



# Release Notes

---

**Lhasa Knowledge Suite - Nexus 2.5**

Leaders in the development of expert chemoinformatic systems and trusted curators of proprietary data.

## Statement of Intended Purpose

The computer programs in the Lhasa Knowledge Suite are intended to be used within a structured decision support system as part of the user's overall risk strategy.

## Limited Warranty

Lhasa Limited makes no warranties, either expressed or implied, regarding the software described in this document or the online help, its merchantability, or its fitness for any particular purpose. In no event will Lhasa Limited be liable for any special, consequential, indirect or similar damages including any loss of profits or lost data arising out of the use of the software or data described in this document.

## © Copyright, Lhasa Limited, 2022. All rights reserved.

The Lhasa Knowledge Suite - Nexus software, the content of reports generated by the use of that software, and the technical documentation relating to that software are proprietary to Lhasa Limited, and are protected by copyrights, database rights and similar intellectual property rights in jurisdictions all over the world.

The use of Lhasa Limited software, and the copying, distribution and other exploitation of the content of reports generated by the use of Lhasa Limited software and associated technical documentation, requires a licence from Lhasa.

Any unlicensed use of those assets will constitute an infringement of Lhasa's copyrights and/or database rights and/or other intellectual property rights in those assets.

Lhasa will take enforcement action in respect of any such intellectual property right infringements.

## Trademarks

Lhasa, Derek, Derek Nexus, Meteor, Meteor Nexus, Vitic, Vitic Nexus and Sarah Nexus are registered trademarks of Lhasa Limited. Microsoft is a registered trademark and Windows is a trademark of the Microsoft Corporation. Other product names mentioned in this document may be trademarks or registered trademarks of their respective companies and are hereby acknowledged.

CAS Registry Numbers® are the intellectual property of the American Chemical Society; and are used by Lhasa Limited with the express permission of CAS. CAS Registry Numbers® have not been verified by CAS and may be inaccurate. Expert data scientists at Lhasa Limited cross reference CAS Registry Numbers® against multiple sources to achieve a high level of accuracy.

## Acknowledgements

Lhasa Limited acknowledges the contributions to the following programs in the Lhasa Knowledge Suite:

For Derek Nexus:

- Members of the Collaborators Group
- Imperial Cancer Research Fund
- Judson Consulting Service
- Logic Programming Associates Limited
- City University (during the StAR project)
- Schering Agrochemicals Limited
- Harvard University
- BABEL developers

For Meteor Nexus:

- Meteor Steering Committee and User Group
- BABEL developers
- Patrick Rydberg, et al, University of Copenhagen (for SMARTCyp)

For Vitic Nexus:

- ITIC SAR Database project

For Sarah Nexus:

- Members who worked closely with us during the development and initial testing phases

## Contact Details

Lhasa Limited  
Granary Wharf House  
2 Canal Wharf  
LEEDS  
LS11 5PS  
United Kingdom

Reception ☎: +44 (0)113 394 6020  
Applied Sciences ☎: +44 (0)113 394 6030  
General ✉: [info@lhasalimited.org](mailto:info@lhasalimited.org)  
Applied Sciences ✉: [hello@lhasalimited.org](mailto:hello@lhasalimited.org)  
Website: [www.lhasalimited.org](http://www.lhasalimited.org)

---

Lhasa Limited is a not-for-profit organisation.

Registered charity number 290866  
(Registered in England and Wales)

## Contents

What's New in the Derek 2022.1.0 Knowledge Base .....	5
New and Modified Endpoints .....	5
New and Modified Rules .....	5
New Species Dependent Variable .....	6
Other Knowledge Base Modifications .....	6
Disabled Alerts .....	6
Disabled Rules .....	6
Disabled Examples .....	6
Deleted Examples .....	6
Knowledge Base Status .....	7
Extrapolated Endpoints .....	7
Validation Comments .....	8
Data Set References for Validation Comments .....	9
Data Set Summary .....	11
What's New in Sarah Nexus 3.2 .....	13
Sarah Model .....	13
Structure Standardisation .....	17
What's New in ICH M7 .....	21
ICH M7 Expert Review (Derek 6.2 and Sarah 3.2) .....	21
ICH M7 Classification (Derek 6.2 and Sarah 3.2) .....	25
Ames Dataset .....	25
ICH M7 Class Assignment .....	25
Permissible Daily Exposure Comments .....	26

# What's New in the Derek 2022.1.0 Knowledge Base

## New and Modified Endpoints

### **New endpoint:** Skin Sensitisation HPC

This endpoint contains alerts describing a skin sensitisation High Potency Category (HPC), consisting of a set of structural features within a mechanistic domain which are likely to be associated with extreme skin sensitisation potential in the Local Lymph Node Assay. These HPC rules were first described in the context of the Dermal Sensitisation Threshold (DST),<sup>1</sup> which is a Threshold of Toxicological Concern for skin sensitisation.<sup>2,3</sup> The presence or absence of alerts for skin sensitisation HPC provides information about which of the DSTs is the most appropriate to use.<sup>4</sup>

### **Modified endpoint:** Irritation (of the Skin)

Endpoint renamed to Skin Irritation/Corrosion.

## New and Modified Rules

Rule 1607 implemented: If a chemical is known to give a corrosive response in a skin irritation study in the rabbit then it is considered certain that the chemical will cause skin irritation/corrosion in rabbits, probable in mammals other than the rabbit and impossible in bacteria.

Rule 1609 implemented: If a chemical is known to give an irritant response in a skin irritation study in rabbits then it is considered certain that the chemical will cause skin irritation/corrosion in rabbits, probable in mammals other than the rabbit and impossible in bacteria.

Rule 1611 implemented: If a chemical is known to be corrosive in a reconstructed human epidermis test method in humans then it is considered certain that the chemical will cause skin irritation/corrosion in humans, probable in mammals and impossible in bacteria.

Rule 1615 implemented: If a chemical is known reconstructed human epidermis test irritant in human then it is considered certain that the chemical will cause skin irritation/corrosion in human, probable in mammals and impossible in bacteria.

Rule 1616 implemented: If the chemical has a LogP value greater than 0.6 then in mammals the variable "Species dependent variable 43" is plausible.

Rule 1619 implemented: If a chemical is known to be corrosive in a membrane barrier test method in humans then it is considered certain that the chemical will cause skin irritation/corrosion in humans, probable in mammals and impossible in bacteria.

Rule 1632 implemented: If a chemical is known to be irritant in human patch test studies then it is considered certain that the chemical will cause skin irritation/corrosion in humans, probable in mammals and impossible in bacteria.

Rule 1645 implemented: If a chemical is known to give an irritant response in a skin irritation study in guinea pigs then it is considered certain that the chemical will cause skin irritation/corrosion in guinea pigs, probable in mammals other than the guinea pig and impossible in bacteria.

---

1. Roberts et al, Regul Toxicol Pharmacol (2015) 72, 683-93.

2. Safford et al, Regul Toxicol and Pharmacol (2011) 60, 218-224.

3. Safford et al, Regul Toxicol Pharmacol (2015), 72, 694-701.

4. Chilton et al, Regul Toxicol Pharmacol (2022), submitted.

## New Species Dependent Variable

Species dependent variable 43: Plausible in mammals if the chemical has a LogP value greater than 0.6. Impossible in bacteria.

Species dependent variable 44: Plausible in mammals if the chemical has a LogP value greater than 2.3. Impossible in bacteria.

Species dependent variable 45: Plausible in mammals if the chemical has a LogP value lower than 7.1. Impossible in bacteria.

Species dependent variable 46: Plausible in mouse. Doubtful in mammals other than the mouse. Impossible in bacteria.

Species dependent variable 47: Plausible in mammals if the chemical has a LogP value greater than 1. Impossible in bacteria.

Species dependent variable 48: Plausible in mammals if the chemical has a LogP value greater than -0.5. Impossible in bacteria.

Species dependent variable 49: Plausible in mammals if a chemical has a LogP value greater than -1. Impossible in bacteria.

Species dependent variable 50: Plausible in mammals if a chemical has a LogP value greater than 2.5. Impossible in bacteria.

Species dependent variable 51: Plausible in mammals if the chemical has a LogP value greater than 0.5. Impossible in bacteria.

Species dependent variable 52: Plausible in mammals if a chemical has a LogP value greater than 1.5. Impossible in bacteria.

Species dependent variable 53: Plausible in mammals if the chemical has a LogP value greater than 3.5. Impossible in bacteria.

## Other Knowledge Base Modifications

No changes from 2020.1.0.

### Disabled Alerts

No changes from 2020.1.0.

### Disabled Rules

No changes from 2020.1.0.

### Disabled Examples

No changes from 2020.1.0.

### Deleted Examples

Five examples have been deleted.

Example Name	CAS	Alert Number	Alert Name
4-nitrophenol	100-02-7	329	Aromatic nitro compound
2-nitrophenol	88-75-5	329	Aromatic nitro compound

Example Name	CAS	Alert Number	Alert Name
ropinirole	91374-21-9	352	Aromatic amine or amide
3,4-dichloroaniline	95-76-1	352	Aromatic amine or amide
(5R)-2,3-dimethyl-5-isopropenyl-2-cyclohexene-1-one	85710-65-2	480	alpha,beta-Unsaturated ketone or precursor

## Knowledge Base Status

There are 938 alerts in the Derek 2022.1.0 Knowledge Base. The following table shows the number of enabled alerts for top level parent endpoints.

Endpoint	Number of Alerts
Carcinogenicity (ALL)	79
Genotoxicity (ALL) including	231
Chromosome damage	105
Mutagenicity	154
Irritation (ALL)	67
Miscellaneous endpoints (ALL)	116
Neurotoxicity (ALL)	10
Organ toxicity (ALL) including	241
Hepatotoxicity	76
Reproductive toxicity (ALL)	61
Respiratory sensitisation (ALL)	13
Skin sensitisation (ALL)	131

## Extrapolated Endpoints

Endpoint	Extrapolated Endpoint
5alpha-Reductase inhibition	Teratogenicity
alpha-2-mu-Globulin nephropathy	Carcinogenicity
Androgen receptor modulation	Teratogenicity
Bladder urothelial hyperplasia	Carcinogenicity
Glucocorticoid receptor agonism	Teratogenicity
Irritation (of the gastrointestinal tract)	Carcinogenicity

Endpoint	Extrapolated Endpoint
Oestrogen receptor modulation	Teratogenicity
Oestrogenicity	Carcinogenicity
Peroxisome proliferation	Carcinogenicity
Thyroid toxicity	Carcinogenicity

## Validation Comments

Validation comments have been updated for the following endpoints:

### Chromosome Damage In Vitro

3 data sets used previously.

- Kirkland et al (CGX)
- US Food and Drug Administration Center for Food Safety and Applied Nutrition (FDA CFSAN)
- Sofuni Database

### Chromosome Damage In Vivo

3 data sets used previously.

- US Food and Drug Administration Center for Food Safety and Applied Nutrition (FDA CFSAN) – *in vivo* chromosome aberration test data
- US Food and Drug Administration Center for Food Safety and Applied Nutrition (FDA CFSAN) – *in vivo* micronucleus test data
- MMS (Mammalian Mutagenicity Study Group)

### Mutagenicity (In Vitro)

1 updated data set and 2 data sets used previously.

Updated

- Proprietary data set 1

Used previously

- US Food and Drug Administration Center for Food Safety and Applied Nutrition (FDA CFSAN)
- Proprietary data set 2

### Skin Sensitisation

3 data sets used previously.

- Contact Dermatitis
- Gerberick et al
- Cronin and Basketter

### Carcinogenicity

3 data sets used previously.



- CPDB
- ToxRefDB
- Brambilla

There is no change to validation comments for HERG channel Inhibition endpoint.

## Data Set References for Validation Comments

This section details the references and composition of the data sets that have been used for the validation comments.

### BRAMBILLA DATA SET

A collection of carcinogenicity data for 537 compounds derived from the following references:

1. Brambilla G and Martelli A. Update on genotoxicity and carcinogenicity testing of 472 marketed pharmaceuticals. *Mutation Research*, 2009, 681, 209-229, available at <http://dx.doi.org/10.1016/j.mrrev.2008.09.002>.
2. Brambilla G, Mattioli F, Robbiano L and Martelli A. Update of carcinogenicity studies in animals and humans of 535 marketed pharmaceuticals. *Mutation Research*, 2012, 750, 1-51, available at <http://dx.doi.org/10.1016/j.mrrev.2011.09.002>.

### CGX DATA SET

A collection of *in vitro* chromosome aberration test data for 488 compounds from the following reference: Kirkland D, Aardema M, Henderson L and Muller L. Evaluation of the ability of a battery of three *in vitro* genotoxicity tests to discriminate rodent carcinogens and non-carcinogens. I. Sensitivity, specificity and relative predictivity. *Mutation Research*, 2005, 584, 1-256, available at <http://dx.doi.org/10.1016/j.mrgentox.2005.02.004>.

### CPDB DATA SET

A collection of carcinogenicity data for 1,547 compounds from the following reference: Carcinogenic Potency Database (CPDB) downloaded from DSSTox (version 5d, revised 20 November 2008), available at [http://www.epa.gov/NCCT/dsstox/sdf\\_cpdbas.html](http://www.epa.gov/NCCT/dsstox/sdf_cpdbas.html).

### CRONIN AND BASKETTER DATA SET

A collection of guinea pig maximisation test data for 216 compounds from the following reference: Cronin MTD and Basketter DA. Multivariate QSAR analysis of a skin sensitization database. SAR and QSAR in Environmental Research, 1994, 2, 159-179, available at <http://dx.doi.org/10.1080/10629369408029901>.

### CONTACT DERMATITIS DATA SET

A collection of local lymph node assay data for 137 compounds published in Contact Dermatitis which have been extracted from Vitic Nexus (13 September 2012).

### DODDAREADY ET AL DATA SET

A collection of HERG Channel inhibition assay data for 607 compounds from following reference: Doddareddy MR, Klaasse EC, Shagufta, Ijzerman AP & Bender A (2010). Prospective Validation of a Comprehensive *in silico* hERG Model and its Applications to Commercial Compound and Drug Databases. *ChemMedChem*, 5, 716–729.

<http://dx.doi.org/10.1002/cmdc.201000024>

### **FDA CFSAN CHROMOSOME DAMAGE DATA SET**

A data set containing *in vitro* chromosome aberration test data for 2,172 compounds, *in vivo* chromosome aberration test data for 449 compounds and *in vivo* micronucleus test data for 1397 compounds, all derived from the FDA/CFSAN/OFAS knowledge base.

### **FDA CFSAN MUTAGENICITY DATA SET**

A collection of Ames test data for 8,421 compounds derived from the FDA/CFSAN/OFAS knowledge base.

### **GERBERICK ET AL DATA SET**

A collection of local lymph node assay data for 318 compounds derived from the following references:

1. Gerberick GF, Ryan CA, Kern PS, Schlatter H, Dearman RJ, Kimber I, Patlewicz GY and Basketter DA. Compilation of historical local lymph node data for evaluation of skin sensitization alternative methods. *Dermatitis*, 2005, 16, 157-202. Dataset downloaded on 3 September 2010, from <http://www.inchemicotox.org/results/>.
2. Kern PS, Gerberick GF, Ryan CA, Kimber I, Aptula A and Basketter DA. Local lymph node data for the evaluation of skin sensitization alternatives: a second compilation. *Dermatitis*, 2010, 21, 8-32, available at <http://dx.doi.org/10.2310/6620.2009.09038>.

### **MMS DATA SET**

A collection of *in vivo* micronucleus test data for 256 compounds from the Kirkland (CGX) data set, collated by the Japanese Mammalian Mutagenesis Study Group (MMS).

### **PROPRIETARY DATA SET 1**

A proprietary collection of Ames test data for 1,812 chemicals.

### **PROPRIETARY DATA SET 2**

A proprietary collection of Ames test data for 475 chemicals contributed by Bayer Schering Pharma AG.

### **PROPRIETARY DATA SET 3**

A proprietary collection of HERG Channel inhibition assay data for 11,630 chemicals contributed by Lhasa Limited member.

### **PROPRIETARY DATA SET 4**

A proprietary collection of HERG Channel inhibition assay data for 1,694 chemicals contributed by Lhasa Limited member.

### **SOFUNI DATA SET**

A collection of *in vitro* chromosome aberration test data for 712 compounds from the following source: Revised Edition 1998 Data Book of Chromosomal Aberration Test *in Vitro*, Sofuni T (editor), Life-Science Information Center, Tokyo, 1999.

### **TOXREFDB DATA SET**

Toxicity Reference Database (ToxRefDB) Chronic & Cancer Endpoints data (downloaded 21 August 2012 from EPA website), which is no longer available. This data set was derived from that described in the following reference: Martin MT, Judson RS, Reif DM, Kavlock RJ and Dix DJ. Profiling

chemicals based on chronic toxicity results from the U.S. EPA ToxRef database. Environmental Health Perspective, 2009, 117, 3, available at <http://dx.doi.org/10.1289/ehp.0800074>.

## Data Set Summary

The following tables show a summary of the data in the data sets referenced above for each endpoint.

### Carcinogenicity

Dataset	Compounds Processed	Positives	Negatives	Other
Brambilla	537	250	191	96 equivocal
CPDB	1,547	806	738	3 unspecified
ToxRefDB	337	179	158	-

### Chromosome Damage In Vitro

Dataset	Compounds Processed	Positives	Negatives	Other
CGX	488	292	168	28 equivocal
FDA CFSAN (CA)	2,172	983	1,105	84 equivocal
Sofuni	712	282	309	121 equivocal

### Chromosome Damage In Vivo

Dataset	Compounds Processed	Positives	Negatives	Other
FDA CFSAN (CA)	449	136	296	17 equivocal
FDA CFSAN (MN)	1,397	419	947	31 equivocal
MMS	256	112	130	14 equivocal

### Mutagenicity

Dataset	Compounds Processed	Positives	Negatives	Other
FDA CFSAN	8,421	4,300	4,115	6 equivocal
Proprietary data set 1	1,812	584	1,164	64 equivocal
Proprietary data set 2	475	56	419	-

### Skin Sensitisation

Dataset	Compounds Processed	Positives	Negatives	Other
Contact Dermatitis	137	59	65	13 equivocal
Cronin and Basketter	216	97	77	42 equivocal
Gerberick et al	318	143	88	87 equivocal

## HERG

Dataset	Compounds Processed	Positives	Negatives	Other
Doddareaddy et al	607	487	120	-
Proprietary data set 3	11,630	5,606	4,833	1,153 inconclusive 38 unknown
Proprietary data set 4	1,694	1,151	543	-

## What's New in Sarah Nexus 3.2

The following sections detail the changes to Sarah Nexus 3.2.

### Sarah Model

A new Sarah Nexus model has been built called "Sarah model – 2022.1", which uses Ames mutagenicity data from an expanded training set of 12195 compounds sourced from data contained in the Vitic database and donated by Lhasa Limited members. The chemical structures of the training set compounds have been standardised and the biological data curated to reach an overall Ames result for each data point. This produces a self-organising hypothesis network (SOHN) model containing 398 hypotheses (305 unique).

### Sarah Training Set Data Records

The training set to prepare Sarah model – 2022.1 has been updated by the addition of Ames test data for several new and existing chemicals within the model.

- New data records = 4,607
- Total data records = 50,682
- New substances = 382
- Total substances = 13,521
  - Number of positive substances = 5,814 (48%)
  - Number of negative substances = 6,381 (52%)
  - Number of additional substances = 1,326\*

\*Additional substances have not been included in the model but may be accessed in the **Additional information** tab. Although these are in the extended training set, they have been rejected by the model because they have neither a positive or negative overall call or for various other reasons during the standardisation process.

### Sarah Training Set Data Sources

The following table shows the number of records contained in the Sarah training set from each data source.

Sources	Total number of records
ACID Halide Mutagenicity Dataset	39
Bursi Mutagenicity Dataset	4,326
Carcinogenic Potency Database (CPDB)	833
CGX Mutagenicity Dataset	709
Derek Example Compounds	373
EURL ECVAM Genotoxicity and Carcinogenicity Consolidated Database	956
Feng Mutagenicity Dataset	1,859

Sources	Total number of records
Hansen Mutagenicity Dataset	6,479
Helma Mutagenicity Dataset	683
ISSSTY Mutagenicity Dataset	7,042
Japan Chemical Industry Ecology-Toxicology and Information Center (JETOC) Mutagenicity Dataset	323
Marketed Pharmaceuticals Database	547
Member data	232
National Institute of Health Sciences Dataset	668
Vitic NTP Table	2,016
Vitic Summary call Table	14,891
US Food and Drug Administration - Center for Drug Evaluation and Research (FDA CDER)	585
US Food and Drug Administration - Center for Food Safety and Applied Nutrition (FDA CFSA)	8,121
Total	50,682

### Internal Validation

5-Fold cross-validation is performed during the Sarah model building process and the results can be assessed within the “Manage prediction models” section of Nexus for “Sarah model – 2022.1”.

### External Validation

6 datasets have been used for external validation:

- DGM/NIHS Dataset
- Proprietary Dataset 1
- Proprietary Dataset 2
- Proprietary Dataset 3
- Proprietary Dataset 4
- Vitic Intermediates Dataset

### DGM/NIHS Dataset

A collection of Ames test data for 12,140 chemicals compiled by The Division of Genetics and Mutagenesis, National Institute of Health Sciences (DMG/NIHS) used in the Ames/QSAR International Challenge Project discussed in the following reference: Honma M et al. Improvement of quantitative structure-activity relationship (QSAR) tools for predicting Ames mutagenicity: outcomes of the Ames/QSAR International Challenge Project. *Mutagenesis*, 2019, 34, 3-16 available at

<http://dx.doi.org/10.1093/mutage/gey031>. Overall, 1,757 (14.4%) compounds have been assigned positive and 10383 (85.6%) have been assigned negative.

#### ***Proprietary Dataset 1***

A proprietary collection of Ames test data for 454 pharmaceutical-related chemicals donated by a Lhasa Limited member. Overall, 54 (11.9%) compounds have been assigned positive and 400 (81.1%) have been assigned negative.

#### ***Proprietary Dataset 2***

A proprietary collection of Ames test data for 507 pharmaceutical-related chemicals donated by a Lhasa Limited member. Overall, 95 (18.7%) compounds have been assigned positive and 412 (81.3%) have been assigned negative.

#### ***Proprietary Dataset 3***

A proprietary collection of Ames test data for 2,865 pharmaceutical-related chemicals donated by a Lhasa Limited member. Overall, 171 (6.0%) compounds have been assigned positive and 2,694 (94.0%) have been assigned negative.

#### ***Proprietary Dataset 4***

A proprietary collection of Ames test data for 1,278 pharmaceutical-related chemicals donated by a Lhasa Limited member. Overall, 259 (20.3%) compounds have been assigned positive and 1,019 (79.7%) have been assigned negative.

#### ***Vitic Intermediates Dataset***

A collection of Ames test data for 1,748 common intermediates shared within the Vitic Intermediates projects. Information about Vitic Intermediates is available at <https://www.lhasalimited.org/Initiatives/vitic-intermediates.htm>. Overall, 584 (33.4%) compounds have been assigned positive and 1,164 (66.6%) have been assigned negative.

## Dataset Performance for External Validation

The following table shows a summary of the data in the datasets used for external validations, as well as performance statistics for Sarah model – 2022.1 against each dataset.

Dataset					Performance Statistics											
Name	Size	Positive	Negative	Bias	BAC	SEN	SPEC	PPV	NPV	COV	TP	TN	FP	FN	EQ	OD
DMG/NIHS Dataset	12,140	1,757	10,383	-0.86	0.7587	0.7167	0.8006	0.3916	0.9404	0.8231	1,088	6,785	1,690	430	1,736	411
Proprietary Dataset 1	454	54	400	-0.88	0.7547	0.6122	0.8971	0.4615	0.9414	0.8568	30	305	35	19	47	18
Proprietary Dataset 2	507	95	412	-0.81	0.7221	0.6625	0.7817	0.4173	0.9075	0.8264	53	265	74	27	66	22
Proprietary Dataset 3	2,865	171	2,694	-0.94	0.6437	0.4167	0.8707	0.1529	0.9638	0.7899	50	1,866	277	70	500	102
Proprietary Dataset 4	1,278	259	1,019	-0.80	0.6705	0.4952	0.8458	0.4421	0.8716	0.8224	103	713	130	105	151	76
Vitic Intermediates Dataset	1,748	584	1,164	-0.67	0.7541	0.7940	0.7141	0.5754	0.8766	0.8129	370	682	273	96	290	37

### Abbreviations

BAC = balanced accuracy:  $(SEN + SPEC) / 2$

SEN = sensitivity:  $TP / (TP + FN)$

SPEC = specificity:  $TN / (TN + FP)$

PPV = positive predictivity:  $TP / (TP + FP)$

NPV = negative predictivity:  $TN / (TN + FN)$

COV = coverage:  $(TP + FP + TN + FN) / (TP + FP + TN + FN + EQ + OD)$

TP = true positive

FP = false positive

TN = true negative

FN = false negative

EQ = equivocal

OD = outside domain



## Structure Standardisation

Nexus standardises all input structures before processing them through its applications. The standardisation techniques Sarah uses to convert user-drawn structures into standard forms employs a set of transform rules which include consideration of mixtures. Whereas all components of a query mixture are assessed by Sarah Nexus, pharmaceutically acceptable salts are removed during standardisation, so they are not considered during the prediction. The list of pharmaceutically accepted salts was developed based on Haynes et al, Journal of Pharmaceutical Sciences, 2005, 94, 2111-2120.

Following an assessment of mixture components which are non-mutagens in the Sarah training set, or have published negative Ames result, the list of salts removed by Nexus during standardisation has been updated in Sarah Nexus 3.2. The following table shows the new salts removed in Sarah Nexus 3.2.

Common Name	SMILES	Reason
Tetrafluoroborate	<chem>[B-](F)(F)(F)F</chem>	Non-mutagen in training set
Hexafluorophosphate	<chem>[P-](F)(F)(F)(F)(F)F</chem>	Negative study for lithium hexafluorophosphate*
Ethanolamine	<chem>NCCO</chem>	Non-mutagen in training set
Triisopropylamine	<chem>N(CC(O)C)(CC(O)C)CC(C)O</chem>	Non-mutagen in training set
Urea	<chem>NC(=O)N</chem>	Non-mutagen in training set
Guanidine	<chem>NC(=N)N</chem>	Non-mutagen in training set
Benzoic acid	<chem>C1=CC=CC=C1C(=O)O</chem>	Non-mutagen in training set
Ethyl sulfate	<chem>O(S([O-])(=O)=O)CC</chem>	Non-mutagen in training set
2-Hydroxyethanesulfonic acid	<chem>O=S(=O)(CCO)O</chem>	Non-mutagen in training set
Salicylic acid	<chem>C1(=CC=CC=C1O)C(=O)O</chem>	Non-mutagen in training set
Glycerophosphoric acid	<chem>OCC(O)COP(O)(O)=O</chem>	Non-mutagen in training set
Beta-Glycerophosphoric acid	<chem>P(O)(=O)(OC(CO)CO)O</chem>	Non-mutagen in training set

\*A negative Ames test has been published for lithium hexafluorophosphate as part of a European Chemicals Agency (ECHA) registration dossier, available at <https://echa.europa.eu/registration-dossier/-/registered-dossier/13201/7/7/2>.

The following screenshots demonstrate the change in behaviour when processing a query containing a salt that is removed during standardisation, using tri(dimethylamino)benzotriazol-1-yl oxyphosphonium hexafluorophosphate (CAS number = 56602-33-6) as an example.

## Behaviour in Nexus 2.4, Sarah Nexus 3.1, Sarah Model – 2020.1

The query is assessed as a mixture to be outside domain, highlighting that the hexafluorophosphate ion is not in the model. The query compound (CAS number = 56602-33-6) is in the additional information tab, as it is in the extended training set but has been rejected by the model as a mixture. The neutral form of the compound (CAS number = 56602-32-5) is also in the additional information tab, as the model treats this as a separate compound.

**Sarah Prediction** For the 'Mutagenicity in vitro' endpoint the prediction is:

# OUTSIDE DOMAIN

'outside domain features' selected, click above to view the original structure

**Prediction Constraints**

- Model: Sarah Model - 2020.1
- Endpoint: Mutagenicity in vitro
- Reasoning type: Weighted
- Equivocal: 8%
- Sensitivity: 8%
- Certified model: Yes
- Prediction date: 09 February 2022 10:30

**Results** Additional Information (2)

Show: All compounds Strain

The compounds below are being shown for additional information. They were not used in the prediction but have a similarity to the query compound of 30% or higher.

- 1 of 2 - 100% (Rejected)
- 2 of 2 - 96% (-Ve)

**Example compound**

Click below to view the standardised structure

**Overall Call:** Rejected  
**Similarity:** 100%

Click on a contribution below to view the original structure

- Source: Vitic Summary Call Table  
Dataset Call: Unreliable  
Source activity call: Negative  
Structure ID: CAS RN® 56602-33-6  
Rejected Reason: Unmapped  
[Reference\(s\)](#)
- Source: ISSSTY Mutagenicity Dataset  
Dataset Call: Unreliable  
Source activity call: Negative  
Structure ID: CAS RN® 56602-33-6  
Rejected Reason: Unmapped  
[Reference\(s\)](#)
- Source: Bursi Mutagenicity Dataset  
Dataset Call: Unreliable  
Source activity call: Negative  
Structure ID: CAS RN® 56602-33-6  
Rejected Reason: Unmapped  
[Reference\(s\)](#)

## Behaviour in Nexus 2.5, Sarah Nexus 3.2, Sarah Model – 2022.1

The query is considered an exact match with the compound in the training set as the hexafluorophosphate ion is not assessed. The model considers the salt (CAS number = 56602-33-6) and neutral form (CAS number = 56602-32-5) to be the same; however, clicking on the contribution for each example compound will show the original structure for that training example to assess whether it was the salt. The user may see the original query entered by clicking on the prediction structure.

The screenshot displays the Sarah Prediction software interface. The main prediction panel on the left shows the text "For the 'Mutagenicity in vitro' endpoint the prediction is: **NEGATIVE** with 100% confidence". Below this is a large chemical structure of a phosphonium salt, with a benzimidazole ring system and a hexafluorophosphate counterion. A dashed green outline highlights the benzimidazole part of the structure. At the bottom of this panel, it says "Click above to view the original structure".

The middle panel, titled "Results", contains the text: "The compound is predicted to be negative with 100% confidence for the 'Mutagenicity in vitro' endpoint in the model: 'Sarah model - 2020.1\_edit 09'. This is based on an exact match with a compound found in the training dataset." Below this text is a section for "Training set example (exact match with query)" which includes a smaller version of the chemical structure and a green bar indicating "Negative 100%".

The right panel, titled "Example compound", contains the text "Click below to view the standardised structure" above a chemical structure of the same compound. Below this, it lists "Overall Call: Negative" and "Similarity: 100%". It also includes a section for "Click on a contribution below to view the original structure" with a list of sources and their corresponding dataset calls and structure IDs:

- Source: Vitic Summary Call Table  
Dataset Call: Negative  
Source activity call: Negative  
Structure ID: CAS RN® 56602-32-5  
[Reference\(s\)](#)
- Source: Vitic Summary Call Table  
Dataset Call: Negative  
Source activity call: Negative  
Structure ID: CAS RN® 56602-33-6  
[Reference\(s\)](#)
- Source: ISSSTY Mutagenicity Dataset  
Dataset Call: Negative  
Source activity call: Negative  
Structure ID: CAS RN® 56602-33-6  
[Reference\(s\)](#)
- Source: Bursi Mutagenicity Dataset  
Dataset Call: Negative  
[Reference\(s\)](#)

Sarah Prediction Results Additional Information (0)

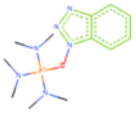
Highlight Hypotheses and Features: [checkbox checked] [checkbox] [checkbox] [checkbox] [checkbox] Strain

For the 'Mutagenicity in vitro' endpoint the prediction is:

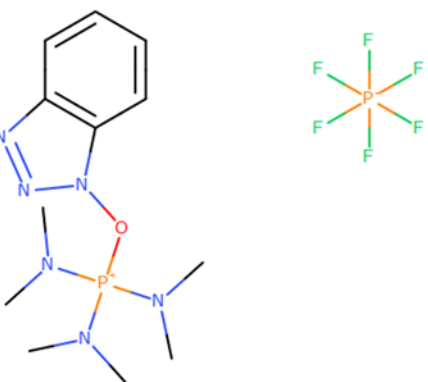
**NEGATIVE**  
with **100%** confidence

The compound is predicted to be negative with 100% confidence for the 'Mutagenicity in vitro' endpoint in the model: 'Sarah model - 2020.1\_edit 09'. This is based on an exact match with a compound found in the training dataset.

Training set example (exact match with query)



Negative 100%



Click below to view the standardised structure

**Overall Call:** Negative  
**Similarity:** 100%

Click on a contribution below to view the original structure

Source: Vitic Summary Call Table  
Dataset Call: Negative  
Source activity call: Negative  
Structure ID: CAS RN® 56602-32-5  
[Reference\(s\)](#)

Source: Vitic Summary Call Table  
Dataset Call: Negative  
Source activity call: Negative  
Structure ID: CAS RN® 56602-33-6  
[Reference\(s\)](#)

Source: ISSSTY Mutagenicity Dataset  
Dataset Call: Negative  
Source activity call: Negative  
Structure ID: CAS RN® 56602-33-6  
[Reference\(s\)](#)

Source: Bursi Mutagenicity Dataset  
Dataset Call: Negative  
[Reference\(s\)](#)

\*Structure\* selected, click above to view the prediction structure

## What's New in ICH M7

The following sections detail the changes to the ICH M7 function in Nexus 2.5.

### ICH M7 Expert Review (Derek 6.2 and Sarah 3.2)

The following changes have been made to the automated expert review arguments displayed following an ICH M7 prediction.

#### Automated Expert Review Arguments

The list of automated arguments has been updated and renumbered to more clearly group arguments based on the same prediction scenario. New arguments have been implemented, and existing arguments updated, relating to the following categories:

- Chemical classes for which the Ames text may not adequately assess the mutagenic hazard
- Chemicals considered to belong to a chemical class listed in the cohort of concern

#### Expert Review Argument Numbers

New Number	Old Number
1.1	60
1.2	61
2.1	2
2.2	13
3.1	14
3.2	3
4.1	39
4.2	38
5.1	26
6.1	36
6.2	37
7.1	27
7.2	28
8.1	29
8.2	30
9.1	31
9.2	32
10.1	4

New Number	Old Number
10.2	12
11.1	6
11.2	7
12.1	11
12.2	40
12.3	41
13.1	5
14.1	9
14.2	8
15.1	46
15.2	42
16.1	10
16.2	43
17.1	15
17.2	53
18.1	16
18.2	54
19.1	17
19.2	55
20.1	44
20.2	45
21.1	56
21.2	57
22.1	19
22.2	18
23.1	21
23.2	20
24.1	47
24.2	48

New Number	Old Number
25.1	59
25.2	58
26.1	50
26.2	33
27.1	52
27.2	51
28.1	22
29.1	24
29.2	23
29.3	35
29.4	25
30.1	62
30.2	63
30.3	
30.4	
30.5	
30.6	
30.7	
31.1	1
31.2	
31.3	
31.4	
31.5	

### **Ames Test Not Adequate**

The following arguments represent a query which belongs to a chemical class for which the related Derek alert description discusses reasons why the standard Ames test may not be adequate to assess the mutagenic hazard presented by the chemical class. They have been implemented with an “Inconclusive” overall *in silico* call to notify the user to assess the relevance of the Derek comments for their review.

Argument Number	Chemical Class
31.1	Carboxylic acid halide, carbamoyl halide, sulfonyl halide, thionyl halide
31.2	Allylbenzene
31.3	N-Methylol
31.4	Alkyl aldehyde
31.5	Benzyl halide

### Cohort of Concern

The following arguments represent a query which belongs to a chemical class suspected of belonging to a cohort of concern as listed within the ICH M7 guideline, and the user is notified of the requirement to conduct a compound-specific risk assessment. 5 new arguments have been implemented to specifically discuss subclasses of N-nitroso compounds for which structural features are expected to reduce the mutagenic potential as discussed in the related Derek alert 007. These have been implemented with either a “Positive”, “Negative” or “Inconclusive” overall *in silico* call.

Argument Number	Argument Summary	Argument Assignment
30.1	Query belongs to a chemical class considered a cohort of concern; a compound-specific risk assessment has concluded positive.	Positive
30.2	Query belongs to a chemical class considered a cohort of concern; a compound-specific risk assessment has concluded negative.	Negative
30.3	Query is an N-nitroso compound that cannot undergo the expected mechanism that leads to potent activity in other N-nitroso compounds due to a lack of alpha-hydrogen.	Negative
30.4	Query is an N-nitroso compound that cannot undergo the expected mechanism that leads to potent activity in other N-nitroso compounds due to a lack of alpha-hydrogen; however, an alternative mechanism may be possible.	Inconclusive
30.5	Query is an N-nitroso compound that may not be able undergo the expected mechanism that leads to potent activity in other N-nitroso compounds due to being in a hindered cyclic compound.	Negative
30.6	Query is an N-nitroso compound with an alpha-heteroatom substituent, examples of which are mutagenic but have been observed to have weak carcinogenic potency.	Positive
30.7	Query is an N-nitroso compound of a primary amine which is expected to undergo formation of a mutagenic diazonium ion; however, stability in vivo is likely to be sufficiently poor and prevent DNA alkylation.	Positive

### Links Between Derek and Sarah

Links between Derek alerts with Sarah hypotheses and training set examples have been updated.



118 Sarah hypotheses have been linked to 55 Derek alerts.

- Sarah hypotheses may be linked to multiple Derek alerts
- Derek alerts may be linked to multiple Sarah hypotheses

6,179 Sarah training examples have been linked to 139 Derek alerts.

- Sarah training examples may be linked to multiple Derek alerts

## ICH M7 Classification (Derek 6.2 and Sarah 3.2)

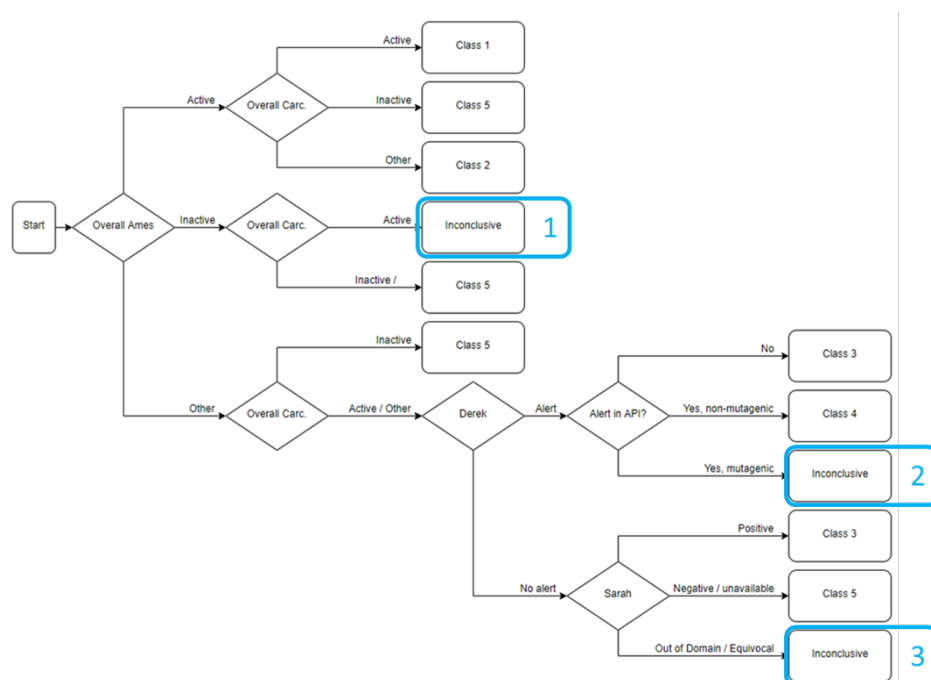
The following changes have been made to the ICH M7 classification tool.

### Ames Dataset

The Lhasa Ames dataset has been updated in line with the Sarah training set and the Mutagenicity Negative Predictions dataset.

### ICH M7 Class Assignment

A predicted ICH M7 class is displayed in the ICH M7 classification table for queries depending on both the predictions provided by Derek and Sarah as well as consideration of any known Ames or carcinogenicity data. The class is assigned according to the following logic process:



There are 3 scenarios where “Inconclusive” may be assigned instead of ICH M7 class 1 – 5:

1. Overall Ames = inactive and overall carcinogenicity = active
  - Query is a known non-mutagenic carcinogen
2. Overall Ames = other and overall carcinogenicity = active/other and Derek = alert activated by query and API
  - Query is an impurity in a mutagenic API
3. Overall Ames = other and overall carcinogenicity = active/other and Derek = no alert and Sarah = outside domain/equivocal

- The *in silico* prediction made by Derek and Sarah is considered inconclusive and requires review

Scenario 3 represents the situation where the calculated *in silico* prediction is inconclusive and requires expert review to determine the ICH M7 class for the query. Instead, scenarios 1 and 2 represent situations where the query should be considered outside ICH M7 as the guideline does not apply to drug products intended for advanced cancer indications or drug substances which are genotoxic at therapeutic concentrations. The workflow has been updated to better represent this, showing “Inconclusive (Outside scope of ICH M7)” for scenarios 1 and 2.

## Permissible Daily Exposure Comments

Comments are displayed in the ICH M7 classification table for queries which have had a permissible daily exposure (PDE) or acceptable intake (AI) published. These comments are displayed for any component of the query that matches a chemical in the published list of PDEs and AIs.

The following updates have been made to the information presented:

- ICH guideline M7 – addendum
  - Entries updated in line with changes in updated publication October 2021
- ICH guideline Q3C
  - Entries updated in line with changes in updated publication (R8) May 2021
- Literature publications
  - New entries added for 5 chemicals listed in the following table

Chemical	CAS	PDE	Reference
Diisopropyl ether	108-20-3	PDE = 980 ug/day (inhalation)	Romanelli and Evandri, Toxicological Research, 2018, 34, 111-125
Irganox 1010	6683-19-8	PDE = 8,000 ug/day (parenteral)	Parris et al, Regulatory Toxicology and Pharmacology, 2020, 118, 104802
Irgafos 168	31570-04-4	PDE = 2,900 ug/day (parenteral)	Parris et al, Regulatory Toxicology and Pharmacology, 2020, 118, 104802
Bisphenol A	80-05-7	PDE = 4.2 ug/day (parenteral)	Parris et al, Regulatory Toxicology and Pharmacology, 2020, 118, 104802
Butylated hydroxytoluene	128-37-0	PDE = 12,500 ug/day (parenteral)	Parris et al, Regulatory Toxicology and Pharmacology, 2020, 118, 104802



## Our Products

---



*Expert knowledge-based toxicity prediction software.*



*Statistical-based software for the prediction of mutagenicity.*



*A tool for assessing the relative purging of mutagenic impurities.*



*Expert decision support software for predicting the forced degradation pathways of organic compounds.*



*The chemical database and information management system, offering researchers and scientists rapid access to searchable toxicological information.*



*Expert decision support software for predicting the metabolic fate of chemicals in mammals.*



*A secondary pharmacology model suite leveraging value from federated learning.*



*A tool to support risk assessment in the context of adverse outcome pathways.*



shared **knowledge** • shared **progress**

Lhasa Limited Registered Office  
Granary Wharf House, 2 Canal Wharf, Leeds LS11 5PS  
Registered Charity (290866)

+44 (0)113 394 6020  
info@lhasalimited.org  
[www.lhasalimited.org](http://www.lhasalimited.org)