Received: 31 July 2024 | Revised: 2 October 2024 | Accepted: 30 October 2024

DOI: 10.1002/em.22640

RESEARCH ARTICLE

Environmental and Molecular Mutagenesis

Environmental Mutagenesis and Genomics Society

WILEY

Evaluation of the standard battery of in vitro genotoxicity tests to predict in vivo genotoxicity through mathematical modeling: A report from the 8th International Workshop on Genotoxicity Testing

Mirjam Luijten[1] | Jan van Benthem[1] | Takeshi Morita[2] | Raffaella Corvi[3] | Patricia A. Escobar[4] | Yurika Fujita[5] | Jennifer Hemmerich[6] | Naveed Honarvar[6] | David Kirkland[7] | Naoki Koyama[8] | David P. Lovell[9] | Miriam Mathea[6] | Andrew Williams[10] | Stephen Dertinger[11] | Stefan Pfuhler[12] | Jeroen L. A. Pennings[1]

[1]Centre for Health Protection, National Institute for Public Health and the Environment (RIVM), Bilthoven, The Netherlands

[2]National Institute of Technology and Evaluation, Tokyo, Japan

[3]European Commission, Joint Research Centre (JRC), Ispra, Italy

[4]Merck & Co., Inc., Rahway, New Jersey, USA

[5]Institute for Protein Research, Osaka University, Osaka, Japan

[6]BASF SE, Ludwigshafen, Germany

[7]Kirkland Consulting, Tadcaster, UK

[8]Chugai Pharmaceutical Co., Ltd., Translational Research Division, Yokohama, Japan

[9]City St George's, University of London, London, UK

[10]Health Canada, Ottawa, Canada

[11]Litron Laboratories, Rochester, New York, USA

[12]Procter & Gamble, Mason, Ohio, USA

Correspondence
Mirjam Luijten, Centre for Health Protection, National Institute for Public Health and the Environment (RIVM), P.O. Box 1, 3720 BA Bilthoven, The Netherlands.
Email: mirjam.luijten@rivm.nl

Funding information
Ministerie van Landbouw, Natuur en Voedselkwaliteit; Ministerie van Volksgezondheid, Welzijn en Sport

Accepted by: A. Zeller

Abstract

In human health risk assessment of chemicals and pharmaceuticals, identification of genotoxicity hazard usually starts with a standard battery of in vitro genotoxicity tests, which is needed to cover all genotoxicity endpoints. The individual tests included in the battery are not designed to pick up all endpoints. This explains why resulting data can appear contradictory, thereby complicating accurate interpretation of the findings. Such interpretation could be improved through application of mathematical modeling. One of the advantages of mathematical modeling is that the strengths and weaknesses of each test are taken into account. Furthermore, the generated predictions are objective and convey the associated uncertainties. This approach was explored by the working group "Predictivity of In Vitro Genotoxicity Testing," convened in the context of the 8th International Workshop on Genotoxicity

Testing (IWGT). Specifically, we applied mathematical modeling to a database with publicly available in vitro and in vivo data for genotoxicity. The results indicate that a mammalian in vitro clastogenicity test and a mammalian cell gene mutation test together provide strong predictive weight-of-evidence for evaluating genotoxic hazard of a substance, although they are better in predicting absence of genotoxic potential than in predicting presence of genotoxic potential. Remarkably, the bacterial reverse mutation (Ames) test did not significantly change these predictions when used in combination with in vitro mutagenicity and clastogenicity tests using cells of mammalian origin. However, in case only data from a bacterial reverse mutation test are available for the assessment of genotoxic potential, these do bear weight of evidence and thus can be used. Genotoxicity assays are generally executed in tiers, in which the bacterial reverse mutation test often is the starting point. Thus, it is reasonable to suspect that early in development test results from the bacterial reverse mutation test have influenced the composition of the database studied here. We performed several tests on the robustness of the database used for the analyses presented here, and the forthcoming results do not indicate a strong bias. Further research comparing in vitro genotoxicity data with in vivo data for additional compounds will provide more insights whether it is indeed time to reconsider the composition of the standard in vitro genotoxicity battery.

KEYWORDS
genetic toxicity, prediction, risk assessment, uncertainty

1 | INTRODUCTION

Genetic toxicity testing, one of the toxicological effects included in the safety evaluation of diverse substances, encompasses three endpoints: gene mutation, chromosomal damage, and aneugenicity (chromosome loss or gain). Assessment of genetic toxicity for regulatory purposes usually involves a tiered approach, starting with a standard battery of in vitro genotoxicity tests, followed by in vivo testing on the same genotoxicity endpoint in case of a positive in vitro test result (Cimino, 2006; Dearfield et al., 2011; Eastmond et al., 2009). Generally, the conclusion from this approach is limited to hazard identification, that is, it relies on a yes/no binary decision. Characterization of the genotoxic hazard using quantitative analyses has been demonstrated to be feasible and of added value (Beal et al., 2023; Chepelev et al., 2023; Luijten et al., 2020; Nicolette et al., 2021; White et al., 2020); however, this is not yet a standard requirement in the regulations currently in use.

The current standard battery of in vitro genotoxicity tests is rather consistent across regulatory jurisdictions and geographical regions; commonly it includes a bacterial mutagenicity test, a mammalian cell gene mutation test, and/or a mammalian cell chromosomal damage (Cimino, 2006; European Commission, 2008, 2009, 2013; U.S. Food and Drug Administration, 2007; Groff et al., 2021; ICH, 2008) test or a bacterial reverse mutation test and a mammalian cell micronucleus test (EFSA Scientific Committee, 2011; Scientific Committee on Consumer Safety [SCCS], 2023). The performance of the battery for detecting genotoxic potential has been evaluated for a large set of chemicals that are known to be carcinogenic and/or to induce genotoxic effects in rodents (Kirkland et al., 2011). The results of this analysis indicated that a combination of the bacterial reverse mutation test, that is, the Ames test (OECD Test Guideline (TG) 471), and a mammalian in vitro test for chromosomal damage, that is, the in vitro mammalian cell micronucleus test (MNvit; OECD TG 487), would be sufficient to reliably predict genotoxic potential of chemicals. The MNvit was preferred over the in vitro mammalian chromosome aberration test (CAvit; OECD TG 473), because the MNvit is capable of detecting both structural chromosomal damage and aneugenicity. The data used in the analysis followed the principles of the OECD test guidelines for the tests listed above, but not all data may have been generated according to the testing recommendations at that time.

A combination of two or more tests for assessing genetic toxicity is considered necessary to cover the different genotoxicity endpoints; hence, the tests included in the battery strategically differ in the biological mechanisms involved. Due to these differences in coverage of biological mechanisms, data resulting from the battery may seem contradictory. In other words, two different tests may give different results for the same test article. This makes accurate interpretation of the findings from multiple tests sometimes challenging. Moreover, this interpretation is often not entirely objective: it is often based on the biological background of each of the tests included in the battery and the perceived relevance of the endpoint, combined with the knowledge of the expert(s) involved in the data evaluation. Mathematical modeling could provide a useful tool to support interpretation of findings from multiple tests, because it takes into consideration in a defined manner all data included in a dataset used for evaluating the predictive performance of combinations of different tests and not only the data for a single test article. In this way, it considers both strengths and weaknesses of a given test. The outcomes of mathematical modeling inform on the predictive weight of evidence (likelihood) for a given result or combination thereof, as well as the contribution of each test involved in the analysis. Furthermore, predictions derived from mathematical modeling are likely to be more objective and can provide insight into the associated uncertainties.

We applied a mathematical modeling approach to a large database of chemicals comprising publicly available genetic toxicity data, with the aim to evaluate the performance of the in vitro genotoxicity test battery. Different combinations of in vitro genotoxicity tests were analyzed, using the results obtained from in vivo genotoxicity test(s) as a reference. The approach used and the forthcoming results were presented and discussed by the "Predictivity of In Vitro Genotoxicity Testing" working group at the 8th International Workshop on Genotoxicity Testing (IWGT) in Ottawa, which took place in August 2022. The discussion was focused on strengths and weaknesses of the modeling approach, the database with in vitro and in vivo data from genetic toxicity tests, and the interpretation of results. This manuscript describes the outcomes for the different combinations of in vitro genotoxicity tests. Additionally, it provides a summary of the working group's evaluation, discussion, and consensus.

2 | METHODS

2.1 | Data collection

For the mathematical modeling presented here, we compiled a database of chemical substances with in vitro and in vivo data from genotoxicity tests that can be used for regulatory toxicology. The database includes the following in vitro genotoxicity tests: the bacterial reverse mutation test (OECD TG 471; [OECD, 2020]), the mouse lymphoma *tk* mutation test (MLA) (OECD TG 490; [OECD, 2016f]), the *hprt* mutation test (OECD TG 476; [OECD, 2016d]), the in vitro mammalian cell micronucleus test (MNvit, OECD TG 487; [OECD, 2023]), and the in vitro mammalian chromosome aberration test (CAvit, OECD TG 473; [OECD, 2016a]). For in vivo genotoxicity, data were collected for the micronucleus (MN) test (OECD TG 474; [OECD, 2016b]), the chromosome aberration (CA) test (OECD TG 475; [OECD, 2016c]), the in vivo transgenic rodent (TGR) gene mutation assay (OECD TG 488; [OECD, 2022b]), the in vivo mammalian alkaline comet assay (OECD TG 489; [OECD, 2016e]) and the mammalian erythrocyte *Pig-a* gene mutation assay (OECD TG 470; [OECD, 2022a]).

The database was constructed using existing databases that were previously reported (Fujita et al., 2016; Kasamatsu et al., 2021; Kirkland et al., 2011, 2016; Kirkland, Zeiger, Madia, & Gooderham et al., 2014; Madia et al., 2020a, 2020b; Morita et al., 2016; Yamada & Honma, 2018), complemented with data from the European Chemicals Agency (ECHA; https://echa.europa.eu/), data from the database of GHS (Globally Harmonized System of Classification and Labelling of Chemicals) classification results in Japan (https://www.nite.go.jp/chem/english/ghs/ghs_download.html), data from the Japanese Ministry of Health, Labour and Welfare (https://anzeninfo.mhlw.go.jp/user/anzen/kag/sokatutbl.htm), and data used to support the *Pig-a* OECD TG 470 [OECD, 2022a]. We adopted the calls as previously reported where appropriate, except for the MLA, where some existing calls for the MLA were replaced by calls resulting from a recent re-evaluation (Schisler et al., 2018). The calls in the database fell into three categories, that is, positive (+), negative (−), and equivocal (E). The categories inconclusive (I) and uninterpretable (U) were not considered in any of the analyses reported in this manuscript. In total the database contains 2239 chemicals with in vitro and/or in vivo data (Table S1).

2.2 | Considerations used for mathematical modeling of the data

The main purpose of the present study was to evaluate the performance of different combinations of in vitro genotoxicity tests. For this, we focused on different classes of in vitro genotoxicity tests instead of individual tests. The rationale for doing so is because the current regulations commonly require data for a specific genetic toxicity endpoint, but do not specifically prescribe which test to use. In other words, for each genetic toxicity endpoint data may be obtained from different tests. We considered the following three classes of in vitro genotoxicity tests: (i) bacterial reverse mutation test (Ames); (ii) mammalian cell gene mutation test (MLA and/or *hprt*; while *hprt* measures only mutation, MLA measures both chromosomal damage and mutations, but for this analysis MLA is considered a mutation endpoint only); and (iii) mammalian in vitro chromosome damage test (MNvit and/or CAvit). For the latter, it should be noted that we did not distinguish between structural (clastogenicity) and numerical (aneuploidy) chromosome aberrations. Putative changes in chromosome number (chromosome loss) can be detected with a variety of methods; however, the number of micronucleus tests for which aneuploidy was investigated and reported is limited.

For each of the three classes, results from tests were combined per chemical available in the database. In cases where data for more than one test with the same substance was available, the calls were merged in such a way that a positive call overruled (a) negative call(s). Equivocal calls were excluded from the main analysis; however, they were used to check for the robustness of the database (Section 2.4.2).

The results obtained for chemicals tested in an in vivo genotoxicity test were used as a reference for evaluating the performance of in vitro genotoxicity tests. For this, we relied on data from in vivo MN, CA, TGR, comet, and *Pig-a* assays. The following approaches for using in vivo genotoxicity data as reference were applied. The first approach, further referred to as "overall in vivo genotoxicity," involved (per chemical) merging of all results for in vivo genotoxicity tests (i.e., MN, CA, TGR, comet, and *Pig-a*). In cases where results for two or more tests were available, a positive result overruled one or more negative results. The resulting overall call was used as an overall reference for the comparison with in vitro data (see Section 2.4 for the various analyses performed). In the second approach, further referred to as "endpoint-specific in vivo genotoxicity," only in vivo tests that are considered appropriate follow-up tests (according to current regulations) for each of the three classes of in vitro genotoxicity tests were used. Thus, data from the bacterial reverse mutation test and mammalian cell gene mutation tests were compared to data from TGR, *Pig-a* and/or comet, while data from mammalian in vitro chromosomal damage tests were compared to in vivo chromosomal damage tests and/or comet assay. The comet assay, which measures single stranded DNA breaks, is thus used for both in vivo mutation and chromosome damage categories.

2.3 | Chemical applicability domain

For the evaluation of the performance of the three classes of in vitro genotoxicity tests a core set of 309 substances was used. This core set was compiled based on the criterion that data should be available for all three classes of in vitro genotoxicity tests (bacterial reverse mutation test, mammalian cell gene mutation test and mammalian in vitro chromosomal damage test; see also Section 3.1, Figure 1). The chemical applicability domain of the 309 substances was determined as follows. Firstly, physico-chemical properties, for example, molecular weight and logK_{ow}, of this core set were compared to those of about 6500 substances tested in the bacterial reverse mutation test (Hansen et al., 2009). Secondly, physico-chemical properties of the core set were compared to those of all substances for which in vivo data were available in the database we constructed (Section 2.1). To highlight differences in chemical space we plotted several physico-chemical properties as histograms. The histograms were created using the R libraries plotly and ggplot2 (version 4.3.1) [Plotly Technologies Inc., 2015; Wickham, 2016]. To calculate the chemical space, the constructed datasets (i.e., Hansen dataset [Hansen et al., 2009], the core set of 309 substances and the full in vivo set) were combined. All calculations were conducted using the KNIME Analytics Platform (version 4.1.2) (Berthold et al., 2008). Morgan fingerprints (radius: 2; 1024 bits) were calculated using the RDKit Fingerprint KNIME node (Landrum, 2015). The fingerprints were then used for a principal component analysis using the PCA node. The first two principal components were then plotted using R and colored by the respective dataset and activity.

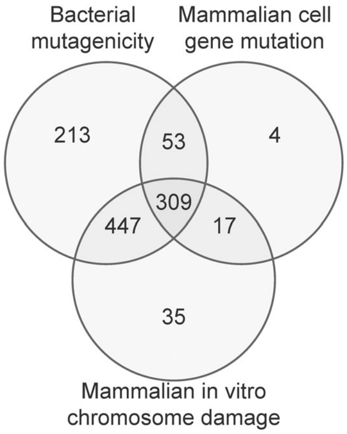
Bacterial mutagenicity
Mammalian cell gene mutation
213
53
4
309
447
17
35
Mammalian in vitro chromosome damage

FIGURE 1 Number of chemicals with data for the three classes of in vitro genotoxicity tests: (i) bacterial reverse mutation test; (ii) mammalian cell gene mutation test; and (iii) mammalian in vitro chromosome damage test.

2.4 | Mathematical modeling

2.4.1 | Predictivity of in vitro genotoxicity tests

All mathematical modeling was performed in R statistical software (version 4.2.0); more details are given below. In accordance with the regulations currently in use for the assessment of the genotoxicity hazard of substances, we conducted the following analyses: (a) prediction of genotoxic potential of each of the three classes of in vitro genotoxicity tests (i.e., bacterial reverse mutation test [Ames], mammalian cell gene mutation tests, or mammalian in vitro chromosomal damage tests); (b) prediction of genotoxic potential using the bacterial reverse mutation test plus a mammalian cell gene mutation test (*hprt* and/or MLA); (c) prediction of genotoxic potential using the bacterial reverse mutation test plus a mammalian cell gene mutation test plus a mammalian in vitro chromosomal damage test; (d) prediction of genotoxic potential using a mammalian cell gene mutation test and a mammalian in vitro chromosomal damage test (MNvit and/or CAvit); (e) prediction of genotoxic potential using the bacterial reverse mutation test plus the MNvit. For these predictions, the "overall in vivo genotoxicity" call based on the outcome of any of the in vivo tests was used as a reference call (Section 2.2). Additionally, combinations of tests were compared to in vivo data from appropriate follow-up tests (Appendices S1 and S2).

We used a workflow based on a combination of methods previously described (Aldenberg & Jaworska, 2010; Jaworska et al., 2010). The methods used include application of Bayes' rule. Bayes' rule, named after the English statistician Thomas Bayes, describes the probability of an event, based on combining prior knowledge of the probability (referred to as the "prior") with data on conditions that might be related to the event, to obtain an updated ("posterior") probability. In the analysis, the prior knowledge is often expressed in the form of a prior probability distribution based on the current knowledge about the mean and variance of the model parameters. There are different approaches to choosing the prior and it should be noted that using a different prior impacts on the posterior probabilities. In our analysis, we assumed for genotoxic potential of chemicals substances having a uniform prior as this is a common approach (Bernardo & Smith, 2000), that is, a 50:50% probability of a substance being negative or positive for in vivo genotoxicity.

The calculations performed are described below. An example of the calculations is provided in Table S2, while the script used for the data analysis is available on GitHub (https://github.com/jlapennings/iwgt_pwoe). First, data for each test combination were taken from the database. Only substances with data for each of the test classes in the combination were used for further calculations.

Next, we used logistic regression to model the probability of an in vivo result, as a function of test combinations:

$$\ln\left(\frac{p(\mathrm{InVivo}^{+})}{p(\mathrm{InVivo}^{-})}\right)=\beta_0+\beta_1\cdot\mathrm{Ames}+\beta_2\cdot\mathrm{MCGM}+\beta_3\cdot\mathrm{Clast}$$

The input test results were *effect-coded*: -1 for negative results and $+1$ for positive results; the in vivo results as (0,1)-coded data. It can be noted here that the subsequent calculations can also be performed if input test results are entered as 0/1; we verified that the final results are identical. However, we prefer to enter input data as $-1/+1$ as this makes it explicit if a test results is positive or negative.

The logistic model fit yielded the binomial probability for an in vivo test result at a given test combination (Table S2, columns J and K). By multiplying these by the number of data points (column I) for a given test battery outcome, we obtained the model predictions (columns L and M) for the number of positives and negatives. Conditional test probabilities (columns N and O) were derived from dividing each model prediction by the totals for in vivo negatives and positives, respectively. These values give the probability for a test (combination) result for a positive or negative substance and therefore generalize the sensitivity and specificity measures for the quality of a single test. It should be noted that these sensitivity and specificity measures differ from the ones traditionally used for genotoxicity testing, where each test carries equal weight. Here, not every test a priori contributes equally to a prediction; the weight of each test is determined by the data in such a way to minimize the overall differences between the predicted probabilities for being positive (for each substance a value between 0 and 1) and the actual in vivo calls. From a regulatory perspective, the value we are most interested in is the probability that a substance is positive or negative given a certain test combination. From the application of Bayes' rule, these are the posterior probabilities (columns P and Q); these are calculated by dividing the conditional positive and negative test probability for each combination by the sum of these two values. Finally, the ratio between the positive and negative posterior probability is expressed as the likelihood ratio (LR; column R), which is converted to the predictive weight of evidence (WoE) as WoE $=10{*}\log_{10}(\mathrm{LR})$. The advantage of a single number implementing WoE is that we can decompose these WoE numbers into contributions from the individual tests. That yields valuable information, because a priori not every test contributes equally to a prediction. We chose to express predictive WoE in units of "deciban," which goes back to Alan Turing (MacKay, 2004) and corresponds to a change in odds of $10^{0.1}$, that is, about 1.26-fold. For example, a WoE value of 2 equals a LR of 1.26^2. Similarly, a WoE value of -10 equals a LR of 0.1 (1.26^{-10}). The WoE numbers are given in column S in Table S2.

It has been shown by Aldenberg and Jaworska (Aldenberg & Jaworska, 2010) that in the case of a uniform prior the predictive weight of evidence can also be calculated as:

$$\mathrm{WoE}=4.343\cdot\Big(\beta_0+(\beta_1\cdot\mathrm{Ames})+(\beta_2\cdot\mathrm{MCGM})+(\beta_3\cdot\mathrm{Clast})-\ln(\mathrm{InVivo}^{+}/\mathrm{InVivo}^{-})\Big)$$

This equation follows from the logistic model. The predictive weight of evidence is a numerical measure based on the log-odds of a positive in vivo result. The log-odds in a logistic model is known to be the linear predictor, the additional correction term, ln(+/−), follows from the test probabilities, correcting for the unbalance between in vivo positives and negatives tested.

Even in a probabilistic estimation procedure we still have parameter uncertainty, due to the model being a simplification of the data as well as the data being inherently limited. Therefore, we simulated the posterior distribution of the WoE by re-running the calculations above on 4000 bootstrap samples. This approach uses a large number of resamplings of the input data to estimate the uncertainty in the WoE outcome at each test battery combination. The resulting credibility limits, given in columns T and U, indicate whether the WoE distribution at each multiple test result can be considered as decisive. Within the context of this study, we considered a predictive WoE as decisive if its absolute value was equal to or greater than 1 and if its 95% credible interval (CI) does not contain zero.

2.4.2 | Robustness checks

The robustness of the database and its impact on the outcomes of the analyses performed was tested in different ways. Firstly, we re-ran the analyses described above, and in this case equivocal data were not ignored but considered as positive test results. Test results were combined as described in Table S3. Secondly, calculations were done with all comet data excluded. Thirdly, various levels of "noise" were introduced. For each of the four parameters used in logistic regression (in vivo genotoxicity, bacterial reverse mutation test, mammalian cell gene mutation, mammalian in vitro chromosome damage), part of the values were replaced by randomly drawn values for that parameter. This was done for 5%, 10%, 15%, and 20% of the data. Finally, we applied five-fold cross-validation. For this purpose, the substances were divided into five groups. Each cross-validation used four out of five groups as a training set to determine WoE values for the remaining substance group. The output for the five substance groups was combined to plot Receiver Operating Characteristic (ROC) curves and calculate the corresponding Area Under the Curve (AUC) using the R package caTools.

3 | RESULTS

3.1 | Composition of the genotoxicity database

The database compiled for this study comprises 2239 chemical substances, that is, industrial chemicals, pesticides, biocides, and cosmetic ingredients as well as pharmaceuticals; collectively these are further referred to as chemicals. Of these, 1078 chemicals have both in vitro and in vivo data for genotoxicity. Of the chemicals with both in vitro and in vivo data for genotoxicity, about 50% (543/1078 chemicals) tested positive in at least one in vivo genotoxicity test, while for the other half (535/1078 chemicals) only negative test results were identified. The distribution of the chemicals across the different mammalian in vitro and in vivo genotoxicity tests is depicted in Figure S1a, b, respectively. For evaluation of the performance of (combinations of) in vitro genotoxicity tests, we focused on different classes of in vitro genotoxicity tests (Section 2): (i) bacterial reverse mutation test (Ames); (ii) mammalian cell gene mutation test (MLA and/or *hprt*); and (iii) mammalian in vitro chromosome damage test (MNvit and/or CAvit). The number of chemicals with data available for each of the three classes of in vitro genotoxicity tests is shown in Figure 1.

As shown in Figure 1, a core set of 309 chemicals had data for all three classes of in vitro genotoxicity tests (bacterial reverse mutation test, mammalian cell gene mutation test and mammalian in vitro chromosome damage test). Of these 309 substances, the majority is considered to be environmental chemicals, while approximately 20% of the substances are considered pharmaceuticals (data not shown). To obtain better insight into the characteristics of the core set, we investigated its chemical applicability domain. Comparison of the physicochemical properties and the chemical fingerprints to a benchmark dataset of approximately 6500 substances developed for in silico prediction of bacterial mutagenicity (Hansen et al., 2009) revealed that the 309 chemicals can be considered representative of chemicals that are typically subjected to genotoxicity safety assessments (Figures 2 and S2).

Similarly, we also compared the core set of 309 chemicals to all chemical substances with in vivo data for genotoxicity in our database, that is, 1078 chemicals. The result of this comparison is shown in Figure 3. Again, the core set was found to be very similar to the total set of chemicals with in vivo data. Although PCA coordinates are based on a large number of chemical fingerprints, it may be of interest that the smaller cluster at the top in Figure 2 contains substances with an (aliphatic or aromatic) nitro group, whereas the substances in the lower cluster lack a nitro group. In Figure 3, division between the clusters on the left and right mostly corresponds to substances that are aliphatic versus those with one or more aromatic rings, respectively.

3.2 | Predicting genotoxic potential using (combinations of) in vitro genotoxicity tests

3.2.1 | Using a single class of in vitro genotoxicity tests for the prediction of genotoxic potential

As a first step we evaluated the performance of each of the classes of in vitro genotoxicity tests individually, using overall in vivo genotoxicity (see Section 2.2 for details) as a reference. For evaluation of the bacterial reverse mutation test, that is, the Ames test, a dataset of 1022 chemical substances was available: 516 substances with a negative Ames test result and 506 substances with a positive Ames test

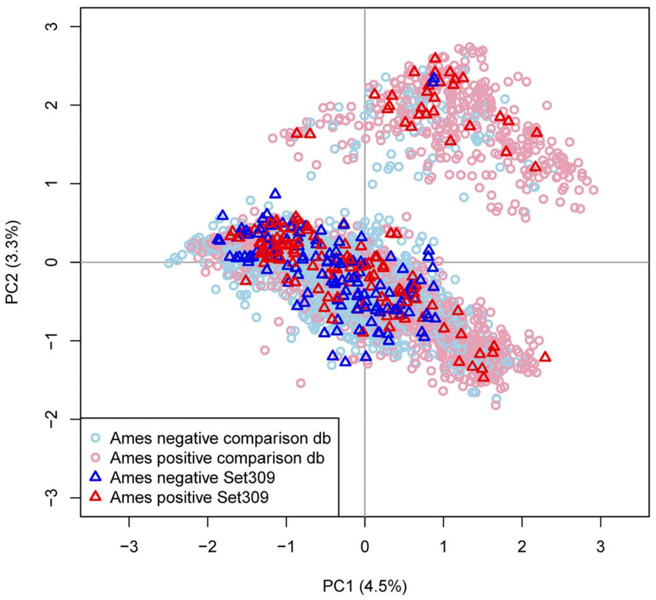

FIGURE 2 Comparison of the chemical applicability domain of the core set of 309 chemicals versus the chemical applicability domain of a large benchmark dataset (Hansen et al., 2009). Blue colors: Chemicals with a negative test result for the bacterial reverse mutation test; red colors: Chemicals with a positive test result for the bacterial reverse mutation test. Substances from the dataset by Hansen et al. are depicted in light-colored circles, while substances from the core set are depicted in dark-colored triangles.

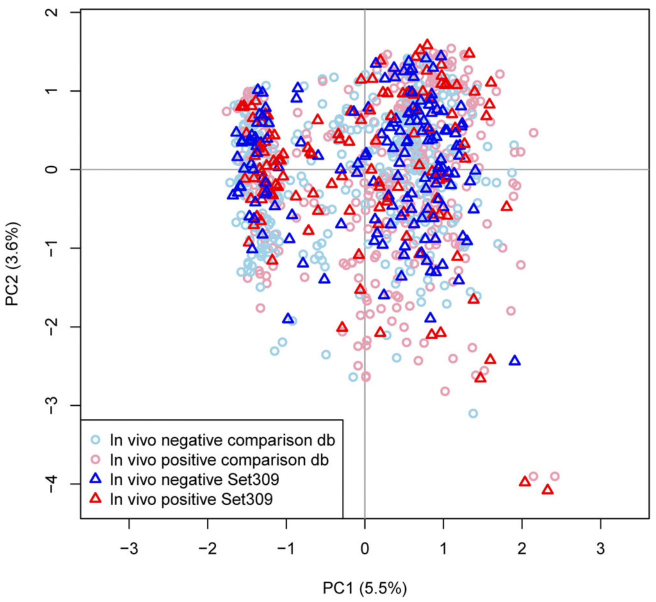

FIGURE 3 Comparison of the chemical applicability domain of the core set of 309 chemicals versus the chemical applicability domain of 1078 chemicals with in vivo data. Blue colors: Chemicals with a negative test result for the bacterial reverse mutation test; red colors: Chemicals with a positive test result for the bacterial reverse mutation test. Substances from the in vivo dataset are depicted in light-colored circles, while substances from the core set are depicted in dark-colored triangles.

result (Table S4). This dataset was used to calculate (in the stepwise fashion described in Section 2) the weight of evidence (WoE) for predicting presence or absence of genotoxic potential and the associated uncertainty when using the Ames test. First, we applied logistic regression to estimate parameters (coefficients) from the data; these parameters are used as input for subsequent steps in the calculations. The estimates were used to calculate the binomial probabilities and the model predictions, followed by the test probabilities as normalized model predictions for the respective outcome (positive or negative). This yielded a sensitivity of 0.632 and a specificity of 0.638 (Table S4, lower panel), which is comparable to previous analyses (Kirkland et al., 2011; Kirkland, Zeiger, Madia, Gooderham, et al., 2014).

Next, posterior distributions were calculated, followed by calculating the predictive WoE, including evaluation of the associated uncertainty by applying bootstrapping (Section 2 for details). The analysis revealed that a bacterial reverse mutation test on its own has a reasonable performance for predicting genotoxic potential of a chemical substance: the WoE values equal −2.39 (95% CI LL [Credible Interval Lower Limit] = −2.98; 95% CI UL [Upper Limit] = −1.83) and 2.42 (LL = 1.87; UL = 3.02) for a negative and a positive result in the bacterial reverse mutation test, respectively (Table 1). Generally, WoE values are considered to be meaningful when exceeding 1 in absolute value (Aldenberg & Jaworska, 2010). The confidence limits indicate whether the WoE distribution at each test result can be considered as decisive. Hence, from a statistical perspective, a negative bacterial reverse mutation test result can be considered equally reliable as a positive bacterial reverse mutation test result. A positive test result gives the same weight to a change from the prior to a posterior probability as a negative test result.

Similarly, the performance of the mammalian cell gene mutation test (MLA and/or *hprt*) was evaluated (Table S5). For the mammalian cell gene mutation test, the values obtained for sensitivity and specificity are 0.889 and 0.616, respectively (Table S5, lower panel). These are comparable to previously reported values (Kirkland et al., 2011; Kirkland, Zeiger, Madia, & Corvi, 2014). The WoE values for the prediction of genotoxic potential using a mammalian cell gene mutation test were somewhat higher than those obtained for the bacterial reverse mutation test, that is, −7.44 (LL = −9.70; UL = −5.75) and 3.64 (LL = 2.90; UL = 4.48) for a negative and a positive result, respectively (Table 1). However, it must be noted that the mammalian cell gene mutation test dataset comprised a smaller number of chemicals compared to the dataset from the bacterial reverse mutation test (i.e., 145 chemicals with a negative test result and 238 chemicals with a positive test result; Table S5).

TABLE 1 Overview of performance of single in vitro genotoxicity tests.

Test; # chemicals[a]	Weight of evidence [95% CI LL; UL][b]
Bacterial reverse mutation test (Ames); *n* = 1022	Negative: −2.39 [−2.98; −1.83]
	Positive: 2.42 [1.87; 3.02]
Mammalian cell gene mutation test; *n* = 383	Negative: −7.44 [−9.70; −5.75]
	Positive: 3.64 [2.90; 4.48]
Mammalian in vitro chromosomal damage test; *n* = 808	Negative: −4.33 [−5.52; −3.31]
	Positive: 1.74 [1.35; 2.17]

Note: The colors used are red and green, where red indicates a positive result and green a negative result.
[a]Number of chemicals included in the analysis.
[b]LL = lower limit, UL = upper limit of 95% credible interval (CI).

For the evaluation of the mammalian in vitro chromosome damage test, a dataset of 808 chemicals was available, with 247 chemicals with a negative test result and 561 chemicals with a positive test result (Table S6). The calculations yielded a sensitivity of 0.838 and a specificity of 0.439 (Table S6, lower panel). The specificity was relatively low, compared to previously published results, while the sensitivity was similar (Kirkland et al., 2011; Kirkland, Zeiger, Madia, Gooderham, et al., 2014). The WoE value for a negative test result using a mammalian in vitro chromosome damage test was −4.33 (LL = −5.52; UL = −3.31), while the WoE value for a positive test result was 1.74 (LL = 1.35; UL = 2.17) (Table 1).

Besides "overall in vivo genotoxicity" data we also used "endpoint-specific in vivo genotoxicity" data as a reference for each of the three classes of in vitro genotoxicity tests (see Section 2.2). The outcomes for the individual comparisons for each of the three classes of in vitro genotoxicity tests are given in Tables S4–S6. Using endpoint-specific in vivo data as a reference yielded almost identical results to those obtained when comparing with overall in vivo genotoxicity for the bacterial mutagenicity test and the in vitro clastogenicity test (Tables S4 and S6; upper panel versus lower panel), while the performance of the mammalian cell gene mutation test was less strong for predicting in vivo mutagenicity compared to overall in vivo genotoxicity (Tables S5, upper panel versus lower panel). It should be noted that the number of chemicals involved in the endpoint-specific comparisons was substantially lower than in the comparisons using overall in vivo genotoxicity data as a reference.

3.2.2 | Predicting genotoxic potential using different combinations of different classes of in vitro genotoxicity tests

Since regulatory safety testing typically considers a combination of in vitro genotoxicity tests to ensure coverage of the three different genetic toxicity endpoints, we also analyzed various combinations of different classes of in vitro genotoxicity tests. In contrast to how sensitivity and specificity values are typically calculated for combinations of results of genotoxicity tests, our approach takes into account the contribution (weight) of each test, which is determined through application of logistic regression to the data (see Section 2.4). Therefore, we refrained from comparing our results to sensitivity and specificity calculations.

The first combination studied was the bacterial reverse mutation test plus the mammalian cell gene mutation test. For the evaluation of

TABLE 2 Evaluation of the bacterial and mammalian cell mutagenicity test combined.

Bacterial reverse mutation	Mammalian cell gene mutation	Negative in vivo	Positive in vivo	Sum	Likelihood ratio	Weight of evidence [95% CI LL; UL][a]
Negative	Negative	79	13	92	0.145	−8.38 [−11.10; −6.34]
Negative	Positive	29	42	71	1.933	2.86 [1.20; 4.73]
Positive	Negative	41	4	45	0.197	−7.05 [−9.73; −4.95]
Positive	Positive	44	110	154	2.630	4.20 [3.18; 5.40]
Total		193	169	362		
		Bacterial reverse mutation				Mammalian cell gene mutation
Estimate β		0.15				1.29
Pr(>|z|)		0.24				2.2E−17

Note: The colors used are red and green, where red indicates a positive result and green a negative result.

[a] LL = lower limit, UL = upper limit of 95% credible interval.

this combination a dataset comprising 362 chemicals was used (Tables 2 and S7). The predictive WoE value for concordant test results for both types of tests was strong: −8.38 (95% CI: −11.10; −6.34) for a negative test result and 4.20 (95% CI: 3.18; 5.40) for a positive test result (Table 2). In case of contradictory test results, the evidence indicates that the outcome is essentially determined by the mammalian cell gene mutation test; the corresponding WoE values are 2.86 (95% CI: 1.20; 4.73) for a negative bacterial reverse mutation test plus a positive mammalian cell gene mutation test and −7.05 (95% CI: −9.73; −4.95) for a positive bacterial reverse mutation test plus a negative mammalian cell gene mutation test. The limited weight of the bacterial reverse mutation test is also reflected in the low coefficient value and the non-significance of the associated probability (estimate β and Pr(>|z|) in Table 2). For the comparison with endpoint-specific data a dataset comprising 133 chemicals was available (Table S7; upper panel). The WoE value for concordant test results for both types of tests remained strong. In the case of a positive bacterial mutagenicity test plus a negative mammalian cell gene mutation test, the evidence is in line with the negative mammalian cell gene mutation test. However, the WoE for a negative bacterial reverse mutation test plus a positive mammalian cell gene mutation test was equivocal (Table S7; upper panel).

The next combination that was evaluated involved all three classes of in vitro genotoxicity tests (Tables 3 and S8). Results from the analysis of the core dataset of 309 chemicals revealed that the WoE for concordant test results was strong, with higher WoE values for all-negative test results (−10.26 [95% CI: −14.46; −7.43]) compared to all-positive test results (4.24; [95% CI: 3.11; 5.58]). Based on this dataset, the mammalian cell gene mutation test is observed to carry most weight, while the bacterial reverse mutation test was assigned almost no weight (estimate β and Pr(>|z|) in Table 3). This is also reflected in Figure 4, in which the WoE values for the predictive performance of a combination of three classes of in vitro genotoxicity tests is depicted.

In the case of discordant results, that is, a positive mammalian cell gene mutation test but a negative bacterial reverse mutation test and a negative mammalian in vitro clastogenicity test, the WoE observed (−0.99 [95% CI: −4.49; 2.07]) was considered equivocal. The same applies for a positive bacterial reverse mutation test and a positive mammalian cell gene mutation test but a negative mammalian in vitro clastogenicity test (WoE = −1.06 [95% CI: −4.70; 2.32]). For the remaining patterns of test results, the absolute WoE values obtained were far larger than 1, and thus considered meaningful. The mammalian cell gene mutation test appeared to be the driver of these outcomes, as reflected in the β estimates and associated probabilities (Table 3).

Given the strong weight assigned to the two classes of mammalian cell in vitro genotoxicity tests, we also evaluated the combination of these two classes. The analysis of data for 326 chemicals showed again a strong WoE for concordant test results, with a higher WoE value for negative test results compared to positive test results (Tables 4 and S8). Similar to the combination of three classes of tests, the mammalian cell gene mutation test derived more weight compared to the mammalian in vitro chromosome damage test. As reflected by the β estimates and associated probability scores in Table 4, both classes of tests contribute significantly to the prediction of genotoxic potential or lack thereof. The WoE was strong for a negative mammalian cell gene mutation test and a positive mammalian in vitro chromosome damage test (−5.07 [95% CI: −7.71; −3.09]), but far from convincing for the reverse situation (−0.50 [95% CI: −3.64; 2.34]; Table 4). The latter result could be perceived as being caused by the low number of chemicals having this combination of test result (i.e., 21 out of 326). This is not necessarily true, as exemplified by the results shown in Table 3. In fact, one of the strengths of our mathematical modeling is that it takes into account all data included in the dataset.

Finally, we also analyzed the combination of the bacterial reverse mutation test plus the MNvit, because this combination is specifically recommended, for example, for food and feed substances (EFSA Scientific Committee, 2011) and cosmetic ingredients (SCCS (Scientific Committee on Consumer Safety), 2023). The estimated coefficients indicated that far more weight should be assigned to the MNvit compared to the bacterial reverse mutation test (Tables 5 and S8). The WoE for concordant test results, based on 236 chemicals, was strong, as was the WoE for a negative MNvit and a positive bacterial reverse

TABLE 3 Evaluation of three classes of in vitro genotoxicity tests combined.

Bacterial reverse mutation	Mammalian cell gene mutation	Mammalian in vitro chromosomal damage	Negative in vivo	Positive in vivo	Sum	Likelihood ratio	Weight of evidence [LL; UL][a]
Negative	Negative	Negative	47	6	53	0.094	−10.26 [−14.46; −7.43]
Negative	Negative	Positive	23	7	30	0.319	−4.96 [−8.07; −2.46]
Negative	Positive	Negative	9	4	13	0.796	−0.99 [−4.49; 2.07]
Negative	Positive	Positive	16	32	48	2.695	4.31 [2.37; 6.65]
Positive	Negative	Negative	13	1	14	0.093	−10.32 [−14.96; −6.90]
Positive	Negative	Positive	20	2	22	0.314	−5.03 [−7.94; −2.66]
Positive	Positive	Negative	5	2	7	0.783	−1.06 [−4.70; 2.32]
Positive	Positive	Positive	34	88	122	2.654	4.24 [3.11; 5.58]
Total			167	142	309		

	Bacterial reverse mutation	Mammalian cell gene mutation	Mammalian in vitro chromosomal damage
Estimate β	−0.01	1.07	0.61
Pr(>|z|)	0.96	2.1E−10	1.4E−3

Note: The colors used are red and green, where red indicates a positive result and green a negative result.
[a]LL = lower limit, UL = upper limit of 95% credible interval.

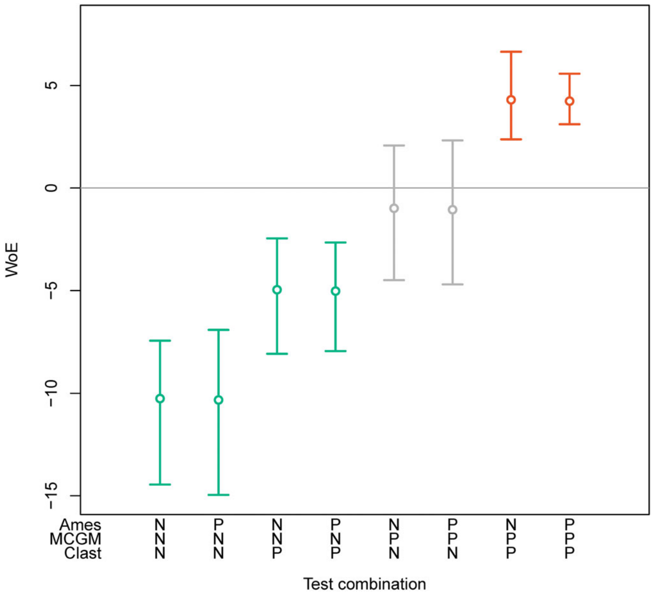

FIGURE 4 Predictive performance of a combination of three classes of in vitro genotoxicity tests, that is, the bacterial reverse mutation test (Ames), the mammalian cell gene mutation test (MCGM) and the mammalian in vitro clastogenicity test (Clast). Predictive performance is expressed as weight of evidence (WoE) values. Analyses were performed for a core set of 309 chemicals, using overall in vivo genotoxicity data as a reference. N = negative; P = positive.

mutation test, which, like the double negative, signaled the absence of a genotoxicity potential in vivo (Table 5). In the case of a positive MNvit and a negative bacterial reverse mutation test, however, the results should be interpreted with care, given that the WoE for this pattern of results was equivocal (1.28 [95% CI: −0.24; 3.03]; Table 5).

3.3 | Robustness checks

For the type of analyses presented above, characteristics of the datasets used may have a significant impact on the outcomes. Datasets that differ strongly in their make-up, for example by each comprising data for a very specific class of chemicals, may yield results that are far from representative for the universe of chemicals. Therefore, we investigated the robustness of the database and its impact on the outcomes of the analyses (see Section 2 for details). Rather than excluding equivocal data from the analyses (as was done for the analyses presented in Section 3.2), equivocal data were considered as positive test results, based on the rationale that for chemicals with an equivocal call genotoxic hazard cannot be excluded. Inclusion of equivocal data as positive did not change our findings: for each of the analyses performed, the results obtained were similar to when excluding equivocal data (Table S9). Next, we analyzed the data with all comet data excluded. This was done because the comet assay is an indicator test for genotoxicity, meaning that it does not detect gene mutation, clastogenicity or aneugenicity directly. The damage inflicted may result in a permanent lesion of the DNA, for example, a gene mutation, but it may also be repaired. Exclusion of the comet data did not appear to impact on the outcomes (i.e., weight of evidence and associated confidence intervals) of the analyses (Table S10); however, it should be noted that the number of chemicals with relevant available results was considerably smaller and we can therefore not draw a firm conclusion. Another approach employed for evaluating robustness was to introduce various levels of "noise" to the database by randomly changing a positive test result into a negative test result and vice versa. Re-analysis of the data revealed that adding up to 20% noise did not affect the overall conclusions (Table S11). Finally, we applied five-fold cross-validation. The resulting Receiver Operating Characteristic (ROC) curves and the corresponding Area Under the Curve (AUC) are shown in Figure 5. It should be noted that the sensitivity and

TABLE 4 Evaluation of the two classes of mammalian in vitro genotoxicity tests combined.

Mammalian cell gene mutation	Mammalian in vitro chromosomal damage	Negative in vivo	Positive in vivo	Sum	Likelihood ratio	Weight of evidence [95% CI LL; UL][a]
Negative	Negative	61	8	69	0.105	−9.78 [−13.62; −7.16]
Negative	Positive	47	10	57	0.311	−5.07 [−7.71; −3.09]
Positive	Negative	14	7	21	0.891	−0.50 [−3.64; 2.34]
Positive	Positive	52	127	179	2.636	4.21 [3.32; 5.26]
Total		174	152	326		

	Mammalian cell gene mutation	Mammalian in vitro chromosomal damage
Estimate β	1.07	0.54
Pr(>|z|)	7.7E−12	2.2E−03

Note: The colors used are red and green, where red indicates a positive result and green a negative result.

[a]LL = lower limit, UL = upper limit of 95% credible interval.

TABLE 5 Evaluation of the bacterial reverse mutation (Ames) test and the in vitro micronucleus (MNvit) test combined.

Bacterial reverse mutation	MNvit	Negative in vivo	Positive in vivo	Sum	Likelihood ratio	Weight of evidence [95% CI LL; UL][a]
Negative	Negative	28	15	43	0.353	−4.52 [−6.72; −2.47]
Negative	Positive	18	58	76	1.341	1.28 [−0.24; 3.03]
Positive	Negative	6	10	16	0.419	−3.78 [−6.64; −1.24]
Positive	Positive	27	74	101	1.591	2.02 [0.62; 3.67]
Total		79	157	236		

	Bacterial reverse mutation	MNvit
Estimate β	0.09	0.67
Pr(>|z|)	0.57	4.2E−05

Note: The colors used are red and green, where red indicates a positive result and green a negative result.

[a]LL = lower limit, UL = upper limit of 95% credible interval.

specificity used in this analysis are similar to those from the analyses presented in previous sections. Conceptually, they are similar to those that are traditionally used for genotoxicity testing. However, in the analyses presented here and in previous sections, not all tests are necessarily attributed equal weights and the weight of each test is determined by the data. The cross-validation analysis confirmed the previous observation that individual classes of genotoxicity tests using mammalian cells have especially good sensitivity, but are less strong in terms of specificity. Combining a mammalian cell gene mutation test with a mammalian in vitro chromosome damage test substantially improves the performance compared to using a single mammalian cell test. Adding the bacterial reverse mutation test to this combination, however, does not further improve the prediction, as reflected in the AUC (0.753 versus 0.754 for three classes of in vitro genotoxicity tests and mammalian cell tests only, respectively; Figure 5).

4 | DISCUSSION AND CONCLUSION

In the present manuscript, we applied mathematical modeling to a genotoxicity database to evaluate the performance of the in vitro test battery to predict in vivo genotoxicity. In other words, the results obtained from in vivo genotoxicity tests were used as a reference. To the best of our knowledge, this is the first time that this approach has been employed for the evaluation of in vitro genotoxicity test to predict genotoxicity; previous work was focused on the prediction of carcinogenicity (Kim & Margolin, 1994; Rosenkranz et al., 1985). Of course, as with all such exercises, the composition of the database used is bound to have an impact on the outcomes of the analyses performed (Burgoon et al., 2023). Therefore, we used well-known datasets on genotoxicity as a basis for the creation of our database. More specifically, we mainly relied on publicly available datasets that have previously been used for evaluating the predictive value of in vitro tests used in the standard battery for genotoxicity (Kirkland et al., 2005; Kirkland et al., 2011; Kirkland, Zeiger, Madia, & Corvi, 2014; Kirkland, Zeiger, Madia, Gooderham, et al., 2014; Madia et al., 2020a; Madia et al., 2020b). We did not analyze the full database in terms of the chemicals categories of intended use (e.g., industrial chemicals, cosmetic ingredients, plant protection products). Therefore, it could very well be that some classes of chemicals (intended use, chemical properties) are somewhat over- or underrepresented. A screen of the core set of 309 chemicals, that was used for

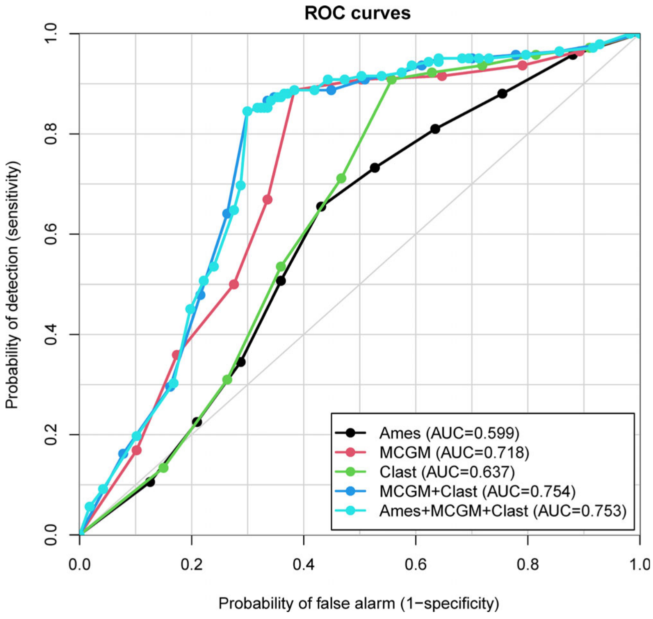

FIGURE 5 Receiver operating characteristic (ROC) curves and the corresponding Area under the curve (AUC) for a five-fold cross-validation performed for the individual classes of in vitro genotoxicity tests (bacterial reverse mutation (Ames) in black; mammalian cell gene mutation (MCGM) in red; mammalian in vitro clastogenicity (Clast) in green), a combination of all three classes (light blue), and mammalian in vitro tests only (dark blue).

the evaluation of the three classes of in vitro genotoxicity tests combined (Table 3), revealed that its majority consists of (industrial) chemicals, followed by pharmaceuticals (∼20%), agrochemicals (∼10%) and other types of substances. Comparison of the chemical applicability domain of this core set to a much larger set of substances (Hansen et al., 2009) revealed that the core set is quite representative of the substances that are typically reviewed for genotoxicity hazard identification. Additionally, we employed various approaches to test the robustness of our database and these analyses confirmed that it is unlikely that the findings obtained are strongly biased by the dataset used. Hence, we consider it justified to conclude that the database is well-suited for the purposes of our study.

As part of the 8th IWGT held in Ottawa, Canada, in August 2022, all available information on the database and the various analyses performed (as described in the previous sections) was presented to the "Predictivity of In Vitro Genotoxicity Testing" working group for discussion. The working group discussed the presented information and delineated six consensus statements. These are described below.

Firstly, the analyses for the individual tests were presented and discussed. Based on the findings obtained (summarized in Table 1) the working group concluded: "*All three types of genotoxicity tests analyzed, i.e., the bacterial reverse mutation test (Ames), the mammalian cell gene mutation test, and the mammalian in vitro chromosome damage test, are individually predictive for what they are supposed to predict: genotoxic potential (or lack thereof) of a chemical. The analyses performed also showed that genotoxicity tests with mammalian cells are better in predicting absence of genotoxic potential than in predicting presence of genotoxic potential.*" This is reflected by the stronger weights of evidence for negative predictions compared to positive predictions (Tables 1 and S4–S6).

Regarding the use of combinations of tests, it was an unexpected finding that the bacterial reverse mutation test, when part of a test battery, did not significantly change the prediction for any of the analyses performed. This observation holds true when the bacterial reverse mutation test is combined with the MNvit (Table 5), but especially when used in combination with a mammalian cell gene mutation test and a mammalian in vitro chromosome damage test (Table 3). This remarkable finding was one of the drivers for further analysis of the composition of the database and its robustness. Given the outcomes of these additional analyses the majority of the working group concluded the following: "*When using a battery of three genotoxicity tests, i.e., a bacterial reverse mutation test (Ames), a mammalian cell gene mutation test, and a mammalian in vitro clastogenicity test, the results of the bacterial reverse mutation test will not contribute much to the final call on genotoxic potential.*" It should be noted that this statement does not apply when the bacterial reverse mutation test is used as a stand-alone test. The latter is important given that under some regulations for chemicals (e.g., REACH 1–10 tpa [European Commission, 2008] or impurities in pharmaceuticals [ICH, 2023]) results from the bacterial reverse mutation test are the only experimental data requested. Therefore, the working group felt the need to stress this point in the following statement: "*In case only bacterial reverse mutation test data are available for the assessment of genotoxic potential, these do bear weight of evidence and thus can be used.*" The discussions held in the context of the 8th IWGT also included the issue of two versus three tests (Kirkland et al., 2005, 2011). Given the limited contribution of the bacterial reverse mutation test to the prediction of genotoxic potential when used in combination with one or more mammalian cell tests, plus the fact that a combination of tests should cover the different endpoints in genetic toxicity, the working group concluded: "*For a battery comprising two genotoxicity tests, a combination of two mammalian cell tests is highly preferred because of their overall high predictive value. However, in the case of a positive mammalian cell gene mutation and a negative chromosomal damage test, the result has little predictive value; this will not be improved by adding a bacterial reverse mutation test.*"

Regarding the methodology used, the working group concluded: "*The mathematical modelling applied is a valuable approach for assessing the predictive value of combinations of toxicity tests.*" Furthermore, this approach can also be used to inform which test to use next, in addition to existing data. In the context of the ongoing paradigm shift in regulatory toxicology and the advances made in the area of next-generation risk assessment, we consider these important findings. However, the working group also noted that "*database(s) required for such analyses should be carefully developed and interpretation of the results should be done with caution.*" For decades now, the bacterial reverse mutation test has played a central role in genotoxic chemical hazard identification. The relative simplicity of the assay, its widespread use, and its key position in regulatory guideline-recommended batteries have all contributed to this. As a practical matter, genotoxicity assays have generally been executed in tiers, with the bacterial reverse mutation test being one of the earliest, if not the very first, benchtop assay. For most product classes, a mutagenicity signal is a severe product liability that is challenging to overcome. Thus, it is reasonable to suspect that early in development bacterial reverse mutation test results have influenced the composition of the database studied here. Similarly, the fact that MLA detects both gene mutation and chromosomal damage, but for this analysis was considered a mutation endpoint, may have had some impact on the results. In ways that are challenging to predict, this may have biased the database we used, and consequently the results of our mathematical modeling exercises, and therefore their generalizability to new, previously untested chemicals. We performed several tests on the robustness of the database used for the analyses presented here, and the forthcoming results do not indicate a strong bias.

Regarding the prior, we consider the use of a 50:50 probability for a chemical being in vivo negative or positive justified, given that of the chemicals with both in vitro and in vivo data in our database about 50% tested positive in at least one in vivo genotoxicity test. For a different database, a different prior might be applicable. In Table S12 we illustrate how using a different prior can affect the WoE outcomes. The top panel shows calculations done for the prior ratio of 1, that is, a 50:50 probability for a substance being in vivo negative or positive, respectively, as used as a default in this manuscript. The second panel shows the same calculations using a 60:40 negative/positive prior ratio and the third panel uses a 40:60 prior ratio. WoE values are somewhat lower and higher, respectively, but the overall calls for a positive/negative/equivocal prediction do not change.

The various discussions held at the 8th IWGT also brought forward the recommendations to look into the chemical applicability domain of the core set of 309 chemicals and to apply cross-validation with the aim to investigate the robustness of the predictions made. Both recommendations have been taken into account by the working group following IWGT. Another recommendation was to verify the outcomes of our analyses using different dataset(s). We fully support this recommendation; however, this was considered out of scope for the present study. In case such a new dataset would be constructed, we suggest to also include data for tests that have been conducted more recently and are therefore likely to comply with current OECD guidelines, while many of the tests in the datasets used herein were from old studies and many did not comply with current testing recommendations. Preferably, model validation techniques should be considered for such exercise (Harrell, 2015). Also, we suggest to include data for tests that are not part (yet) of the standard battery for genotoxicity, e.g. MultiFlow (Bryce et al., 2017; Dertinger et al., 2019) or ToxTracker (Hendriks et al., 2012, 2016). This would allow for a wider discussion on the preferred composition of a standard battery of in vitro tests for genotoxic potential. Such discussions should be initiated sooner rather than later with representatives of relevant organizations, in order to ensure that different stakeholders benefit from the relevant insights from the current and future analyses.

AUTHOR CONTRIBUTIONS

Conceptualization of the analyses performed, J.L.A.P., J.v.B. and M.L.; initial discussions, J.v.B., M.L., J.L.A.P, S.P., D.K., D.L.; methodology, J.L.A.P, D.L., and A.W.; analysis for the studies described, J.L.A.P and S.D.; data contributions, D.K., N.K., T.M., Y.F., and S.D.; curation for the database, J.L.A.P; chemical space analysis, J.H., M.M., N.H.; writing—original draft preparation, M.L., J.L.A.P, J.v.B.; writing—review and editing, T.M., R.C., P.A.E., Y.F., J.H., N.H., D.K., N.K., D.L., M.M., A.W., S.D., and S.P. All authors have read and agreed to the published version of the manuscript.

ACKNOWLEDGMENTS

We would like to acknowledge Frank Le Curieux (European Chemicals Agency, ECHA) for his help in identifying relevant in vitro and in vivo data from the ECHA disseminated database. We thank the IWGT organizing committee for all their efforts and are grateful for the valuable input provided by Tom Aldenberg, Tim McGovern, Minne B. Heringa, Kerry L. Dearfield, and the members of the audience of the 8th IWGT.

FUNDING INFORMATION

The research described in this manuscript was financially supported by the Dutch Ministry of Health, Welfare and Sport and the Dutch Ministry of Agriculture, Nature and Food Quality.

CONFLICT OF INTEREST STATEMENT

This publication stems from a working group of the IWGT. The authors declare no conflict of interest. The funders had no role in the design of this study; in the collection, analyses or interpretation of data; in the writing of the manuscript or in the decision to publish the results.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available in Zenodo at https://zenodo.org, reference number 8363786.

DISCLAIMER

This document represents the consensus of the authors' views expressed as individual scientists and does not necessarily represent the policies and procedures of their respective institutions.

ORCID

Mirjam Luijten https://orcid.org/0000-0002-5277-1443
Jennifer Hemmerich https://orcid.org/0000-0003-0372-8956
Miriam Mathea https://orcid.org/0000-0002-3214-1487
Stefan Pfuhler https://orcid.org/0000-0001-8869-5975
Jeroen L. A. Pennings https://orcid.org/0000-0002-9188-6358

REFERENCES

Aldenberg, T. & Jaworska, J.S. (2010) Multiple test in silico weight-of-evidence for toxicological endpoints. In: Cronin, M.T.D. & Madden, J.C. (Eds.) *Silico toxicology: principles and applications*. Cambridge: Royal Society of Chemistry (RSC Publishing), pp. 558–583.

Beal, M.A., Audebert, M., Barton-Maclaren, T., Battaion, H., Bemis, J.C., Cao, X. et al. (2023) Quantitative in vitro to in vivo extrapolation of genotoxicity data provides protective estimates of in vivo dose. *Environmental and Molecular Mutagenesis*, 64, 105–122.

Bernardo, J.M. & Smith, A.F.M. (2000) *Bayesian theory*. Chichester: Wiley.

Berthold, M.R., Cebron, N., Dill, F., Gabriel, T.R., Kötter, T., Meinl, T. et al. (2008) *Knime: the konstanz information miner. Data analysis, machine learning and applications*. Berlin, Heidelberg: Springer Berlin Heidelberg.

Bryce, S.M., Bernacki, D.T., Bemis, J.C., Spellman, R.A., Engel, M.E., Schuler, M. et al. (2017) Interlaboratory evaluation of a multiplexed high information content in vitro genotoxicity assay. *Environmental and Molecular Mutagenesis*, 58, 146–161.

Burgoon, L.D., Kluxen, F.M. & Frericks, M. (2023) Understanding and overcoming the technical challenges in using in silico predictions in regulatory decisions of complex toxicological endpoints—a pesticide perspective for regulatory toxicologists with a focus on machine learning models. *Regulatory Toxicology and Pharmacology: RTP*, 137, 105311.

Chepelev, N., Long, A.S., Beal, M., Barton-Maclaren, T., Johnson, G., Dearfield, K.L. et al. (2023) Establishing a quantitative framework for regulatory interpretation of genetic toxicity dose-response data: margin of exposure case study of 48 compounds with both in vivo mutagenicity and carcinogenicity dose-response data. *Environmental and Molecular Mutagenesis*, 64, 4–15.

Cimino, M.C. (2006) Comparative overview of current international strategies and guidelines for genetic toxicology testing for regulatory purposes. *Environmental and Molecular Mutagenesis*, 47, 362–390.

Dearfield, K.L., Thybaud, V., Cimino, M.C., Custer, L., Czich, A., Harvey, J.S. et al. (2011) Follow-up actions from positive results of in vitro genetic toxicity testing. *Environmental and Molecular Mutagenesis*, 52, 177–204.

Dertinger, S.D., Kraynak, A.R., Wheeldon, R.P., Bernacki, D.T., Bryce, S.M., Hall, N. et al. (2019) Predictions of genotoxic potential, mode of action, molecular targets, and potency via a tiered multiflow(r) assay data analysis strategy. *Environmental and Molecular Mutagenesis*, 60, 513–533.

Eastmond, D.A., Hartwig, A., Anderson, D., Anwar, W.A., Cimino, M.C., Dobrev, I. et al. (2009) Mutagenicity testing for chemical risk assessment: update of the who/ipcs harmonized scheme. *Mutagenesis*, 24, 341–349.

EFSA Scientific Committee. (2011) Scientific opinion on genotoxicity testing strategies applicable to food and feed safety assessment. *EFSA Journal*, 9, 2379.

European Commission. (2008) Council regulation (ec) no 440/2008 of 30 may 2008 laying down test methods pursuant to regulation (ec) no 1907/2006 of the european parliament and of the council on the registration, evaluation, authorisation and restriction of chemicals (reach). *Official Journal of the European Union L*, 142, 1–729.

European Commission. (2009) Regulation (ec) no 1107/2009 of the european parliament and of the council of 21 october 2009 concerning the placing of plant protection products on the market and repealing council directives 79/117/eec and 91/414/eec. *Official Journal of the European Union L*, 309/1, 1–50.

European Commission. (2013) Commission regulation (eu) no 284/2013 of 1 march 2013 setting out the data requirements for plant protection products, in accordance with regulation (ec) no 1107/2009 of the european parliament and of the council concerning the placing of plant protection products on the market. *Official Journal of the European Union L*, 93/85, 85–152.

Fujita, Y., Morita, T., Matsumura, S., Kawamoto, T., Ito, Y., Nishiyama, N. et al. (2016) Comprehensive retrospective evaluation of existing in vitro chromosomal aberration test data by cytotoxicity index transformation. *Mutation Research, Genetic Toxicology and Environmental Mutagenesis*, 802, 38–49.

Groff, K., Evans, S.J., Doak, S.H., Pfuhler, S., Corvi, R., Saunders, S. et al. (2021) In vitro and integrated in vivo strategies to reduce animal use in genotoxicity testing. *Mutagenesis*, 36, 389–400.

Hansen, K., Mika, S., Schroeter, T., Sutter, A., ter Laak, A., Steger-Hartmann, T. et al. (2009) Benchmark data set for in silico prediction of ames mutagenicity. *Journal of Chemical Information and Modeling*, 49, 2077–2081.

Harrell, F. (2015) *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis*. Cham: Springer.

Hendriks, G., Atallah, M., Morolli, B., Calleja, F., Ras-Verloop, N., Huijskens, I. et al. (2012) The toxtracker assay: novel gfp reporter systems that provide mechanistic insight into the genotoxic properties of chemicals. *Toxicological Sciences*, 125, 285–298.

Hendriks, G., Derr, R.S., Misovic, B., Morolli, B., Calleja, F.M. & Vrieling, H. (2016) The extended toxtracker assay discriminates between induction of DNA damage, oxidative stress, and protein misfolding. *Toxicological Sciences*, 150, 190–203.

ICH. (2008) Guideline s2 (r1) on genotoxicity testing and data interpretation for pharmaceuticals intended for human use.

ICH. (2023) Ich m7(r2) guideline on assessment and control of DNA reactive (mutagenic) impurities in pharmaceuticals to limit potential carcinogenic risk. EMA/CHMP/ICH/83812/2013.

Jaworska, J., Gabbert, S. & Aldenberg, T. (2010) Towards optimization of chemical testing under reach: a bayesian network approach to integrated testing strategies. *Regulatory Toxicology and Pharmacology*, 57, 157–167.

Kasamatsu, T., Kitazawa, A., Tajima, S., Kaneko, M., Sugiyama, K.I., Yamada, M. et al. (2021) Development of a new quantitative structure-activity relationship model for predicting ames mutagenicity of food flavor chemicals using stardrop auto-modeller. *Genes and Environment*, 43, 16.

Kim, B.S. & Margolin, B.H. (1994) Predicting carcinogenicity by using batteries of dependent short-term tests. *Environmental Health Perspectives*, 102(Suppl 1), 127–130.

Kirkland, D., Aardema, M., Henderson, L. & Muller, L. (2005) Evaluation of the ability of a battery of three in vitro genotoxicity tests to discriminate rodent carcinogens and non-carcinogens i. Sensitivity, specificity and relative predictivity. *Mutation Research*, 584, 1–256.

Kirkland, D., Kasper, P., Martus, H.J., Muller, L., van Benthem, J., Madia, F. et al. (2016) Updated recommended lists of genotoxic and non-genotoxic chemicals for assessment of the performance of new or improved genotoxicity tests. *Mutation Research, Genetic Toxicology and Environmental Mutagenesis*, 795, 7–30.

Kirkland, D., Reeve, L., Gatehouse, D. & Vanparys, P. (2011) A core in vitro genotoxicity battery comprising the ames test plus the in vitro micronucleus test is sufficient to detect rodent carcinogens and in vivo genotoxins. *Mutation Research*, 721, 27–73.

Kirkland, D., Zeiger, E., Madia, F. & Corvi, R. (2014) Can in vitro mammalian cell genotoxicity test results be used to complement positive results in the ames test and help predict carcinogenic or in vivo genotoxic activity? II. Construction and analysis of a consolidated database. *Mutation Research, Genetic Toxicology and Environmental Mutagenesis*, 775–776, 69–80.

Kirkland, D., Zeiger, E., Madia, F., Gooderham, N., Kasper, P., Lynch, A. et al. (2014) Can in vitro mammalian cell genotoxicity test results be used to complement positive results in the ames test and help predict carcinogenic or in vivo genotoxic activity? I. Reports of individual databases presented at an eurl ecvam workshop. *Mutation Research, Genetic Toxicology and Environmental Mutagenesis*, 775–776, 55–68.

Landrum, G. (2015) Rdkit: open-source cheminformatics.

Luijten, M., Ball, N.S., Dearfield, K.L., Gollapudi, B.B., Johnson, G.E., Madia, F. et al. (2020) Utility of a next generation framework for assessment of genomic damage: a case study using the industrial chemical benzene. *Environmental and Molecular Mutagenesis*, 61, 94–113.

MacKay, D.J.C. (2004) *Information theory, inference, and learning algorithms*. Cambridge: Cambridge University Press.

Madia, F., Kirkland, D., Morita, T., White, P., Asturiol, D. & Corvi, R. (2020a) Corrigendum to "EURL ECVAM genotoxicity and carcinogenicity database of substances eliciting negative results in the ames test: Construction of the database" [Mutat. Res. 854–855 June–July (2020) 503199]. *Mutation Research/Genetic Toxicology and Environmental Mutagenesis*, 858–860, 503274.

Madia, F., Kirkland, D., Morita, T., White, P., Asturiol, D. & Corvi, R. (2020b) Eurl ecvam genotoxicity and carcinogenicity database of substances eliciting negative results in the ames test: construction of the database. *Mutation Research, Genetic Toxicology and Environmental Mutagenesis*, 854–855, 503199.

Morita, T., Hamada, S., Masumura, K., Wakata, A., Maniwa, J., Takasawa, H. et al. (2016) Evaluation of the sensitivity and specificity of in vivo erythrocyte micronucleus and transgenic rodent gene mutation tests to detect rodent carcinogens. *Mutation Research, Genetic Toxicology and Environmental Mutagenesis*, 802, 1–29.

Nicolette, J., Luijten, M., Sasaki, J.C., Custer, L., Embry, M., Froetschl, R. et al. (2021) Utility of a next-generation framework for assessment of genomic damage: a case study using the pharmaceutical drug candidate etoposide. *Environmental and Molecular Mutagenesis*, 62, 512–525.

OECD. (2016a) Test no. 473: in vitro mammalian chromosomal aberration test.

OECD. (2016b) Test no. 474: mammalian erythrocyte micronucleus test.

OECD. (2016c) Test no. 475: mammalian bone marrow chromosomal aberration test.

OECD. (2016d) Test no. 476: In vitro mammalian cell gene mutation tests using the hprt and xprt genes.

OECD. (2016e) Test no. 489: In vivo mammalian alkaline comet assay.

OECD. (2016f) Test no. 490: In vitro mammalian cell gene mutation tests using the thymidine kinase gene.

OECD. (2020) Test no. 471: bacterial reverse mutation test.

OECD. (2022a) Test no. 470: mammalian erythrocyte pig-a gene mutation assay.

OECD. (2022b) Test no. 488: transgenic rodent somatic and germ cell gene mutation assays.

OECD. (2023) Test no. 487: in vitro mammalian cell micronucleus test.

Plotly Technologies Inc. (2015) *Collaborative data science*. Montréal, QC: Plotly Technologies Inc.

Rosenkranz, H.S., Mitchell, C.S. & Klopman, G. (1985) Artificial intelligence and bayesian decision theory in the prediction of chemical carcinogens. *Mutation Research*, 150, 1–11.

Schisler, M.R., Gollapudi, B.B. & Moore, M.M. (2018) Evaluation of u. S. National toxicology program (ntp) mouse lymphoma assay data using international workshop on genotoxicity tests (iwgt) and the organization for economic co-operation and development (oecd) criteria. *Environmental and Molecular Mutagenesis*, 59, 829–841.

Scientific Committee on Consumer Safety (SCCS). (2023) SCCS notes of guidance for the testing of cosmetic ingredients and their safety evaluation 12th revision. SCCS/1647/22.

U.S. Food and Drug Administration. (2007) Chapter iv.C.1. Short-term tests for genetic toxicity. In: *Redbook 2000: Toxicological principles for the safety assessment of food ingredients*. Rockville: U.S. Food and Drug Administration.

White, P.A., Long, A.S. & Johnson, G.E. (2020) Quantitative interpretation of genetic toxicity dose-response data for risk assessment and regulatory decision-making: current status and emerging priorities. *Environmental and Molecular Mutagenesis*, 61, 66–83.

Wickham, H. (2016) *Ggplot2: elegant graphics for data analysis*. New York: Springer-Verlag.

Yamada, M. & Honma, M. (2018) Summarized data of genotoxicity tests for designated food additives in Japan. *Genes and Environment*, 40, 27.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Luijten, M., van Benthem, J., Morita, T., Corvi, R., Escobar, P.A., Fujita, Y. et al. (2024) Evaluation of the standard battery of in vitro genotoxicity tests to predict in vivo genotoxicity through mathematical modeling: A report from the 8th International Workshop on Genotoxicity Testing. *Environmental and Molecular Mutagenesis*, 1–14. Available from: https://doi.org/10.1002/em.22640