Regulatory Toxicology and Pharmacology 171 (2026) 106153

Contents lists available at ScienceDirect

Regulatory Toxicology and Pharmacology

journal homepage: www.elsevier.com/locate/yrtph

In silico prediction of Ames mutagenicity for organosilicon compounds: Exploring and enhancing chemical space boundaries

Barbara G. Schmitt [a,*], Emilio Benfenati [b], Robert S. Foster [c], Grace Kocks [c], Giuseppa Raitano [b], Rachael E. Tennant [c], Carole Forlini [d], Eckart Gura [e], Shawn Seidel [f], Farah Koraichi-Emeriau [g]

[a] *Evonik Operations GmbH, Essen, Germany*
[b] *Istituto di Ricerche Farmacologiche Mario Negri, IRCCS, Milan, Italy*
[c] *Lhasa Limited, Granary Wharf House, Leeds, UK*
[d] *Elkem Silicones France SAS, St Fons, France*
[e] *Wacker Chemie AG, München, Germany*
[f] *The Dow Chemical Company, Midland, MI, United States*
[g] *Equitox, Lyon, France*

ARTICLE INFO

Handling Editor: Dr. Martin van den Berg

Keywords:
Organosilicon substances
Mutagenicity
(Q)SAR
Sarah nexus
Derek nexus
VEGA mutagenicity models
New approach methodologies (NAMs)

ABSTRACT

Organosilicon chemistry, a subsection of organic chemistry with unique chemical properties, is typically underrepresented in the training sets of (Q)SAR models and often outside the applicability domain.

Using bacterial reverse mutation as an endpoint for a proof-of-concept study, we compiled a peer-reviewed reference dataset with publicly available reliable Ames tests of more than 100 organosilicon substances to assess the predictive performance of fourteen *in silico* methods, including mechanistic profilers, expert systems, hybrid and statistical models with varying underlying algorithms (Derek Nexus, Sarah Nexus, VEGA mutagenicity models, TEST mutagenicity models, OECD QSAR Toolbox profilers).

The assessment showed that most, but not all, tools predict mutagenic potential of organosilicon chemistry as accurately as for other organic datasets, with six tools providing balanced accuracies ≥80%. The nature of the algorithm was identified as the key driver for accuracy. The presence of the silicon atom in a molecule sometimes resulted in substances being outside model-specific applicability domains, but this did not necessarily correlate with predictive performance. Furthermore, the assessment showed that organosilicon chemistry does not possess intrinsic gene mutation potential. All positive Ames test results in the dataset could be linked to organofunctional structural alerts known to be related to mutagenicity.

1. Introduction

The reliable determination of genotoxicity potential is a key part of toxicological safety assessment, with the Bacterial Reverse Mutation (Ames) test (OECD Test Guideline (TG) 471; OECD, 2020) forming part of core testing batteries under various regulatory frameworks. The Ames test determines the potential of a substance to chemically induce reverse mutations that restore the functional capability of the bacteria to synthesize essential amino acids.

The simple and well-established link between the structure of a chemical and its ability to covalently bind to DNA nucleotides via electrophilic attack allows the development of comparably robust computational applications to predict the outcome of the Ames test. (Quantitative) Structure-Activity Relationship ((Q)SAR) models can be grouped into rule-based expert systems and statistical models. Expert systems screen for the presence of known structural alerts (SA) in a molecule. Statistical QSAR models are agnostic to mechanistic knowledge and employ computational algorithms to assess chemical structures

Abbreviations: AD, Applicability Domain; ADI, Applicability Domain Index; AP, Alert Performance; CAS RN, CAS registry number; NP, Negative Predictivity; PP, Positive Predictivity; (Q)SAR, (Quantitative) Structure Activity Relationship; QAF, (Q)SAR Assessment Framework; SA, Structural Alert(s); TB, OECD QSAR Toolbox; TEST, Toxicity Estimation Software Tool; TG, Test Guideline; TTC, Threshold of Toxicological Concern; WoE, Weight of Evidence.

* Corresponding author. Evonik Operations GmbH, Goldschmidtstraße 100, D-45127, Essen, Germany.
E-mail address: barbara1.schmitt@evonik.com (B.G. Schmitt).

https://doi.org/10.1016/j.yrtph.2026.106153
Received 24 January 2026; Received in revised form 29 May 2026; Accepted 4 June 2026
Available online 7 June 2026

and/or molecular/physicochemical descriptors from a reference dataset ("training set"). Hybrid models combine expert systems and statistical approaches. While not (Q)SAR models by definition, mechanistic profilers (primarily found in the OECD QSAR Toolbox (TB)) are also useful *in silico* tools to detect relevant structural alerts. Mechanistic profilers derive their logic from chemistry expert knowledge (*e.g.* electrophilicity) rather than test results and – due to the absence of a training set - do not have a defined applicability domain in the same sense as conventional (Q)SAR models.

Current high-quality (Q)SAR models match the accuracy of the Ames test itself (Benigni et al., 2019; Honma et al., 2019) and form an integral part as toxicological screening tools during research and development and in chemical safety assessment of impurities or constituents of complex substance compositions, especially in cases where isolation of single chemical entities is not practical. Regulatory applications include assessment of impurities in food (EFSA, 2019), pharmaceuticals (ICH M7, 2023) and cosmetics (SCCS, 2023). The recently published (Q)SAR Assessment Framework (QAF, OECD, 2024) allows a standardised and semi-quantitative assessment of the reliability of (Q)SAR predictions and is expected to increase their acceptance and use in regulatory contexts (Gissi et al., 2024).

A defined Applicability Domain (AD) constitutes one of the five principles of (Q)SAR validation (OECD, 2004). Whether or not a chemical is considered within the AD is defined by a model-specific assessment of its similarity to the training set (OECD, 2007). In certain regulatory settings, the reliability of out-of-domain predictions can be improved if predictions with structurally similar chemicals provide sufficient performance (ECHA, 2016; OECD, 2024).

Organosilicon compounds (*e.g.* silicone polymers, cyclic and short-chain linear siloxanes and structurally diverse silanes) form a distinct area of the chemical space of organic chemistry, containing one or more covalent silicon-carbon bonds in the molecule. The chemistry is used in a variety of industrial, medical and consumer applications. The silicon atom is more electropositive than carbon, resulting in specific differences from carbon-based molecular structures and reactivities. Organosilicon compounds are typically not well represented in training sets of (Q)SAR models (ISS, 2025; Hansen et al., 2009) and often outside their AD. Until now, *in silico* methods have not been systematically assessed to verify their applicability to this chemistry.

The formation of covalent bonds between electrophilic agents and nucleophilic DNA bases is a key driver of chemically-induced mutagenesis. Due to the electropositive nature of the silicon atom, and based on previous test results, we hypothesised that electrophilic organofunctional groups but not the silicon-based moieties have the ability to form covalent bonds with DNA, and that organosilicon compounds can therefore be predicted with existing (Q)SAR models. We therefore compiled curated datasets of Ames tests with organosilicon compounds and assessed coverage and performance of fourteen different *in silico* tools (statistical models, expert systems and mechanistic profilers) against these datasets. In addition, the training set of Sarah Nexus (Lhasa, 2026) was bolstered with an organosilicon reference dataset, aiming to increase coverage and performance. The results of these investigations are presented below.

2. Materials and methods

2.1. Ames tests

The bacterial reverse mutation assays were conducted in accordance with or similar to OECD TG 471 (OECD, 2020). Five auxotrophic bacterial tester strains (typically *Salmonella typhimurium* TA 98, TA100, TA1535, TA1537 and *Escherichia coli* WP2 uvrA) were incubated with test substance at different concentrations without and with metabolic activation in media lacking histidine or tryptophan. Mutagenic test substances revert point mutations, enabling the bacteria to grow.

2.2. Compilation of the Si-Ames dataset and datasets A and B

The Si-Ames dataset comprises Ames tests that were submitted in REACH registrations dossiers by the REACH consortium RECONSILE, comprising the major manufacturers of organosilicon compounds. Details on the study design, as provided in Robust Study Summaries, are publicly available. Only studies with a Klimisch score of 1 or 2 were included in the dataset (Klimisch et al., 1997). Non-/pre-GLP (OECD, 1998) studies were included if the documentation indicated GLP-like reliability. Only studies on discrete (mono-constituent) chemicals were considered for inclusion.

The Si-Ames dataset was highly imbalanced towards negative test results. To bolster the dataset with additional positive Ames assays, a property search was conducted via the OECD eChemPortal, querying the following information: Genetic toxicity *in vitro*, experimental study, Reliability: 1 (reliable without restrictions) or 2 (reliable with restrictions), Test guideline: OECD TG 471 (Bacterial Reverse Mutation Assay), test results, genotoxicity: positive.

The list was then exported into Excel and searched for mono-constituent organosilicon compounds. The endpoint study records were manually reviewed to verify the results, assess the robustness of the documentation and to confirm guideline compliance.

This combined dataset underwent data curation as shown in Fig. 1. Substances with negative test results but insufficient strain coverage (as per OECD TG 471) throughout all available studies were removed from the assessment. If both positive and negative studies were available for the same substance, the endpoint study records and – if available - the study reports of the Ames tests were reviewed and re-assessed in detail, with a focus on validity and study design, test item identity and biological plausibility in accordance with OECD TG 471 and public literature (Cariello and Piegorsch, 1996; Levy et al., 2019; Prival and Zeiger, 1998). The assessment was intentionally agnostic to the presence of well-known SA for DNA-reactivity in the molecule and of the outcome of other genotoxicity studies.

The outcome of this assessment resulted in two separate datasets:

- Dataset A (Conservative approach): Equivocal study outcomes were counted as positive, and any increased mutation frequency observed in a tester strain throughout all studies with one individual substance led to an overall substance call as positive.
- Dataset B (Refined approach): Any substance with equivocal or conflicting study outcomes was removed from the dataset.

2.3. In silico mutagenicity predictions

Datasets A and B were assessed with fourteen different *in silico* applications for the prediction of Ames mutagenicity, comprising four TB profilers and ten (Q)SAR models (Table 1). For each prediction, the entry was checked manually to verify correct matching of chemical structure/SMILES and CAS registry number (RN). Where necessary, SMILES notations were desalted and neutralised. Where applicable, both in- and out-of-domain predictions were obtained and labelled accordingly.

Derek Nexus (Lhasa, 2026) was used with the default settings except for the following: Species = bacterium; Endpoints = mutagenicity *in vitro*; Perceive tautomers = no; Perceive mixtures = no. Predictions at the reasoning level of "PROBABLE", "PLAUSIBLE" or "EQUIVOCAL" were considered positive. Predictions at the reasoning level of "INACTIVE" were considered negative. Predictions with unclassified/misclassified feature(s) (Williams et al., 2016) were considered as in-domain negative predictions for the purpose of this study; however, it should be noted that predictions with these features have higher uncertainty than predictions without and expert assessment of the identified features is required to conclude negative for regulatory use.

Mutagenicity predictions were obtained using the Sarah Nexus mutagenicity model with the default settings. Sarah Nexus provides a

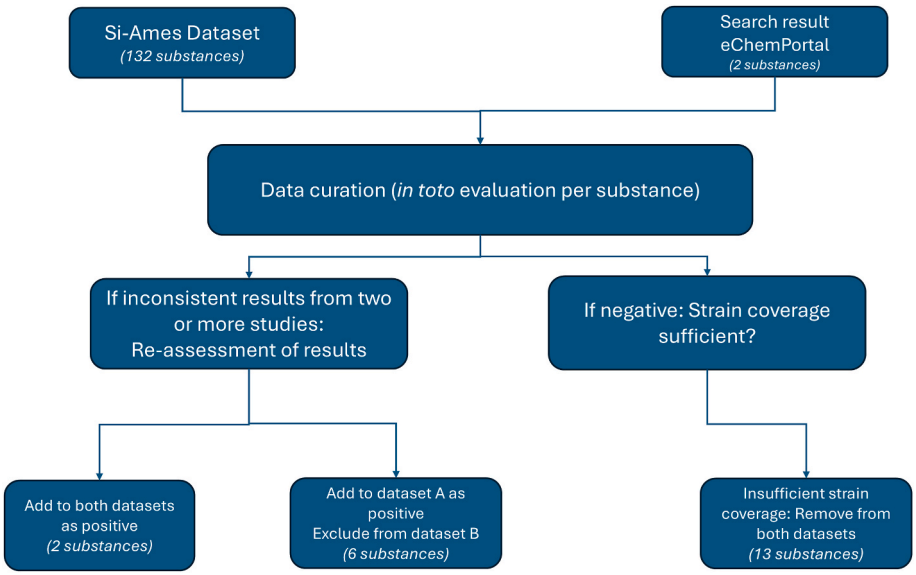

Fig. 1. Data curation workflow for the compilation of datasets A and B.

prediction result ("positive" or "negative") alongside a confidence score (since v.5.0.0: probability score). If either the score is below the threshold to make a prediction ("equivocal") or the structure is outside the model's AD ("outside domain"), no prediction can be made.

The VEGA platform (VEGAHUB, 2026) provides four individual models and one consensus model to predict Ames mutagenicity, based on the predictions and the reliability of the four individual models. The individual models (Caesar, ISS, SarPy/IRFMN, KNN/Read-Across) are trained with different datasets, covering several types of substances and their experimental results from the Ames test. Each prediction is linked to a reliability assessment, based on a quantitative appreciation (Applicability Domain Index (ADI), ranging between 0 and 1) of the model predictivity for the structurally most similar substances in the training set. In addition, any potential inconsistencies are indicated to the user. Accordingly, predictions are given with low, moderate and high reliability. The metrics for the applicability domain within VEGA do not define a binary score (in- or out-of-domain). The higher reliability is associated to higher statistical values, but even predictions with "low reliability" provide acceptable statistical value for mutagenicity (Danieli et al., 2023). Aiming for simplified language and definitions, we define VEGA predictions with low reliability as out-of-domain, and predictions with moderate or high predictivity as in-domain for the purpose of this project.

The three TEST mutagenicity QSAR models comprised two statistical models (Hierarchical Clustering and Nearest Neighbour) and a consensus model based on the two individual models (Martin, 2020). The models were used with default settings. The models only provide results for in-domain predictions.

The list of *in silico* applications included four TB profilers that are flagged as "suitable" for the prediction of Ames mutagenicity (Table 1). TB profilers screen for SA either solely based on knowledge of chemical reactivity ("general mechanistic") or by taking toxicological data into consideration ("endpoint specific"). TB profilers were not developed to serve as validated stand-alone expert systems and lack a defined AD. However, their ability to identify SA resembles expert systems, and, for endpoint specific profilers, the underlying reference data may be identical. Rarely, TB profilers are accepted as SAR models in regulatory settings (EFSA, 2021).

2.4. Inclusion in Sarah Nexus

After the performance assessment with Sarah Nexus, the initial Si-Ames dataset was donated to Lhasa Limited for peer review of the experimental data and for inclusion in the model, either directly into the training set, or – in case of conflicted or equivocal results – as additional information. The dataset is available in Sarah Nexus v.3.3.0 and subsequent versions.

2.5. OECD QSAR toolbox profilers - alert performance

To assess the overall relevance of the TB profilers, the Alert Performance (AP) was assessed for the result "no alert found" for each profiler (Yordanova et al., 2019), using the following TB databases: Genotoxicity OASIS, Bacterial mutagenicity ISSSTY, Genotoxicity & Carcinogenicity ECVAM, and ECHA REACH. The AP was then checked against both Gene Mutation I and Mutagenicity I (ECVAM) databases, with "maximal" selected for the Categorical Scale. Only results based on ten or more substances were considered. The substance output was then searched for chemical structures containing a silicon atom.

2.6. Statistics

The coverage was calculated as follows:

$$\%\ Coverage = \frac{in-domain\ predictions}{all\ substances}\ x\ 100$$

Model Performance Statistics were calculated as introduced by Cooper et al. (1979), based on the following definitions and equations (*TP*: True Positives, *FN*: False Negatives, *TN*: True Negatives, *FP*: False Positives):

$$\%\ Sensitivity = \frac{TP}{TP+FN}\ x\ 100$$

$$\%\ Specificity = \frac{TN}{TN+FP}\ x\ 100$$

$$\%\ Accuracy = \frac{TP+TN}{TP+FN+TN+FP}\ x\ 100$$

$$\%\ Balanced\ Accuracy = \frac{\%Sensitivity + \%Specificity}{2}$$

$$\%\ Positive\ Predictivity\ (PP) = \frac{TP}{TP+FP}\ x\ 100$$

Table 1
Assessed in silico applications.

Application	Type	Description/References
Derek Nexus (Lhasa Ltd.), version 6.2.1 (Derek KB, 2022 2.0, Nexus v.2.5.2)	Expert system	Rule-based expert system with endpoint-specific structural alerts QMRF: Lhasa Ltd. (2026)
Sarah Nexus (Lhasa Ltd.), version 3.2.1 (Sarah model – 2022.2) (Nexus v.2.5.2)	Statistical model	Weighted average of contributing hypotheses in self-organising hypothesis network (SOHN) (Hanser et al., 2014) QMRF: Lhasa Ltd. (2026)
VEGA Mutagenicity Consensus (version 1.0.4)	Consensus model	Based on prediction and reliability of the Vega mutagenicity models Caesar, ISS, SarPy/IRFMN and kNN/Read-Across QMRF: Gamba A and Raitano G (2022)
VEGA Mutagenicity Caesar (version 2.1.14)	Hybrid model	Based on data mining with support vector machines and expert knowledge coded as structural alerts (Ferrari and Gini, 2010) QMRF: Gamba and Benfenati (2022)
VEGA Mutagenicity ISS (version 1.0.3)	Expert system	Implementation of the Toxtree (Toxtree, 2025) Module "In vitro mutagenicity (Ames test) alerts by ISS" (Istituto Superiore di Sanita) (Benigni and Bossa, 2008, 2011; Benigni et al., 2008) QMRF: Benfenati et al. (2022)
VEGA Mutagenicity SarPy/IRFMN (version 1.0.8)	Statistical Model	List of automatically extracted rules (active and inactive), using the SarPy algorithm (Ferrari et al., 2013) QMRF: Benfenati and Raitano (2022)
VEGA Mutagenicity KNN/Read-Across (version 1.0.1)	Statistical Model	Automated read-across (Flores et al., 2014) QMRF: Golbamaki and Benfenati (2022)
TEST version 5.1.2., Mutagenicity Consensus	Consensus Model	Average of individual model predictions (Hierarchical Clustering and Nearest Neighbour) Martin (2020)
TEST Mutagenicity Hierarchical Clustering 5.1.2	Statistical Model	Weighted average of the predictions from several different cluster models Martin (2020)
TEST Mutagenicity Nearest Neighbour 5.1.2	Statistical Model	Average of the predictions from the most similar 3 chemicals in the training set Martin (2020)
TB Profiler: DNA binding by OASIS v2.10	Mechanistic Alert	General mechanistic profiler. 244 structural alerts for presumed DNA reactivity based on theoretical considerations
TB Profiler: DNA binding by OECD v2.3	Mechanistic Alert	General mechanistic profiler. Structural alerts for covalent DNA binding of small organic molecules (Enoch and Cronin, 2010; OECD, 2010), based on theoretical considerations. The initial version of this profiler sourced, among others, alerts developed by Benigni et al. (2007), which are also used in Toxtree (OECD, 2010); therefore an overlap between this profiler and the *in vitro* mutagenicity (Ames test) alerts by ISS is expected. Metabolic activation is considered in this profiler (OECD, 2010).
TB Profiler: DNA alerts for Ames, CA and MNT by OASIS v3.10	Mechanistic Alert/Expert system	Endpoint-specific profiler Identifies structural alerts, informed by Ames test results. Takes into account steric and electronic factors that hinder DNA binding of a reactive functional group.
TB Profiler: *In vitro* mutagenicity (Ames test) alerts by ISS v2.7	Mechanistic Alert/Expert system	Endpoint-specific profiler Implementation of the Toxtree (Toxtree, 2025) Module "*In vitro* mutagenicity (Ames test) alerts by ISS" (Istituto Superiore di Sanita) (Benigni and Bossa, 2008, 2011; Benigni et al., 2008). Informed by Ames test results.

$$\%\ Negative\ Predictivity\ (NP) = \frac{TN}{TN + FN} x\ 100$$

3. Results

3.1. Dataset characterisation

After removal of non-discrete substances and inorganics, the Si-Ames dataset comprised 132 registered substances with one or more reliable Ames tests.

The search via eChemPortal gave 1640 hits, including four additional mono-constituent organosilicon compounds. A subsequent review revealed that for one substance, read-across was applied from two substances that were already part of the Si-Ames dataset. A second substance showed cytotoxicity but no mutation. Consequently, both hits were excluded, and the two remaining substances with positive Ames tests were added.

The combined dataset contained 134 substances with a total of 207 studies. 13 non-mutagenic substances were removed due to insufficient coverage of tester strains throughout all available studies. The Ames tests of nine substances with contradicting outcomes were re-assessed: For two substances (3-chloropropyltrimethoxysilane (CAS RN 2530-87-2) and dichloro (3-chloropropyl)methylsilane (CAS RN 7787-93-1)), the positive outcome was confirmed. The data of six substances were considered conflicted, and the substances were included in dataset A but not B (Supplementary Material S1 and Fig. 2). For the ninth substance (3-trimethoxysilylpropane-1-thiol (CAS RN 4420-74-0)), the positive study was considered inadequate and removed from both final datasets. The decision logic for each substance as well as the characteristics of the final composition of both datasets are described in Supplementary Material S1 and S2.

Both datasets comprised more than 100 substances, with 12% mutagens in dataset A and 8% in dataset B. Alkoxysilanes (mono-, di- and trialkoxysilanes, primarily methoxy- and ethoxy-functionalised) formed the biggest part of both datasets (A:50%, B:49%), followed by chlorosilanes (18% each) and siloxanes (A:17%, B:18%).

3.2. In silico predictions

With a few exceptions, SMILES notations and/or CAS RN could be processed by all applications. The OECD QSAR Toolbox required manual verification for both the input as SMILES notation and CAS RN due to several errors and mismatches. Likewise, infrequent mismatches were identified with the TEST software. The VEGA KNN/Read-Across model could not process three data entries.

In most cases, the VEGA Consensus model used all four models for the prediction. For six substances, only one model was considered, and

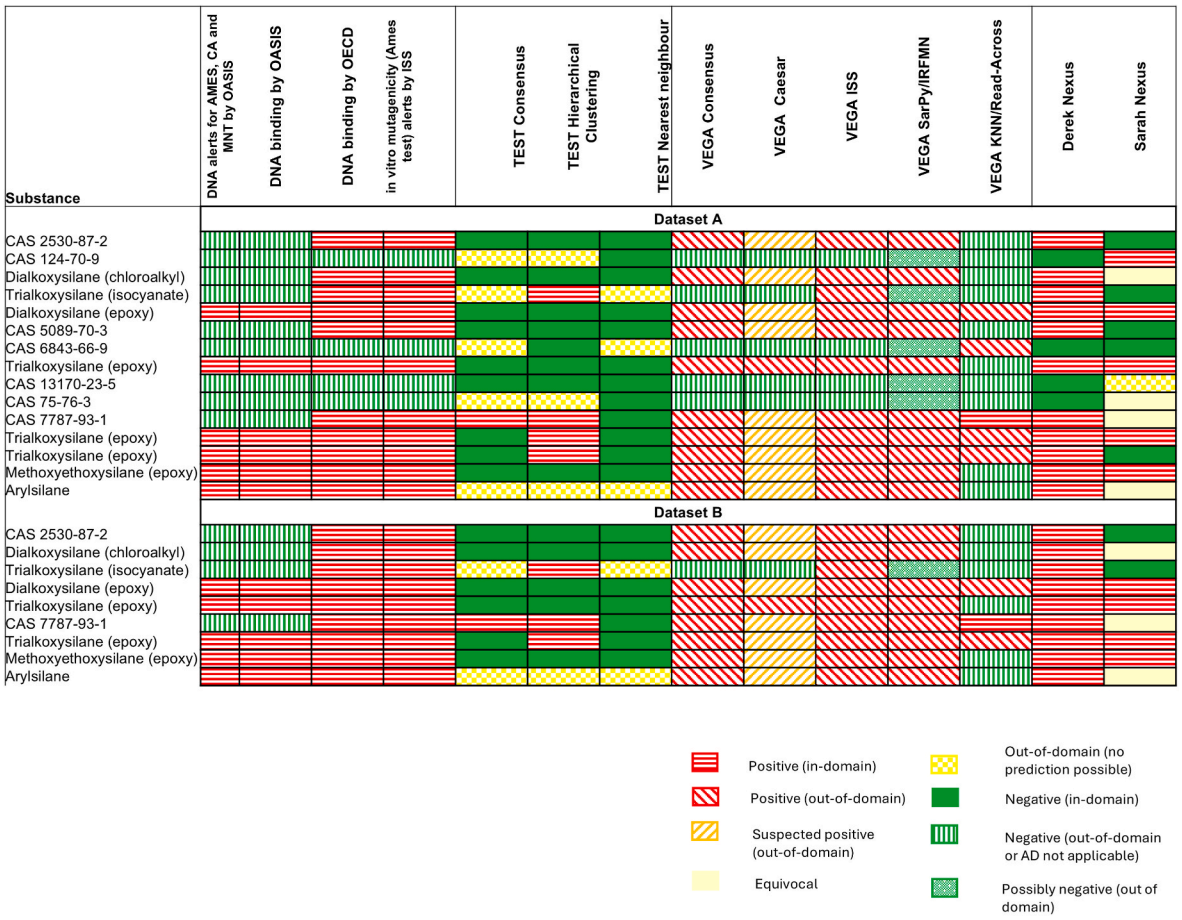

Fig. 2. Prediction Heatmaps for Ames-positive substances in datasets A and B. Substances are anonymised by default but identified by CAS RN if discussed in Supplementary Material S1.

for three substances, the results from three models were used.

The assessment revealed that some representatives of our organosilicon datasets were already part of the training set of some individual models (Sarah Nexus, VEGA KNN/Read-Across and the two individual TEST models). None of the substance were included in the training sets of the VEGA CAESAR, SarPy/IRFMN and ISS models. Verification was not possible in Derek Nexus or the TB profilers. In conclusion, the assessment did not constitute a purely external validation.

The performance metrics of each application for datasets A and B are shown in Table 2.

Specificity was high for all tools, ranging between 85% and 100%. Sensitivity varied significantly across all assessed applications (A: 0-73%, B: 0-100%) and was equal or higher in dataset B in comparison with dataset A for all models (Table 2). Except for the VEGA SarPy/IRFMN model, dataset B showed an equal or lower positive predictivity compared to dataset A. Overall, VEGA ISS, VEGA SarPy/IRFMN and Derek Nexus performed best, each reaching sensitivity of ≥89%, specificity of ≥96% and balanced accuracy of ≥94% in dataset B, in combination with moderate to high (69%-100%) positive and negative predictivity.

3.2.1. *OECD QSAR toolbox profilers*

While the two TB profilers "DNA binding by OECD" and "*in vitro* mutagenicity (Ames test) alerts by ISS" showed good (73% each for dataset A) to perfect (100% each for dataset B) sensitivity, the two OASIS-generated profilers predicted mutagens with significantly lower sensitivity (40% and 56% for dataset A and B, respectively) (Table 2). The positive predictivity of "DNA binding by OECD" was below 50% in each dataset. Overall, the "*in vitro* mutagenicity (Ames Test) alerts by ISS" profiler performed best of the four profilers.

Results of the alert performance (AP) verification of "No alert found" for each of the four profilers are shown in the Supplementary Material, S3. For all four profilers, AP values of 99% were calculated based on data from at least 265 substances, each containing fourteen organosilicon compounds. No false negative prediction was identified from this external dataset for organosilicon chemistry.

3.2.2. *TEST models*

TEST comprises a training set of 6512 chemicals (Hansen et al., 2009), including fourteen silicon-based substances, translating to 0.2% in the training set. The coverage of the three TEST models ranged between 73% and 88%, with the coverage of the consensus model being lower than of the two individual models. In-domain predictions of TEST consensus models require a minimum of two in-domain predictions of individual models; since for mutagenicity, only two individual models are available, substances were only considered within the AD of the consensus model if covered by both individual ones.

The model is not designed to provide out-of-domain predictions; performance metrics were only based on predictions within the AD. Sensitivity was low across all models, with the Nearest Neighbour model not being able to identify any organosilicon mutagens. The prediction reports revealed that the algorithm of this model consistently identified (non-mutagenic) organosilicon substances from the training set as nearest neighbours while ignoring the presence of organofunctional groups that were identified as SA by other models.

3.2.3. *VEGA models*

The QMRFs of the VEGA models do not foresee applicability to

Table 2
Performance metrics of assessed in silico tools for datasets A and B.

Dataset A

	nf*	TP	TN	FP	FN	OOD	% Cov	% Sens	% Spec	% BA	% Acc	% PP	% NP
DNA alerts for AMES, CA and MNT by OASIS*	0	6	106	0	9			40	100	70	93	100	92
DNA binding by OASIS*	0	6	104	2	9			40	98	69	91	75	92
DNA binding by OECD*	0	11	93	13	4			73	88	81	86	46	96
in vitro mutagenicity (Ames test) alerts by ISS*	0	11	100	6	4			73	94	84	92	65	96
TEST Consensus	0	1	77	1	9	33	73	10	99	54	89	50	90
TEST Hierarchical Clustering	0	4	78	8	8	23	81	33	91	62	84	33	91
TEST Nearest Neighbour	0	0	93	1	12	15	88	0	99	49	88	0	89
Vega Consensus**	0	10	101	5	5	121	5	67	95	81	92	67	95
Vega Caesar**	0	10	92	14	5	121	0	67	87	77	84	42	95
Vega ISS**	0	11	102	4	4	121	0	73	96	85	93	73	96
Vega SarPy/IRFMN**	0	10	103	3	5	121	0	67	97	82	93	77	95
Vega KNN/Read-Across**	3	5	88	15	10	91	22	33	85	59	79	25	90
Derek Nexus	0	11	103	3	2	0	100	73	97	85	94	79	96
Sarah Nexus	27	5	67	2	5	15	65	50	97	74	91	71	93

Dataset B

	nf*	TP	TN	FP	FN	OOD	% Cov	% Sens	% Spec	% BA	% Acc	% PP	% NP
DNA alerts for AMES, CA and MNT by OASIS*	0	5	106	0	4			56	100	78	97	100	96
DNA binding by OASIS*	0	5	104	2	4			56	98	77	95	71	96
DNA binding by OECD*	0	9	93	13	0			100	88	94	89	41	100
in vitro mutagenicity (Ames test) alerts by ISS*	0	9	100	6	0			100	94	97	95	60	100
TEST Consensus	0	1	77	1	6	30	74	14	99	57	92	50	93
TEST Hierarchical Clustering	0	3	78	8	5	21	82	38	91	64	86	27	94
TEST Nearest Neighbour	0	0	93	1	7	14	88	0	99	49	92	0	93
Vega Consensus**	0	8	101	5	1	109	10	89	95	92	95	62	99
Vega Caesar**	0	8	92	14	1	115	0	89	87	88	87	36	99
Vega ISS**	0	9	102	4	0	115	0	100	96	98	97	69	100
Vega SarPy/IRFMN**	0	8	106	0	1	115	0	89	100	94	99	100	99
Vega KNN/Read-Across**	3	3	88	15	6	85	27	33	85	59	81	17	94
Derek Nexus	0	9	103	3	0	0	100	100	97	99	97	75	100
Sarah Nexus	26	4	67	2	2	14	67	67	97	82	95	67	97

<65 65≤x<85 ≥85

nf: prediction technically not feasible (VEGA KNN/Read-Across) or equivocal (Sarah Nexus), TP/TN: true positive/true negative, FP/FN: false positive/false negative, OOD: out-of-domain predictions, Cov: Coverage, Sens: Sensitivity, Spec: Specificity, BA: Balanced Accuracy, Acc: Accuracy, PP: positive predictivity, NP: negative predictivity. Colour codings indicate low (red), moderate (yellow) or high (green) predictive performance. Sarah Nexus: before inclusion of Si-Ames dataset. Performance assessments of TEST, Derek Nexus and Sarah Nexus only considered in-domain predictions.
*For TB profilers, no applicability domain assessment is provided and performance metrics are based on all predictions.
** For the VEGA models, performance metrics are based on all predictions (including OOD).

organosilicon chemistry (Gamba and Raitano, 2022; Gamba and Benfenati, 2022; Benfenati and Raitano, 2022; Benfenati et al., 2022; Golbamaki and Benfenati, 2022). This was confirmed by 0% coverage for CAESAR, ISS and SarPy/IRFMN and 22%/27% (dataset A/B) for the KNN/Read-Across Model, which identified six substances as part of its training set. Consequently, those six substances were the only predictions that were assigned high reliability by the consensus model. In order to nonetheless understand the predictive performance of the VEGA models for this specific dataset, the performance metrics listed in Table 2 include predictions of low reliability (here defined as out-of-domain, see section 2.3).

Except for the KNN/Read-Across model (33%), VEGA models nevertheless showed high sensitivity for the prediction of this out-of-domain dataset compared to the other *in silico* tools, reaching 89-100% in dataset B.

3.2.4. Derek Nexus and Sarah Nexus

Derek Nexus was among the models showing the highest sensitivity, reaching 100% for dataset B. Only 2 out of the 103 negative predictions reported unclassified features, which were considered in-domain and included in the performance assessment.

Prior to inclusion of the Si-Ames dataset, Sarah Nexus model comprised a training set of 12,199 chemicals, including 64 silicon-based substances. The coverage of Sarah Nexus was 65/67% (dataset A/B), with the remaining chemicals being equivocal or outside domain, respectively. Sarah Nexus has settings to adjust the sensitivity of the model, reducing the potential for false negative results at the expense of lowering coverage. The default settings used in this analysis cause negative predictions with low confidence (0% - 16%) to be called equivocal; adjusting these settings can increase coverage. 25 of the chemicals in the dataset were also present in the Sarah training set, with all except one having concordant Ames results between Sarah and the Si-Ames dataset.

3.3. Prediction of Ames-positive organosilicon substances

Across the fourteen *in silico* applications, all Ames positive substances from dataset B were correctly identified by five or more models (Fig. 2), both expert systems and statistical QSAR models. VEGA ISS, Derek Nexus and the two TB profilers "DNA binding by OECD" and "*in vitro* mutagenicity (Ames test) alerts by ISS" identified all Ames-positive substances correctly, flagging the following organofunctional groups in the molecules as SA for Ames mutagenicity: alkyl halide, isocyanate, epoxide, nitro-aromatic. No silicon-specific SA was identified.

Also, in the more conservative dataset A, most Ames positive substances were predicted by five or more models (Fig. 2). However, two Ames-positive substances (Diacetoxydi-tert-butoxysilane (CAS RN 13170-23-5) and tetramethylsilane (CAS RN 75-76-3)) were not predicted as mutagenic by any model. Two substances were predicted as positive for bacterial gene mutation by one model only: Dimethoxy (diphenyl)silane (CAS 6843-66-9) was identified by the VEGA KNN/Read-Across model as Ames-positive (out-of-domain, low reliability). Dichloro (methyl) (vinyl)silane (CAS RN 124-70-9) was identified as positive by Sarah Nexus.

4. Discussion

We assessed the predictive performance of fourteen (Q)SAR applications for a total of 121 organosilicon substances with Ames mutagenicity data, to address the following question: Can *in silico* tools be used to predict Ames mutagenicity for organosilicon chemistry?

4.1. The overall predictive performance is comparable to classic organic chemistry

Sensitivity, the ability to correctly identify true positives, *i.e.* Ames mutagens, is the most important performance metric to ensure safety and gain sufficient confidence for regulatory uses. With some exceptions (TEST models, VEGA KNN/Read-Across and the two OASIS profilers), sensitivity values of dataset A are within the range of benchmark assessments for eight of the fourteen tools. Even so, the dataset B sensitivity outperforms benchmark values for several models (Fig. 3) (Kemény et al., 2025; Aljallal et al., 2024; Martin, 2020; Honma et al., 2019; Fischer et al., 2024; Furuhama et al., 2023; Golbamaki and Benfenati, 2022; Benfenati and Raitano, 2022; Ferrari et al., 2013; Gamba and Benfenati, 2022; Benfenati et al., 2022; Gamba A and Raitano G, 2022). The wide range between performance metrics for the same models can be explained by version differences as well as different characteristics of the assessed datasets. Fischer et al. (2024), for example, focused on pesticides and biocides.

The intrinsic experimental variability of the Ames test averages 87% (Piegorsch and Zeiger, 1991); reproducibility is likely to be lower in case of weak responses (Cronin et al., 2025; Zeiger et al., 2024). For eight out of the fourteen assessed applications, the balanced accuracies of (Q)SAR predictions against organosilicon chemistry of dataset B can be considered to be in the same order of magnitude as the experimental error of the Ames test ($\geq$82%).

Mitigating factors can modulate mutagenic potential, *i.e.* by steric, electronic or detoxifying effects (OECD, 2010). There was no indication that the presence of the silicon atom or silicon-based fragments resulted in an increase of mutagenic potential. Within a molecule, the silicon atom is more electropositive than carbon. While, in theory, this could result in an increase of the electronegativity and resulting DNA-reactivity of a neighbouring organofunctional group, it is noteworthy that in nearly all Ames-positive substances in our dataset, the identified DNA-reactive organofunctional group does not directly neighbour the silicon atom but terminates a propyl group that is covalently bound to the silicon atom. The electropositive effect is known to weaken with increasing length of the alkyl chain and is negligible if the functionality is in gamma position; the propyl chain, therefore, places the functional group exactly at or beyond this gamma position, effectively shielding it from the silicon atom's electronic influence (Smith, 2020). The connecting propyl group is also hypothesised to minimise steric effects.

The assessed organosilicon datasets are comparably small and imbalanced in comparison with similar assessments described in Fig. 3, and small changes in the reference data may drive significant changes in the performance metrics. This specifically applies to the low number of Ames-positive substances (which was limited by the availability of public data) in the context of sensitivity assessment of the models. However, the high quality of the data due to the aforementioned additional peer review and curation steps, as well as the consistency of the outcome across multiple models increase the reliability and relevance of the observed performance metrics.

4.2. No silicon-based structural alert was identified

Overall, the assessment confirmed that most *in silico* tools can predict bacterial reverse mutations for organosilicon chemistry. In this assessment, rule-based expert SAR systems performed better than statistical models. However, expert systems can only recognise previously established SA, while statistical models can identify novel SA from the training set. Since, except for recent versions of Sarah Nexus (v.3.3.0 onwards), organosilicon compounds are not well represented in the training set of statistical models, potential silicon-based SA might remain undetected.

For each Ames-positive substance in dataset B, an established SA for mutagenicity based on known organofunctional but not silicon-based chemistry could be identified (Fig. 2). The four substances in dataset A without or with only one corresponding positive prediction (diacetoxydi-tert-butoxysilane (CAS RN 13170-23-5), tetramethylsilane (CAS RN 75-76-3), dimethoxy (diphenyl)silane (CAS RN 6843-66-9) and dichloro (methyl) (vinyl)silane (CAS RN 124-70-9)) need a case-by-case assessment to assess their true mutagenic potential:

OECD Test Guideline 471 considers a concentration-dependent increase of revertant colonies of one out of the five strains as an indication for a positive response; however, the guideline text highlights the need for an assessment of the biological relevance. While most assays show clear binary results, the interpretation of weak responses remains challenging (Omori and Hayashi, 2024; Levy et al., 2019).

For all four substances, one out of two or more Ames tests was considered positive due to an increase of revertant mutants with TA1535 but not TA100 or any other tester strain (Supplementary Material, S1). TA100, derived from TA1535 and addressing base-pair substitutions at the same target site, typically responds similarly to TA1535 and is overall considered more sensitive (Prival and Zeiger, 1998; Williams et al., 2019). Exceptions have been reported; about 2% of mutagens cause a selective response in TA1535 but not TA100 (Prival and Zeiger, 1998; Williams et al., 2019). However, this finding is usually supported by positive findings in concomitant *in vitro* genotoxicity assays (Williams et al., 2019), which was not the case for these four substances.

While OECD TG 471 (OECD, 2020) does not provide clear criteria to determine a positive test outcome, the duplication of colonies (two-fold rule) is often considered an indicator for mutagenic activity. With the exception of dichloro(methyl)(vinyl)silane (up to four-fold, after metabolic activation), the number of TA1535 revertants barely doubled in the four positive assays. Due to a low number of background colonies of TA1535, the application of the two-fold rule may produce false positive outcomes for this tester strain (Cariello and Piegorsch, 1996; Levy et al., 2019).

Tetramethylsilane specifically is a gas, which adds technical challenges to the conduct of an Ames test. Exposure to the test item was not analytically verified and the potential for exceedance of the limit dose in the positive study cannot be excluded. Tetramethylsilane is globally used as an internal standard in Nuclear Magnetic Resonance (NMR) spectroscopy (IUPAC, 2001) and well known to be chemically inert.

The sole positive prediction for dimethoxy(diphenyl)silane was found with the VEGA KNN/Read-across model; however, the prediction output was linked to low reliability. The overall low performance of this model for organosilicon chemistry is summarised above. In addition to the low sensitivity, the model also provided a positive predictivity of only 25% and 17% for datasets A and B, respectively, *i.e.*, only one in four or six positive predictions are matched with a positive test result (Table 2).

The positive prediction of dichloro(methyl)(vinyl)silane (CAS RN 124-70-9) was provided by Sarah Nexus v3.2.1 and linked to the presence of the positive Ames test on this specific substance in its training set. The study is summarised in public literature (Seifried et al., 2006; Hansen et al., 2009) and the ISSSTY database (Benigni et al., 2013). The consideration of the Si-Ames dataset in v.3.3.0 resulted in a re-evaluation of the overall substance call and lead to the exclusion of this substance from the training set; the test results are still accessible as "additional data". The current version of Sarah Nexus considers this substance as outside domain. To verify suspected mutagenic potential for this substance, a Compliance Check under REACH (EU, 2006) is foreseen (ECHA, 2024).

Furthermore, the four silanes do not share any common structural feature that might indicate the existence of an unidentified SA. Therefore, overall, there is no indication that Ames-positive organosilicon substances are routinely missed in existing *in silico* models. In the contrary, for all four substances, the overall outcome of the conducted *in silico* predictions suggests the positive test outcomes of the legacy test data to be artifacts.

Lastly, the updated model of Sarah Nexus (after inclusion of the Si-

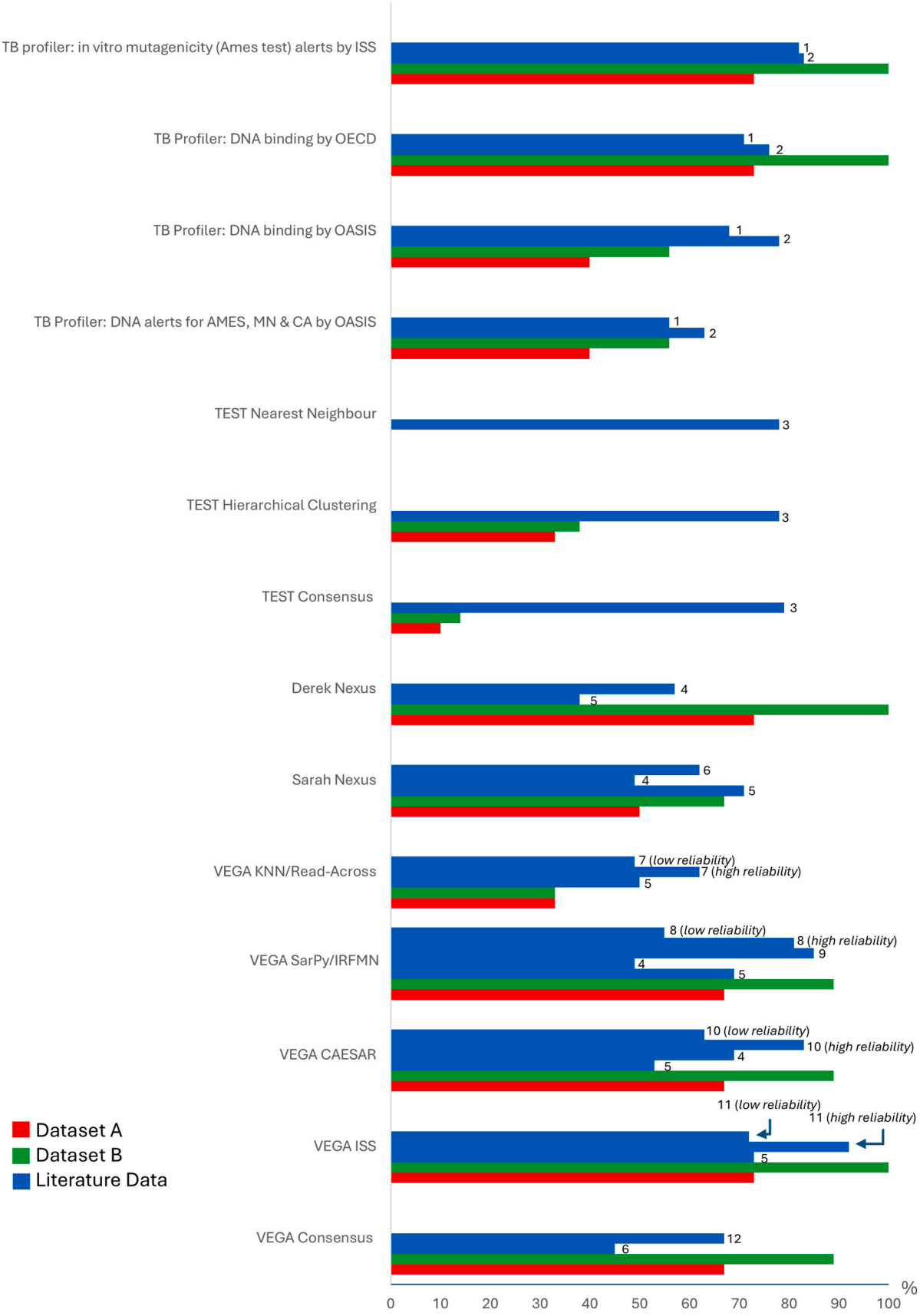

Fig. 3. Sensitivity comparisons across benchmark assessments from external validation and public literature (1: Kemény et al. (2025), 2: Aljallal et al. (2024), 3: Martin (2020), 4: Honma et al. (2019), 5: Fischer et al. (2024) 6: Furuhama et al. (2023), 7: Golbamaki and Benfenati (2022), 8: Benfenati and Raitano (2022), 9: Ferrari et al. (2013), 10: Gamba and Benfenati (2022), 11: Benfenati et al. (2022), 12: Gamba A and Raitano G, 2022).

Ames dataset) serves as an additional indicator of the absence of Si-specific mutagenicity potential: Sarah Nexus uses a proprietary self-organising hypothesis network (SOHN) algorithm. This methodology identifies positive or negative hypotheses based on the presence of structural features in the query which have been associated with activity or inactivity in the training set (Hanser et al., 2014). After inclusion of the Si-Ames dataset, no Si-specific positive but three silicon-based negative hypotheses were added (structures containing Si(OH)2, Si(OH)3 or SiCl2). This targeted injection of data increases the coverage of Sarah Nexus for silicon-based substances, ensuring that newer versions identify more relevant nearest neighbours to support predictions.

4.3. Model performance is driven by its algorithm

The results indicate significant performance differences between the fourteen *in silico* tools. Expert systems, and, by extension, TB profilers, show comparably high performance, often reaching the predictivity of the Ames test itself. Three of these tools (QSAR TB profilers DNA binding by OECD, *In vitro* mutagenicity (Ames test) alerts by ISS, and the VEGA Mutagenicity ISS model) are largely sourced from the Benigni-Bossa Rulebase (Benigni et al., 2008). The original version was already shown to provide high predictive performance for the Bacterial Reverse Mutation assay, reaching 85% sensitivity and 72% specificity (Benigni and Bossa, 2008) and considered the most comprehensive attempt at the time (OECD, 2010). Consequently, the prediction output and performance of the three applications is similar (albeit not identical, due to individual updates of each model).

The assessed statistical and hybrid models can be grouped into two clusters based on their underlying algorithm:

Within the target substance, the algorithms of Sarah Nexus, VEGA SarPy and VEGA CAESAR identify molecular fragments with higher probability to cause DNA mutations (Hanser et al., 2014; Ferrari et al., 2013; Ferrari and Gini, 2010). The performance of all three models was comparable to published data on organic chemistry, even if predictions were considered out-of-domain. The importance of selecting molecular descriptors with relevance to the modelled endpoint has been raised by Cronin et al. (2025). Before inclusion of the Si-Ames dataset, the training set of Sarah Nexus already contained 0.5% of silicon-based chemistry, while organosilicon compounds are not included in the current training sets of VEGA SarPy and VEGA CAESAR. It is believed that inclusion of additional organosilicon compounds may increase both coverage and performance of these models. Therefore, inclusion of the Si-Ames dataset is planned for the upcoming update of the VEGA mutagenicity models as well.

The algorithms of TEST Hierarchical Clustering, TEST Nearest Neighbour and VEGA KNN/Read-Across identify structurally similar substances in the training set and build their predictions on the structural similarity and outcome of test data of these substances. They can therefore be considered endpoint-agnostic and are used for a broad range of toxicological and environmental endpoints (Martin, 2020; Benfenati, 2023). The low sensitivity observed with organosilicon chemistry can largely be attributed to the identification of other (typically non-mutagenic) organosilicon substances from the training set as closest structural analogues. The similarity assessment is driven by the presence or absence of the silicon atom rather than by potentially DNA-reactive organofunctional groups (Martin and Harten, 2008; Flores et al., 2014). All three QSAR models include a small number of organosilicon substances in their training set. While the TEST predictions were considered in-domain, the VEGA KNN/Read-Across model largely indicated low reliability of its predictions.

4.4. Model-specific applicability domain assessments do not necessarily correlate with overall predictive performance

Both the algorithm for the similarity assessment and the definition of the borders of the AD are subject to the decision of the model developer. Multiple methodologies to define the AD exist, and there is no harmonised approach. Out-of-domain predictions are generally understood to be less reliable and to introduce a higher degree of uncertainty into toxicological assessments (EFSA, 2017; OECD, 2024). However, it is well acknowledged that in certain cases, they may provide equally reliable information as in-domain predictions (OECD, 2007; Benigni et al., 2019). Different degrees of confidence across the chemical space of a model are generally possible even if for in-domain predictions (OECD, 2007; Cronin et al., 2025).

For each (Q)SAR model, VEGA uses quantitative values to address the AD of the prediction. The predictive performance for the most similar substances is used to quantify the reliability via the ADI. The ADI of the four individual models ranges from 0 to 1. Each individual model has a specific ADI threshold to classify the reliability of the prediction. The reliability assessment of the consensus model is driven by the reliability of the individual models. Within a given model, the ADI correlates well with the predictive performance (Danieli et al., 2023). It is advised to apply expert judgement when interpreting result and reliability (Danieli et al., 2023). With the exception of the KNN/Read-Across model, the VEGA models predict organosilicon chemistry with the same accuracy as classic organics, even though the predictions are provided with low reliability. The predictions with moderate reliability in the KNN/Read-Across model – albeit insufficient in number for a quantitative performance assessment – indicate potential for improved predictivity.

The predictive performance of the TEST models, on the other hand, was much lower, even though the predicted chemicals were considered in-domain by the model. This shows that, for organosilicon chemistry, the model-driven AD assessment is not sufficient to allow a conclusion on the predictive performance of the model.

The four TB profilers are not designed to provide information on the AD; therefore, the coverage could not be calculated but was assessed indirectly, using the "Alert Performance" functionality for the alert "No alert found". The AP (identical to the NP) can be used to assess the predictive power of an alert; with high predictivity allowing similar confidence as an SAR expert system (Yordanova et al., 2019). The AP indicates similar predictivity as the assessed organosilicon data. The presence of organosilicon substances in the reference data utilised to calculate the AP adds to the confidence of the predictivity for organosilicon substances.

As a consequence, for organosilicon compounds, instead of reliance on model-specific AD assessments, a thorough assessment of the local performance (equivalent to the work presented here) is advised to assess the applicability of a model to this chemistry. The results should be utilised to inform the AD- and reliability-related assessment elements in the context of the QAF (OECD, 2024).

4.5. Model performance benefits from peer review of reference data

The overall substance calls were based on a review of the Ames tests themselves and agnostic to the nature of the test item, *i.e.* not considering test results from structurally-similar substances or known SA. Similarly, the outcome of other *in vitro* or *in vivo* genotoxicity data was not taken into account. A subsequent assessment of SA and concomitant genotoxicity data indicated that two substances excluded from dataset B may well be true (but presumably weak) Ames mutagens: Nearly all tools predicted 2-(3,4-epoxycyclohexyl)ethyltrimethoxysilane (CAS RN 3388-04-3) to be mutagenic due to the presence of an epoxy functionality; likewise, (3-chloropropyl)triethoxysilane (CAS RN 5089-70-3) was predicted to be a potential Ames mutagen in half of the tools.

For the training and validation of (Q)SAR models, removal of contradictory and doubtful results in a reference dataset is recommended to remove unnecessary noise (Honma et al., 2019; Herrmann et al., 2020). In this external validation, the curation exercise improved the overall predictive performance of the models. While a thorough data curation that includes peer review of study design and outcome clearly improves

the reliability of *in silico* models, big data assessments often do not allow this resource-intensive step, and this is not always applied.

The data review prior to inclusion into Sarah Nexus, resulting in exclusion of some substances with conflicted or equivocal data from the training set, is therefore believed to result into improved predictive performance. To date, due to the lack of positive Ames data with organosilicon substances outside the Si-Ames dataset, an external validation is not possible. The accessibility of the training set and excluded data as "additional information" addresses the need for transparency (OECD, 2007; Alves et al., 2021) and increases the trust in the model.

4.6. *In silico tools can support regulatory applications of organosilicon substances*

The successful confirmation of the applicability of some *in silico* tools to predict Ames mutagenicity for organosilicon chemistry opens the door for several regulatory uses. Under REACH, (Q)SARs can aid identifying false positive outcomes of Ames assays and add a robust line of evidence in a Weight-of-Evidence (WoE) assessment (EU, 2006; Benigni et al., 2019), to avoid unnecessary animal testing.

Previous work has shown that the Threshold of Toxicological Concern (TTC), a pragmatic approach for risk assessment of chemicals with low exposure and inadequate toxicological information, can be applied to organosilicon compounds (Schmitt et al., 2021) to identify safe levels for non-genotoxic substances. In the absence of data obtained from testing or read-across, (Q)SAR modelling is suggested as a suitable tool to discriminate between genotoxic and non-genotoxic compounds (EFSA, 2019). The confirmation that some *in silico* tools allow reliable prediction of the DNA-reactivity of organosilicon compounds now allows full access to the TTC approach for this chemistry, for example in cosmetic (SCCS, 2023), food contact materials (Serafimova et al., 2021) or medical devices (Patlewicz et al., 2022; U.S. FDA, 2023).

Lastly, organosilicon chemistry has recently become a relevant contribute to drug discovery (Panayides et al., 2024). The ICH M7 Guideline (ICH M7, 2023) allows *in silico* assessment of impurities in pharmaceutical active ingredients, requesting a combination of expert systems and statistical models. The same approach is adopted for medical devices (U.S. FDA, 2023) to assess extractables and leachables, now allowing *in silico* assessment of silicone polymers in medical use.

5. Conclusions

We assessed the utility of fourteen *in silico* tools with varying algorithms and training sets to predict the outcome of the Ames test for organosilicon substances. Eleven out of fourteen tools showed similar predictivity towards this chemistry as described elsewhere for organics. The predictive performance did not necessarily correspond to model-specific AD assessments. No silicon-related SA could be identified; in all cases, mutagenic potential was linked to the presence of an organofunctional group with known potential for DNA reactivity. Several *in silico* tools could be identified to predict the outcome of the Ames test with equal reliability as the Ames test itself and can be utilised to support hazard (WoE, read-across) and risk assessment (TTC) in various regulatory settings, ultimately reducing unnecessary animal testing. Highest predictivity was obtained with VEGA ISS, VEGA SarPy/IRFMN and Derek Nexus.

This exercise also served as a pilot project to improve the understanding of the applicability of *in silico* applications to predict organosilicon chemistry. Each model requires a robust local performance assessment with an extended reference dataset to inform its applicability to this chemistry.

Lastly, we showed the benefits of careful curation and peer review of reference data and – with Sarah Nexus - the increase of coverage and reliability of *in silico* predictions by successful inclusion of silicon-based reference data into the training set via data sharing.

Funding

The authors thank Silicones Europe (www.silicones.eu) for administrative support and funding.

CRediT authorship contribution statement

Barbara G. Schmitt: Conceptualization, Data curation, Formal analysis, Writing – original draft. **Emilio Benfenati:** Data curation, Formal analysis, Writing – review & editing. **Robert S. Foster:** Data curation, Formal analysis, Writing – review & editing. **Grace Kocks:** Data curation, Formal analysis, Writing – review & editing. **Giuseppa Raitano:** Data curation, Formal analysis, Writing – review & editing. **Rachael E. Tennant:** Data curation, Formal analysis, Writing – review & editing. **Carole Forlini:** Data curation, Writing – review & editing. **Eckart Gura:** Data curation, Writing – review & editing. **Shawn Seidel:** Data curation, Writing – review & editing. **Farah Koraichi-Emeriau:** Data curation, Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Barbara Schmitt reports financial support, administrative support, and article publishing charges were provided by Silicones Europe. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We thank the RECONSILE REACH consortium, Vitis Regulatory and Knoell for facilitating the compilation of the reference dataset as well as for their valuable input during the review process.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.yrtph.2026.106153.

Data availability

Data will be made available on request.

References

Aljallal, M.A., Chaudhry, Q., Price, N.R., 2024. Assessment of performance of the profilers provided in the OECD QSAR toolbox for category formation of chemicals. Sci. Rep. 14 (1), 18330. https://doi.org/10.1038/s41598-024-69157-1.

Alves, V.M., Auerbach, S.S., Kleinstreuer, N., Rooney, J.P., Muratov, E.N., Rusyn, I., Tropsha, A., Schmitt, C., 2021. Curated data in - trustworthy *in silico* models out: the impact of data quality on the reliability of artificial intelligence models as alternatives to animal testing. Altern. Lab. Anim. 49 (3), 73–82. https://doi.org/10.1177/02611929211029635.

Benfenati, E., 2023. In Silico Models: Theory, Guidance and Applications Within VEGAHUB. Chapter B.3. The read-across tools in VEGAHUB. Ebook. First edition. LIFE17 GIE/IT/000461.

Benfenati, E., Golbamaki, A., Vukovic, K., 2022. Mutagenicity ISS model - v. 1.0.3 – QMRF. Publicly available at: https://www.vegahub.eu/vegahub-dwn/qmrf/QMRF_MUTA_ISS.pdf. (Accessed 10 December 2024).

Benfenati, E., Raitano, G., 2022. Mutagenicity (Ames test) model (SarPy/IRFMN) (version 1.0.8) – QMRF. Publicly available at: https://www.vegahub.eu/vegahub-dwn/qmrf/QMRF_MUTA_SARpy.pdf. (Accessed 10 December 2024).

Benigni, R., Battistelli, C.L., Bossa, C., Giuliani, A., Fioravanzo, E., Bassan, A., Gatnik, M. F., Rathman, J., Yang, C., Tcheremenskaia, O., 2019. Evaluation of the applicability of existing (Q)SAR models for predicting the genotoxicity of pesticides and similarity analysis related with genotoxicity of pesticides for facilitating of grouping and read across. EFSA Supporting publication 2019. https://doi.org/10.2903/sp.efsa.2019.EN-1598. EN-1598.

Benigni, R., Battistelli, C.L., Bossa, C., Tcheremenskaia, O., Crettaz, P., 2013. New perspectives in toxicological information management, and the role of ISSTOX

databases in assessing chemical mutagenicity and carcinogenicity. Mutagenesis 28 (4), 401–409. https://doi.org/10.1093/mutage/get016.
Benigni, R., Bossa, C., Jeliazkova, N., Netzeva, T., Worth, A., 2008. The Benigni/Bossa rulebase for mutagenicity and carcinogenicity – a module of Toxtree. EUR 23241 EN.
Benigni, R., Bossa, C., 2008. Structure alerts for carcinogenicity, and the Salmonella test system: a novel insight through the chemical relational databases technology. Mutat. Res. 659, 248–261. https://doi.org/10.1016/j.mrrev.2008.05.003.
Benigni, R., Bossa, A., 2011. Mechanisms of chemical carcinogenicity and mutagenicity: a review with implications for predictive toxicology. Chem. Rev. 111, 2507–2536. https://doi.org/10.1021/cr100222q.
Cariello, N.F., Piegorsch, W.W., 1996. The Ames test: the two-fold rule revisited. Mutat. Res. 369, 23–31. https://doi.org/10.1016/S0165-1218(96)90044-0.
Cooper, J.A., Saracci, R., Cole, P., 1979. Describing the validity of carcinogen screening tests. Br. J. Cancer 39 (1), 87–89. https://doi.org/10.1038/bjc.1979.10. https://www.nature.com/articles/bjc197910.
Cronin, M.T.D., Basiri, H., Chrysochoou, G., Enoch, S.J., Firman, J.W., Spînu, N., Madden, J.C., 2025. The predictivity of QSARs for toxicity: recommendations for improving model performance. Comput. Toxicol. 33, 100338. https://doi.org/10.1016/j.comtox.2024.100338.
Danieli, A., Colombo, E., Raitano, G., Lombardo, A., Roncaglioni, A., Manganaro, A., Sommovigo, A., Carnesecchi, E., Dorne, J.C.M., Benfenati, E., 2023. The VEGA tool to check the applicability Domain gives greater confidence in the prediction of in silico models. Int. J. Mol. Sci. 24 (12), 9894. https://doi.org/10.3390/ijms24129894.
ECHA, 2016. Practical guide. How to use and report (Q)SARs. Version 3.1. ECHA-16-B-09-EN. https://doi.org/10.2823/81818.
ECHA, 2024. Assessment of regulatory needs. Group Name: Hydrocarbylhalosilanes. Version 1.0. Publicly available at: https://echa.europa.eu/documents/10162/4ab642b9-ca5d-592b-276b-5eb75ced9c4b. (Accessed 3 July 2025).
EFSA Scientific Committee, Hardy, A., Benford, D., Halldorsson, T., Jeger, M.J., Knutsen, H.K., More, S., Naegeli, H., Noteborn, H., Ockleford, C., Ricci, A., Rychen, G., Schlatter, J.R., Silano, V., Solecki, R., Turck, D., Benfenati, E., Chaudhry, Q.M., Craig, P., Frampton, G., Greiner, M., Hart, A., Hogstrand, C., Lambre, C., Luttik, R., Makowski, D., Siani, A., Wahlstroem, H., Aguilera, J., Dorne, J.-L., Fernandez Dumont, A., Hempen, M., Valtueña Martínez, S., Martino, L., Smeraldi, C., Terron, A., Georgiadis, N., Younes, M., 2017. Scientific Opinion on the guidance on the use of the weight of evidence approach in scientific assessments. EFSA J. 15 (8), 4971. https://doi.org/10.2903/j.efsa.2017.4971.
EFSA Scientific Committee, More, S.J., Bampidis, V., Benford, D., Bragard, C., Halldorsson, T.I., Hernández-Jerez, A.F., Hougaard Bennekou, S., Koutsoumanis, K. P., Machera, K., Naegeli, H., Nielsen, S.S., Schlatter, J.R., Schrenk, D., Silano, V., Turck, D., Younes, M., Gundert-Remy, U., Kass, G.E.N., Kleiner, J., Rossi, A.M., Serafimova, R., Reilly, L., Wallace, H.M., 2019. Guidance on the use of the Threshold of Toxicological Concern approach in food safety assessment. EFSA J. 17 (6), 5708. https://doi.org/10.2903/j.efsa.2019.5708.
EFSA Panel on Food Additives and Flavourings (FAF), Younes, M., Aquilina, G., Castle, L., Fowler, P., Frutos Fernandez, M.J., Fürst, P., Gundert-Remy, U., Gürtler, R., Husøy, T., Manco, M., Mennes, W., Moldeus, P., Passamonti, S., Shah, R., Waalkens-Berendsen, I., Wölfle, D., Wright, M., Benigni, R., Bolognesi, C., Boon, P., Chipman, K., De Knecht, J., Sahlin, U., Arcella, D., Barmaz, S., Carfì, M., Martino, C., Tard, A., Vianello, G., Engel, K.H., 2021. Scientific guidance for the preparation of applications on smoke flavouring primary products. EFSA J. 19 (3), e06435. https://doi.org/10.2903/j.efsa.2021.6435.
Enoch, S.J., Cronin, M.T.D., 2010. A review of the electrophilic reaction chemistry involved in covalent DNA binding. Crit. Rev. Toxicol. 40, 728–748. https://doi.org/10.3109/10408444.2010.494175.
European Union, 2006. Regulation (EC) No 1907/2006 of the European Parliament and of the Council of 18 December 2006 Concerning the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH), Establishing a European Chemicals Agency, Amending Directive 1999/45/EC and Repealing Council Regulation (EEC) No 793/93 and Commission Regulation (EC) No 1488/94 as Well as Council Directive 76/769/EEC and Commission Directives 91/155/EEC, 93/67/EEC, 93/105/EC and 2000/21/EC.
Ferrari, T., Cattaneo, D., Gini, G., Golbamaki Bakhtyari, N., Manganaro, A., Benfenati, E., 2013. Automatic knowledge extraction from chemical structures: the case of mutagenicity prediction. SAR QSAR Environ. Res. 24 (5), 365–383. https://doi.org/10.1080/1062936X.2013.773376.
Ferrari, T., Gini, G., 2010. An open source multistep model to predict mutagenicity from statistical analysis and relevant structural alerts. Chem. Cent. J. 4 (Suppl. 1), S2.
Fischer, B.C., Foil, D.H., Kadic, A., Kneuer, C., Koenig, J., Herrmann, K., 2024. Pobody's nerfect: (Q)SAR works well for predicting bacterial mutagenicity of pesticides and their metabolites, but predictions for clastogenicity in vitro have room for improvement. Comput. Toxicol. 30, 100318. https://doi.org/10.1016/j.comtox.2024.100318.
Flores, M., Manganaro, A., Nicolotti, O., Medda, R., Mangiatordi, G.F., Benfenati, E., 2014. A generalizable definition of chemical similarity for read-across. J. Cheminf. 6, 39. https://doi.org/10.1186/s13321-014-0039-1.
Furuhama, A., Kitazawa, A., Yao, J., Matos Dos Santos, C.E., Rathman, J., Yang, C., Ribeiro, J.V., Cross, K., Myatt, G., Raitano, G., Benfenati, E., Jeliazkova, N., Saiakhov, R., Chakravarti, S., Foster, R.S., Bossa, C., Battistelli, C.L., Benigni, R., Sawada, T., Wasada, H., Hashimoto, T., Wu, M., Barzilay, R., Daga, P.R., Clark, R.D., Mestres, J., Montero, A., Gregori-Puigjané, E., Petkov, P., Ivanova, H., Mekenyan, O., Matthews, S., Guan, D., Spicer, J., Lui, R., Uesawa, Y., Kurosaki, K., Matsuzaka, Y., Sasaki, S., Cronin, M.T.D., Belfield, S.J., Firman, J.W., Spînu, N., Qiu, M., Keca, J.M., Gini, G., Li, T., Tong, W., Hong, H., Liu, Z., Igarashi, Y., Yamada, H., Sugiyama, K.I., Honma, M., 2023. Evaluation of QSAR models for predicting mutagenicity: outcome of the Second Ames/QSAR international challenge project. SAR QSAR Environ. Res. 34, 983–1001. https://doi.org/10.1080/1062936X.2023.2284902.
Gamba, A., Benfenati, E., 2022. Mutagenicity (Ames test) model (CAESAR) - version 2.1.13 – QMRF. Publicly available at: https://www.vegahub.eu/vegahub-dwn/qmrf/QMRF_MutaAmes_CAESAR.pdf. (Accessed 10 December 2024).
Gamba, A., Raitano, G., 2022. Mutagenicity (Ames test) CONSENSUS model - v. 1.0.4 – QMRF. Publicly available at: https://www.vegahub.eu/vegahub-dwn/qmrf/QMRF_MUTA_CONSENSUS.pdf. (Accessed 10 December 2024).
Gissi, A., Tcheremenskaia, O., Bossa, C., Battistelli, C.L., Browne, P., 2024. The OECD (Q)SAR assessment framework: a tool for increasing regulatory uptake of computational approaches. Comput. Toxicol. 31, 100326.
Golbamaki, A., Benfenati, E., 2022. Mutagenicity (ames test) model (KNN/Read-Across) - v. 1.0.1. – QMRF. Publicly available at: https://www.vegahub.eu/vegahub-dwn/qmrf/QMRF_KNN_MUTA.pdf. (Accessed 10 December 2024).
Hansen, K., Mika, S., Schroeter, T., Sutter, A., ter Laak, A., Steger-Hartmann, T., Heinrich, N., Müller, K.R., 2009. Benchmark data set for in silico prediction of ames mutagenicity. Journal of Chemical Information and Modelling 49, 2077–2081. https://doi.org/10.1021/ci900161g.
Hanser, T., Barber, C., Rosser, E., Vesser, J.D., Webb, S.J., Werner, S., 2014. Self organising hypothesis networks: a new approach for representing and structuring SAR knowledge. J. Cheminf. 6, 21. https://doi.org/10.1186/1758-2946-6-21.
Herrmann, K., Holzwarth, A., Rime, S., Fischer, B.C., Kneuer, C., 2020. (Q)SAR tools for the prediction of mutagenic properties: are they ready for application in pesticide regulation? Pest Manag. Sci. 76 (10), 3316–3325. https://doi.org/10.1002/ps.5828.
Honma, M., Kitazawa, A., Cayley, A., Williams, R.V., Barber, C., Hansen, T., Saiakhov, R., Chakravarti, S., Myatt, G.J., Cross, K.P., Benfenati, E., Raitano, G., Mekenyan, O., Petkov, P., Bossa, C., Benigni, R., Battistelli, C.L., Giuliani, A., Tcheremenskaia, O., DeMeo, C., Norinder, U., Koga, H., Jose, C., Jeliazkova, N., Kochev, N., Paskaleva, V., Yang, C., Daga, P.R., Clark, R.D., Rathman, J., 2019. Improvement of quantitative structure–activity relationship (QSAR) tools for predicting ames mutagenicity: outcomes of the Ames/QSAR international challenge project. Mutagenesis 34, 3–16. https://doi.org/10.1093/mutage/gey031.
International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH) M7, 2023. ICH harmonised guideline M7(R2). Assessment and control of DNA reactive (mutagenic) impurities in pharmaceuticals to limit potential carcinogenic risk. Publicly available at: https://database.ich.org/sites/default/files/ICH_M7%28R2%29_Guideline_Step4_2023_0216_0.pdf.
ISS, 2025. ISSSTY dataset. Publicly available at: https://www.iss.it/en/isstox.
International Union of Pure and Applied Chemistry (IUPAC). Physical chemistry division COmmision on molecular structure and spectroscopy 2001. NMR nomenclature, nuclear spin properties and conventions for chemical shifts (IUPAC recommendations 2001). Pure Appl. Chem. 73 (11): 1795–1818. DOI: 10.1002/mrc.1042.
Kemény, M., North, C.M., Kluxen, F.M., Frericks, M., Vukelic, D., Cao, S., Saiakhov, R., Girireddy, M., 2025. Validation of OECD QSAR toolbox profilers for genotoxicity assessment of pesticides using the MultiCase genotoxicity database. Comput. Toxicol. 34, 100356. https://doi.org/10.1016/j.comtox.2025.100356.
Klimisch, H.J., Andreae, M., Tillmann, U., 1997. A systematic approach for evaluating the quality of experimental toxicological and ecotoxicological data. Regul. Toxicol. Pharmacol. 25, 1–5. https://doi.org/10.1006/rtph.1996.1076.
Levy, D.D., Zeiger, E., Escobar, P.A., Hakura, A., van der Leede, B.J., Kato, M., Moore, M. M., Sugiyama, K., 2019. Recommended criteria for the evaluation of bacterial mutagenicity data (Ames test). Mutat. Res. Genet. Toxicol. Environ. Mutagen 848, 403074. https://doi.org/10.1016/j.mrgentox.2019.07.004.
Lhasa, 2026. Lhasa. https://www.lhasalimited.org.
Martin, T.M., 2020. User's Guide for T. E. S. T. (Toxicity Estimation Software Tool) Version 5.1. U.S. Environmental Protection Agency. Publicly available at: https://www.epa.gov/sites/default/files/2016-05/documents/600r16058.pdf.
Martin, T.M., Harten, P., 2008. A hierarchical clustering method for the estimation of toxicity. Toxicol. Mech. Methods 18 (2), 251–266. https://doi.org/10.1080/15376510701857353.
OECD, 1998. OECD principles on good laboratory practice. OECD Series on Principles of Good Laboratory Practice and Compliance Monitoring No. 1. OECD Publishing Paris. https://doi.org/10.1787/9789264078536-en.
OECD Environment Directorate, 2004. Joint meeting of the chemicals committee and the working party on chemicals, pesticides and biotechnology, the report from the expert group on (quantitative) structure-activity relationships on the principles for the validation of (Q)SARs. Series on Testing and Assessment No. 49. ENV/JM/MONO, p. 24, 2004.
OECD Environment Directorate, 2007. Joint meeting of the chemicals committee and the working party on chemicals, pesticides and biotechnology, guidance document on the validation of (quantitative) structure-activity relationships [(Q)SAR] models. Series on Testing and Assessment No. 69. ENV/JM/MONO 2, 2007.
OECD Environment Directorate, 2010. Joint meeting of the chemicals committee and the working party on chemicals, pesticides and biotechnology, report of the expert consultation on scientific and regulatory evaluation of organic chemistry mechanism-based structural alerts for the identification of DNA binding chemicals. Series on Testing and Assessment No. 120 Part 1. ENV/JM/MONO(2010)8/PART1.
OECD, 2020. Guidelines for the Testing of Chemicals. Test Guideline No. 471: Bacterial Reverse Mutation Test. Adopted: 21 July 1997, corrected. (Accessed 26 June 2020).
OECD, 2024. (Q)SAR assessment framework: guidance for the regulatory assessment of (quantitative) structure activity relationship models and predictions. OECD Series on Testing and Assessment No. 405, second ed. OECD Publishing, Paris.
Omori, T., Hayashi, M., 2024. The assessment and communication of genotoxicity test results: moving beyond binary. Mutation Research - Genetic Toxicology and

Environmental Mutagenesis 893, 503722. https://doi.org/10.1016/j.mrgentox.2023.503722.
Panayides, J.-L., Riley, D.L., Hasenmaile, F., van Otterlo, W.A.L., 2024. The role of silicon in drug discovery: a review. RSC Med. Chem. 15, 3286–3344. https://doi.org/10.1039/D4MD00169A.
Patlewicz, G., Nelms, M., Rua, D., 2022. Evaluating the utility of the threshold of toxicological concern (TTC) and its exclusions in the biocompatibility assessment of extractable chemical substances from medical devices. Comput. Toxicol. 24, 1–11. https://doi.org/10.1016/j.comtox.2022.100246.
Piegorsch, W.W., Zeiger, E., 1991. Measuring intra-assay agreement for the ames salmonella assay. In: Hothorn, L. (Ed.), Lecture Notes in Medical Informatics, vol. 43. Springer, Heidelberg, Germany, pp. 35–41.
Prival, M.J., Zeiger, E., 1998. Chemicals mutagenic in Salmonella typhimurium strain TA1535 but not in TA100. Mutation research. Genetic Toxicology and Environmental Mutagenesis 412, 251–260. https://doi.org/10.1016/S1383-5718(97)00196-4.
Scientific Committee on Consumer Safety (SCCS), 2023. The SCCS Notes of Guidance for the Testing of Cosmetic Ingredients and Their Safety Evaluation 12th Revision. SCCS/1647/22 Corrigendum 2.
Schmitt, B.G., Jensen, E., Laufersweiler, M.C., Rose, J.L., 2021. Threshold of toxicological concern: extending the chemical space by inclusion of a highly curated dataset for organosilicon compounds. Regul. Toxicol. Pharmacol. 127, 105074. https://doi.org/10.1016/j.yrtph.2021.105074.
Seifried, H.E., Seifried, R.M., Clarke, J.J., Junghans, T.B., San, R.H.C., 2006. A compilation of two decades of mutagenicity test results with the ames *Salmonella typhimurium* and L5178Y mouse lymphoma cell mutation assays. Chem. Res. Toxicol. 19 (5), 627–644. https://doi.org/10.1021/tx0503552.
Serafimova, R., Coja, T., Kass, G.E.N., 2021. Application of the threshold of toxicological concern (TTC) in food safety: challenges and opportunities. Front Toxicol 3, 655951. https://doi.org/10.3389/ftox.2021.655951.
Smith, M.B., 2020. March's advanced organic chemistry - reactions, mechanisms, and structure. Chapter 1.I: Inductive and Field Effects, eighth ed. John Wiley & Sons, p. 21. ISBN 978-1-119-37180-9.
Toxtree, 2025. Toxtree. https://toxtree.sourceforge.net/index.html.
U.S. Food and Drug Administration (US FDA), 2023. Use of international standard ISO 10993-1, "Biological evaluation of medical devices - part 1: evaluation and testing within a risk management process" guidance for industry and food and drug administration staff. Publicly available at: https://www.fda.gov/media/142959/download. (Accessed 21 March 2026).
VEGAHUB, 2026. Vegahub. https://www.vegahub.eu/portfolio-item/vega-qsar/.
Williams, R.V., Amberg, A., Brigo, A., Coquin, L., Giddings, A., Glowienke, S., Greene, N., Jolly, R., Kemper, R., O'Leary-Steele, C., Parenty, A., Spirkl, H.P., Stalford, S.A., Weiner, S.K., Wichard, J., 2016. It's difficult, but important, to make negative predictions. Regul. Toxicol. Pharmacol. 76, 79–86. https://doi.org/10.1016/j.yrtph.2016.01.008.
Williams, R.V., DeMarini, D.M., Stankowski, L.F., Escobar, P.A., Zeiger, E., Howe, H., Elespuru, R., Cross, K.P., 2019. Are all bacterial strains required by OECD mutagenicity test guideline TG471 needed? Mutation research. Genetic Toxicology and Environmental Mutagenesis 848, 503081.
Yordanova, D., Schultz, T.W., Kuseva, C., Ivanova, H., Pavlov, T., Chankov, G., Karakolev, Y., Gissi, A., Sobanski, T., Mekenyan, O.G., 2019. Alert performance: a new functionality in the OECD QSAR toolbox. Comput. Toxicol. 10, 26–37. https://doi.org/10.1016/j.comtox.2018.12.003.
Zeiger, E., Mitchell, C.A., Pfuhler, S., Liao, Y., Witt, K.L., 2024. Within-laboratory reproducibility of ames test results: are repeat tests necessary? Environ. Mol. Mutagen. 65, 116–120. https://doi.org/10.1002/em.22597.