 (OECD, 1997) with and without metabolic activation were included in the final list.

Finally, mutagenicity measured values were available for about half of the compounds in the dataset (97 compounds). All details are reported in Table S2⁹. Therefore, we used them to validate the predictions of the three-consensus models moving towards a weight of evidence approach.

2.3 Consensus models

All chemicals in the dataset were then processed using the following battery of models: Robust hybrid classifier (RHC),

¹ MarvinView 16.1.18.0, <http://www.chemaxon.com> (accessed July 2017)

² JChem for Office 16.5.1600.806, <http://www.chemaxon.com> (accessed July 2017)

³ <http://www.chemspider.com/> (accessed July 2017)

⁴ <https://chem.nlm.nih.gov/chemidplus/> (accessed July 2017)

⁵ <https://pubchem.ncbi.nlm.nih.gov/> (accessed July 2017)

⁶ IstMolBase v.1.0.2, <https://www.kode-solutions.net/> (accessed July 2017)

⁷ doi:10.14573/altex.1707171s1

⁸ <http://echa.europa.eu/web/guest/information-on-chemicals/registered-substances> (accessed July 2017)

⁹ doi:10.14573/altex.1707171s2



VEGA and T.E.S.T. consensus models, each based on the combination of different algorithms. Hence, predictions were evaluated based on information on the applicability domain and reliability of each model, and the related compounds used to build the model. A brief description of the mutagenicity models and of the parameters considered to assign reliabilities to their predictions is provided below.

Robust hybrid classifier (RHC)

The RHC model, developed by Mazzatorta et al. (2007), integrates (i) the Structural Alerts model (SA_m), including the list of improved structural alerts (SA) gathered by Kazius et al. (2005), and (ii) the Artificial Intelligence model (AI_m), which is a modified *k*-nearest neighbor based on the LAZAR system developed by Helma (2004) (Mazzatorta et al., 2007). The training set of 4,337 substances used for building RHC was collected by Kazius et al. (2005), and the test set of 753 chemicals used for its validation was assembled and curated by Young et al. (2002), as described in detail by Mazzatorta et al. (2007).

RHC returns the Ames prediction together with a confidence level, which depends on the ratio between the number of mutagens containing a given toxicophore and the total number of compounds in the test set with that moiety and takes into account the error associated with the prediction of each SA (Mazzatorta et al., 2007). If both models predict the compound as non-mutagenic, RHC considers it negative with a confidence equal to 0.85, which refers to the overall specificity of the system; if there is a convergence regarding the mutagenicity, RHC considers it as mutagenic and the confidence is equal to the sensitivity of RHC weighted by the product of the individual error associated with the SAs present in the compound. In case of non-consensus prediction, SA_m prevails, because it is based on well-documented experimental evidence and has a superior accuracy, but the confidence of the prediction is accordingly lowered.

Based on criteria chosen by Mazzatorta et al. (2007) to define different levels of confidence, 0.65 was chosen as cutoff value for prediction reliability; estimations with a confidence level greater than or equal to 0.65 were considered reliable, otherwise they were associated with low reliability. The model does not indicate if the predicted chemical is included in its training/test sets.

VEGA consensus model

The VEGA consensus model¹⁰ integrates predictions from the following (Q)SAR models:

- CAESAR, which integrates a support vector machine (SVM) algorithm coupled with two sets of structural alerts aimed to reduce the false negative rate (Ferrari and Gini, 2010);
- SARpy (SAR in python), which extracts a set of structural alerts related to a specific activity from data without any *a priori* knowledge (Ferrari et al., 2013);

- ISS-VEGA, which is based on a series of rules defined by Benigni and Bossa detecting mutagenic chemicals originally implemented within the Toxtree application (Benigni et al., 2008; Benigni and Bossa, 2011);
- *k*-NN, which performs a *k*-nearest neighbors with a weighted integration of the experimental values of the four chemicals most similar to the target (Manganaro et al., 2016).

The CAESAR and SARpy VEGA models were developed based on 4,204 chemicals extracted from the Bursi dataset (Hansen et al., 2009). The *k*-NN VEGA model was built on a dataset of 5,770 chemicals from the Hansen dataset (Hansen et al., 2009) and from data produced within the Ames QSAR project organized by the National Institute of Health Sciences of Japan. The training set of the ISS VEGA model was extracted from the Toxtree software (v 2.6), and consists of 670 compounds.

Predictions from single VEGA models are associated with three possible levels of reliability, based on the definition of their applicability domains: low, moderate and high. The consensus algorithm gives toxicity estimations based on these levels of reliability. It also assigns a numerical score (ranging from 0 to 1) to each estimation, which depends on the number of convergent predictions and on their reliability. If experimental value is provided (because the target molecule has been found in the training/test set of a model) at least by one model, it is kept as final consensus result.

In our evaluation, we considered predictions with a consensus score higher than 0.3 as reliable, else we assigned low reliability to them. Indeed, the cutoff value of 0.3 was able to discard consensus estimations based on the prevalence of predictions associated with low reliability.

We preferred the consensus to the single VEGA models to estimate mutagenicity since its algorithm produces a final assessment influenced by the most reliable individual predictions.

T.E.S.T. consensus model

T.E.S.T.¹¹ estimates Ames mutagenicity using four QSAR methods: the hierarchical method, the Food and Drug Administration (FDA) method, the nearest neighbor method and the consensus method. The consensus method takes an average of the predicted toxicities from the above QSAR methods (taking into account the AD of each method). The dataset of T.E.S.T. mutagenicity models is taken from the Hansen dataset (Hansen et al., 2009). T.E.S.T. provides continuous prediction values to be interpreted as follows:

$$Class = \begin{cases} \text{Mutagenic} & \geq 0.5 \\ \text{Non-mutagenic} & < 0.5 \end{cases}$$

We used this value as an indicator for predictive relevance, assigning the highest uncertainty to predictions equal or close to the 0.5 cutoffs. We considered predicted values greater than

¹⁰ VEGA v1.4. <https://www.vegahub.eu/> (accessed July 2017)

¹¹ T.E.S.T. (Toxicity Estimation Software Tool) v4.2.1, <https://www.epa.gov/chemical-research/toxicity-estimation-software-tool-test> (accessed July 2017)

0.7 as mutagenic with high reliability and lower than 0.3 as non-mutagenic with high reliability. Prediction values between 0.3 and 0.7 were considered uncertain. In this case, if the experimental value was present in the model's dataset, it superseded the predicted value in the final assessment.

We evaluated the results obtained by *in silico* predictions based on the information on the applicability domain and the uncertainty provided by the models.

2.4 Algorithms for evaluation of classification models

The performance of the three consensus models was evaluated using Cooper's parameters (Cooper et al., 1979), which include accuracy, sensitivity and specificity. These parameters take into account the number of correctly classified mutagens (true positive = TP) and non-mutagens (true negative = TN) and the number of misclassified mutagenic (false positive = FP) and non-mutagenic (false negative = FN) compounds. Matthew's Correlation Coefficient (MCC) was also assessed.

These are calculated as follows:

$$\text{Accuracy} = \frac{TP + TN}{\text{Total}}$$

$$\text{Sensitivity} = \frac{TP}{(TP + FN)}$$

$$\text{Specificity} = \frac{TN}{(TN + FP)}$$

Accuracy (concordance or "Q") measures the total errors, while models with high sensitivity produce fewer false negatives, i.e., mutagenic compounds that are predicted as non-mutagenic. Models with high specificity give fewer false positives (non-mutagenic chemicals incorrectly predicted as mutagens).

The Matthews Correlation Coefficient (MCC) evaluates the quality of binary classifications and is generally considered a balanced measure, which can be used even for classes of very different sizes. (Matthews et al., 1975). This parameter prevails over any imbalance in the data classes, which may lead to unfair values of accuracy. It is calculated as follows:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

MCC values vary between -1 and +1: +1 indicates exact classification, -1 results from complete misclassification and 0 implies a random result.

3 Results

Overall, the three consensus models gave convergent predictions for 144 out of 183 compounds, corresponding to 79% of the total dataset: 21 were mutagenic and 123 were non-mutagenic.

First, we compared predicted and experimental values where available and evaluated the statistical performance of the three

models. All models correctly classified 75 out of 97 experimentally known chemicals, 19 as mutagenic and 56 as non-mutagenic. However, 17 out of the 19 mutagenic ones were included in the training/test sets of at least one model; 22 chemicals were not correctly predicted by at least one model. Table S2⁹ contains the list of 183 chemicals with predicted values from the three models with their "reliability/confidence scores" and experimental values.

RHC generated nine false negatives; one was in common with T.E.S.T. However, these inaccurate predictions were overridden by the correct ones from the VEGA consensus model based on the presence of experimental values. T.E.S.T. produced two false negatives (one in common with RHC), which were overridden by positive predictions of VEGA, which contains the experimental data.

Overall, the models gave 11 false positives: four were generated by VEGA, two by T.E.S.T., and seven were misclassified by RHC, all but one with high uncertainty. Three of these chemicals were alkyl phenyl sulfonates, suggesting that the RHC model may encounter problems when predicting this chemical class.

Based on the available measured mutagenicity values, we examined the possibility to combine and validate the predictions of the three-consensus models moving towards a weight of evidence approach. We drew up a strategy to combine predictions from the individual consensus models. Essentially this integration scheme first checks the presence of experimental data and then the prediction reliabilities for each model.

The algorithm we developed involves the following steps:

1. If an experimental value is present in the dataset(s) of at least one model, it takes the place of the predicted one(s) in the final assessment.
2. If there is no experimental value, processing a molecule by the three models gives rise to different possibilities:
 - (a) All estimations are convergent (all mutagenicity positives or negatives); in this case, these become the final prediction, regardless of their reliability.
 - (b) Two predictions are convergent (both associated with mutagenicity/non-mutagenicity) and one is divergent; the convergent estimations are used as final prediction if at least one of them is reliable, otherwise the divergent one supersedes them if it is reliable. If both convergent and divergent estimations are uncertain, they are discarded and the model is unable to estimate the molecule.
 - (c) One of the three models cannot provide any prediction. In this case, if the estimations from the other models are convergent and at least one of them is reliable, this is kept as final assessment; otherwise the integrated model does not provide any prediction. If the other two predictions are divergent, the one with high reliability is taken.
 - (d) Two models are unable to predict the molecule. In this case, the only available estimation is kept as final assessment only if it is reliable; otherwise it is discarded.

Table 1 lists the statistical performance of the three consensus models plus the combined one. We calculated statistics on curated experimental data with and without information about reliability of prediction.

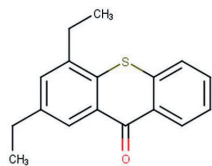
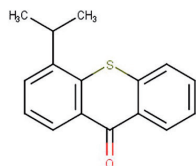
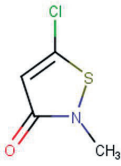


Tab. 1: Statistical performance of VEGA, RHC and T.E.S.T models for mutagenicity evaluated using accuracy, sensitivity, specificity and the Matthew's Correlation Coefficient (MCC)

These parameters show the number of correctly classified mutagens (true positive = TP) and non-mutagens (true negative = TN) and the number of misclassified mutagenic (false positive = FP) and non-mutagenic (false negative = FN) compounds.

	VEGA (all predictions)	RHC (all predictions)	T.E.S.T. (all predictions)	VEGA (reliable predictions)	RHC (reliable predictions)	T.E.S.T. (reliable predictions)	Combined model
TP	29	20	27	28	15	21	29
TN	64	61	65	57	58	58	65
FN	0	9	2	0	7	0	0
FP	4	7	2	0	1	0	0
Total	97	97	96	85	81	80	95
Accuracy	0.96	0.84	0.96	1.00	0.90	0.99	0.99
Specificity	0.94	0.90	0.97	1.00	0.98	0.98	0.98
Sensitivity	1.00	0.69	0.93	1.00	0.68	1.00	1.00
MCC	0.93	0.60	0.90	1.00	0.74	0.97	0.98

Tab. 2: Positively predicted food contact chemicals lacking experimental information

ID	Name	Structure
3	2,4-Diethyl-9H-thioxanthen-9-one	
17_p	4-Isopropylthioxanthone (4-ITX)	
123	5-Chloro-2-methyl-2H-isothiazol-3-one (CIT)	

All the models showed good statistical performance. The use of information about reliability based on the selected cutoffs led to enhancement of statistical parameters. Moreover, data curation allowed fixing a number of misclassifications of the used models.

Besides the statistical improvement, the use of the integrated strategy provided higher prediction coverage compared to sin-

gle prediction taking into account the information on reliability. In addition, measured data available in VEGA and T.E.S.T. training/test sets filled the gap of experimental information from the RHC model within this integrated scheme. The combined model was unable to assess two chemicals, bis(2,6-diisopropylphenyl)carbodiimide and 2,2,3-trifluoro-3-(trifluoromethyl)oxirane, reported as negative in the ECHA CHEM database⁷

with reliability 1 according to the Klimisch score (Klimisch et al., 1997). In the framework of developing a strategy of toxicity assessment for food contact chemicals, experimental data supersedes predicted data.

The high level of accuracy of the integrated strategy provided a rationale to apply it to evaluate the remaining 86 experimentally untested compounds in our dataset to identify mutagenic chemicals. The new integrated model gave 83 non-mutagenic and three mutagenic predictions. Nine out of the 83 substances predicted as non-mutagenic were part of the training/test sets of the model(s). The chemicals predicted positive were not included in the models' databases (training and/or test sets) and were among the possible migrating substances. These three positively classified food contact chemicals lacking experimental information are shown in Table 2. All of them contain structural alerts, which have been associated with mutagenic activity based on mechanism of toxicity (Benigni, 2008) or on statistical evidence (Benfenati et al., 2015). In particular, these include thioxanones, which are present in 2,4-diethyl-9H-thioxanthen-9-one and 4-isopropylthioxanthone (4-ITX), and an α,β unsaturated carbonyl moiety, which occurs in 5-chloro-2-methyl-2H-isothiazol-3-one (CIT).

Both 4-ITX and 2,4-diethyl-9H-thioxanthen-9-one are 2-ITX structural analogues (contained in our dataset). Consequently, the reasoning on mutagenicity for 2-ITX can be extended to these chemicals through a read-across approach, because of the high structural similarity and the presence of the thioxanone ring as structural alert shared by the three molecules. As in the case of 2-ITX, the two related chemicals might exhibit their mutagenic potential *in vitro* but not *in vivo* (as explained in Section 2.2).

CIT is a component of a biocide with CAS number 55965-84-9, mixture 3:1 with 2-methyl-2H-isothiazol-3-one (MIT). According to EFSA Scientific Opinion (EFSA, 2010), the biocide gave positive results in genotoxicity tests *in vitro* in bacteria, while no significant genotoxicity was observed *in vivo*. The other component of the mixture, MIT, was predicted as non-mutagenic by the combined model. Based on these estimations, CIT might be considered as responsible for positive *in vitro* results of the CIT/MIT mixture.

4 Discussion

In this study, an integrated strategy for mutagenicity prediction was developed and validated on about a hundred experimentally known chemicals, including mostly non-mutagenic migrating substances from FCM plus their positive analogues. Even if our aim is to integrate QSAR and read-across in the frame of the WoE approach, in the present study, we focus only on the integration of QSAR models because read-across is quite difficult to automate. Comparing the results obtained by the QSAR consensus model and read-across approach can surely increase the accuracy of the final prediction. Such a procedure is strongly recommended for compounds predicted with low reliability. This study highlighted some other key aspects to take into account in the evaluation of predictions from *in silico* models.

First, the performance of the individual predictive models was affected by the quality of experimental data and by the information on prediction uncertainty, where available. The information about prediction reliability improved all the statistical parameters. Moreover, a fair comparison between measured and estimated values is not a simple matter. Indeed, it is important to identify the experimental protocol used to measure or estimate the endpoint that is being examined, in accordance with OECD principles (OECD, 2014). This example illustrates that the use of data curation gives a more objective estimation of actual predictive power of the models, often accompanied by an improvement of their statistical behavior.

It is increasingly recommended to combine models based on complementary algorithms (ICH, 2017). This is often considered a default option to minimize the risk of producing false negatives and therefore to ensure optimal consumer protection. However, this may potentially be at the expense of generating numerous false positive predictions and reducing overall accuracy. This could potentially result in over-conservative and non-discriminative predictions, preventing their most efficient use for decision-making.

In the present study, enhanced (Q)SAR model performance was observed by applying a new algorithm for model integration, taking into account the different reliabilities of orthogonal methods. A possible drawback of using a highly accurate model may be a loss of chemical structure coverage. The use of the combination approach developed in this study together with the use of the available information on applicability domain/reliability provided a higher prediction coverage compared to single model estimations. Indeed, the aim of the paper was to study the strengths of a new consensus model based on easy rules, taking into account the convergence and the reliability of predictions. In this way, our strategy relates more strongly to the strength of the more reliable models. It shows that most Ames mutagenicity (Q)SAR models already perform quite well (the error of the models is very close to the experimental one) and the benefit of our approach is mainly represented in the increase of the applicability domain. As a first application, the consensus model was applied to a limited number of chemicals. In the near future we are planning to test it on bigger datasets and to include other kinds of applications.

Finally, the integrated strategy was applied to 86 chemicals. All of them could be predicted. Three were considered genotoxic (reported in Tab. 2) and were analyzed more deeply. The results obtained allow further assessment of the safety of these toxicologically untested molecules through the application of the TTC approach (Kroes et al., 2004) and/or the conceptual scheme developed by Schilter et al. (2014).

5 Conclusions

In the framework of understanding and managing risks for consumer health posed by untested food contact chemicals such as NIAS, the present study provides an algorithm combining existing models for a time- and cost-efficient evaluation of Ames



mutagenicity. The integration scheme resulted in an increased domain of applicability. Moreover, we are planning to test the model in the near future on a bigger number of chemicals and including other kinds of applications (not only food contact chemicals). These results will improve the implementation of a tool, such as VEGA, to integrate predictions from different models. Indeed, we believe that such a strategy may be applied as the first step of a more complex screening strategy aiming to establish the level of safety concern of experimentally untested substances according to the broadly accepted TTC concept (Kroes et al., 2004) and/or other more recently developed non-testing approaches (Schilter et al., 2014) aiming at identifying a dose to be compared with estimated level of exposure within a Margin of Exposure (MoE) approach. Each step of the overall scheme will be sequentially automated through implementation in the publicly available VEGA platform in the near future. This will not only provide time-saving, but also the advantage of minimizing inconsistencies and errors due to the manual building, validation and application of *in silico* methods.

References

- Amaury, N., Benfenati, E., Bumbaru, S. et al. (2007). Hybrid systems. In E. Benfenati (ed.), *Quantitative Structure-Activity Relationships (QSAR) for Pesticide Regulatory Purposes* (149-183). Amsterdam, NL: Elsevier Science Ltd. doi:10.1016/B978-044452710-3/50007-0
- Benfenati, E., Benigni, R., Marini, D. et al. (2009). Predictive models for carcinogenicity and mutagenicity. Frameworks, state-of-the art and perspectives. *J Environ Sci Health Part C* 27, 57-90. doi:10.1080/10590500902885593
- Benfenati, E., Manganello, S., Giordano, S. et al. (2015). Hierarchical rules for read-across and *in silico* models of mutagenicity. *J Environ Sci Health Part C* 33, 385-403. doi:10.1080/10590501.2015.1096881
- Benfenati, E., Belli, M., Borges, T. et al. (2016). Results of a round-robin exercise on read-across. *SAR QSAR Environ Res* 27, 371-384. doi:10.1080/1062936X.2016.1178171
- Benigni, R. (2008). The Benigni/Bossa rulebase for mutagenicity and carcinogenicity – A module of Toxtree. *EUR 23241*, 1-70. https://eurl-ecvam.jrc.ec.europa.eu/laboratories-research/predictive_toxicology/doc/EUR_23241_EN.pdf
- Benigni, R. and Bossa, C. (2011). Mechanisms of chemical carcinogenicity and mutagenicity: A review with implications for predictive toxicology. *Chem Rev* 111, 2507-2536. doi:10.1021/cr100222q
- Cassano, A., Raitano, G., Mombelli, E. et al. (2014). Evaluation of QSAR models for the prediction of Ames genotoxicity: A retrospective exercise on the chemical substances registered under the EU REACH regulation. *J Environ Sci Health Part C* 32, 273-298. doi:10.1080/10590501.2014.938955
- Cooper, J. A., Saracci, R. and Cole, P. (1979). Describing the validity of carcinogen screening tests. *Br J Cancer* 39, 87-89. doi:10.1038/bjc.1979.10
- Cox, R., Green, D. V. S., Luscombe, C. N. et al. (2013). QSAR workbench: Automating QSAR modeling to drive compound design. *J Comput Aided Mol Des* 27, 321-336. doi:10.1007/s10822-013-9648-4
- Dixon, S. L., Duan, J., Smith, E. et al. (2016). AutoQSAR: An automated machine learning tool for best-practice QSAR modeling. *Future Med Chem* 8, 1825-1839. doi:10.4155/fmc-2016-0093
- EC – European Commission (2004). Regulation (EC) No 1935/2004 of the European Parliament and of the Council of 27 October 2004 on materials and articles intended to come into contact with food and repealing Directives 80/590/EEC and 89/109/EEC. <http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2004:338:0004:0017:en:PDF>
- EC (2011). Regulation (EC) No 10/2011 of the European Parliament and of the Council of 14 January 2011 on plastic materials and articles intended to come into contact with food. <http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32011R0010&from=DE>
- EFSA – European Food Safety Authority (2005). Opinion of the Scientific Panel on Food Additives, Flavourings, Processing Aids and Materials in Contact with Food on a request from the Commission related to 2-Isopropyl thioxanthone (ITX) and 2-ethylhexyl-4-dimethylaminobenzoate (EHDAB) in food contact materials. *EFSA J* 3, 293 (1-15). doi:10.2903/j.efsa.2005.293
- EFSA (2008). Flavouring Group Evaluation 5, Revision 1 (FGE.05Rev1): Esters of branched- and straight-chain aliphatic saturated primary alcohols and of one secondary alcohol, and branched- and straight-chain unsaturated carboxylic acids from chemical groups 1, 2, and 5 (Commission Regulation (EC) No 1565/2000 of 18 July 2000). *EFSA J* 6, 643 (1-81). doi:10.2903/j.efsa.2008.643
- EFSA (2009). Scientific Opinion of the Panel on food contact materials, enzymes, flavourings and processing aids (CEF) on 24th list of substances for food contact materials. *EFSA J* 7, 1157-1163. doi:10.2903/j.efsa.2009.1157
- EFSA (2010). Scientific Opinion on the safety evaluation of the substance, 5-chloro-2-methyl-2H-isothiazol-3-one, mixture with 2-methyl-2H-isothiazol-3-one (3:1), CAS No. 55965-84-9, as a biocide for processing coatings and paper and boards. *EFSA J* 8, 1541. doi:10.2903/j.efsa.2010.1541
- EFSA (2012). Food contact materials, flavouring substances and smoke flavourings. *EFSA J* 10, Spec Issue, s1007. doi:10.2903/j.efsa.2012.s1007
- EFSA (2014). Scientific Opinion on the re-evaluation of propionic acid (E 280), sodium propionate (E 281), calcium propionate (E 282) and potassium propionate (E 283) as food additives. *EFSA J* 12, 3779 (1-45). doi:10.2903/j.efsa.2014.3779
- EFSA (2015). Food Contact Materials. <http://www.efsa.europa.eu/en/topics/topic/foodcontactmaterials>
- EFSA (2017). Guidance on The Use of the Weight of Evidence Approach in Scientific Assessments. *EFSA J* 15, 4971. <https://www.efsa.europa.eu/sites/default/files/consultation/170306-0.pdf>

- Ferrari, T. and Gini, G. (2010). An open source multistep model to predict mutagenicity from statistical analysis and relevant structural alerts. *Chem Cent J* 4, Suppl 1, S2. doi:10.1186/1752-153X-4-S1-S2
- Ferrari, T., Cattaneo, D., Gini, G. et al. (2013). Automatic knowledge extraction from chemical structures: The case of mutagenicity prediction. *SAR QSAR Environ Res* 24, 365-383. doi:10.1080/1062936X.2013.773376
- Floris, M., Manganaro, A., Nicolotti, O. et al. (2014). A generalizable definition of chemical similarity for read-across. *J Cheminform* 6, 39. doi:10.1186/s13321-014-0039-1
- Gini, G., Franchi, A. M., Manganaro, A. et al. (2014). ToxRead: A tool to assist in read across and its use to assess mutagenicity of chemicals. *SAR QSAR Environ Res* 25, 999-1011. doi:10.1080/1062936X.2014.976267
- Hansen, K., Mika, S., Schroeter, T. et al. (2009). Benchmark data set for in silico prediction of Ames mutagenicity. *J Chem Inf Model* 49, 2077-2081. doi:10.1021/ci900161g
- Hearty, A., Gibney, M. J., Vin, K. et al. (2011). The FACET project: A chemical exposure surveillance system for Europe. *Food Sci Technol* 25, 26-29.
- Helma, C., Cramer, T., Kramer, S. and De Raedt, L. (2004). Data mining and machine learning techniques for the identification of mutagenicity inducing substructures and structure activity relationships of noncongeneric compounds. *J Chem Inf Comput Sci* 44, 1402-1411. doi:10.1021/ci034254q
- ICH (2017). ICH Harmonised Guideline, M7 (R1). Assessment and control of DNA reactive (mutagenic) impurities in pharmaceuticals to limit potential carcinogenic risk. http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Multidisciplinary/M7/M7_R1_Addendum_Step_4_31Mar2017.pdf
- Jacobs, M. N., Colacci, A., Louekari, K. et al. (2015). International regulatory needs for development of an IATA for non-genotoxic carcinogenic chemical substances. *ALTEX* 33, 359-392. doi:10.14573/altex.1601201
- JRC (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 Supporting Publication* 7, EN-50 (1-311). doi:10.2903/sp.efsa.2010.EN-50
- Kazius, J., McGuire, R. and Bursi, R. (2005). Derivation and validation of toxicophores for mutagenicity prediction. *J Med Chem* 48, 312-320. doi:10.1021/jm040835a
- Klimisch, H. J., Andreae, M. and Tillmann, U. (1997). A systematic approach for evaluating the quality of experimental toxicological and ecotoxicological data. *Regul Toxicol Pharmacol* 25, 1-5. doi:10.1006/rtp.1996.1076
- Kroes, R., Renwick, A. G., Cheeseman, M. et al. (2004). Structure-based thresholds of toxicological concern (TTC): Guidance for application to substances present at low levels in the diet. *Food Chem Toxicol* 42, 65-83. doi:10.1016/j.fct.2003.08.006
- Kulkarni, S. A., Benfenati, E. and Barton-Maclaren, T. (2016). Improving confidence in (Q)SAR predictions under Canada's Chemicals Management Plan – A chemical space approach. *SAR QSAR Environ Res* 27, 851-863. doi:10.1080/1062936X.2016.1243152
- Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim Biophys Acta* 405, 442-451. doi:10.1016/0005-2795(75)90109-9
- Manganaro, A., Pizzo, F., Lombardo, A. et al. (2016). Predicting persistence in the sediment compartment with a new automatic software based on the k-Nearest Neighbor (k-NN) algorithm. *Chemosphere* 144, 1624-1630. doi:10.1016/j.chemosphere.2015.10.054
- Manganelli, S., Benfenati, E., Manganaro, A. et al. (2016). New quantitative structure-activity relationship models improve predictability of Ames mutagenicity for aromatic azo compounds. *Toxicol Sci* 153, 316-326. doi:10.1093/toxsci/kfw125
- Mazzatorta, P., Tran, L., Schilter, B. and Grigorov, M. (2007). Integration of structure-activity relationship and artificial intelligence systems to improve in silico prediction of Ames test mutagenicity. *J Chem Inf Model* 47, 34-38. doi:10.1021/ci600411v
- Mombelli, E., Raitano, G. and Benfenati, E. (2016). In silico prediction of chemically induced mutagenicity: How to use QSAR models and interpret their results. In E. Benfenati (ed.), *In Silico Methods for Predicting Drug Toxicity. Methods in Molecular Biology* (87-105). New York, USA: Springer. doi:10.1007/978-1-4939-3609-0_5
- OECD (1997). Test No. 471: Bacterial Reverse Mutation Test, Section 4, OECD Publishing. doi:10.1787/9789264071247-en
- OECD (2014). Guidance Document on the Validation of (Quantitative) Structure-Activity Relationship [(Q)SAR] Models, OECD Publishing. doi:10.1787/9789264085442-en
- OECD (2017). The OECD QSAR Toolbox v4.0. <http://www.oecd.org/chemicalsafety/risk-assessment/oecd-qsar-toolbox.htm> (accessed July 2017).
- Price, N. and Chaudhry, Q. (2014). Application of in silico modelling to estimate toxicity of migrating substances from food packaging. *Food Chem Toxicol* 71, 136-141. doi:10.1016/j.fct.2014.05.022
- Romano, P. (2008). Automation of in-silico data analysis processes through workflow management systems. *Brief Bioinform* 9, 57-68. doi:10.1093/bib/bbm056
- Schilter, B., Benigni, R., Boobis, A. et al. (2014). Establishing the level of safety concern for chemicals in food without the need for toxicity testing. *Regul Toxicol Pharm* 68, 275-296. doi:10.1016/j.yrtph.2013.08.018
- Sutter, A., Amberg, A., Boyer, S. et al. (2013). Use of in silico systems and expert knowledge for structure-based assessment of potential mutagenic impurities. *Regul Toxicol Pharmacol* 67, 39-52. doi:10.1016/j.yrtph.2013.05.001
- Van Bossuyt, M., Van Hoeck, E., Vanhaecke, T. et al. (2016). Printed paper and board food contact materials as a potential source of food contamination. *Regul Toxicol Pharm* 81, 10-19. doi:10.1016/j.yrtph.2016.06.025
- Van Bossuyt, M., Van Hoeck, E., Raitano, G. et al. (2017). (Q)



SAR tools for priority setting: A case study with printed paper and board food contact material substances. *Food Chem Toxicol* 102, 109-119. doi:10.1016/j.fct.2017.02.002

Warr, W. A. (2012). Scientific workflow systems: Pipeline Pilot and KNIME. *J Comput Aided Mol Des* 26, 801-804. doi:10.1007/s10822-012-9577-7

Weininger, D. (1988). SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J Chem Inf Comput Sci* 281, 31-36. doi:10.1021/ci00057a005

Young, S. S., Gombar, V. K., Emptage, M. R. et al. (2002). Mixture deconvolution and analysis of Ames mutagenicity data. *Chem Int Lab Syst* 60, 5-11. doi:10.1016/S0169-7439(01)00181-2

Zhang, S., Golbraikh, A., Oloff, S. et al. (2006). A novel automated lazy learning QSAR (ALL-QSAR) approach: Method development, applications, and virtual screening of chemical databases using validated ALL-QSAR models. *J Chem Inf Model* 46, 1984-1995. doi:10.1021/ci060132x

Conflict of interest

Authors declare no potential conflict of interests.

Correspondence to

Elena Lo Piparo, PhD
Chemical Food Safety Group
Nestlé Research Center
PO Box 44, CH-1000 Lausanne 26
Lausanne, Switzerland
Phone: +41 21 785 82 94
e-mail: elena.lopiparo@rdls.nestle.com

Das 3R-Kompetenzzentrum (3RCC)

– bessere Forschung
mit weniger Tierversuchen?

11. Tierversuchstagung des Schweizer Tierschutz STS

Freitag, 18. Mai 2018

Beginn: 08:45 Uhr

Kongresszentrum Hotel Arte
Riggenbachstrasse 10
4600 Olten

Tagungsgebühr

(inkl Verpflegung und Tagungsunterlagen)

Vollzahler(in) CHF 180.–

Student(in) CHF 90.–

Tagungssprache: Hochdeutsch, Französisch

Simultanübersetzung: Deutsch-Französisch und
Französisch-Deutsch

Anmeldung

Schweizer Tierschutz STS

Geschäftsstelle

Dornacherstrasse 101, Postfach
4018 Basel

Tel. 0041-(0)61-365 99 99

www.tierschutz.com



Die Tagung wird von der Vereinigung der Schweizer Kantonstierärztinnen und Kantonstierärzte (VSKT) den zuständigen kantonalen Behörden zur Anerkennung im Rahmen der Aus- und Weiterbildung von Fachpersonal für Tierversuche empfohlen sowie von der Gesellschaft Schweizer Tierärztinnen und Tierärzte GST mit zwei Bildungspunkten anerkannt.