

Evaluation of QSAR models for predicting mutagenicity: outcome of the Second Ames/QSAR international challenge project

A. Furuham, A. Kitazawa, J. Yao, C.E. Matos dos Santos, J. Rathman, C. Yang, J.V. Ribeiro, K. Cross, G. Myatt, G. Raitano, E. Benfenati, N. Jeliaskova, R. Saiakhov, S. Chakravarti, R.S. Foster, C. Bossa, C. Laura Battistelli, R. Benigni, T. Sawada, H. Wasada, T. Hashimoto, M. Wu, R. Barzilay, P.R. Daga, R.D. Clark, J. Mestres, A. Montero, E. Gregori-Puigjané, P. Petkov, H. Ivanova, O. Mekenyan, S. Matthews, D. Guan, J. Spicer, R. Lui, Y. Uesawa, K. Kurosaki, Y. Matsuzaka, S. Sasaki, M.T.D. Cronin, S.J. Belfield, J.W. Firman, N. Spînu, M. Qiu, J.M. Keca, G. Gini, T. Li, W. Tong, H. Hong, Z. Liu, Y. Igarashi, H. Yamada, K.-I. Sugiyama & M. Honma

To cite this article: A. Furuham, A. Kitazawa, J. Yao, C.E. Matos dos Santos, J. Rathman, C. Yang, J.V. Ribeiro, K. Cross, G. Myatt, G. Raitano, E. Benfenati, N. Jeliaskova, R. Saiakhov, S. Chakravarti, R.S. Foster, C. Bossa, C. Laura Battistelli, R. Benigni, T. Sawada, H. Wasada, T. Hashimoto, M. Wu, R. Barzilay, P.R. Daga, R.D. Clark, J. Mestres, A. Montero, E. Gregori-Puigjané, P. Petkov, H. Ivanova, O. Mekenyan, S. Matthews, D. Guan, J. Spicer, R. Lui, Y. Uesawa, K. Kurosaki, Y. Matsuzaka, S. Sasaki, M.T.D. Cronin, S.J. Belfield, J.W. Firman, N. Spînu, M. Qiu, J.M. Keca, G. Gini, T. Li, W. Tong, H. Hong, Z. Liu, Y. Igarashi, H. Yamada, K.-I. Sugiyama & M. Honma (2023) Evaluation of QSAR models for predicting mutagenicity: outcome of the Second Ames/QSAR international challenge project, *SAR and QSAR in Environmental Research*, 34:12, 983-1001, DOI: [10.1080/1062936X.2023.2284902](https://doi.org/10.1080/1062936X.2023.2284902)

To link to this article: <https://doi.org/10.1080/1062936X.2023.2284902>



© 2023 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.



[View supplementary material](#)



Published online: 04 Dec 2023.



[Submit your article to this journal](#)



[View related articles](#)



View Crossmark data [↗](#)

Evaluation of QSAR models for predicting mutagenicity: outcome of the Second Ames/QSAR international challenge project

A. Furuham^a, A. Kitazawa^a, J. Yao^b, C.E. Matos dos Santos^c, J. Rathman^d, C. Yang^d, J.V. Ribeiro^d, K. Cross^e, G. Myatt^e, G. Raitano^f, E. Benfenati^f, N. Jeliakova^g, R. Saiakhov^h, S. Chakravarti^h, R.S. Fosterⁱ, C. Bossaⁱ, C. Laura Battistelli^j, R. Benigni^{j,k}, T. Sawada^{l,m}, H. Wasada^l, T. Hashimoto^l, M. Wuⁿ, R. Barzilayⁿ, P.R. Daga^o, R.D. Clark^o, J. Mestres^p, A. Montero^p, E. Gregori-Puigjané^p, P. Petkov^q, H. Ivanova^q, O. Mekenyani^q, S. Matthews^r, D. Guan^r, J. Spicer^r, R. Lui^r, Y. Uesawa^s, K. Kurosaki^s, Y. Matsuzaka^s, S. Sasaki^s, M.T.D. Cronin^t, S. J. Belfield^t, J.W. Firman^t, N. Spînu^t, M. Qiu^u, J.M. Keca^u, G. Gini^v, T. Li^w, W. Tong^w, H. Hong^w, Z. Liu^{w,x}, Y. Igarashi^y, H. Yamada^y, K.-I. Sugiyama^a and M. Honma^a

^aDivision of Genetics and Mutagenesis (DGM), National Institute of Health Sciences (NIHS), Kawasaki, Japan; ^bKey Laboratory of Fluorine and Nitrogen Chemistry and Advanced Materials (Chinese Academy of Sciences), Shanghai Institute of Organic Chemistry, Chinese Academy of Sciences (SIOC, CAS), Shanghai, China; ^cDepartment of Computational Toxicology and In Silico Innovations, Alttox Ltd, São Paulo-SP, Brazil; ^dMN-AM, Nuremberg, Germany/Columbus, OH, USA; ^eIn Silico Department, Instem, Conshohocken, PA, USA; ^fLaboratory of Environmental Toxicology and Chemistry, Department of Environmental Health Sciences, Istituto di Ricerche Farmacologiche Mario Negri IRCCS (IRFMN), Milano, Italy; ^gIdeaConsult Ltd, Sofia, Bulgaria; ^hMultiCASE Inc, Mayfield Height, OH, USA; ⁱLhasa Ltd, Leeds, UK; ^jEnvironment and Health Department, Istituto Superiore di Sanità (ISS), Rome, Italy; ^kAlpha-PreTox, Rome, Italy; ^lFaculty of Regional Studies, Gifu University, Gifu, Japan; ^mxenoBiotic Inc, Gifu, Japan; ⁿMassachusetts Institute of Technology, Cambridge, MA, USA; ^oSimulations Plus, Lancaster, CA, USA; ^pChemotargets, Barcelona, Spain; ^qLMC - Bourgas University, Bourgas, Bulgaria; ^rComputational Pharmacology & Toxicology Laboratory, Discipline of Pharmacology, School of Pharmacy, Faculty of Medicine and Health, The University of Sydney, Sydney, Australia; ^sDepartment of Medical Molecular Informatics, Meiji Pharmaceutical University, Tokyo, Japan; ^tSchool of Pharmacy and Biomolecular Sciences, Liverpool John Moores University, Liverpool, UK; ^uEvergreen AI, Inc, Toronto, Canada; ^vDepartment of Electronics, Information and Bioengineering (DEIB), Politecnico di Milano, Milano, Italy; ^wDivision of Bioinformatics and Biostatistics, National Center for Toxicological Research, U.S. Food and Drug Administration (NCTR/FDA), Jefferson, AR, USA; ^xIntegrative Toxicology, Nonclinical Drug Safety, Boehringer Ingelheim Pharmaceuticals, Inc, Ridgefield, CT, USA; ^yArtificial Intelligence Center for Health and Biomedical Research, National Institutes of Biomedical Innovation, Health and Nutrition (NIBIOHN), Osaka, Japan

ABSTRACT

Quantitative structure–activity relationship (QSAR) models are powerful *in silico* tools for predicting the mutagenicity of unstable compounds, impurities and metabolites that are difficult to examine using the Ames test. Ideally, Ames/QSAR models for regulatory use should demonstrate high sensitivity, low false-negative rate and wide coverage of chemical space. To promote superior model development, the Division of Genetics and Mutagenesis, National Institute of Health Sciences, Japan (DGM/NIHS), conducted the Second Ames/QSAR


ARTICLE HISTORY

Received 11 September 2023
Accepted 13 November 2023

KEYWORDS

Ames mutagenicity prediction; ANEI-HOU new chemical; sensitivity; Ames/QSAR International Challenge Projects;

CONTACT A. Furuham  ayako_furuham@nihs.go.jp

 Supplemental data for this article can be accessed at: <https://doi.org/10.1080/1062936X.2023.2284902>.

© 2023 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (<http://creativecommons.org/licenses/by-nc-nd/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

International Challenge Project (2020–2022) as a successor to the First Project (2014–2017), with 21 teams from 11 countries participating. The DGM/NIHS provided a curated training dataset of approximately 12,000 chemicals and a trial dataset of approximately 1,600 chemicals, and each participating team predicted the Ames mutagenicity of each trial chemical using various Ames/QSAR models. The DGM/NIHS then provided the Ames test results for trial chemicals to assist in model improvement. Although overall model performance on the Second Project was not superior to that on the First, models from the eight teams participating in both projects achieved higher sensitivity than models from teams participating in only the Second Project. Thus, these evaluations have facilitated the development of QSAR models.

imbalanced toxicity data;
model performance

Introduction

Regulatory bodies are interested in using *in silico* methods to address animal welfare issues, reduce costs, and obtain information regarding chemicals that are difficult to purify for *in vivo* and *in vitro* tests. *In silico* models, such as quantitative structure–activity relationships (QSARs), can be used to predict the biological activities of chemicals from their structure. A type of QSARs is Ames/QSAR, which is based on the Ames test data and can predict the mutagenicity of a chemical. Ames/QSAR models are currently used to predict the mutagenicity of impurities in pharmaceuticals [1] and other chemicals, such as pesticides and their metabolites [2].

Many of the currently available Ames/QSAR models were developed using publicly available Ames datasets, such as that of Hansen et al. [3], which includes data on more than 5,000 chemicals, the EURL ECVAM Genotoxicity and Carcinogenicity Consolidated Database of Ames-Positive Chemicals [4,5] and Ames-Negative Chemicals [6], European Food Safety Authority Genotoxicity Database [7]; other various other genotoxicity datasets [8–10]. Although these Ames/QSAR models have high accuracy for predicting the mutagenicity of existing chemicals (in the public domain), they have lower accuracy [11], and particularly sensitivity, for detecting new Ames-positive compounds. The use of imbalanced datasets with unequal numbers of Ames-positive and -negative chemicals [12] has contributed to this lack of model performance; balanced accuracy ((sensitivity + specificity)/2) may be a better measure of performance against unbalanced proprietary datasets [13]. However, the primary factor limiting model performance is insufficient coverage of the chemical space and mutagenic mechanisms in the training sets or mutagenicity expert rules. Thus, improved Ames/QSAR models with high sensitivity, low false-negative rate and wide coverage of chemical space are needed in the regulatory setting.

To improve the predictivity of Ames/QSAR models, especially model sensitivity, the Division of Genetics and Mutagenesis, National Institute of Health Sciences, Japan (DGM/NIHS), conducted the First Ames/QSAR International Challenge Project (hereafter ‘the First Project’) from 2014 to 2017 [14]. The project involved 12 teams, mainly QSAR model vendors, from seven countries who were asked to use their Ames/QSAR models to predict the mutagenicity of approximately 12,000 new chemicals, after which the DGM/NIHS compared predicted results with the results of the actual Ames test data to derive various model performance metrics. While the ability of these QSAR models to predict the mutagenicity of

new chemicals was improved by this project [14], the extent was limited because the participants were not provided adequate Ames test information by the DGM/NIHS.

From 2020 to 2022, the DGM/NIHS conducted the Second Ames/QSAR International Challenge Project (hereafter ‘the Second Project’), again with the aim of evaluating and improving the performance of Ames/QSAR models. This time several changes were made. More academic and other non-commercial entities took part, and deep-learning models, as well as conventional QSAR models, were examined. The training dataset included the approximately 12,000 chemicals used in the First Project, and a trial dataset comprised of 1,589 new chemicals. The training dataset were curated and was also provided in addition to mutagenicity results for multiple test strains without and with metabolic activation, the solvents used and the test chemical purity. By making these changes, we expect the Second Project to facilitate further improvements in the predictive ability of currently available Ames/QSAR models.

Here we summarize the results from the Second Project and describe the outstanding issues to be solved for the successful use of Ames/QSAR models in the regulatory setting.

Materials and methods

Overview of the First and Second Project

The essential characteristics of the First and Second Projects are summarized in Table 1. The First Project was conducted from 2014 to 2017 and involved 12 teams from seven countries. The study comprised three phases, an initial trial phase (Phase I) and two training and trial phases (Phase II and Phase III). A dataset of approximately 12,000 new chemicals was used, of which 4,000 were used in each phase. Most Ames/QSAR models were categorized as statistical or rule-based. The Second Project was conducted from 2020 to 2022, with 19 teams participating in 2020 and two additional teams joining in 2021. The teams were from academia and non-commercial institutions in addition to QSAR vendors. Nine teams that participated in the First Project also participated in the Second Project. All teams participating in the Second Project are listed in Table 2. The Second Project involved one phase using both a training dataset comprised of the approximately 12,000 chemicals from the First Project and a trial dataset of 1,589 new chemicals.

Table 1. Overview of the First and Second Ames/QSAR International Challenge Projects.

	<i>First Project</i>	<i>Second Project</i>
Aim	QSAR tool improvements	
Date	2014–2017	2020–2022
Participants	12 teams (7 countries) Mainly QSAR vendors	21 teams (11 countries) QSAR vendors/academia/non-commercial entities
Training dataset	Phase I: none Phase II: 3,902 chemicals Phase III: 3,902 + 3,802 chemicals	Chemicals used the First Project
Trial dataset	Phase I: 3,902 chemicals Phase II: 3,802 chemicals Phase III: 4,409 chemicals	1,589 new chemicals
Models analysed	Statistical and rule-based models	Statistical, rule-based and machine-learning models

Table 2. Teams participating in the Second Project.

<i>Team no.</i>	<i>Team name</i>	<i>Country</i>	<i>Note*</i>
1	Shanghai Institute of Organic Chemistry, Chinese Academy of Sciences (SIOC, CAS)	China	
2	Alttox Ltd.	Brazil	
3	MN-AM	Germany/ USA	(i)
4	Instem	USA	(i)
5	Istituto di Ricerche Farmacologiche Mario Negri IRCCS (IRFMN)	Italy	(i)
6	IdeaConsult Ltd.	Bulgaria	(i)
7	MultiCASE Inc.	USA	(i)
8	Lhasa Ltd.	UK	(i)
9	Istituto Superiore di Sanità (ISS)	Italy	(i)
10	Gifu University	Japan	
11	Massachusetts Institute of Technology	USA	
12	Simulations Plus, Inc.	USA	(i)
13	Chemotargets	Spain	
14	LMC – Bourgas University	Bulgaria	(i)
15	The University of Sydney	Australia	
16	Meiji Pharmaceutical University	Japan	
17	Liverpool John Moores University	UK	
18	Evergreen AI, Inc.	Canada	
19	Politecnico di Milano	Italy	
20	National Center for Toxicological Research U.S. Food and Drug Administration (NCTR/FDA)	USA	
21	National Institutes of Biomedical Innovation, Health and Nutrition (NIBIOHN)	Japan	

*(i): team also participated in the First Project.

Training and trial datasets

Data source

Since 1979, the Ministry of Health, Labour and Welfare (MHLW) in Japan has stipulated under the Industrial Safety and Health Act (ANEI-HOU) that producers of new chemical substances and importers of chemicals in amounts greater than 100 kg per year must conduct hazard investigations prior to manufacture or import as part of these investigations, and therefore Ames tests must be conducted and reported [15]. The Ames data used in both the First Project and the Second Project were obtained from the MHLW and comprised Ames class (A, B or C), chemical name and molecular structure. In the First Project, no other information was provided, such as bacterial strain, solvent or cytotoxicity. Class A or ‘strong positive’ indicates that the chemical induces more than 1,000 revertant colonies per milligram of at least one Ames test strain in the presence or absence of metabolic activation. Class B indicates that the tested chemical induces at least a 2-fold increase in revertant colonies but fewer than induced by class A compounds compared to the negative control in at least one Ames strain with or without metabolic activation. Finally, class C or ‘negative’ indicates a < 2-fold increase in revertant colonies (non-mutagenic). In Japan, Ames test data are confidential for chemicals in class B or C, while a list of class A chemicals is publicly available [16] All participants agreed in writing to uphold the confidentiality of the results. Additional details are provided in the report from the First Project [14].

Training dataset originated from the First Project

The training dataset used in the Second Project was created by combining the three subsets used in the three phases of the First Project. This combined dataset contained

Table 3. Number of chemicals in the Ames classes in the trial set for the Second Project.

<i>Class A</i>	<i>Class B</i>	<i>Class C</i>	<i>Total</i>
80 (5.0)	156 (9.8)	1,353 (85.1)	1,589 (100)

Expressed as numbers (%).

chemical structure information on approximately 12,000 new chemicals as SDF files with a corresponding list of SMILES notations [17]. The chemical structures in the training dataset included salts (e.g. [Na+]). The list was curated and additional information related to the Ames test was introduced to improve QSAR models (See section ‘Data curation’). If the mutagenicity (class A, B or C) of duplicated chemicals in the training set was the same, only the older Ames data was listed as a part of the training set. If the mutagenicity (class A, B or C) of duplicated chemicals in the training set was different, we did not use the results of such a chemical as part of the training set. Stereoisomers were treated as different chemicals in the training set as, on some occasions, the mutagenicity is not always same in the stereoisomers. Similarly, if the mutagenicity (class A, B or C) of matching pairs of parent compounds and their salt in the training set was the same, only the older Ames test data was listed as a part of the training set. If the mutagenicity (class A, B or C) of such matching pairs in the training set was different, we did not use these chemicals as part of the training set.

Trial dataset generated from new chemicals

By the onset of the Second Project, 5,303 additional compounds were registered as ANEI-HOU new chemicals by the MHLW as a result of Ames class and chemical structure information being available. This list was curated to exclude chemicals unsuitable for evaluating Ames mutagenicity by chemical structure-based QSAR predictions. After structural curation, chemicals with undefined SMILES notations (e.g. oils, extracts, polymers), duplicate chemicals, metal ions and mixtures with undefined components were also removed. Finally, 1,589 ANEI-HOU new chemicals were included as the trial dataset for the Second Project. In addition, minor components (e.g. salts, counterions and solvent molecules) were removed from the 1,589 chemical structures.

Table 3 shows the proportions of trial dataset chemicals in each Ames class. Like the training dataset, the majority of chemical (about 85%) were class C (non-mutagenic). The training dataset including Ames class was sent to each participating team for the development of their QSAR models. Simultaneously the trial dataset was sent to each participating teams without the Ames class results. The results of all predictions (for trial chemicals) were then reported to the DGM/NIHS. The DGM/NIHS calculated the performance metrics of each QSAR model and disclosed the actual Ames test data of the trial chemicals to the participating teams.

Ames/QSAR performance evaluations

As in the First Project [14], the prediction from the QSAR models were compared to the actual Ames test data, generating the prediction performance metrics defined in Table 4. Hereafter, the predicted results are classified as ‘true positive’ or TP when the measured Ames result is either class A or B (positive) and the

Table 4. Performance metrics used to evaluate Ames/QSAR model performance.

Performance metric	Calculation and description
A-Sensitivity (A-Sens.)	$TPA/(TPA + FN)$ Measures the ability to correctly predict strongly Ames-positive compounds (class A).
Sensitivity (Sens.)	$TP/(TP + FN)$ Measures the ability to correctly predict Ames-positive compounds (class A or B).
Specificity (Spec.)	$TN/(FP + TN)$ Measures the ability to predict Ames-negative compounds (class C).
Accuracy (Acc.)	$(TP + TN)/(TP + TN + FP + FN)$ Assesses overall prediction performance by returning the fraction of compounds that were correctly classified.
Balanced Accuracy (BA)	$(Sens. + Spec.)/2$ Assesses overall model performance while giving each class equal weight.
Mathews Correlation Coefficient (MCC)	$[(TP * TN) - (FP * FN)] / [(TP + FP) * (TP + FN) * (TN + FP) * (TN + FN)]^{1/2}$ Assesses overall model performance. Values can range from -1 to 1, unlike the other metrics in this table that range from 0 to 1.
Coverage (Cov.)	$(TP + TN + FP + FN)/Total$ Assesses the proportion of compounds for which the model can make positive or negative prediction.
Positive Prediction Value (PPV)	$TP/(TP + FP)$ Indicates how frequently positive predictions are correct.
Negative Prediction Value (NPV)	$TN/(TN + FN)$ Indicates how frequently negative predictions are correct.
F1-Score	$2 * (Recall * Precision) / (Recall + Precision)$ Indicates the harmonic mean of Recall and Precision, where Recall = Sensitivity and Precision = PPV.

TPA: True positive for class A (prediction is positive and Ames test result is class A); TP: True positive (prediction is positive and Ames test result is class A or B); TN: True negative (prediction is negative and Ames test result is class C). FN: False-negative (prediction is negative and Ames test result is class A or B); FP: False positive (prediction is positive and Ames test results is class C).

model prediction is positive, 'true negative' or TN when the measured Ames result is class C (negative) and the model prediction is negative, 'false positive' or FP when the measured Ames result is class C (negative) and the model prediction is positive and 'false-negative' (FN) when the measured Ames result is either class A or B (positive) and the model prediction is negative. For the evaluation, Ames classes A and B were combined into a single 'positive' class to calculate all performance metrics except for sensitivity, which was divided into A-sensitivity, the ability to detect strong Ames-positive (class A) compounds, and sensitivity, the ability to detect Ames-positive (class A or B) compounds.

Results and discussion

Data curation

In the First Project, we only provided the Ames test results (class A, B or C) for the chemicals. In the Second Project, teams were also provided with the molecular weight, purity and solvent used as well as the Ames test results for five bacterial strains with and without metabolic activation (Table 5) to aid in the development of QSAR models. As in the First Project, cytotoxicity and dose – response data were not provided.

Table 6 listed the curation of the training dataset (see also [18] for an overview). First, the DGM/NIHS confirmed all class A and B chemicals, including those with FN predictions as well as class C chemicals with FP predictions by most teams in the First Project. Of the

Table 5. Sample of training data provided to the participants of the Second Project*.

Class	Chemical name	Chemical structure	Mol. weight	Purity (%)	Solvent	-S9					+S9				
						TA100	TA1535	WP2uvrA	TA98	TA1537	TA100	TA1535	WP2uvrA	TA98	TA1537
C			163.2												
C			216.4	98.3	THF	C	C	C	C	C	C	C	C	C	C
B			420.1	99.4	Acetone	C	C	C	C	C	B	B	C	C	C
C			515.7												
C			2366.2												
B			148.0	99.7	DMSO	C	C	C	C	C	B	C	B	C	C
C			264.1												
A			203.0	99.8	DMSO	C	C	C	C	C	A	C	C	B	C

*The chemical names and structures were provided to the participants under a confidentiality agreement. Mol. weight, purity and solvent indicate the molecular weight of the training chemical, the chemical purity of the sample used in the Ames test and the solvent used for the Ames test, respectively. -S9 and +S9 indicate that the five strains were tested without (-) and with (+) metabolic activation. Ames class A is strongly positive, class B is positive and class C is negative for mutagenicity.

Table 6. Curation of the training dataset used in the Second Project.

	Class A	Class B	Class C	Total
(I) Chemicals used in the First Project	672	1,085	10,383	12,140
(II) Ames study report missing	44	164	1,238	1,446
(III) Reports for curation (= I – II)	628	921	9,145	10,694
(IV) Curation completed on March 2020	628	921	16	1,565
Curation not completed (= III – IV) ^{*1}	0	0	9,129	9,129
Number of modified reports	26	18	16	
Change after modification ^{*2}	–23	15		–6
Provided training dataset (March 2020)	649	1,100	10,385	12,134

^{*1}Curation was performed by the DGM/NIHS members. The participating teams were not provided with detailed information on the noncurated chemicals.

^{*2}Six chemicals were not assigned to class A, B, or C.

10,694 test reports available to the DGM/NIHS for the First Project, 1,565 (about 15%) were included for the Second Project. Of these, 60 changed (these were chemical reassigned to other classes) according to expert reviews by the DGM/NIHS. Thus, the final training dataset for the Second Project comprised 12,134 chemicals, including 649 class A chemicals, 1,100 class B chemicals and 10,385 class C chemicals. It is important to note that this curation did not change the class determination under ANEI-HOU, and the decision to use the curated data for model development was left to each participating team. The DGM/NIHS also provided the teams with the list of class changes and the underlying reasons.

Model diversity

The 21 participating teams submitted from one to five sets of predicted results for each of the 1,589 chemicals in the trial dataset (positive, negative, equivocal, out-of-domain or no call, etc). Basically, DGM/NIHS did not make any restriction to the models which the 21 teams developed and used to predict the mutagenicity of the 1,598 chemicals. Only a positive versus negative classification was requested (A/B versus C rather than A versus B versus C) because teams were able to use other training data containing only positive versus negative information as well as training data provided by the DGM/NIHS. In addition to the predicted results, the teams were asked to enter details of their models on a model information sheet with the following fields: Date, QSAR Builder name, QSAR model name, QSAR model version, training data (required), training data description, explicit model algorithm and Notes (optional). The response in the 'Training data' field was selected from (1) release model (training data provided in the Second Project were not used for model development), (2) release model + all NIHS data (all training data were used for model development) and (3) release model + selected NIHS data (only some of the NIHS training data were used for model development). Additionally, in Appendix I (Supplementary material), each model is fully described, in addition providing more details of the training data as well as model algorithms. Furthermore, each team was asked to select the model with the best performance metrics. Hereafter, this model is referred to as that selected before access to the Ames test results. Once the teams were informed of the Ames test results, they were again asked to select the most predictive model. Hereafter, this model is referred to as that selected after access to the Ames test results.

The 21 teams submitted predicted Ames results using a total of 50 Ames/QSAR models (the summarized model information sheets are available in Appendix I). One of the 21 teams (Laboratory of Mathematical Chemistry, Bourgas University) submitted positive versus negative and in-domain versus out-of-domain data separately, thus generating two sets of performance metrics (one with in-domain versus out-of-domain data, and one including only in-domain data). Therefore, 51 sets of predicted results were considered in the analyses. The performance metrics of these 51 datasets are presented in Appendix II (Supplementary material).

The models were categorized as expert rule-based or statistical. Some teams provided the predicted results for statistical models and/or rule-based models that are currently used under the ICH M7 guideline. Machine-learning (ML) models, including deep-learning artificial neural networks (ANNs), were categorized as statistical models and considered they are not based in conventional statistical methods but considered a family of statistical learning algorithms that emulates the learning pattern in the human brain with trained neurons by statistical algorithms [19].

ANNs and other ML algorithms (K-Nearest Neighbours, Genetic algorithms-Gas etc.) are cited in Chapter 3 - Unambiguous Algorithms in the guideline 'OECD principles for the Validation for Regulatory Purposes of (Q)SAR Models' [20]. Although a neural network is one example of a larger class of ML algorithms [21], 'deep-learning' was a keyword frequently used in the model information sheets completed by the participating teams. This subject will not be discussed further here, as the intention is only to provide an overview of the Second Project. In addition, some teams developed more than one model using the same algorithm but different training data or introduced additional techniques for managing unbalanced genotoxicity data (see model information sheets).

The models selected by each team before and after access to the Ames test results (Tables 7 and 8) were evaluated in this article because the 51 models (as shown in Appendices I and II) were too diverse for analysis of overall performance. As explained in the footnotes of Table 7 and Appendix II, three teams (Alttox Ltd., Simulations Plus Inc. and NCTR/FDA) did not select a single model before access to the Ames test results. Rather, Alttox Ltd. selected models only after access to the results, while Simulations Plus Inc. selected three models (S+MUT_NIHS_ABC, S+MUT_NIHS_AC and S+MUT_NIHS) and used two (S+MUT_NIHS_ABC and S+MUT_NIHS_AC) to generate a single set of performance metrics by averaging. Finally, NCTR/FDA selected two models developed by two independent groups. Thus, only 21 ($= 21 - 1 + 1$) models selected before accessing the Ames test results were evaluated. The names of the 21 models and corresponding performance metrics are summarized in Tables 7 and 9. As explained in the footnotes of Table 8 and Appendix II, NCTR/FDA selected two models after access to the Ames test results, while all other teams selected only one, so 22 models ($= 21 + 1$) were evaluated after Ames tests results were made available. The names of the 22 models and corresponding performance metrics are summarized in Tables 8 and 10. The model evaluations included comparisons of those selected before or after access to the Ames test results as well as comparisons with models reported in the First Project (listed in Table 11).

Table 7. Models selected by each team BEFORE access to Ames test results for the trial chemicals.

<i>Model no.^{*1}</i>	<i>Team name</i>	<i>Before being informed of the results</i>
1	SIOC, CAS	CISOC-PSMT (SIOC, CAS, China)
3	MN-AM	ChemTunes. ToxGPS Ames NIHS_v2
4	Instem	Leadscope 2nd QSAR Challenge Consensus Model
5–1	IRFMN	Mutagenicity (Ames test) CONSENSUS model (18k) version 0.9.1
6	IdeaConsult Ltd.	AMBIT DeepN v4.85
7	MultiCASE Inc.	PHARM_BMUT model version (1.8.0.0.17691.350)
8	Lhasa Ltd.	Sarah Nexus v.3.0.1 with 2068 NIHS chemicals
9 ^{*2}	ISS	in vitro Mutagenicity (Ames test) by ISS- modified2020
10	Gifu University	xenoBiotic 0.9q
11	Massachusetts Institute of Technology	Chemprop
12–1 ^{*3}	Simulations Plus Inc.	Average of S+MUT_NIHS_ABC model and S+MUT_NIHS_AC model
13	Chemotargets	CHMT_GBoostSC
14	LMC – Bourgas University	TIMES_AMES 17.17.3 (in domain TIMES model)
15–1	The University of Sydney	DRSpicySTiM-Ensemble
16–1	Meiji Pharmaceutical University	MMI-STK2
17–1	Liverpool John Moores University	DL
18	Evergreen AI, Inc.	Avalon
19	Politecnico di Milano	GCN
20-a	NCTR/FDA (one of two best models)	DeepAmes
20-b	NCTR/FDA (one of two best models)	Decision Forest
21	NIBIOHN	GNN(kMoL)_bestbalanced

^{*1}The model number encodes the team number shown in Table 2 and whether the model was selected before (1) or after (2) access to the Ames test results. Letters indicate that more than one model was selected by a team (see NCTR/FDA). Team no. 2 selected a model only after access to the Ames test results.

^{*2}in vitro Mutagenicity (Ames test) by ISS- modified2020 is not publicly available.

^{*3}Performance metrics were calculated as the average of values generated by two models (S+MUT_NIHS_ABC and S+MUT_NIHS_AC) among the three submitted models (S+MUT_NIHS_ABC, S+MUT_NIHS_AC and S+MUT_NIHS).

Performance of the selected models

In addition to the performance metrics derived in the First Project, the harmonic mean of recall (sensitivity) and precision (positive prediction value, PPV), or F1-score, was introduced in the Second Project (see Table 4). For a high F1-score, both FN and FP must be low [12]. The ratio of class A and B positives to class C negatives was around 15:85 for both the training dataset (1749:10385) and trial dataset (238:1353), comparable to the First Project (1757:10383) [14]. Due to this imbalance in Ames test results, however, accuracy alone ($(TP + TN)/\text{all classifications}$) cannot be used as a measure of model performance; therefore, additional metrics are needed. Compared to TN and FP, the counts and ratios of TP and FN are always smaller in a dataset containing predominantly negative results. In such cases, changing TP has little effect on accuracy. Balanced accuracy also remains stable when the number of FN results is much higher than the number of TP results [12], although balanced accuracy does not depend on the balance of the dataset. Thus, the introduction of the F1-score allowed us to focus on TP when evaluating the models. The F1-score can also be calculated using sensitivity and PPV, so F1-scores were also calculated for the First Project using the values from Tables 5–7 in Honma et al. [14] and compared to those obtained for the Second Project.

Together with the performance metrics in Tables 9 and 10, receiver operating characteristic (ROC) graphs (Figure 1) indicated that sensitivity and specificity were correlated, except for the values obtained using Model no. 17–1 (Table 9), for which sensitivity was high when specificity was low and *vice versa*. In addition, the A-sensitivity was higher than

Table 8. Models selected by each team AFTER access to Ames test results for the trial chemicals.

<i>Model no.^{*1}</i>	<i>Team name</i>	<i>After being informed of the results</i>
1	SIOC, CAS	CISOC-PSMT (SIOC, CAS, China)
2–2 ^{*2}	Altox Ltd.	GeneTox-iS – Prototype - Decision tree core v.1.0
3	MN-AM	ChemTunes. ToxGPS Ames NIHS_v2
4	Instem	Leadscope 2nd QSAR Challenge Consensus Model
5–2	IRFMN	NCSTOXVega-0.18 version 0.18
6	IdeaConsult Ltd.	AMBIT DeepN v4.85
7	MultiCASE Inc.	PHARM_BMUT model version (1.8.0.0.17691.350)
8	Lhasa Ltd.	Sarah Nexus v.3.0.1 with 2068 NIHS chemicals
9 ^{*3}	ISS	in vitro Mutagenicity (Ames test) by ISS- modified2020
10	Gifu University	xenoBiotic 0.9q
11	Massachusetts Institute of Technology	Chemprop
12–2	Simulations Plus Inc.	S+MUT_NIHS
13	Chemotargets	CHMT_GBoostSC
14	LMC – Bourgas University	TIMES_AMES 17.17.3 (in domain TIMES model)
15–2	The University of Sydney	GreedyMBAK
16–2	Meiji Pharmaceutical University	MMI-VOTE1
17–2	Liverpool John Moores University	RF model2
18	Evergreen AI, Inc.	Avalon
19	Politecnico di Milano	GCN
20-a	NCTR/FDA (one of two best models)	DeepAmes
20-b	NCTR/FDA (one of two best models)	Decision Forest
21	NIBIOHN	GNN(kMol)_bestbalanced

^{*1}Model number encodes team number shown in Table 2 and whether the model was selected before (1) or after (2) access to Ames test results. Letters indicate that more than one model was selected by a team (see NCTR/FDA).

^{*2}A model was selected only after access to the Ames test results.

^{*3}in vitro Mutagenicity (Ames test) by ISS- modified2020 is not publicly available.

the general sensitivity for all models except Model no.17–1. Greater A-sensitivity is expected because strong positive mutagenicity (class A) can be predicted with greater sensitivity and lower error than positive mutagenicity (class A + B). However, A-sensitivity and sensitivity were still highly correlated for the selected models. The F1-score was also strongly correlated with Matthew's correlation coefficient (MCC), which assesses overall model performance. Thus, we concluded that the F1-score could be used as an alternative to MCC for assessing the performance of Ames/QSAR models. In addition, specificity and accuracy were strongly correlated, which is expected given the high proportion of FN results generated from the negative data-dominant trial dataset.

Performance evaluations of the models selected before access to the Ames test results (Table 9) were used for external validation. The three well-adjusted models were Model no. 7 (F1-score of 49.7%, A-sensitivity of 71.4%, sensitivity of 50.8% and specificity of 91.5%), 14 (51.1%, 95.7%, 58.0% and 88.6%, respectively) and 16–1 (51.6%, 71.3%, 53.4% and 90.7%, respectively). In this article, the well-adjusted models are defined as the F1-scores, A-sensitivities, sensitivities and specificities of these three models were substantially higher than the averages of all 21 models (F-score = 42.2%, A-sensitivity = 62.2%, sensitivity = 46.3%, specificity = 87.5%) but not including MCC at this stage. Further, these three models yielded higher accuracy and balanced accuracy, although not always greater coverage (Table 9). Equivocal predictions are not treated here. The coverage of QSAR models should be discussed if we plan further Projects.

The three teams reported using different training sets (Model no. 7, release model + selected NIHS data; Model no. 14, release model only; Model no. 16–1, release model + selected NIHS data) and methodologies for developing the models (Model no. 7 and 16–1, statistical; Model no. 14, rule-based).

Table 9. Performance metrics for the models selected BEFORE access to the Ames test results^{*1}.

Model no.	A-Sens. (%)	Sens. (%)	Spec. (%)	Acc. (%)	BA (%)	MCC	Cov. (%)	F1- score (%)	Training data ^{*2}	Model category ^{*3}
1	68.8	54.0	78.8	75.1	66.4	0.27	99.3	39.3	(1)	Rule-based + Statistical
3	89.3	76.9	80.0	79.5	78.5	0.45	80.9	53.8	(2)	Rule-based + Deep-learning (statistical)
4	88.5	62.7	84.6	81.4	73.7	0.40	97.9	49.7	(2)	Statistical
5–1	58.2	44.7	89.6	82.9	67.1	0.34	99.9	43.7	(2)	Rule-based + Statistical
6	60.0	40.7	90.5	83.5	65.6	0.31	82.2	40.8	(2)	Deep-learning (statistical)
7	71.4	50.8	91.5	85.9	71.2	0.42	86.8	49.7	(3)	Statistical
8	74.3	62.0	75.9	73.9	69.0	0.29	80.4	41.0	(3)	Statistical
9	60.0	47.5	78.1	73.6	62.8	0.21	100	34.8	(2) ^{*4}	Rule-based
10	33.3	23.7	96.9	86.0	60.3	0.30	98.7	33.4	(3)	Statistical
11	48.8	32.2	96.3	86.8	64.3	0.38	100	42.0	(2)	Statistical
12–1	72.7	62.8	76.1	74.2	69.5	0.30	94.1	41.2	(3) ^{*5}	Deep-learning (statistical)
13	50.0	33.5	95.0	85.9	64.3	0.35	100	41.4	(2)	Statistical
14	95.7	58.0	88.6	84.2	73.3	0.42	35.9	51.1	(1) ^{*6}	Rule-based
15–1	52.5	35.6	91.7	83.4	63.7	0.30	100	38.9	(1) ^{*7}	Statistical
16–1	71.3	53.4	90.7	85.1	72.0	0.43	100	51.6	(2)	Statistical
17–1	19.0	20.0	79.3	70.5	49.6	–0.01	99.5	16.8	(2)	Deep-learning (statistical)
18	76.3	58.1	85.8	81.7	71.9	0.38	100	48.5	(2)	No information
19	58.8	36.4	95.2	86.5	65.8	0.38	100	44.4	(2)	Deep-learning (statistical)
20-a	58.4	47.4	90.8	84.3	69.1	0.38	97.1	47.6	(3)	Statistical
20-b	45.0	31.8	86.5	78.4	59.2	0.18	100	30.4	(2)	Statistical
21	55.0	39.4	95.0	86.8	67.2	0.41	100	47.0	(3)	Deep-learning (statistical)

^{*1}Model no. is shown in Table 7. Abbreviations as in Table 4. All metrics range from 0% to 100% except MCC, which ranges from –1 to 1.

^{*2}Training data category: (1) release model (excluding data provided by the Second Project), (2) release model + all NIHS data, (3) release model + selected NIHS data.

^{*3}Models were categorized as rule-based, statistical or deep-learning.

^{*4}In the case of (1), A-Sens. = 78.8%, Sens. = 58.5%, Spec. = 73.2%, Acc. = 71.1%, BA = 65.9%, MCC = 0.24, Cov. = 100.0%, F1-score = 37.5%.

^{*5}Values were the average of two models (S+MUT_NIHS_ABC and S+MUT_NIHS_AC) of the three models submitted (S+MUT_NIHS_ABC, S+MUT_NIHS_AC and S+MUT_NIHS).

^{*6}In the case of all chemicals, A-Sens. = 76.3%, Sens. = 52.1%, Spec. = 83.0%, Acc. = 78.4%, BA = 67.5%, MCC = 0.30, Cov. = 100.0%, F1-score = 41.8%.

^{*7}This team reported the training data as (1), but it appears to be (2) according to the additional comments entered into the model information sheet.

Among the 21 models selected before access to the Ames test results, Model no. 3, 4, 8, and 12–1 demonstrated > 70% A-sensitivity (89.3, 88.5, 74.3, and 72.7%, respectively) and > 60% sensitivity (76.9, 62.7, 62.0, and 62.8%, respectively). Model no. 3 and 4 used release model + all NIHS data for training, while Model no. 8 and 12–1 used release model + selected NIHS data for training. Model no. 3 was categorized as a rule-based + deep-learning model, Model no. 4 and 8 as statistical models and Model no. 12–1 as a deep-learning model. It should be noted that the four teams (Team no. 3, 4, 8, 12) as well as Team no. 7 and 14 also participated in the First Project, suggesting that participation in the both projects facilitated the development of models with improved sensitivity.

Examination of the sensitivity of the 22 models selected after access to the results of the Ames test results (Table 10) is an evaluation positioned somewhere between an internal and an external validation because the model is assigned after access to the Ames test results but developed before access to the Ames test results. Models no. 3, 4, 5–2, 8, 12–2, 15–2, 16–2 and 17–2 yielded > 70% A-sensitivity (89.3, 88.5, 75.0, 74.3, 77.3, 75.0, 82.5, and 75.9%, respectively) and > 60% sensitivity (76.9, 62.7, 61.4, 62.0, 69.3, 64.4, 72.0, and 64.7%, respectively). Three of the eight teams (Team no. 15, 16 and 17) that developed these models did not participate in

Table 10. Performance metrics for the models selected AFTER access to the Ames test results^{*1}.

Model no.	A-Sens. (%)	Sens. (%)	Spec. (%)	Acc. (%)	BA (%)	MCC	Cov. (%)	F1- score (%)	Training data ^{*2}	Model category ^{*3}
1	68.8	54	78.8	75.1	66.4	0.27	99.3	39.3	(1)	Rule-based + Statistical
2–2	83.1	57.9	87.2	82.9	72.6	0.40	77.8	50.0	(1)	Statistical
3	89.3	76.9	80.0	79.5	78.5	0.45	80.9	53.8	(2)	Rule-based + Deep-learning (statistical)
4	88.5	62.7	84.6	81.4	73.7	0.40	97.9	49.7	(2)	Statistical
5–2	75.0	61.4	74.9	72.9	68.1	0.28	99.9	40.2	(3)	Rule-based + Statistical
6	60.0	40.7	90.5	83.5	65.6	0.31	82.2	40.8	(2)	Deep-learning (statistical)
7	71.4	50.8	91.5	85.9	71.2	0.42	86.8	49.7	(3)	Statistical
8	74.3	62.0	75.9	73.9	69.0	0.29	80.4	41.0	(3)	Statistical
9	60.0	47.5	78.1	73.6	62.8	0.21	100	34.8	(2)	Rule-based
10	33.3	23.7	96.9	86.0	60.3	0.30	98.7	33.4	(3)	Statistical
11	48.8	32.2	96.3	86.8	64.3	0.38	100	42.0	(2)	Statistical
12–2	77.3	69.3	73.1	72.6	71.2	0.32	94.1	42.1	(1)	Deep-learning (statistical)
13	50.0	33.5	95.0	85.9	64.3	0.35	100	41.4	(2)	Statistical
14	95.7	58.0	88.6	84.2	73.3	0.42	35.9	51.1	(1)	Rule-based
15–2	75.0	64.4	75.8	74.1	70.1	0.31	100	42.5	(1)	Deep-learning (statistical)
16–2	82.5	72.0	82.0	80.6	77.0	0.44	100	52.4	(2)	Statistical
17–2	75.9	64.7	72.7	71.5	68.7	0.28	99.5	40.3	(3)	Statistical
18	76.3	58.1	85.8	81.7	71.9	0.38	100	48.5	(2)	No information
19	58.8	36.4	95.2	86.5	65.8	0.38	100	44.4	(2)	Deep-learning (statistical)
20a	58.4	47.4	90.8	84.3	69.1	0.38	97.1	47.6	(3)	Statistical
20b	45.0	31.8	86.5	78.4	59.2	0.18	100	30.4	(2)	Statistical
21	55.0	39.4	95.0	86.8	67.2	0.41	100	47.0	(3)	Deep-learning (statistical)

^{*1}Model no. is shown in Table 8. Abbreviations as in Table 4.

^{*2}Training data category: (1) release model (excluding data provided by the Second Project) (2) release model + all NIHS data (3) release model + selected NIHS data.

^{*3}Models were categorized as rule-based, statistical or deep-learning.

the First Project. Nonetheless, the DGM/NIHS expects that the experience will help all teams improve model performance, particularly to reduce the FN rate.

From the complete set of models (Tables 7–10) selected before and after access to the results of the actual Ames test data, those with the highest MCC values and F1-scores were Model no. 2–2, 3, 14, 16–1, and 16–2. All five demonstrated MCC values ≥ 0.4 (0.40, 0.45, 0.42, 0.43, and 0.44%), F1-scores $\geq 50.0\%$ (50.0, 53.8, 51.1, 51.6, and 52.4%) and A-sensitivity $\geq 70\%$ (ranging from 71.3% and 95.7%). These groups included both rule-based and statistical models, and all were developed using different training datasets. These model performances were unrelated to the rough categories of the training dataset or development methodology. Thus, careful selection of training data and development methodology may partially help to provide improved Ames/QSAR models. In addition, given that the chemical space of the dataset used in the First Project was not the same as that used in the Second Project, expert knowledge of both mutagenicity and the chemical spaces of new chemicals may be necessary for developing improved models.

Range of performances

The ranges of performance metrics (minimum – maximum) and averages for the selected models tested in the Second Project are listed in Table 11 together with corresponding values from the First Project. The ranges and averages of performance metrics for models from the nine teams who also participated in the First Project are presented separately to assess the potential cumulative benefits of participation. Indeed, average A-sensitivity, sensitivity,

Table 11. Averages (and ranges) of performance metrics for all models in the Second Project versus the First Project^{*1}.

	Second Project			First Project ^{*2}		
	Selected models BEFORE access to Ames test results	Selected models BEFORE access to Ames test results (First Project participates only) ^{*3}	Selected models AFTER access to Ames test results	Phase I	Phase II	Phase III
A-Sens. (%)	62.2 (19.0–95.7)	74.5 (58.2–95.7)	68.3 (33.3–95.7)	68.9 (51.4–82.8)	72.4 (55.3–89.5)	71.1 (42.7–85.7)
Sens. (%)	46.3 (20.0–76.9)	56.2 (40.7–76.9)	52.0 (23.7–76.9)	55.8 (38.6–70.0)	56.6 (41.6–72.1)	56.6 (31.7–70.4)
Spec. (%)	87.5 (75.9–96.9)	83.9 (75.9–91.5)	85.2 (72.7–96.9)	78.7 (62.5–91.5)	85.4 (64.9–93.5)	80.1 (60.7–93.0)
Acc. (%)	81.4 (70.5–86.8)	79.9 (73.6–85.9)	80.4 (71.5–86.8)	75.3 (63.6–83.9)	81.2 (65.8–87.7)	76.8 (61.1–87.3)
BA (%)	66.9 (49.6–78.5)	70.1 (62.8–78.5)	68.6 (59.2–78.5)	67.2 (62.1–72.5)	71.0 (64.0–78.9)	68.4 (62.0–74.4)
PPV (%)	42.5 (14.4–60.3)	38.8 (27.5–48.7)	41.7 (27.5–60.3)	32.3 (23.8–46.1)	42.6 (27.4–58.2)	34.1 (21.1–51.0)
NPV (%)	90.5 (85.0–95.0)	92.0 (89.5–95.0)	91.3 (87.9–95.0)	91.4 (89.4–93.4)	91.9 (88.1–94.2)	91.9 (89.1–93.6)
MCC	0.33 (–0.01–0.45)	0.35 (0.21–0.45)	0.34 (0.18–0.45)	0.28 (0.20–0.39)	0.38 (0.24–0.50)	0.31 (0.17–0.44)
Cover. (%)	93.0 (35.9–100)	84.2 (35.9–100)	92.3 (35.9–100)	86.7 (14.5–100)	85.5 (18.0–100)	86.0 (9.7–100)
F1-score (%)	42.2 (16.8–53.8)	45.1 (34.8–53.8)	43.7 (30.4–53.8)	40.0 (31.8–48.9)	47.8 (36.9–57.9)	41.5 (31.7–51.5)
Number of models	21	9	22	18	21	19
No. of chemicals	1,589	1,589	1,589	3,902	3,829	4,409

^{*1}Abbreviations as in Table 4. Selected model information (names and training data) before and after access to Ames test results are shown in Tables 7–10.

^{*2}All models from Tables 5–7 of Honma et al. [14].

^{*3}Nine teams that participated in both the First Project and the Second Projects are indicated in Table 2.

balanced accuracy, negative prediction value, MCC and F1-score for the models developed by these nine teams (selected before access to the Ames test results) were higher than the averages of all 21 models before access to the Ames test results. These findings indicate that model performance for predicting positives was enhanced by participating in both the First and Second Projects. Similarly, average specificity, accuracy, positive prediction value and coverage for the 21 models before access to the Ames test results were higher than for the 22 models selected after access to the Ames test results, while average A-sensitivity, sensitivity, balanced accuracy, negative prediction value, MCC and F1-score were higher for the 22 models selected after access to the Ames test results. The ROC graphs (Figure 1) revealed that some low sensitivity models were replaced by models with higher sensitivity after the Ames

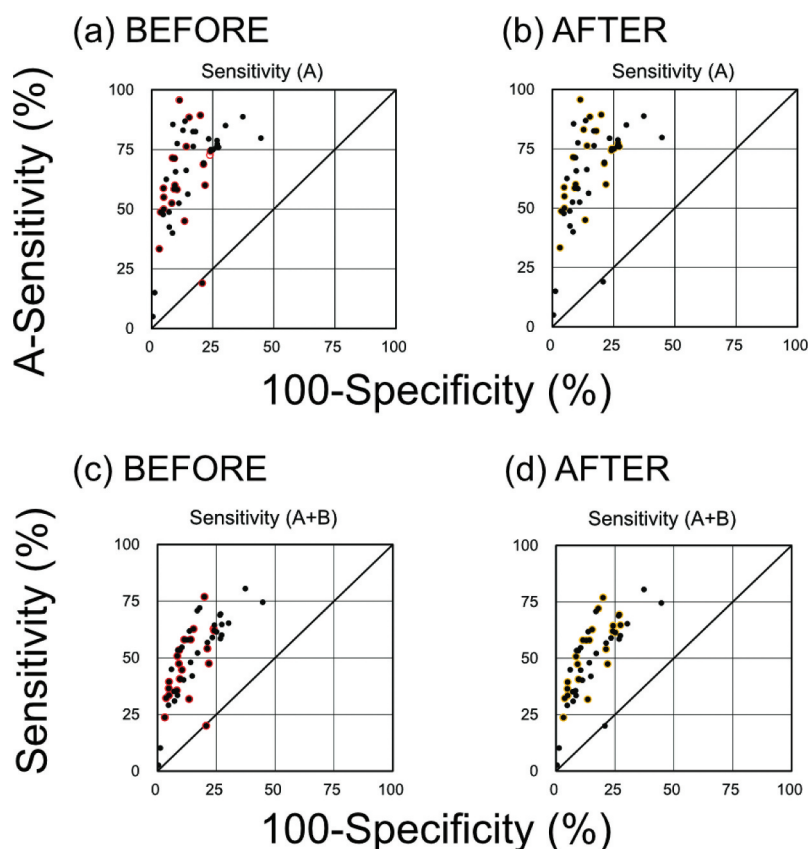


Figure 1. Receiver operating characteristic graphs of Ames mutagenicity prediction for the QSAR models. (a, b) Sensitivity to class A chemicals (A-sensitivity) versus specificity to class C chemicals. (c, d) Sensitivity to class A + B chemicals versus specificity to class C chemicals. Each black dot represents a single QSAR model from one of the participating teams. Red and yellow circles indicate the models selected before and after access to the Ames test data, respectively.

test results were available. Thus, participating teams tended to select models with superior performance for predicting positives (class A or B versus C) during the Second Project.

The 51 models examined in the Second Project demonstrated relatively high specificity but low sensitivity (see Appendices I and II) compared to the First Project. The average sensitivity values in all three phases of the First Project exceeded 55% (Table 11), whereas that for the 21 models selected before accessing the Ames test results in the Second Project was only 46.3% and the average for the 22 models selected after accessing the Ames test results was 52.0%. For these 21 and 22 models, respectively, selected before and after access to the Ames test results in the Second Project, average A-sensitivity was lower than in the First Project. In addition, the ranges (minimum – maximum) of A-sensitivity, sensitivity, and specificity for the 51 Ames/QSAR models in the Second Project were much wider than in the First Project (see Appendix II), which likely reflected the greater diversity of model types and training datasets (Figure 1).

Nonetheless, the ability to detect positives was higher for the models from teams participating in both First and Second Projects. Also, the average sensitivity (56.2%), MCC (0.35) and F1-score (45.1%) of all models were higher compared to Phases I and III of the First Project, while the average A-sensitivity (74.5%) of the models from teams participating in both was higher than the overall average in the First Project. These comparisons further suggest that participation in both projects improved model performance for predicting positives.

Conclusion

To improve QSAR models for predicting Ames mutagenicity, the Second Ames/QSAR International Challenge Project was conducted from 2020 to 2022. Overall, 21 teams from 11 countries participated in the project, with the DGM/NIHS providing the teams with curated training and trial datasets comprising data on 12,134 and 1,589 chemicals, respectively. After training, the teams were asked to use their models to predict the Ames mutagenicity of the trial chemicals and to report their predicted results to the DGM/NIHS. The DGM/NIHS then provided the teams with the results of the actual Ames test data for the trial chemicals to help the teams improve their models. To analyse the performance metrics of the models, each team was asked to select their best model before and after access to the results of the Ames test data. This aspect is only related to the spontaneous selection of models by the teams after the challenge results, however, all models cited at this work were developed, validated and challenged without previous access to the Ames data. Generally, the models included in the Second Project demonstrated high specificity but low sensitivity. Although the model performances were not as high as those reported from the First Project in this series, we expect that the experience of participating in the study will help the teams in their future model building. Actually, the nine teams who attended both the First and Second Projects showed improved sensitivity. We would like to emphasize again that the purpose of these projects is not to promote competition but to improve the model development skills of the participating teams.

Acknowledgments

The authors report there are no competing interests to declare. The authors express their gratitude to the Chemical Hazards Control Division, Industrial Safety and Health Department, MHLW, for allowing us to use ANEI-HOU Ames data in these projects. The authors express their acknowledgement to Dr. Toshio Kasamatsu for Ames data curation. This article reflects the views of the authors and does not necessarily reflect those of the U.S. Food and Drug Administration.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by the Ministry of Health, Labour and Welfare (MHLW) of Japan grant numbers H27-Chemistry-Designation-005, H28-Food-General-001 and H30-Chemistry-Destination-005, [21KD2005 and 21KA1001].

ORCID

A. Furuhamu  <http://orcid.org/0000-0003-4145-9590>
 C.E. Matos dos Santos  <http://orcid.org/0000-0002-3267-6297>
 J. Rathman  <http://orcid.org/0000-0002-3910-6844>
 C. Yang  <http://orcid.org/0000-0003-2529-866X>
 K. Cross  <http://orcid.org/0000-0002-2462-0533>
 G. Raitano  <http://orcid.org/0000-0001-6802-5595>
 E. Benfenati  <http://orcid.org/0000-0002-3976-5989>
 N. Jeliaskova  <http://orcid.org/0000-0002-4322-6179>
 R. Saiakhov  <http://orcid.org/0009-0002-2219-585X>
 S. Chakravarti  <http://orcid.org/0000-0001-7745-8747>
 R.S. Foster  <http://orcid.org/0000-0003-0618-9046>
 C. Bossa  <http://orcid.org/0000-0003-2084-2902>
 C. Laura Battistelli  <http://orcid.org/0000-0003-2386-0727>
 R. Benigni  <http://orcid.org/0000-0003-0943-1811>
 T. Sawada  <http://orcid.org/0009-0005-3629-7234>
 H. Wasada  <http://orcid.org/0000-0003-1694-9619>
 T. Hashimoto  <http://orcid.org/0009-0001-8582-5807>
 P.R. Daga  <http://orcid.org/0000-0002-2508-0903>
 R.D. Clark  <http://orcid.org/0000-0001-9509-8132>
 J. Mestres  <http://orcid.org/0000-0002-5202-4501>
 S. Matthews  <http://orcid.org/0000-0002-1652-543X>
 D. Guan  <http://orcid.org/0000-0001-6290-3166>
 J. Spicer  <http://orcid.org/0009-0005-6890-2128>
 R. Lui  <http://orcid.org/0000-0003-4673-9030>
 Y. Uesawa  <http://orcid.org/0000-0002-5773-991X>
 M.T.D. Cronin  <http://orcid.org/0000-0002-6207-4158>
 S.J. Belfield  <http://orcid.org/0000-0002-6532-2532>
 J.W. Firman  <http://orcid.org/0000-0003-0319-1407>
 N. Spînu  <http://orcid.org/0000-0002-9465-3090>
 G. Gini  <http://orcid.org/0000-0002-0334-420X>
 H. Hong  <http://orcid.org/0000-0001-8087-3968>
 Y. Igarashi  <http://orcid.org/0000-0001-8636-796X>

References

- [1] ICH-M7 (R1), ICH Harmonized Guideline, *Assessment and control Of DNA reactive (Mutagenic) impurities in pharmaceuticals to limit potential carcinogenic risk Current step 4 version dated 31 March 2017*, 2017. Available at https://database.ich.org/sites/default/files/M7_R1_Guideline.pdf.
- [2] R. Benigni, C. Laura Battistelli, C. Bossa, A. Giuliani, E. Fioravanzo, A. Bassan, M. Fuat Gatnik, J. Rathman, C. Yang, and O. Tcheremenskaia, *Evaluation of the applicability of existing (Q)SAR models for predicting the genotoxicity of pesticides and similarity analysis related with genotoxicity of pesticides for facilitating of grouping and read across*, EFSA Support. Publ. 16 (2019), pp. 1598E. doi:10.2903/sp.efsa.2019.EN-1598.
- [3] K. Hansen, S. Mika, T. Schroeter, A. Sutter, A. Ter Laak, T. Steger-Hartmann, N. Heinrich, and K.-R. Müller, *Benchmark data set for in silico prediction of Ames mutagenicity*, J. Chem. Inf. Model. 49 (2009), pp. 2077–2081. doi:10.1021/ci900161g.
- [4] D. Kirkland, E. Zeiger, F. Madia, N. Gooderham, P. Kasper, A. Lynch, T. Morita, G. Ouedraogo, J. M. Parra Morte, S. Pfuhler, V. Rogiers, M. Schulz, V. Thybaud, J. van Benthem, P. Vanparys, A. Worth, and R. Corvi, *Can in vitro mammalian cell genotoxicity test results be used to complement positive results in the Ames test and help predict caCan 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*, *Mutat. Res. Genet. Toxicol. Environ. Mutag.* 775-776 (2014), pp. 55–68.
- [5] D. Kirkland, E. Zeiger, F. Madia, and R. Corvi, *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*, *Mutat. Res. Genet. Toxicol. Environ. Mutag.* 775-776 (2014), pp. 69–80. doi:[10.1016/j.mrgentox.2014.10.006](https://doi.org/10.1016/j.mrgentox.2014.10.006).
 - [6] F. Madia, D. Kirkland, T. Morita, P. White, D. Asturiol, and R. Corvi, *EURL ECVAM genotoxicity and carcinogenicity database of substances eliciting negative results in the Ames test: Construction of the database*, *Mutat. Res. Genet. Toxicol. Environ. Mutag.* 854-855 (2020), pp. 503199. doi:[10.1016/j.mrgentox.2020.503199](https://doi.org/10.1016/j.mrgentox.2020.503199).
 - [7] F. Metruccio, I. Castelli, C. Civitella, C. Galbusera, F. Galimberti, L. Tosti, and A. Moretto, *Compilation of a database, specific for the pesticide active substance and their metabolites, comprising the main genotoxicity endpoints*, *EFSA Support. Publ.* 14 (2017), pp. 1229E. doi:[10.2903/sp.efsa.2017.EN-1229](https://doi.org/10.2903/sp.efsa.2017.EN-1229).
 - [8] P. Pradeep, R. Judson, D.M. DeMarini, N. Keshava, T.M. Martin, J. Dean, C.F. Gibbons, A. Simha, S.H. Warren, M.R. Gwinn, and G. Patlewicz, *An evaluation of existing QSAR models and structural alerts and development of new ensemble models for genotoxicity using a newly compiled experimental dataset*, *Comput. Toxicol.* 18 (2021), pp. 100167. doi:[10.1016/j.comtox.2021.100167](https://doi.org/10.1016/j.comtox.2021.100167).
 - [9] P. Pradeep, R. Judson, D.M. DeMarini, N. Keshava, M. Todd, J. Dean, C. Gibbons, A. Simha, S. Warren, M. Gwinn, and G. Patlewicz, *Evaluation of existing QSAR models and structural alerts and development of new ensemble models for genotoxicity using a newly compiled experimental dataset*, The United States Environmental Protection Agency's Center for Computational Toxicology and Exposure, 2021. Available at https://gaftp.epa.gov/COMPTOX/CCTE_Publication_Data/CCED_Publication_Data/PatlewiczGrace/CompTox-genetox/.
 - [10] R. Benigni, C.L. Battistelli, C. Bossa, O. Tcheremenskaia, and P. Crettaz, *New perspectives in toxicological information management, and the role of ISSTOX databases in assessing chemical mutagenicity and carcinogenicity*, *Mutagenesis* 28 (2013), pp. 401–409. doi:[10.1093/mutage/get016](https://doi.org/10.1093/mutage/get016).
 - [11] A. Hillebrecht, W. Muster, A. Brigo, M. Kansy, T. Weiser, and T. Singer, *Comparative evaluation of in silico systems for Ames test mutagenicity prediction: Scope and limitations*, *Chem. Res. Toxicol.* 24 (2011), pp. 843–854. doi:[10.1021/tx2000398](https://doi.org/10.1021/tx2000398).
 - [12] S.-Y. Bae, J. Lee, J. Jeong, C. Lim, and J. Choi, *Effective data-balancing methods for class-imbalanced genotoxicity datasets using machine learning algorithms and molecular fingerprints*, *Comput. Toxicol.* 20 (2021), pp. 100178. doi:[10.1016/j.comtox.2021.100178](https://doi.org/10.1016/j.comtox.2021.100178).
 - [13] C. Barber, A. Cayley, T. Hanser, A. Harding, C. Heghes, J.D. Vessey, S. Werner, S.K. Weiner, J. Wichard, A. Giddings, S. Glowienke, A. Parenty, A. Brigo, H.-P. Spirkel, A. Amberg, R. Kemper, and N. Greene, *Evaluation of a statistics-based Ames mutagenicity QSAR model and interpretation of the results obtained*, *Regul. Toxicol. Pharmacol.* 76 (2016), pp. 7–20. doi:[10.1016/j.yrtph.2015.12.006](https://doi.org/10.1016/j.yrtph.2015.12.006).
 - [14] M. Honma, A. Kitazawa, A. Cayley, R.V. Williams, C. Barber, T. Hanser, R. Saiakhov, S. Chakravarti, G.J. Myatt, K.P. Cross, E. Benfenati, G. Raitano, O. Mekenyan, P. Petkov, C. Bossa, R. Benigni, C.L. Battistelli, A. Giuliani, O. Tcheremenskaia, C. DeMeo, U. Norinder, H. Koga, C. Jose, N. Jeliaskova, N. Kochev, V. Paskaleva, C. Yang, P.R. Daga, R.D. Clark, and J. Rathman, *Improvement of quantitative structure–activity relationship (QSAR) tools for predicting Ames mutagenicity: Outcomes of the Ames/QSAR International Challenge Project*, *Mutagenesis* 34 (2019), pp. 3–16. doi:[10.1093/mutage/gey031](https://doi.org/10.1093/mutage/gey031).
 - [15] Japan Ministry of Health Labour and Welfare (MHLW), *Industrial Safety and Health Act of Japan*, Japan Ministry of Health Labour and Welfare (MHLW), ed., 1972.
 - [16] The Division of Genetics and Mutagenesis/National Institute of Health Sciences (DGM/NIHS), *AMES/QSAR international collaborative study*, 2019. Available at <https://www.nihs.go.jp/dgm/amesqsar.html>.

- [17] D. Weininger, *SMILES, a chemical language and information-system. 1. Introduction to methodology and encoding rules*, J. Chem. Inf. Comput. Model. 28 (1988), pp. 31–36. doi:[10.1021/ci00057a005](https://doi.org/10.1021/ci00057a005).
- [18] A. Furuhashi, T. Kasamatsu, K. Sugiyama, and M. Honma, *Curation of more than 10,000 Ames test data used in the Ames/QSAR international challenge projects*, in *QSAR in Safety Evaluation and Risk Assessment*, H. Hong, ed., Academic Press, 2023, pp. 365–372. doi:[10.1016/B978-0-443-15339-6.00022-9](https://doi.org/10.1016/B978-0-443-15339-6.00022-9).
- [19] H.H. Gul, E. Egrioglu, and E. Bas, *Statistical learning algorithms for dendritic neuron model artificial neural network based on sine cosine algorithm*, Inf. Sci. 629 (2023), pp. 398–412. doi:[10.1016/j.ins.2023.02.008](https://doi.org/10.1016/j.ins.2023.02.008).
- [20] OECD, *OECD Series on Testing and Assessment No. 69, Guidance Document on the Validation of (Quantitative) Structure-Activity Relationships [(Q)SAR] Models*, OECD, Paris, 2007. Available at <http://www.oecd.org/dataoecd/55/35/38130292.pdf>.
- [21] W.F. Schneider and H. Guo, *Machine learning*, J. Phys. Chem A 122 (2018), pp. 879–879. doi:[10.1021/acs.jpca.8b00034](https://doi.org/10.1021/acs.jpca.8b00034).