TOXICOLOGICAL SCIENCES, 163(2), 2018, 632–638

doi: 10.1093/toxsci/kfy057
Advance Access Publication Date: March 20, 2018
Research Article

Performance of *In Silico* Models for Mutagenicity Prediction of Food Contact Materials

Melissa Van Bossuyt,*,†,[1] Els Van Hoeck,* Giuseppa Raitano,‡ Tamara Vanhaecke,† Emilio Benfenati,‡ Birgit Mertens,*,[2] and Vera Rogiers†,[2]

*Department of Food, Medicines and Consumer Safety, Scientific Institute of Public Health, 1050 Brussels, Belgium; †Department of In Vitro Toxicology and Dermato-Cosmetology, Vrije Universiteit Brussel, 1090 Brussels, Belgium; and ‡Department of Environmental Health Sciences, IRCCS-Istituto di Ricerche Farmacologiche Mario Negri, Milan 20156, Italy

[1]To whom correspondence should be addressed at Department of Food, Medicines and Consumer Safety, Scientific Institute of Public Health, J. Wytsmanstraat 14, 1050 Brussels, Belgium. Fax: +32 2 642 52 24; E-mail: melissa.vanbossuyt@wiv-isp.be.

[2]These authors contributed equally to this study.

ABSTRACT

In silico methodologies, such as (quantitative) structure-activity relationships ([Q]SARs), are available to predict a wide variety of toxicological properties and biological activities for structurally diverse substances. To obtain insights in the scientific value of these predictions, the capacity of the prediction models to generate (sufficiently) reliable results for a particular type of compounds needs to be evaluated. In the current study, performance parameters to predict the endpoint "bacterial mutagenicity" were calculated for a battery of common (Q)SAR tools, namely Toxtree, Derek Nexus, VEGA Consensus, and Sarah Nexus. Printed paper and board food contact material (FCM) constituents were chosen as study substances because many of these lack experimental data, making them an interesting group for *in silico* screening. Accuracy, sensitivity, specificity, positive predictivity, negative predictivity, and Matthews correlation coefficient for the individual models and for the combination of VEGA Consensus and Sarah Nexus were determined and compared. Our results demonstrate that performance varies among the four models, but can be increased by applying a combination strategy. Furthermore, the importance of the applicability domain is illustrated. Limited performance to predict the mutagenic potential of substances that are new to the model (ie, not included in the training set) is reported. In this context, the generally poor sensitivity for these new substances is also addressed.

Key words: *in silico*; (Q)SAR; mutagenicity; food contact materials; validation.

In silico methods are among the most suitable tools for the initial safety screening of chemicals. They allow predictions to be made for a large amount of structurally characterized substances, in a fast, reproducible and relatively straightforward manner. Due to their resource- and time-saving characteristics, they are now recognized as helpful toxicity screening tools (Golbamaki and Benfenati, 2016). Throughout various legislations, the application of these computer-based techniques is increasingly supported by regulatory authorities (Raunio, 2011).

It is generally agreed, however, that *in silico* methods are not yet capable of fully replacing *in vitro* and *in vivo* testing. A number of obstacles hamper their widespread application, even for screening purposes. One problem frequently encountered is the lack of knowledge with respect to the applicability domain (AD) of the model (Cherkasov *et al.*, 2014). A computer model is by definition built from a structurally limited training set of compounds. As a result, predictions will only be sufficiently reliable for test compounds that sufficiently represent structural similarity with the training set compounds. Hence the applicability of an *in silico* method must be investigated in order to assure satisfactory predictive capacity for a given group of substances.

The development of *in silico* models is largely induced by the paradigm shift in toxicology, favouring the use of alternative (nonanimal) testing methods. For example, animal testing of cosmetic ingredients is prohibited in the European Union since March 2013 (European Union, 2009). The sector of industrial chemicals is another example where, due to the enormous number of substances concerned, *in silico* models are gaining importance and their performance is increasingly studied (European Union, 2006). However, other regulatory fields may also benefit from information on which *in silico* model performs best for what type of compounds. In this context, food contact materials (FCM) represent a growing source of concern with regard to human health, due to the possible migration of the substances into food and drinks. In particular, no harmonized European legislation is in place for nonplastic FCM and thousands of substances potentially used have not been officially evaluated for their safety (Van Bossuyt *et al.*, 2016).

Among the nonplastic FCM, printed paper and board constitute a major category and have been the subject of several contamination issues in the recent past (Van Hoeck *et al.*, 2017). Genotoxicity data for these compounds are urgently needed as this toxicological endpoint has been associated with important adverse health effects including cancer (Ames *et al.*, 1975). Investigation of the genotoxic potential of a substance generally starts by running two *in vitro* assays: A gene mutation test in bacteria (Ames test) and a mammalian cell micronucleus test (EFSA, 2016). Performing a two-test *in vitro* battery for all printed paper and board substances is not feasible, as thousands of substances are concerned. To prioritize nonevaluated printed paper and board substances for in-depth safety studies, we recently predicted the bacterial mutagenicity of 1723 substances using a battery of (quantitative) structure-activity relationship ([Q]SAR) methods (Van Bossuyt *et al.*, 2017).

(Q)SARs are popular *in silico* methods that determine the activity (eg, toxicity) of a test compound based on the activity of structurally similar compounds included in the model training set. The prediction is either (1) rule-based through the use of expert knowledge and called SAR, or (2) statistically based through the application of a mathematical algorithm and called QSAR. A large variety of free as well as commercial (Q)SARs are available. In the latter study, a battery consisting of two SAR methods (Toxtree [free] and Derek Nexus [commercial]) and two QSAR methods (VEGA [free] and Sarah Nexus [commercial]) was used (Van Bossuyt *et al.*, 2017). This is in agreement with current international guidelines advising the combined use of at least two complementary (Q)SARs (ICH, 2017). Bacterial mutagenicity was investigated as, in contrast to other genotoxicity endpoints, this is one of the most modeled endpoints by (Q)SAR tools. Because the aim of the study was to prioritize compounds for further investigation, the highest priority was assigned to substances positive in all four (Q)SAR methods.

To know which (Q)SAR model performs best with regard to predicting the endpoint "bacterial mutagenicity" for printed paper and board FCM substances, in the present study, the prediction performance of the abovementioned (Q)SAR methods is determined for printed paper and board substances. First, existing experimental bacterial mutagenicity data were collected for substances used in the manufacture of printed paper and board FCM. Subsequently, the substance structures were processed in the (Q)SAR models and performance parameters including accuracy, sensitivity, specificity, positive predictivity, negative predictivity, and Matthews correlation coefficient (MCC) were calculated. The prediction performance was evaluated at different levels: (1) total set, (2) subset of compounds inside the AD, (3) subset of compounds outside the training set, and (4) subset of compounds inside the AD and outside the training set. Indeed, to estimate the prediction performance of a (Q)SAR model for new compounds with unknown activity, an evaluation set of compounds inside the AD, but outside the training set is required.

MATERIALS AND METHODS

Reference Dataset

The reference dataset was obtained by investigating the overlap between a previously constructed inventory of substances that can be used in printed paper and board FCM (Van Bossuyt *et al.*, 2016) and six inventories containing chemical substances with experimental Ames mutagenicity data including (1) Hansen benchmark dataset (Hansen *et al.*, 2009), (2) Leadscope toxicity database (Leadscope, 2017), (3) Japan's Health Ministry dataset (National Institute of Health Sciences of Japan, 2017), (4) Scientific Committee on Consumer Safety evaluation dossiers on hair dyes (Ates *et al.*, 2016), (5) PROSIL dataset, consisting of data on dyes from source (3) and Kulkarni *et al.* (Kulkarni and Barton-Maclaren, 2014), and (6) CALEIDOS ECHA CHEM dataset (Cassano *et al.*, 2014). Printed paper and board FCM substances included in one or more of the six inventories were binary classified as mutagenic or nonmutagenic. Consequently, substances appearing in more than one inventory were only retained in case the experimental outcome was consistent in the different inventories. The final reference dataset contained 875 substances, of which 729 (83%) were classified as nonmutagenic and 146 (17%) as mutagenic. A broad range of substance types (aldehydes, alcohols, phenols, ketones, amides, amines, etc.) was covered.

(Q)SAR Models

Ames mutagenicity models were selected from two SARs, ie, Toxtree and Derek Nexus, and two QSARs, ie, VEGA Consensus and Sarah Nexus. These tools are briefly described below, whereas further details can be found in Van Bossuyt *et al.* (2017).

Toxtree. Toxtree (www.toxtree.sourceforge.net) is freely available toxicity prediction software established by the Joint Research Centre of the European Union (European Commission, 2016). In the current study, the *In vitro mutagenicity alerts (Ames test) by ISS* module of Toxtree version 2.6.0 was used. This SAR model provides binary classification (mutagenic/nonmutagenic), but does not include an AD functionality and the list of training set compounds is not publically available.

Derek Nexus. Derek Nexus is part of the Lhasa Knowledge Suite and commercially available from Lhasa Limited (Lhasa Limited, 2016a). The *Mutagenicity in vitro* model of Derek Nexus version 4.1.0 was used in the present study. Prediction results ranging from *equivocal* to *certain* were classified as mutagenic, whereas results from *inactive* to *doubted* were classified as nonmutagenic. The latter classification is based on a previous assessment that measured prediction confidence against reliability for expert systems (Judson *et al.*, 2003). This SAR model does not provide information regarding training set compounds nor AD, although the reliability of negative predictions is increased by a recently implemented feature that identifies misclassified and unclassified structures (Williams *et al.*, 2016).

VEGA Consensus. The Istituto di Ricerche Farmacologiche Mario Negri (IRFMN) develops free prediction models that are available through the VEGA HUB (www.vegahub.eu). Three Ames mutagenicity models; *Mutagenicity (Ames test) model (CAESAR) version 2.1.13*, *Mutagenicity (Ames test) model (SarPy/IRFMN) version 1.0.7*, and *Mutagenicity (Ames test) model (ISS) version 1.0.2* were used for the current evaluation (IRFMN, 2016). Because combination of the three individual results into one VEGA Consensus result was previously reported to increase the prediction performance, the same approach is followed here (Cassano *et al.*, 2014). The weighted consensus result is obtained by taking into account the result of each of the three models and their associated compound-specific AD index. This is done as follows:

$$\text{CONSENSUS} = \frac{(\pm 1) * \text{ADI}_{\text{CAESAR}} + (\pm 1) * \text{ADI}_{\text{SarPy}} + (\pm 1) * \text{ADI}_{\text{ISS}}}{\text{ADI}_{\text{CAESAR}} + \text{ADI}_{\text{SarPy}} + \text{ADI}_{\text{ISS}}}$$

ADI = AD index
±1 = prediction result, ie, +1 for mutagenic and −1 for nonmutagenic substances

Only compounds with an AD index ≥ 0.75 in all three models were considered to fall into the AD (Cassano *et al.*, 2014). Compounds included in the training set are automatically labelled as *experimental mutagen* or *experimental nonmutagen* by the software.

Sarah Nexus. Sarah Nexus is a QSAR model designed to make Ames mutagenicity predictions. It is part of the same Lhasa Knowledge Suite in which Derek Nexus is also incorporated (Lhasa Limited, 2016b). Here, the *Ames mutagenicity* model of Sarah Nexus version 1.2.0 was used. It applies a binary classification (mutagenic/nonmutagenic), unless a compound is outside the AD, in which case this is indicated as such. Equivocal results are possible but were not encountered in the present study. The training set is not publically available, however a prediction confidence score of 100% demonstrates the inclusion of the substance under evaluation among the training data.

Prediction Performance Evaluation

The performance of the four methods was assessed according to the internationally accepted guidance document of the Organisation for Economic Co-operation and Development (OECD) (OECD, 2014). The statistical parameters described for classification-based (Q)SAR models include accuracy, sensitivity, specificity, positive predictivity, and negative predictivity through considering the number of true positives (= mutagenic), true negatives (= nonmutagenic), false positives (= misclassified mutagenic), and false negatives (= misclassified nonmutagenic):

$$\text{Accuracy} = \frac{(\text{True positives} + \text{True negatives})}{\text{Total}}$$

$$\text{Sensitivity} = \frac{\text{True positives}}{(\text{True positives} + \text{False negatives})}$$

$$\text{Specificity} = \frac{\text{True negatives}}{(\text{True negatives} + \text{False positives})}$$

$$\text{Positive predictivity} = \frac{\text{True positives}}{(\text{True positives} + \text{False positives})}$$

$$\text{Negative predictivity} = \frac{\text{True negatives}}{(\text{True negatives} + \text{False negatives})}$$

The current dataset contains a high number of nonmutagens and a clearly lower number of mutagens. Although accuracy is generally a suitable parameter to evaluate balanced binary data, it may not be a proper measure to assess numerically skewed data, as it treats true and false negatives and positives equally. Therefore, the MCC (Matthews, 1975) was included as an additional performance parameter. This metric is compatible with imbalanced data and represents a widely used performance measure in biomedical research (Boughorbel *et al.*, 2017). Furthermore, the MCC metric has also been introduced to evaluate the performance of *in silico* mutagenicity models (Cassano *et al.*, 2014; Gadaleta *et al.*, 2016; Manganelli *et al.*, 2016). The MCC is calculated as follows:

$$\text{MCC} = \frac{(\text{True positives} * \text{True negatives}) - (\text{False positives} * \text{False negatives})}{\sqrt{(\text{True positives} + \text{False positives})(\text{True positives} + \text{False negatives})(\text{True negatives} + \text{False positives})(\text{True negatives} + \text{False negatives})}}$$

The MCC ranges from −1 (= total disagreement of experimental and predicted result) to +1 (= full agreement of experimental and predicted result), implying that 0 represents a random result.

In order to evaluate the performance at different levels, all six performance metrics were calculated for the reference dataset and for different subsets:

- Global performance: Analysis of the total reference dataset of 875 compounds to evaluate the overall performance of the mutagenicity models.
- Performance in AD: Analysis of the subset of compounds that are situated in the AD of the model. This evaluation cannot be carried out for Toxtree and Derek Nexus because their AD is not defined.
- Performance for new substances: Analysis of the subset of compounds that are new to the model, meaning they are not included in the training set of known mutagens and nonmutagens. This evaluation cannot be carried out for Toxtree and Derek Nexus because their training set is not defined.
- Performance for new substances in the AD: Analysis of the subset of compounds that are new to the model and inside the AD. This evaluation cannot be carried out for Toxtree and Derek Nexus because their AD and training set are not defined.

The global performance was thus evaluated for all methods, whereas the subset evaluations could only be carried out for VEGA Consensus and Sarah Nexus.

From a regulatory perspective, sensitivity is more important than specificity because substances should not be labelled safe whereas in reality they are toxic (Fjodorova *et al.*, 2010). With the aim of increasing sensitivity, the prediction results of compounds overlapping between subsets of the models with a defined training set and AD, namely VEGA Consensus and Sarah Nexus, were paired after which the performance of this combined prediction system was assessed. In this strategy, substances were considered positive if they were positive in at least one of both QSARs.

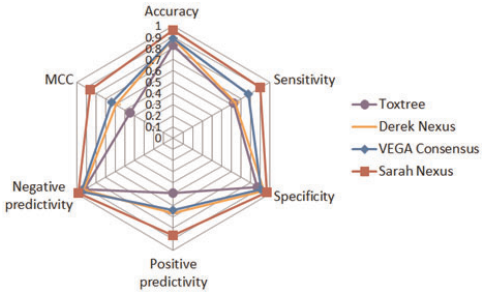
RESULTS AND DISCUSSION

Global Performance

First, the overall prediction performance of the bacterial mutagenicity models was evaluated using all FCM compounds of the reference dataset (Table 1). Sarah Nexus does not provide predictions for compounds labeled as outside AD by the software.

Table 1. Global Prediction Performance of (Q)SAR Models for Ames Mutagenicity

	Toxtree	Derek Nexus	VEGA Consensus	Sarah Nexus	Radar Plot Summary
Global Performance					
Number of compounds	875	875	875	863	
Accuracy (%)	83	89	89	96	
Sensitivity (%)	62	64	78	90	
Specificity (%)	87	94	91	97	
Positive predictivity (%)	49	67	64	87	
Negative predictivity (%)	92	93	95	98	
MCC	0.45	0.59	0.64	0.86	



MCC, Matthews correlation coefficient.

Consequently, 863 compounds were considered for the evaluation of Sarah Nexus, whereas for the other three models the complete reference dataset of 875 compounds could be used.

Accuracy is high for all models, varying from 83% for Toxtree up to 96% for Sarah Nexus. This can, to a great extent, be attributed to the high specificity reported for all models (between 87% and 97%), as also illustrated by their high negative predictivity (between 92% and 98%). In contrast, a large gap is seen in terms of sensitivity (ranging from 62% to 90%) and positive predictivity (ranging from 49% to 87%). Clearly, Sarah Nexus produces the best overall performance. This is also reflected by the MCC value of 0.86. It must be noted, however, that compounds outside the AD of Sarah Nexus are not considered in this evaluation as they are automatically excluded by the model software.

Interestingly, the global performance evaluation suggests that QSARs (i.e., VEGA Consensus and Sarah Nexus) have a higher sensitivity and are thus better than SARs (i.e., Toxtree and Derek Nexus) in identifying FCM substances that are mutagenic *in vitro*. High sensitivity is especially important in a regulatory context, as regulators are most concerned to overlook a substance being mutagenic. However, the ideal model also demonstrates high specificity so that good candidate molecules are not abandoned unnecessarily. To further investigate prediction capacity, AD and training set information play an important role. QSAR models such as VEGA Consensus and Sarah Nexus are based on a mathematical algorithm and therefore, they can be allocated an AD. SAR models such as Toxtree and Derek Nexus, on the other hand, are built from expert rules and intrinsically, a formal AD cannot be defined. Furthermore, the training set of SARs is often not fully characterized, which is also the case in the current study. Consequently, the performance of the four models to predict Ames mutagenicity of FCM substances could only be compared on the first and most simple level.

Performance in AD

In the next step of our study, the models with an AD functionality, namely VEGA Consensus and Sarah Nexus, were selected and performance values were determined using only those FCM reference compounds that were inside the AD of the respective model (Table 2). In general, model performance will increase when considering only substances that are in the AD. For Sarah Nexus, all performance values are identical to the values reported for the global evaluation (Table 1). Indeed, mutagenicity of substances with structures outside the predefined AD of this model will not be predicted. In contrast, VEGA models generate a prediction result for each compound analyzed with the software. The prediction outcome of these models is associated with an AD index ranging from 0 (= fully outside domain) to 1 (= training set compound), providing a measure of the applicability of each model for a particular compound. Thus, the user chooses the threshold that determines whether a test compound is inside or outside the AD. In the current study, the approach of Cassano et al. (2014) was followed and therefore, a substance was considered to be inside the AD of VEGA Consensus if its AD index is ≥ 0.75 for all three individual VEGA models. As a result, only 440 reference compounds could be used for the VEGA Consensus evaluation, which is substantially lower than the 863 reference compounds that were analyzed for Sarah Nexus.

Table 2 illustrates that by setting requirements for the AD index the performance of VEGA Consensus to predict bacterial mutagenicity can be drastically increased, approaching the performance values obtained for Sarah Nexus. Indeed, accuracy is equal and the MCC value is nearly equal for both models. Compared with the results reported for the global performance (Table 1), the gap between the sensitivities of both software programs becomes smaller when only compounds in the AD are considered. Interestingly, specificity is slightly higher in VEGA Consensus than in Sarah Nexus.

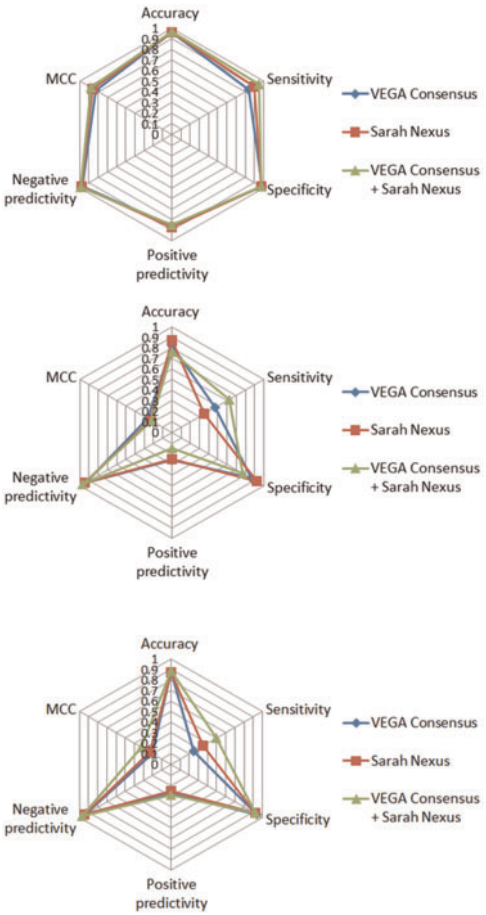
In the combined approach, the 440 substances covered by the AD of VEGA Consensus and Sarah Nexus were processed in both models. A substance was considered positive when a positive result was obtained in at least one of them. The performance statistics reveal the utility of the latter combination strategy, as sensitivity (95%), specificity (97%), accuracy (97%), and MCC (0.88) are high.

Performance for New Substances

Another part of the study consisted of evaluating the performance of the bacterial mutagenicity models when predicting compounds that are "new" to VEGA Consensus or Sarah Nexus. Hereto, only FCM reference compounds that are not part of the training set of the model were used (Table 2). Three hundred and eighty five compounds of the reference dataset are not part of the VEGA Consensus training set. Likewise, 230 of the reference compounds are not found in the Sarah Nexus training set. In contrast to the findings reported for the global performance and the in-AD performance, it is observed that for new compounds, the sensitivity for predicting bacterial mutagenicity is higher in VEGA Consensus as opposed to Sarah Nexus, whereas

Table 2. Prediction Performance for Ames Mutagenicity in VEGA Consensus, Sarah Nexus, and Combined

	VEGA Consensus	Sarah Nexus	VEGA Consensus + Sarah Nexus[a]	Radar Plot Summary
Performance in the AD				
Number of compounds	440	863	440	
Accuracy (%)	96	96	97	
Sensitivity (%)	84	90	95	
Specificity (%)	98	97	97	
Positive predictivity (%)	86	87	85	
Negative predictivity (%)	97	98	99	
MCC	0.83	0.86	0.88	
Performance for New Substances				
Number of compounds	385	230	217	
Accuracy (%)	82	87	76	
Sensitivity (%)	47	35	62	
Specificity (%)	86	92	77	
Positive predictivity (%)	26	25	15	
Negative predictivity (%)	94	95	97	
MCC	0.26	0.23	0.22	
Performance for New Substances in the AD				
Number of compounds	141	230	66	
Accuracy (%)	88	87	89	
Sensitivity (%)	25	35	50	
Specificity (%)	94	92	92	
Positive predictivity (%)	27	25	29	
Negative predictivity (%)	93	95	97	
MCC	0.20	0.23	0.32	



MCC, Matthews correlation coefficient.
[a]Substances were considered positive when positive in at least 1 model.

specificity is lower. Although accuracy is highest for Sarah Nexus, the MCC value for the VEGA Consensus method indicates the best performance when corrected for dataset imbalance.

With values as low as 35% and 47%, sensitivity is poor for both models individually. In the combined approach, the 217 substances included in neither of the VEGA Consensus nor Sarah Nexus training set were analyzed and a substance was considered positive when a positive result was obtained in at least one of them. Model combination produces a sensitivity raise up to 62%, however specificity declines to 77%. Moreover, accuracy (76%) and MCC (0.22) are below those registered for the individual models.

Performance for New Substances in the AD

Ultimately, only those compounds in the AD of the model that are not part of the training set were selected from the FCM reference dataset and used for analysis; 141 were found for VEGA Consensus and 230 for Sarah Nexus (Table 2). For the VEGA Consensus model, a remarkable difference exists between its capacity to identify true positives on the one hand and true negatives on the other hand. Indeed, sensitivity is very low (25%), whereas specificity is particularly high (94%). The resulting accuracy is also high (88%), however the balanced performance is poor as illustrated by an MCC of 0.20. As previously mentioned, substances outside the AD are not predicted in Sarah Nexus and therefore, the results for this model are identical to the results reported for the analysis with new substances as such.

Only 66 substances are in the AD of both VEGA Consensus and Sarah Nexus but not in their training set. The combined results demonstrate a sensitivity of 50%, specificity of 92%, accuracy of 89% and MCC of 0.32. Although combining the two models improves the prediction performance, it is important to note that still only one out of two mutagens is picked up.

Limitations and Future Perspectives

Based on the current results, it is found that the free VEGA Consensus method as well as the commercial Sarah Nexus application provide good overall performance with respect to the bacterial mutagenicity prediction of printed paper and board FCM. Sarah Nexus performs slightly better, although the commercial aspect can be an obstacle. Ideally the methods are used in combination, especially because this increases sensitivity up to 95%. Regarding new substances, however, poor performance

is demonstrated by both models and even by their combination. When experimental data become available for new substances, they are often integrated in the training set and the prediction model is updated. This not only explains the relatively low number of new compounds, but also implies that the performance might change with each update. Because numerous other free as well as commercial QSARs exist besides those studied here, it is also of interest to evaluate their performance, to ultimately identify the best model (combination) for future mutagenicity predictions of FCMs. In a very recent study, the performance of VEGA and two other freely available QSARs for mutagenicity prediction was evaluated on a limited set of 97 compounds potentially migrating from plastic FCM (Manganelli *et al.*, 2017). For VEGA Consensus, an extended version with additional training set compounds was used that likely contributed to its high performance. Remarkably, the VEGA Consensus results for the 85 reliable predictions were 100% accurate and the MCC value was equal to 1. It should however be noted that the authors did not indicate whether the study substances are training set compounds. As illustrated in the present study, this information has a considerable impact on the outcome of the evaluation.

Although the current FCM reference dataset consists of a rather large number of compounds, evidently this number becomes smaller when evaluating the performance in the AD, for new substances, or the combination of both. The last scenario is the strictest, especially in the combined approach where only compounds found to be in the AD as well as new to both models are considered (Table 2). Importantly, at the same time the imbalance of experimental mutagens and nonmutagens further deteriorates (results not shown): Whereas the initial distribution of the complete reference dataset is 17% mutagens and 83% nonmutagens, for the most severe evaluation, i.e., for new substances in the AD only 6% of the reference compounds are mutagens and 94% are nonmutagens. It can be questioned whether the performance can be determined with sufficient reliability when the reference dataset (1) contains less than 100 compounds and (2) is extremely unbalanced. However, most studies that are available on (Q)SAR performance are based on a much smaller reference dataset than the one used in our study, only calculate global performance and do not calculate the MCC value for unbalanced datasets.

One might reflect on the necessity of going as far as taking into account whether or not a test substance is also included in the model training set. Indeed, the true performance is calculated from new substances not previously known to the model. In practice, however, the user does not make this distinction and collects prediction results for all test substances in the AD, regardless of their new or known status. At most, if a test substance appears to be in the training set, the prediction result is granted a higher degree of certainty. The majority of FCM reference dataset compounds used in the present study are not new to the models, hence their experimental mutagenicity data are already available, however not necessarily in the public domain.

Moreover, it is not surprising that most of the study compounds are nonmutagenic, as it concerns substances that can be used in FCM, hence to which consumers can be exposed when migration occurs. It must furthermore be noted that, in addition to printed paper and board FCM substances, plastic FCM substances are also included in the inventory consulted for the present evaluation (Van Bossuyt *et al.*, 2016). Namely, for plastic FCM, a harmonized European regulation exists (European Union, 2011) and it is recognized that the latter materials and substances can also be used in other FCM types, eg, printing inks, paper (board), and coatings. Many study substances are thus not exclusively applied in printed paper and board but can also be found in several other FCM types. Consequently, the current performance results of the (Q)SAR models for bacterial mutagenicity will be representative for another FCM type proportionally to the overlap of its associated substances with the ones studied here.

CONCLUSION

The performance evaluation of four (Q)SAR models reveals a variable capacity to predict bacterial mutagenicity of printed paper and board FCM substances. Sarah Nexus displays the best global performance, followed by VEGA Consensus and Derek Nexus, whereas Toxtree ranks last. Statistics calculated for the QSARs with AD and training set information indicate that the performance drops significantly for new compounds, however it may be sufficient to focus on the performance in the AD. In the latter case, both VEGA Consensus and Sarah Nexus individually as well as combined perform remarkably good. It must be further explored to what extent these findings are similar for other FCM types and chemicals in general.

FUNDING

This work was supported by the Belgian Scientific Institute of Public Health [WIV-ISP, P2026.0510.1] and the European Commission [LIFE16ENV/IT/000167].

REFERENCES

Ames, B. N., McCann, J., and Yamasaki, E. (1975). Methods for detecting carcinogens and mutagens with the salmonella/mammalian-microsome mutagenicity test. *Mutat. Res.* **31**, 347–364.

Ates, G., Raitano, G., Heymans, A., Van Bossuyt, M., Vanparys, P., Mertens, B., Chesne, C., Roncaglioni, A., Milushev, D., and Benfenati, E. (2016). In silico tools and transcriptomics analyses in the mutagenicity assessment of cosmetic ingredients: A proof-of-principle on how to add weight to the evidence. *Mutagenesis* **31**, 453–461.

Boughorbel, S., Jarray, F., and El-Anbari, M. (2017). Optimal classifier for imbalanced data using Matthews correlation coefficient metric. *PLoS One* **12**, e0177678.

Cassano, A., Raitano, G., Mombelli, E., Fernández, A., Cester, J., Roncaglioni, A., and Benfenati, E. (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 C Environ. Carcinog. Ecotoxical. Rev.* **32**, 273–298.

Cherkasov, A., Muratov, E. N., Fourches, D., Varnek, A., Baskin, I. I., Cronin, M., Dearden, J., Gramatica, P., Martin, Y. C., Todeschini, R., *et al.* (2014). QSAR modeling: Where have you been? Where are you going to? *J. Med. Chem.* **57**, 4977–5010.

EFSA. (2012). Special issue: Food contact materials, flavouring substances and smoke flavourings. *EFSA J.* **10**, s1007.

EFSA. (2016). Recent developments in the risk assessment of chemicals in food and their potential impact on the safety assessment of substances used in food contact materials. *EFSA J.* **14**, 4357.

European Commission. (2016). *Toxtree Software*. Available at: http://toxtree.sourceforge.net/; last accessed December 5, 2016.

European Union. (2006). Regulation (EC) No 1907/2006 of the European Parliament and of the Council of 18 December 2006 concerning the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH). *Off. J. Eur. Union* **L396**, 1–503.

European Union. (2009). Regulation (EC) No 1223/2009 of the European Parliament and of the Council of 30 November 2009 on cosmetic products. *Off. J. Eur. Union* **L342**, 1–345.

European Union. (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. *Off. J. Eur. Union* **L12**, 1138.

Fjodorova, N., Vracko, M., Novic, M., Roncaglioni, A., and Benfenati, E. (2010). New public QSAR model for carcinogenicity. *Chem. Cent. J.* **4**, S3.

Gadaleta, D., Manganelli, S., Manganaro, A., Porta, N., and Benfenati, E. (2016). A knowledge-based expert rule system for predicting mutagenicity (Ames test) of aromatic amines and azo compounds. *Toxicology* **370**, 20–30.

Golbamaki, A., and Benfenati, E. (2016). In silico methods for carcinogenicity assessment. In *In Silico Methods for Predicting Drug Toxicity* (E. Benfenati, Ed.), pp. 107–119. Springer, New York, NY.

Hansen, K., Mika, S., Schroeter, T., Sutter, A., ter Laak, A., Steger-Hartmann, T., Heinrich, N., and Müller, K. R. (2009). Benchmark data set for in silico prediction of Ames mutagenicity. *J. Chem. Inf. Model.* **49**, 2077–2081.

ICH. (2017). Assessment and control of DNA Reactive (mutagenic) impurities in pharmaceuticals to limit potential carcinogenic risk M7(R1).

IRFMN. (2016). *VEGA Software*. Available at: http://www.vega-qsar.eu/, last accessed December 6, 2016.

Judson, P. N., Marchant, C. A., and Vessey, J. D. (2003). Using argumentation for absolute reasoning about the potential toxicity of chemicals. *J. Chem. Inf. Comput. Sci.* **43**, 1364–1370.

Kulkarni, S. A., and Barton-Maclaren, T. S. (2014). Performance of (Q) SAR models for predicting Ames mutagenicity of aryl azo and benzidine based compounds. *J. Environ. Sci. Health C Environ. Carcinog. Ecotoxicol. Rev.* **32**, 46–82.

Leadscope. (2016). *Leadscope Toxicity Database*. Available at: http://www.leadscope.com/toxicity_databases/, last accessed February 28, 2016.

Lhasa Limited. (2016a). *Derek Nexus® Software*. Available at: http://www.lhasalimited.org/products/derek-nexus.htm; last accessed December 5, 2016.

Lhasa Limited. (2016b). *Sarah Nexus® Software*. Available at: http://www.lhasalimited.org/products/sarah-nexus.htm; last accessed December 5, 2016.

Manganelli, S., Benfenati, E., Manganaro, A., Kulkarni, S., Barton-Maclaren, T. S., and Honma, M. (2016). New quantitative structure-activity relationship models improve predictability of ames mutagenicity for aromatic azo compounds. *Toxicol. Sci.* **153**, 316–326.

Manganelli, S., Schilter, B., Benfenati, E., Manganaro, A., and Lo Piparo, E. (2017). Integrated strategy for mutagenicity prediction applied to food contact chemicals. *Altex*. doi: 10.14573/altex.1707171. [Epub ahead of print]

Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta* **405**, 442–451.

National Institute of Health Sciences of Japan. (2016). *AMES/QSAR International Collaborative Study*. Available at: http://www.nihs.go.jp/dgm/amesqsar.html, last accessed February 22, 2016.

OECD. (2014). *Guidance Document on the Validation of (Quantitative) Structure-Activity Relationship [(Q)SAR] Models*. OECD Publishing, Paris.

Raunio, H. (2011). *In silico* toxicology – Non-testing methods. *Front. Pharmacol.* **2**, 33.

Van Bossuyt, M., Van Hoeck, E., Raitano, G., Manganelli, S., Braeken, E., Ates, G., Vanhaecke, T., Van Miert, S., Benfenati, E., Mertens, B., *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.

Van Bossuyt, M., Van Hoeck, E., Vanhaecke, T., Rogiers, V., and Mertens, B. (2016). Printed paper and board food contact materials as a potential source of food contamination. *Regul. Toxicol. Pharmacol.* **81**, 10–19.

Van Hoeck, E., Van Den Houwe, K., Van Bossuyt, M., Vanhaecke, T., Rogiers, V., and Mertens, B. (2017). A safety evaluation of printed paper and board contaminants: Photo-initiators as a case study. *Ref. Module Food Sci.* Elsevier, pp. 1–13.

Williams, R. V., Amberg, A., Brigo, A., Coquin, L., Giddings, A., Glowienke, S., Greene, N., Jolly, R., Kemper, R., O'Leary-Steele, C., *et al.* (2016). It's difficult, but important, to make negative predictions. *Regul. Toxicol. Pharmacol.* **76**, 79–86.