Organisation for Economic Co-operation and Development

ENV/CBC/MONO(2025)2/ANN9

Unclassified

English - Or. English

11 September 2025

ENVIRONMENT DIRECTORATE
CHEMICALS AND BIOTECHNOLOGY COMMITTEE

Annex 9: DASS Similar Methods Performance Evaluation - Supporting document to Test Guideline 497 on Defined Approaches for Skin Sensitisation

Series on Testing and Assessment
No. 336

(Second edition)

E-mail: ehscont@oecd.org.

JT03570723

Please cite this publication as:

OECD (2025), *Supporting document to Test Guideline 497 on Defined Approaches for Skin Sensitisation, Annex 1: Evaluation Framework*, OECD Series on Testing and Assessment, No. 336, OECD Environment, Health and Safety, Paris, https://one.oecd.org/document/ENV/CBC/MONO(2025)2/ANN1/en/pdf

Contact us

OECD Environment Directorate,
Environment, Health and Safety Division
2 rue André-Pascal
75775 Paris Cedex 16
France

E-mail: ehscont@oecd.org

About the OECD

The Organisation for Economic Co-operation and Development (OECD) is an intergovernmental organisation in which representatives of 38 countries in North and South America, Europe and the Asia and Pacific region, as well as the European Union, meet to co-ordinate and harmonise policies, discuss issues of mutual concern, and work together to respond to international problems. Most of the OECD's work is carried out by more than 200 specialised committees and working groups composed of member country delegates. Observers from several Partner countries and from interested international organisations attend many of the OECD's workshops and other meetings. Committees and working groups are served by the OECD Secretariat, located in Paris, France, which is organised into directorates and divisions.

The Environment, Health and Safety Division publishes free-of-charge documents in twelve different series: **Testing and Assessment**; **Good Laboratory Practice and Compliance Monitoring**; **Pesticides**; **Biocides**; **Risk Management**; **Harmonisation of Regulatory Oversight in Biotechnology**; **Safety of Novel Foods and Feeds; Chemical Accidents; Pollutant Release and Transfer Registers; Emission Scenario Documents; Safety of Manufactured Nanomaterials;** and **Adverse Outcome Pathways.** More information about the Environment, Health and Safety Programme and EHS publications is available on the OECD's World Wide Web site (https://www.oecd.org/en/topics/chemical-safety-and-biosafety.html).

This publication was developed in the IOMC context. The contents do not necessarily reflect the views or stated policies of individual IOMC Participating Organizations.

The Inter-Organisation Programme for the Sound Management of Chemicals (IOMC) was established in 1995 following recommendations made by the 1992 UN Conference on Environment and Development to strengthen co-operation and increase international co-ordination in the field of chemical safety. The Participating Organisations are FAO, ILO, UNDP, UNEP, UNIDO, UNITAR, WHO, World Bank, Basel, Rotterdam and Stockholm Conventions and OECD. The purpose of the IOMC is to promote co-ordination of the policies and activities pursued by the Participating Organisations, jointly or separately, to achieve the sound management of chemicals in relation to human health and the environment.

Foreword

This summary provides information on, and analysis of, the performance of two Defined Approaches (DAs) in GL 497, where similar[1], key event (KE)-based methods have been substituted for the original assays adopted within the 2o3 and ITS DAs to create alternate[2] DAs. All supporting data can be found in Annex 2 of the GL497 supporting documents.

[1] Similar in this case refers to methods which measure the same key event (KE) as the original method, and is included in OECD TG 442C-E.
[2] Alternate DAs refer to the same Defined Approaches as originally published, with one or more original methods substituted with an alternate similar method(s). The data interpretation procedures do not change.

Table of contents

FIGURES

TABLES

1 Datasets extracted from Annex 2

In this analysis, we evaluated the performance impact of substituting six similar *in chemico* and *in vitro* methods (ADRA, LuSens, U-SENS, IL-8Luc, EpiSensA, and GARDskin) in place of the original methods included in the classic[3] 2o3 (original: DPRA, KeratinoSens, h-CLAT) and ITS (original: DPRA, h-CLAT, Derek Nexus (ITSv1)/OECD (Q)SAR Toolbox (ITSv2)) DAs. Only one version of the classic ITS was chosen as a comparator to simplify the ITS analyses (ITSv1) and due to the similar performance of the ITSv1 and v2.

One of the main challenges of the analysis is the diversity of data available for the individual methods, with respect to number of data points. Each method has a different set of data available, making performance comparisons across the various permutations of the DAs challenging.

To compare the performance of each permutation utilising a similar method to the classic permutation of the DA, we conducted the following analyses using the data in Annex 2:

- **Key-Event Anchored Dataset** – This analysis focuses on the impact of substituting a single similar method within the relevant DA. A KE-anchored dataset was derived for each alternate DA permutation and its classic DA. To ensure that the comparison of the alternate DA vs the classic DA was on the same dataset (an “apples to apples” comparison), any chemicals lacking data for the specific substitute method, or chemicals with an inconclusive DA prediction (for either the alternate or classic DA) were filtered out. This created an identical dataset to more easily evaluate the impact of a single substitution of the original classic methods in the DAs on the performance. ***Note: This means that every chemical used in the individual evaluations has a data point for the similar method under evaluation in the specific permutation. e.g. 2o3 using DPRA, KeratinoSens and U-SENS: if no U-SENS data point was available then the 2o3 prediction was not used in the analysis.***
- **Maximal Common Dataset** – This analysis is based on datasets which use all available data points for a specific permutation of the DAs, even if data for the similar method under evaluation is missing because a decision can be made in the DAs using partial information sources (under specific circumstances detailed in GL 497). The permutation can have up to three similar methods or a single alternate within the relevant DA. A maximal dataset was derived for each DA permutation and its classic DA. To ensure the comparison of alternate DA vs classic DA on the same database with maximal size, only chemicals with an inconclusive DA prediction (for both the alternate and classic DA) were filtered out (irrespective of the presence of a data point for the specific method(s) being evaluated), creating a common dataset for each comparison. ***Note: This means that some chemicals used in the evaluations may lack a data point for the similar method under evaluation in the specific***

[3] Hereafter classic refers to the three DAs adopted in the original Defined Approach for Skin Sensitisation Guideline 497, namely the DPRA, KeratinoSens, h-CLAT, Derek Nexus, and OECD (Q)SAR Toolbox.

permutation. e.g. 2o3 using DPRA, KeratinoSens and U-SENS: if no U-SENS data point was available but a 2o3 prediction could be made based on DPRA and KeratinoSens then this data point was still used in the analysis.

New data/analyses provided in Annex 2 Spreadsheet

- 2o3
  - The summary data for each permutation, including performance (balanced accuracy, accuracy, sensitivity, specificity), the total number of data points, and the true/false positive/negative counts
    - 2o3_KE-anchored vs LLNA (Table 1)
    - 2o3_KE-anchored vs human (Table 2)
    - 2o3_Maximal common vs LLNA (Table 3)
    - 2o3_Maximal common vs human (Table 4)
- ITS
  - The summary data for each permutation, including performance (accuracy/balanced accuracy, sensitivity, specificity, over/under-predictions (for potency only), the total number of data points, and the true/false positive/negative counts.
    - ITS_KE-anchored vs LLNAhaz (Table 5)
    - ITS_KE-anchored vs humanhaz (Table 6)
    - ITS_KE-anchored vs LLNApot (Table 7)
    - ITS_KE-anchored vs humanpot (Table 8)
    - ITS_maximal vs LLNAhaz (Table 9)
    - ITS_maximal vs humanhaz (Table 10)
    - ITS_maximal vs LLNApot (Table 11)
    - ITS_maximal vs humanpot (Table 12)

2 2o3 DA

All 2o3 permutations predict skin sensitisation hazard comparably to the classic DA permutation, when compared to the LLNA. This is true for both the KE-anchored and maximal common datasets, for both single-similar-method drop-in replacements and for all alternate DA permutations, when multiple methods are replaced. Compared to the LLNA, the BA for all permutations ranges from 81-95%, which is similar or higher to the BA for the classic 2o3 (84%). When compared to human data, the BA performance is more variable, but for all permutations remains above 72%. Because there is a limited amount of human data, fewer chemicals can be evaluated, giving one or two misclassifications a large effect on the overall performance statistics.
When considering these datasets, the majority of 2o3 permutations predict skin sensitisation hazard as well as or better than the classic permutation. Regardless of the dataset used as the basis for comparison, all of the evaluated 2o3 permutations outperform the LLNA when compared to human hazard reference data.
See tabs in the Annex 2 Excel file for full tables of data:

- 2o3_KE-anchored vs LLNA (Table 1)
- 2o3_KE-anchored vs human (Table 2)
- 2o3_Maximal common vs LLNA (Table 3)
- 2o3_Maximal common vs human (Table 4)

KE-Anchored Dataset

When evaluating permutations where only one similar method was replaced in the DA for the 2o3 (e.g., ADRA in place of DPRA), in most cases the BA remains within +/- 2% of the BA of the classic DA, and is always greater than 81% and 77%, when compared to both the LLNA and human datasets, respectively. Generally, the alternate DAs using similar methods also had similar sensitivity and specificity when compared to both the LLNA and human reference datasets, and all performed better than the LLNA against the human reference data.
Small datasets, as is the case with the human data, can result in large impacts on performance metrics by single misclassifications. For example, when compared against the human dataset and EpiSensA replaces KeratinoSens in the 2o3 DA (DPRA, EpiSensA, h-CLAT), the BA is reduced by 8% (to 81.9%) when compared against the human dataset. It is very important to note that this 8% difference in BA is caused by a single chemical being predicted as a false positive (FP) in the EpiSensA permutation and being correctly predicted as a negative (True Negative, TN) by the classic 2o3 DA.

LLNA

Take home message: All permutations evaluated here performed similarly to the classic 2o3 (provided as a comparison for each permutation; Table 1) when evaluated using a common dataset. The percentage difference in performance between the KE-anchored and the classic 2o3 against the LLNA reference dataset was generally within two for all performance metrics,

with many permutations performing slightly better than the classic 2o3. A few 2o3 DA permutations had lower specificity (EpiSensA, and U-SENS); however, it should be remembered that one or two misclassifications (false positive) can have a significant impact on the specificity due to the low numbers of true negatives (TN) and false positives (FP). As a reminder, the chemicals evaluated in these datasets all had a datapoint for the method under consideration.

Figure 1. Comparison of balanced accuracy against LLNA for single-method (KE-anchored) replacements (blue) to the same chemical list for the classic 2o3 (orange).

There is minimal difference between the two DAs for each comparison, showing that similar methods can be used within the 2o3 DA for hazard assessment.

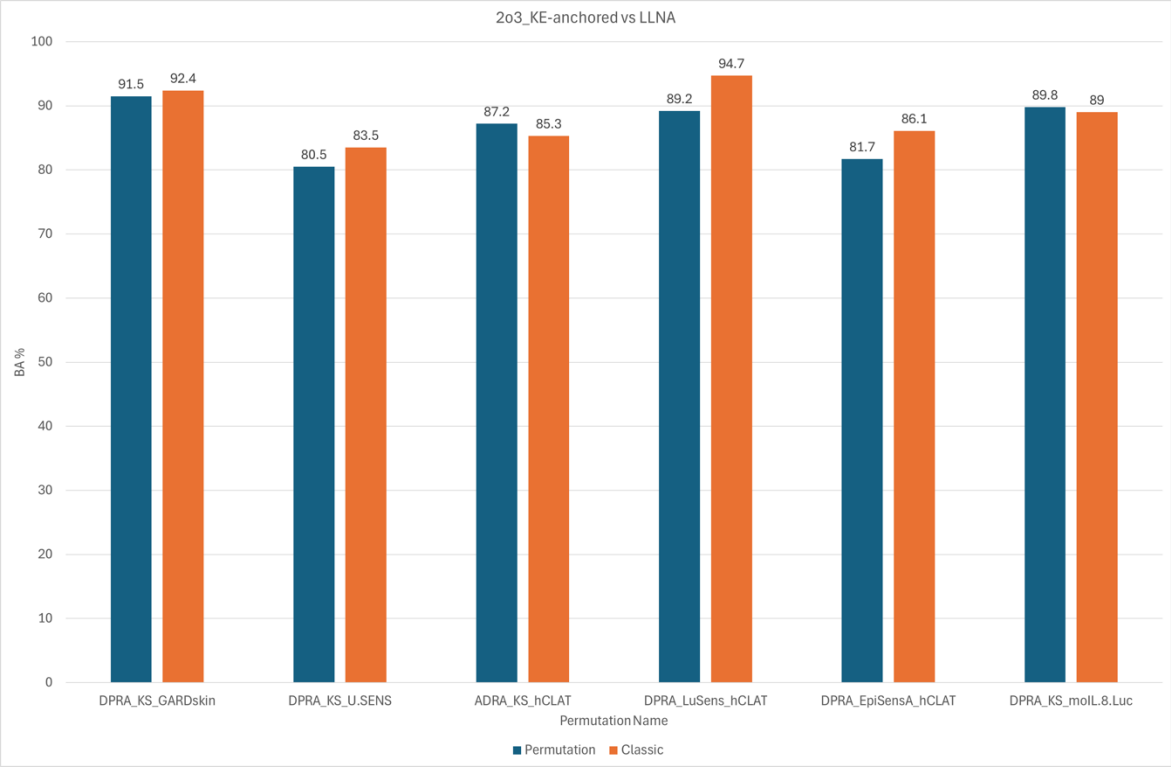

Table 1. Performance for each of the KE-anchored DA permutations compared to the classic 2o3 when evaluated against the LLNA reference dataset.

Iteration	Dataset	BA (%)	Acc (%)	Sens (%)	Spec (%)	Total	TP	FP	TN	FN
1	Classic 2o3 (all data from GL 497 2023)	83.5	82.8	82.4	84.6	134	89	4	22	19
2	2o3[DPRA, KeratinoSens, h-CLAT]	85.3	84.3	83.7	87	127	87	3	20	17
2	2o3_[ADRA, KeratinoSens, h-CLAT]	87.2	87.4	87.5	87	127	91	3	20	13
3	2o3_[DPRA, KeratinoSens, h-CLAT]	94.7	91.5	89.5	100	47	34	0	9	4
3	2o3_[DPRA, LuSens, h-CLAT]	89.2	89.4	89.5	88.9	47	34	1	8	4
4	2o3_[DPRA, KeratinoSens, h-CLAT]	86.1	83.6	82.2	90	55	37	1	9	8
4	2o3_[DPRA, EpiSensA, h-CLAT]	81.7	89.1	93.3	70	55	42	3	7	3
5	2o3_[DPRA, KeratinoSens, h-CLAT]	92.4	87	84.7	100	69	50	0	10	9
5	2o3_[DPRA, KeratinoSens, GARDskin]	91.5	85.5	83.1	100	69	49	0	10	10
6	2o3_[DPRA, KeratinoSens, h-CLAT]	89	88.3	88.1	90	77	59	1	9	8
6	2o3_[DPRA, KeratinoSens, IL-8]	89.8	89.6	89.6	90	77	60	1	9	7
7	2o3_[DPRA, KeratinoSens, h-CLAT]	83.5	85.8	87.1	80	113	81	4	16	12
7	2o3_[DPRA, KeratinoSens, U-SENS]	80.5	84.1	86	75	113	80	5	15	13

BA = balanced accuracy = (Sens + Spec) / 2
Acc = Accuracy = (TP + TN) / Total
Sens = Sensitivity = TP / ((TP + FN)
Spec = Specificity = TN / (TN + FP)
TP = True Positives, FP = False Positives, TN = True Negatives, FN = False Negatives
Red text = For comparison - results as reported in GL 497 in 2023 for the classic 2o3. Iteration = repeating instruction (loop) from algorithm used to calculate each permutation in turn. Allows for easy comparison of permutation against 'classic' dataset.

Human

Take home message: All permutations evaluated here performed similarly to the classic 2o3 (provided as a comparison for each permutation; Table 2) when evaluated using a common dataset. The percentage difference in performance between the KE-anchored and the classic against the human reference dataset was generally within 2% for all performance metrics, with many permutations performing slightly better than the classic 2o3. A few permutations had lower specificity (EpiSensA); however, one or two misclassifications (false positive) can have a significant impact on the specificity due to the low numbers of true negatives (TN) and false positives (FP) in these small human datasets. As a reminder, the chemicals evaluated in these

datasets all had a datapoint for the method under consideration. The comparison of the LLNA to the human reference data (maximal number of chemicals) is also provided. **It should be noted that the balanced accuracy (58%) and specificity (22%) of the LLNA is significantly lower than any of the evaluated permutations.**

Figure 2. Comparison balanced accuracy against human data for single-method (KE-anchored) replacements (blue) to the same chemical list for the classic 2o3 (orange).

There is minimal difference between the two DAs for each comparison, showing that similar methods can be used within the 2o3 DA for hazard assessment. LLNA against human for the maximal dataset is also included for reference (green).

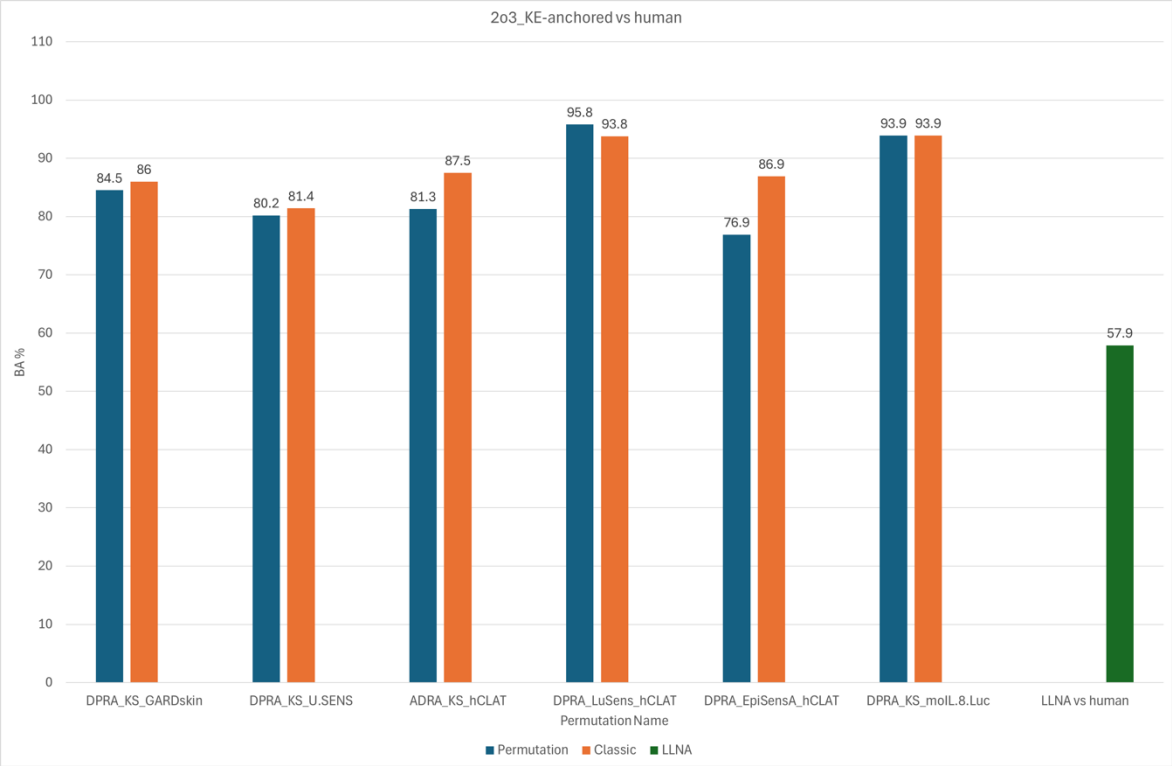

Table 2. Performance for each of the KE-anchored DA permutations compared to the classic 2o3 when evaluated against the human reference dataset. LLNA against the human for the maximal dataset is also included for reference.

Iteration	2o3 permutation	BA (%)	Acc (%)	Sens (%)	Spec (%)	Total	TP	FP	TN	FN
1	Classic 2o3 (all data from GL 497 2023)	88.4	89.1	89.4	87.5	55	42	1	7	5
2	2o3_[DPRA, KeratinoSens, h-CLAT]	87.5	87.5	87.5	87.5	48	35	1	7	5
2	2o3_[ADRA, KeratinoSens, h-CLAT]	81.3	85.4	87.5	75	48	35	2	6	5
3	2o3_[DPRA, KeratinoSens, h-CLAT]	93.8	89.3	87.5	100	28	21	0	4	3
3	2o3_[DPRA, LuSens, h-CLAT]	95.8	92.9	91.7	100	28	22	0	4	2
4	2o3_[DPRA, KeratinoSens, h-CLAT]	86.9	90.5	93.8	80	21	15	1	4	1
4	2o3_[DPRA, EpiSensA, h-CLAT]	76.9	85.7	93.8	60	21	15	2	3	1
5	2o3_[DPRA, KeratinoSens, h-CLAT]	86	87.8	88.6	83.3	41	31	1	5	4
5	2o3_[DPRA, KeratinoSens, GARDskin]	84.5	85.4	85.7	83.3	41	30	1	5	5
6	2o3_[DPRA, KeratinoSens, h-CLAT]	93.9	89.2	87.9	100	37	29	0	4	4
6	2o3_[DPRA, KeratinoSens, IL-8]	93.9	89.2	87.9	100	37	29	0	4	4
7	2o3_[DPRA, KeratinoSens, h-CLAT]	81.4	86.7	87.8	75	45	36	1	3	5
7	2o3_[DPRA, KeratinoSens, U-SENS]	80.2	84.4	85.4	75	45	35	1	3	6
	LLNA vs human	**57.9**	**82.1**	**93.6**	**22.2**	**56**	**44**	**7**	**2**	**3**

BA = balanced accuracy = (Sens + Spec) / 2
Acc = Accuracy = (TP + TN) / Total
Sens = Sensitivity = TP / ((TP + FN)
Spec = Specificity = TN / (TN + FP)
TP = True Positives, FP = False Positives, TN = True Negatives, FN = False Negatives
LLNA vs human added at bottom row for comparison against DAs
Red text = For comparison - results as reported in GL 497 in 2023 for the classic 2o3. Iteration = repeating instruction (loop) from algorithm used to calculate each permutation in turn. Allows for easy comparison of permutation against 'classic' dataset.

Maximal Common Dataset

As a reminder, for the maximal common dataset, if there lacked a data point for the method under evaluation, these chemicals were not filtered out as was done for the KE-anchored datasets. This was to avoid creating datasets that were too small for adequate analyses to be conducted. This resulted in slight differences in numbers for similar permutations in the KE-anchored datasets, as a decision could be made on the two classic methods without the input of the replacement similar method. It is again noted that overall BA was comparable across the comparisons, with no more than +/- 6% between the classic and the permutation of interest. All permutations had a BA greater than 80% (LLNA) and 72% (human).

LLNA

Take home message: All alternate DAs evaluated here performed comparatively to the classic 2o3 (provided as a comparison for each permutation; Table 3) when evaluated using a common dataset. The percentage difference in performance between the permutation and the classic 2o3 against the LLNA reference dataset was generally within +/- 5% for all performance metrics, with many permutations performing slightly better than the classic 2o3.

A few 2o3 DA permutations had lower sensitivity or specificity. It should be remembered that one or two misclassifications (false positive) can have a significant impact on the specificity due to the low numbers of true negatives (TN) and false positives (FP).

Table 3. Performance for all DA permutations compared to the classic 2o3 when evaluated against the LLNA reference dataset, for maximal available data.

Iteration	2o3 permutation	BA (%)	Acc (%)	Sens (%)	Spec (%)	Total	TP	FP	TN	FN
1	Classic 2o3 (all data from GL 497 2023)	83.5	82.8	82.4	84.6	134	89	4	22	19
5	2o3_[DPRA, KeratinoSens, h-CLAT]	85.8	84.8	84.2	87.5	125	85	3	21	16
5	2o3_[DPRA, KeratinoSens, GARDskin]	85.3	84	83.2	87.5	125	84	3	21	17
7	2o3_[DPRA, KeratinoSens, h-CLAT]	83.3	82.9	82.7	84	129	86	4	21	18
7	2o3_[DPRA, KeratinoSens, U-SENS]	80.9	81.4	81.7	80	129	85	5	20	19
2	2o3_[DPRA, KeratinoSens, h-CLAT]	85.4	84.4	83.8	87	128	88	3	20	17
2	2o3_[ADRA, KeratinoSens, h-CLAT]	87.3	87.5	87.6	87	128	92	3	20	13
3	2o3_[DPRA, KeratinoSens, h-CLAT]	96.7	94.4	93.3	100	90	70	0	15	5
3	2o3_[DPRA, LuSens, h-CLAT]	93.3	93.3	93.3	93.3	90	70	1	14	5
4	2o3_[DPRA, KeratinoSens, h-CLAT]	92.3	91.1	90.5	94.1	101	76	1	16	8
4	2o3_[DPRA, EpiSensA, h-CLAT]	89.4	94.1	96.4	82.4	101	81	3	14	3
6	2o3_[DPRA, KeratinoSens, h-CLAT]	84.8	83.7	83.2	86.4	123	84	3	19	17
6	2o3_[DPRA, KeratinoSens, IL-8]	85.3	84.6	84.2	86.4	123	85	3	19	16
8	2o3_[DPRA, KeratinoSens, h-CLAT]	93.9	89.6	87.9	100	77	58	0	11	8
8	2o3_[ADRA, EpiSensA, GARDskin]	93.2	94.8	95.5	90.9	77	63	1	10	3
9	2o3_[DPRA, KeratinoSens, h-CLAT]	86.3	87.6	88.4	84.2	105	76	3	16	10
9	2o3_[ADRA, EpiSensA, U-SENS]	83.9	90.5	94.2	73.7	105	81	5	14	5
10	2o3_[DPRA, KeratinoSens, h-CLAT]	93.7	89.4	87.4	100	104	76	0	17	11
10	2o3_[ADRA, EpiSensA, h-CLAT]	88.3	92.3	94.3	82.4	104	82	3	14	5
11	2o3_[DPRA, KeratinoSens, h-CLAT]	93.8	89.2	87.7	100	83	64	0	10	9
11	2o3_[ADRA, EpiSensA, IL-8]	87.9	94	95.9	80	83	70	2	8	3
12	2o3_[DPRA, KeratinoSens, h-CLAT]	86.9	84.9	83.8	90	119	83	2	18	16
12	2o3_[ADRA, KeratinoSens, GARDskin]	88.4	87.4	86.9	90	119	86	2	18	13
13	2o3_[DPRA, KeratinoSens, h-CLAT]	84	84.8	85.3	82.6	125	87	4	19	15
13	2o3_[ADRA, KeratinoSens, U-SENS]	80.1	84	86.3	73.9	125	88	6	17	14
14	2o3_[DPRA, KeratinoSens, h-CLAT]	86.8	85	84.2	89.5	120	85	2	17	16
14	2o3_[ADRA, KeratinoSens, IL-8]	88.3	87.5	87.1	89.5	120	88	2	17	13
15	2o3_[DPRA, KeratinoSens, h-CLAT]	92.8	87.5	85.5	100	80	59	0	11	10
15	2o3_[ADRA, LuSens, GARDskin]	94.9	91.3	89.9	100	80	62	0	11	7
16	2o3_[DPRA, KeratinoSens, h-CLAT]	88.3	88.3	88.4	88.2	103	76	2	15	10
16	2o3_[ADRA, LuSens, U-SENS]	83.6	88.3	90.7	76.5	103	78	4	13	8
17	2o3_[DPRA, KeratinoSens, h-CLAT]	94.9	91	89.9	100	78	62	0	9	7
17	2o3_[ADRA, LuSens, IL-8]	90.8	92.3	92.8	88.9	78	64	1	8	5
18	2o3_[DPRA, KeratinoSens, h-CLAT]	89.7	89.5	89.4	90	76	59	1	9	7
18	2o3_[DPRA, EpiSensA, GARDskin]	92.7	94.7	95.5	90	76	63	1	9	3
19	2o3_[DPRA, KeratinoSens, h-CLAT]	87	89.4	90.7	83.3	104	78	3	15	8
19	2o3_[DPRA, EpiSensA, U-SENS]	83.8	91.3	95.3	72.2	104	82	5	13	4
20	2o3_[DPRA, KeratinoSens, h-CLAT]	90.2	90.4	90.4	90	83	66	1	9	7
20	2o3_[DPRA, EpiSensA, IL-8]	88.6	95.2	97.3	80	83	71	2	8	2

21	2o3_[DPRA, KeratinoSens, h-CLAT]	95.4	92	90.8	100	75	59	0	10	6
21	2o3_[DPRA, LuSens, GARDskin]	94.6	90.7	89.2	100	75	58	0	10	7
22	2o3_[DPRA, KeratinoSens, h-CLAT]	86.5	90	91.7	81.3	100	77	3	13	7
22	2o3_[DPRA, LuSens, U-SENS]	83.3	89	91.7	75	100	77	4	12	7
23	2o3_[DPRA, KeratinoSens, h-CLAT]	96.9	94.6	93.8	100	74	61	0	9	4
23	2o3_[DPRA, LuSens, IL-8]	91.4	93.2	93.8	88.9	74	61	1	8	4
24	2o3_[DPRA, KeratinoSens, h-CLAT]	93.1	88.4	86.3	100	95	69	0	15	11
24	2o3_[ADRA, LuSens, h-CLAT]	92.3	91.6	91.3	93.3	95	73	1	14	7

BA = Balanced accuracy = (Sens + Spec) / 2
Acc = Accuracy = (TP + TN) / Total
Sens = Sensitivity = TP / ((TP + FN)
Spec = Specificity = TN / (TN + FP)
TP = True Positives, FP = False Positives, TN = True Negatives, FN = False Negatives.
Red text = For comparison - results as reported in GL 497 in 2023 for the classic 2o3. Iteration = repeating instruction (loop) from algorithm used to calculate each permutation in turn. Allows for easy comparison of permutation against 'classic' dataset.

Human

Take home message: All alternate DAs evaluated here performed similarly to the classic 2o3 (provided as a comparison for each permutation; Table 4) when evaluated using a common dataset. The percentage difference in performance between the permutations and the classic against the human reference dataset was generally within +/- 5% for all performance metrics, with many permutations performing slightly better than the classic 2o3. Two combinations containing EpiSensA and IL-8 had much greater variability between the permutation and classic DAs (iterations 23 and 24). A few permutations had lower sensitivity or specificity; however, it should be remembered that one or two misclassifications (false positive) can have a significant impact on the specificity due to the low numbers of true negatives (TN) and false positives (FP). As a reminder, the chemicals evaluated in these datasets all had a datapoint for the method under consideration. The comparison of the LLNA to the human reference data (maximal number of chemicals) is also provided. **It should be noted that the balanced accuracy (58%) and specificity of the LLNA (22%) is significantly lower than any of the evaluated permutations.**

Table 4. Performance for all DA permutations compared to the classic 2o3 when evaluated against the human reference dataset, for maximal available data. LLNA vs human is provided as a reference.

Iteration	2o3 permutation	BA (%)	Acc (%)	Sens (%)	Spec (%)	Total	TP	FP	TN	FN
1	Classic 2o3 (all data from GL 497 2023)	88.4	89.1	89.4	87.5	55	42	1	7	5
5	2o3_[DPRA, KeratinoSens, h-CLAT]	88.2	88.7	88.9	87.5	53	40	1	7	5
5	2o3_[DPRA, KeratinoSens, GARDskin]	87.1	86.8	86.7	87.5	53	39	1	7	6
7	2o3_[DPRA, KeratinoSens, h-CLAT]	88.1	88.5	88.6	87.5	52	39	1	7	5
7	2o3_[DPRA, KeratinoSens, U-SENS]	86.9	86.5	86.4	87.5	52	38	1	7	6
2	2o3_[DPRA, KeratinoSens, h-CLAT]	88.2	88.7	88.9	87.5	53	40	1	7	5
2	2o3_[ADRA, KeratinoSens, h-CLAT]	81.9	86.8	88.9	75	53	40	2	6	5

3	2o3_[DPRA, KeratinoSens, h-CLAT]	96.1	92.9	92.1	100	42	35	0	4	3
3	2o3_[DPRA, LuSens, h-CLAT]	97.4	95.2	94.7	100	42	36	0	4	2
4	2o3_[DPRA, KeratinoSens, h-CLAT]	90.3	95.2	97.2	83.3	42	35	1	5	1
4	2o3_[DPRA, EpiSensA, h-CLAT]	81.9	92.9	97.2	66.7	42	35	2	4	1
6	2o3_[DPRA, KeratinoSens, h-CLAT]	89	90	90.5	87.5	50	38	1	7	4
6	2o3_[DPRA, KeratinoSens, IL-8]	89	90	90.5	87.5	50	38	1	7	4
8	2o3_[DPRA, KeratinoSens, h-CLAT]	87.1	89.7	90.9	83.3	39	30	1	5	3
8	2o3_[ADRA, EpiSensA, GARDskin]	80.3	89.7	93.9	66.7	39	31	2	4	2
9	2o3_[DPRA, KeratinoSens, h-CLAT]	84.7	85.7	86.1	83.3	42	31	1	5	5
9	2o3_[ADRA, EpiSensA, U-SENS]	86.1	88.1	88.9	83.3	42	32	1	5	4
10	2o3_[DPRA, KeratinoSens, h-CLAT]	88.4	90.2	91.2	85.7	41	31	1	6	3
10	2o3_[ADRA, EpiSensA, h-CLAT]	82.8	90.2	94.1	71.4	41	32	2	5	2
11	2o3_[DPRA, KeratinoSens, h-CLAT]	95	91.4	90	100	35	27	0	5	3
11	2o3_[ADRA, EpiSensA, IL-8]	76.7	88.6	93.3	60	35	28	2	3	2
12	2o3_[DPRA, KeratinoSens, h-CLAT]	87.9	88.2	88.4	87.5	51	38	1	7	5
12	2o3_[ADRA, KeratinoSens, GARDskin]	80.5	84.3	86	75	51	37	2	6	6
13	2o3_[DPRA, KeratinoSens, h-CLAT]	87	88	88.4	85.7	50	38	1	6	5
13	2o3_[ADRA, KeratinoSens, U-SENS]	85.9	86	86	85.7	50	37	1	6	6
14	2o3_[DPRA, KeratinoSens, h-CLAT]	87.5	87.5	87.5	87.5	48	35	1	7	5
14	2o3_[ADRA, KeratinoSens, IL-8]	81.3	85.4	87.5	75	48	35	2	6	5
15	2o3_[DPRA, KeratinoSens, h-CLAT]	86.4	87	87.2	85.7	46	34	1	6	5
15	2o3_[ADRA, LuSens, GARDskin]	78	82.6	84.6	71.4	46	33	2	5	6
16	2o3_[DPRA, KeratinoSens, h-CLAT]	93.4	88.4	86.8	100	43	33	0	5	5
16	2o3_[ADRA, LuSens, U-SENS]	92.1	86	84.2	100	43	32	0	5	6
17	2o3_[DPRA, KeratinoSens, h-CLAT]	94.1	89.7	88.2	100	39	30	0	5	4
17	2o3_[ADRA, LuSens, IL-8]	85.6	89.7	91.2	80	39	31	1	4	3
18	2o3_[DPRA, KeratinoSens, h-CLAT]	88.5	94.7	97	80	38	32	1	4	1
18	2o3_[DPRA, EpiSensA, GARDskin]	78.5	92.1	97	60	38	32	2	3	1
19	2o3_[DPRA, KeratinoSens, h-CLAT]	83.7	90.7	92.3	75	43	36	1	3	3
19	2o3_[DPRA, EpiSensA, U-SENS]	96.2	93	92.3	100	43	36	0	4	3
20	2o3_[DPRA, KeratinoSens, h-CLAT]	96.9	94.4	93.8	100	36	30	0	4	2
20	2o3_[DPRA, EpiSensA, IL-8]	71.9	88.9	93.8	50	36	30	2	2	2
21	2o3_[DPRA, KeratinoSens, h-CLAT]	87.7	90.9	92.1	83.3	44	35	1	5	3
21	2o3_[DPRA, LuSens, GARDskin]	86.4	88.6	89.5	83.3	44	34	1	5	4
22	2o3_[DPRA, KeratinoSens, h-CLAT]	96.1	93	92.1	100	43	35	0	5	3
22	2o3_[DPRA, LuSens, U-SENS]	94.7	90.7	89.5	100	43	34	0	5	4
Iteration	**2o3 permutation**	**BA (%)**	**Acc (%)**	**Sens (%)**	**Spec (%)**	**Total**	**TP**	**FP**	**TN**	**FN**
23	2o3_[DPRA, KeratinoSens, h-CLAT]	95.5	91.9	90.9	100	37	30	0	4	3
23	2o3_[DPRA, LuSens, IL-8]	97	94.6	93.9	100	37	31	0	4	2
24	2o3_[DPRA, KeratinoSens, h-CLAT]	92.9	87.5	85.7	100	40	30	0	5	5
24	2o3_[ADRA, LuSens, h-CLAT]	84.3	87.5	88.6	80	40	31	1	4	4
	LLNA vs human	**57.9**	**82.1**	**93.6**	**22.2**	**56**	**44**	**7**	**2**	**3**

BA = Balanced accuracy = (Sens + Spec) / 2

Acc = Accuracy = (TP + TN) / Total
Sens = Sensitivity = TP / ((TP + FN)
Spec = Specificity = TN / (TN + FP)
TP = True Positives, FP = False Positives, TN = True Negatives, FN = False Negatives
LLNA vs human added at bottom row for comparison against DAs. Red text = For comparison - results as reported in GL 497 in 2023 for the classic 2o3. Iteration = repeating instruction (loop) from algorithm used to calculate each permutation in turn. Allows for easy comparison of permutation against 'classic' dataset.

3 ITS DA

The majority of the alternate ITS DAs predict skin sensitisation hazard comparably to the classic DAs, when using a common dataset between each permutation. When evaluating all of the different permutations, for comparisons to LLNA, the balanced accuracy was above 78%, with no more than +/- 5% when comparing to the classic ITS. When evaluating for potency, the accuracy is generally above 70% for most permutations, with no more than +/- 3% when comparing to the classic ITS, with the lowest accuracy 66% (DPRA, U-SENS, OECD Toolbox).

When compared to human data, the accuracy is a little more variable, with most permutations having +/- 5% difference between the permutation and the classic ITS. When evaluating for balanced accuracy (potency), most permutations are above 60%, equivalent to the performance of the LLNA. With the limited amount of human data, these comparisons have a small total number of chemicals evaluated. Thus, one or two misclassifications can greatly affect the performance. This should be considered with the overall performance of the different permutations.

When considering these datasets, the majority of ITS permutations predict skin sensitisation hazard as well as or better than the classic permutation. See tabs in the Annex 2 Excel file for full tables of data:

- ITS_KE-anchored vs LLNAhaz (Table 5)
- ITS_KE-anchored vs humanhaz (Table 6)
- ITS_maximal vs LLNAhaz (Table 9)
- ITS_maximal vs humanhaz (Table 10)

The majority of the ITS permutations predict skin sensitisation GHS sub-categories with high accuracy (most >70% for LLNA and > 65% for human) and are similar to the ITSv1 GHS sub-category predictions for the same dataset as well. See tabs in the relevant Excel file for full tables of data:

- ITS_KE-anchored vs LLNApot (Table 7)
- ITS_KE-anchored vs humanpot (Table 8)
- ITS_maximal vs LLNApot (Table 11)
- ITS_maximal vs humanpot (Table 12)

Regardless of the dataset used as the basis for comparison, all of the evaluated ITS permutations outperform the LLNA when compared to human reference data for hazard and the vast majority outperform the LLNA when compared to human reference data for potency.

KE-Anchored Dataset

When evaluating permutations where only one similar method was replaced in the DA for the ITS, there is equivalent performance of the DA. When considering hazard, generally, the similar methods also had similar sensitivity and specificity when compared to both the LLNA

and human reference datasets, and all performed better than the LLNA against the human reference data. For potency, the percentage of under- and over-predictions was consistent between the permutations and the classic.

Overall impacts on balanced accuracy (BA) (for hazard), and accuracy (Acc) (for potency) were low when compared to both the LLNA and human reference datasets with, in most cases, less than 5% difference between the permutation and classic.

LLNA Hazard

Take home message: All alternate DAs evaluated here performed similarly to the classic ITS (provided as a comparison for each permutation; Table 5) when evaluated using a common dataset. The percentage difference in performance between the KE-anchored and the classic ITS against the LLNA reference dataset was generally within 3% for all performance metrics, with one permutation performing slightly better than the classic ITS. Overall, performance of the alternate DAs with a similar method was equivalent if not better across the permutations.

Figure 3. Comparison balanced accuracy against LLNA for single-method (KE-anchored) replacements (blue) to the same chemical list for the classic 2o3 (orange).

There is minimal difference between the two DAs for each comparison, showing that similar methods can be used within the ITS DA for hazard assessment.

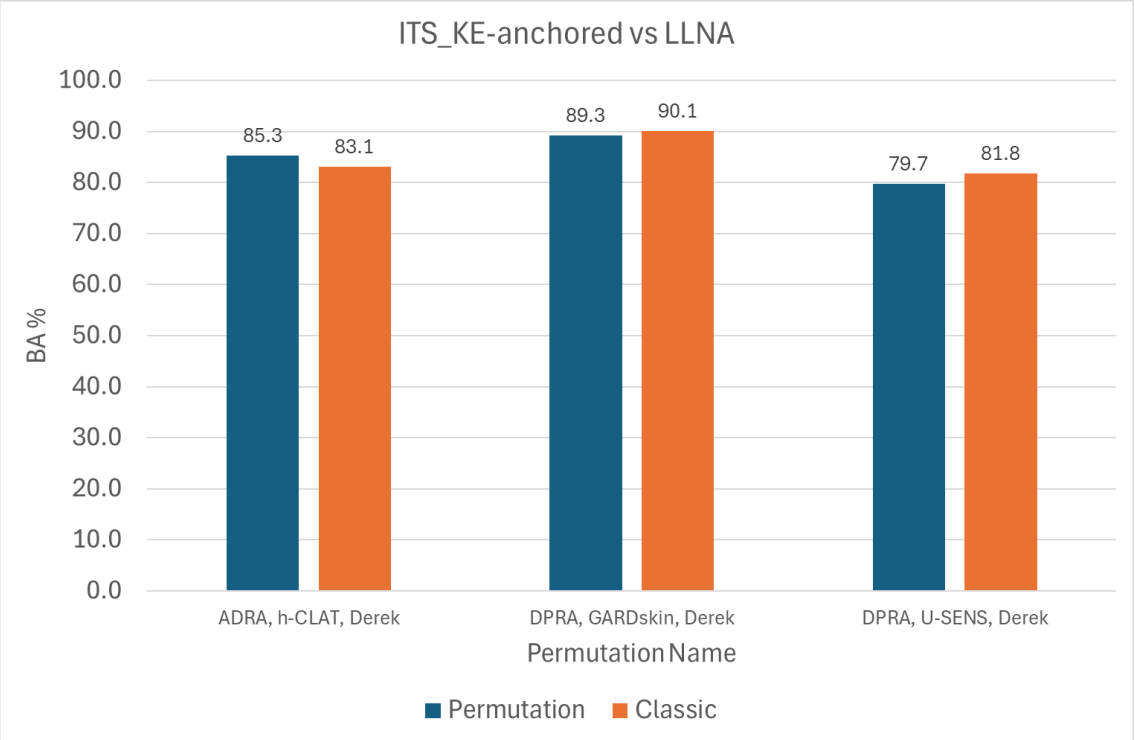

Table 5. Performance for each of the KE-anchored DA permutations compared to the ITSv1 when evaluated against the LLNA reference dataset.

Iteration	ITS permutation	BA (%)	Acc (%)	Sens (%)	Spec (%)	Total	TP	FP	TN	FN
-	Classic ITSv1 (all data from GL 497 2023)	80.7	87.4	91.5	70	159	118	9	21	11
-	Classic ITSv2 (all data from GL 497 2023)	79.8	87.8	92.9	66.7	156	117	10	20	9
5	DPRA, h-CLAT, Derek	83.1	88.3	91.3	75.0	154	115	7	21	11
5	ADRA, h-CLAT, Derek	85.3	89.6	92.1	78.6	154	116	6	22	10
10	DPRA, h-CLAT, Derek	90.1	90.1	90.2	90.0	71	55	1	9	6
10	DPRA, GARDskin, Derek	89.3	88.7	88.5	90.0	71	54	1	9	7
11	DPRA, h-CLAT, Derek	81.8	87.7	91.3	72.4	155	115	8	21	11
11	DPRA, U-SENS, Derek	79.7	86.5	90.5	69.0	155	114	9	20	12

BA = Balanced accuracy = (Sens + Spec) / 2
Acc = Accuracy = (TP + TN) / Total
Sens = Sensitivity = TP / ((TP + FN)
Spec = Specificity = TN / (TN + FP)
TP = True Positives, FP = False Positives, TN = True Negatives, FN = False Negatives
Red text = For comparison - results as reported in GL 497 in 2023 for the classic ITS. Iteration = repeating instruction (loop) from algorithm used to calculate each permutation in turn. Allows for easy comparison of permutation against 'classic' dataset.

Human Hazard

Take home message: All alternate DAs evaluated here performed similarly to the classic ITS (provided as a comparison for each permutation; Table 6) when evaluated using a common dataset. The percentage difference in performance between the permutation and the classic against the human reference dataset was generally within +/- 3% for all performance metrics, with one permutation performing slightly better than the classic ITS. Specificity was generally low across all permutations and the classic ITS. **It should be noted that the balanced accuracy and specificity of the LLNA is significantly lower than any of the evaluated permutations.**

Figure 4. Comparison balanced accuracy against human data for single-method (KE-anchored) replacements (blue) to the same chemical list for the classic 2o3 (orange).

There is minimal difference between the two DAs for each comparison, showing that similar methods can be used within the ITS DA for hazard assessment. LLNA against human for the maximal dataset is also included for reference (green).

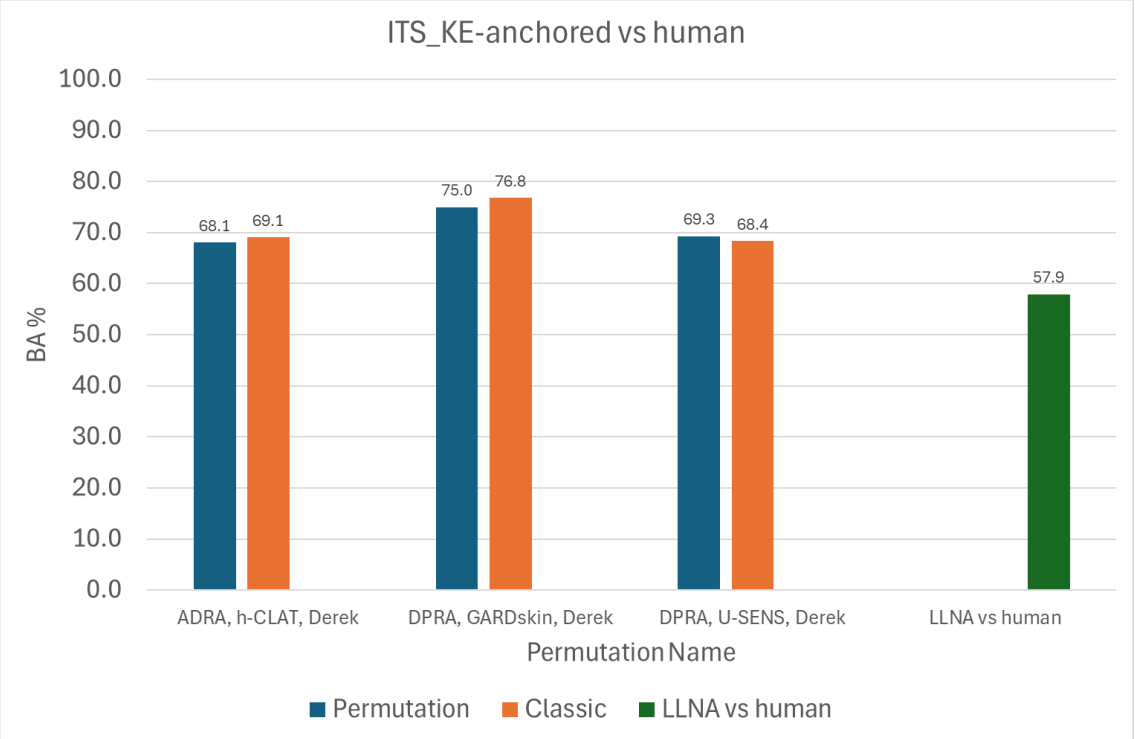

Table 6. Performance for each of the KE-anchored DA permutations compared to the ITSv1 when evaluated against the human reference dataset. LLNA against the human for the maximal dataset is also included for reference.

Iteration	ITS permutation	BA (%)	Acc (%)	Sens (%)	Spec (%)	Total	TP	FP	TN	FN
-	Classic' ITSv1 (all data from GL 497 2023)	68.6	85.9	92.7	44.4	64	51	5	4	4
-	Classic' ITSv2 (all data from GL 497 2023)	69.4	87.1	94.3	44.4	62	50	5	4	3
5	DPRA, h-CLAT, Derek	69.1	86.0	93.8	44.4	57	45	5	4	3
5	ADRA, h-CLAT, Derek	68.1	84.2	91.7	44.4	57	44	5	4	4
10	DPRA, h-CLAT, Derek	76.8	88.6	96.4	57.1	35	27	3	4	1
10	DPRA, GARDskin, Derek	75.0	85.7	92.9	57.1	35	26	3	4	2
11	DPRA, h-CLAT, Derek	68.4	85.2	92.3	44.4	61	48	5	4	4
11	DPRA, U-SENS, Derek	69.3	86.7	94.1	44.4	60	48	5	4	3
na	**LLNA vs human**	**57.9**	**82.1**	**93.6**	**22.2**	**56**	**44**	**7**	**2**	**3**

BA = Balanced accuracy = (Sens + Spec) / 2
Acc = Accuracy = (TP + TN) / Total
Sens = Sensitivity = TP / ((TP + FN)
Spec = Specificity = TN / (TN + FP)
TP = True Positives, FP = False Positives, TN = True Negatives, FN = False Negatives
LLNA vs human added at bottom row for comparison against DAs
Red text = For comparison - results as reported in GL 497 in 2023 for the classic ITS. Iteration = repeating instruction (loop) from algorithm used to calculate each permutation in turn. Allows for easy comparison of permutation against 'classic' dataset.

LLNA Potency

Take home message: All alternate DAs evaluated here performed similarly to the classic ITS (provided as a comparison for each permutation; Table 7) when evaluated using a common dataset. The percentage difference in performance between the KE-anchored and the classic ITS against the LLNA reference dataset was generally within 3% for accuracy, with one permutation performing slightly better than the classic ITS. Where larger performance gaps between the permutation and the classic are noted, the permutation performed better in most cases. It should be remembered that one or two misclassifications can also result in these percentage differences. The permutations were consistently around 70% for accuracy, with typically no more than 15% over/under-predicted. Overall, performance with a similar method was equivalent if not better across the permutations.

Figure 5. Comparison of accuracy for potency estimates against LLNA for single-method (KE-anchored) replacements (blue) to the same chemical list for the ITSv1 (orange).

There is minimal difference between the two DAs for each comparison, showing that similar methods can be used within the ITS DA for potency assessment.

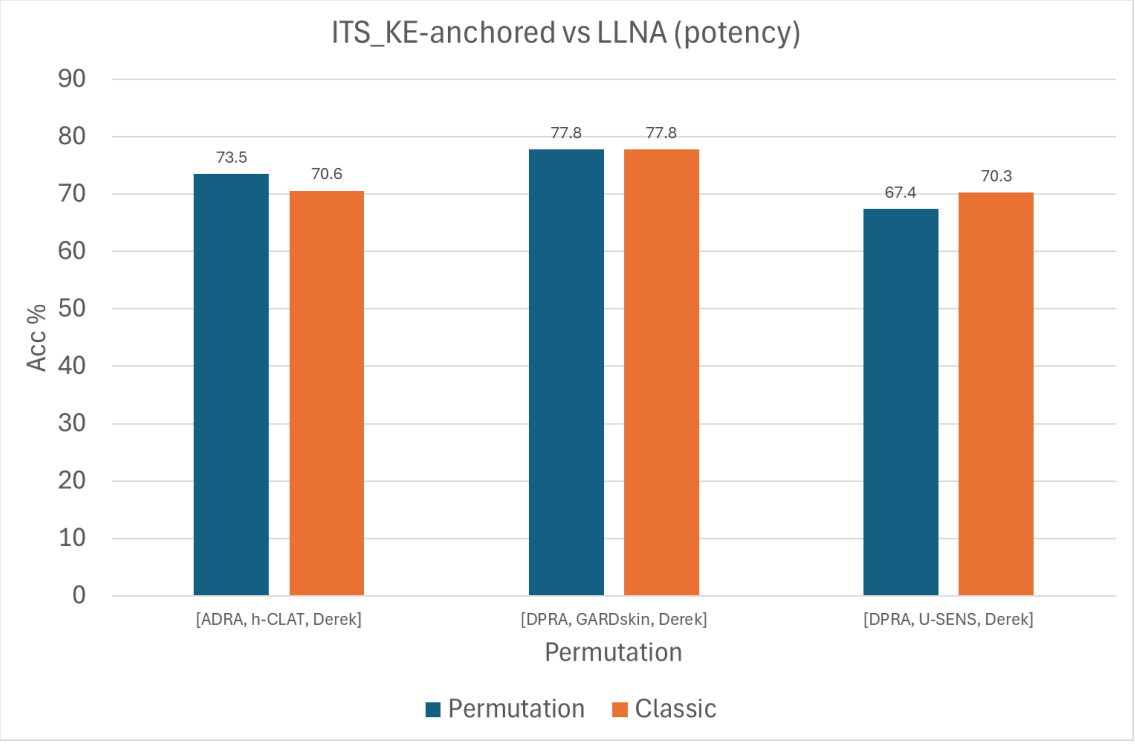

Table 7. Performance for each of the KE-anchored DA permutations compared to the ITSv1 when evaluated against the LLNA reference dataset when referencing potency.

Iteration	ITS permutation	Acc	Total	Correct.1A	Correct.1B	Correct.NC	Over (1B pred as 1A)	Over (NC pred as 1B)	Over (NC pred as 1A)	Under (1A pred as 1B)	Under (1B pred as NC)	Under (1A pred as NC)	Under.%	Over.%	Correct.%	Correct.n	Under.n	Over.n
-	Classic ITSv1 (all data from GL 497 2023)	70.2	141	28	50	21	12	9	0	10	11	0	14.9	14.9	70.2	99	21	21
-	Classic ITSv2 (all data from GL 497 2023)	70.6	136	26	49	21	12	9	0	10	9	0	14	15.4	70.6	96	19	21
5	DPRA, h-CLAT, Derek	70.6	136	28	47	21	12	7	0	10	11	0	15.4	14	70.6	96	21	19
5	ADRA, h-CLAT, Derek	73.5	136	28	50	22	11	6	0	10	9	0	14	12.5	73.5	100	19	17
10	DPRA, h-CLAT, Derek	77.8	63	19	21	9	3	1	0	4	6	0	15.9	6.3	77.8	49	10	4
10	DPRA, GARDskin, Derek	77.8	63	21	19	9	4	1	0	2	7	0	14.3	7.9	77.8	49	9	5
11	DPRA, h-CLAT, Derek	70.3	138	27	49	21	12	8	0	10	11	0	15.2	14.5	70.3	97	21	20
11	DPRA, U-SENS, Derek	67.4	138	26	47	20	14	9	0	11	11	0	15.9	16.7	67.4	93	22	23

Acc = Accuracy = (Correct predictions) / Total
Correct = predictions correctly predicted by the DA as 1A/1B/Not Classified (NC)
Over = predictions misclassified by the DA, NC predicted to be 1B or 1A, 1B predicted to be 1A
Under = predictions misclassified by the DA, 1A predicted to be 1B or NC, 1B predicted to be NC
Red text = For comparison - results as reported in GL 497 in 2023 for the classic ITS. Iteration = repeating instruction (loop) from algorithm used to calculate each permutation in turn. Allows for easy comparison of permutation against 'classic' dataset

Human Potency

Take home message: All alternate DAs evaluated here performed similarly to the classic ITS (provided as a comparison for each permutation; Table 8) when evaluated using a common dataset. The percentage difference in performance between the permutation and the classic against the human reference dataset was generally within +/- 5% for accuracy. All permutations were greater than 65% accuracy except for where ADRA replaced DPRA, which was 63.5%, however, this is due to 1-2 misclassifications – and is still higher accuracy than that of the LLNA. Importantly, over/under-predictions are similar between alternate DAs and the classic ITS. Overall, performance with a similar method was equivalent across the permutations. The comparison of the LLNA to the human reference data (maximal number of chemicals) is also provided. **It should be noted that the accuracy of the LLNA is lower than any of the evaluated permutations.**

Figure 6. Comparison of accuracy for potency estimates against human for single-method (KE-anchored) replacements (blue) to the same chemical list for the ITSv1 (orange).

There is minimal difference between the two DAs for each comparison, showing that similar methods can be used within the ITS DA for potency assessment. Accuracy for LLNA vs human is provided as a reference (green).

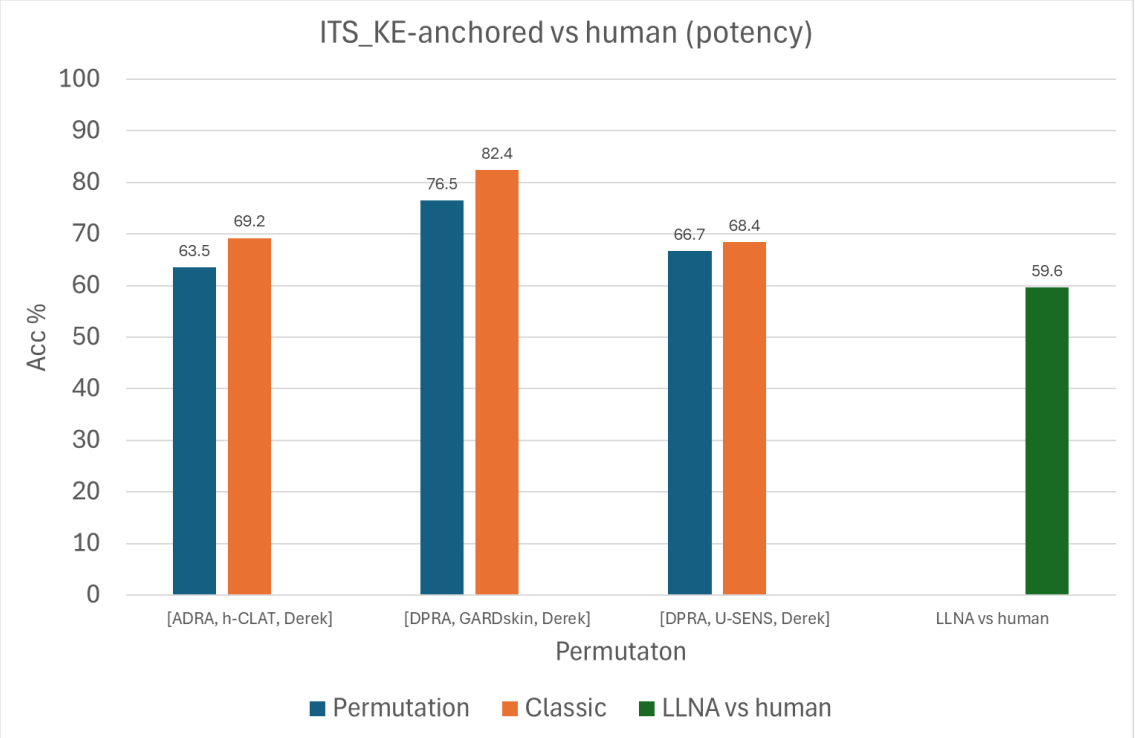

Table 8. Performance for each of the KE-anchored DA permutations compared to the ITSv1 when evaluated against the human reference dataset when estimating potency. LLNA against the human for the maximal dataset is also included for reference.

Iteration	ITS permutation	Acc	Total	Correct. 1A	Correct. 1B	Correct. NC	Over (1B pred as 1A)	Over (NC pred as 1B)	Over (NC pred as 1A)	Under (1A pred as 1B)	Under (1B pred as NC)	Under (1A pred as NC)	Under.%	Over. %	Correct. %	Correct.n	Under. n	Over .n
-	Classic' ITSv1 (all data from GL 497 2023)	68.3	60	13	24	4	3	5	0	7	4	0	18.3	13.3	68.3	41	11	8
-	Classic' ITSv2 (all data from GL 497 2023)	70.2	57	12	24	4	3	5	0	6	3	0	15.8	14	70.2	40	9	8
5	DPRA, h-CLAT, Derek	69.2	52	13	19	4	3	5	0	5	3	0	15.4	15.4	69.2	36	8	8
5	ADRA, h-CLAT, Derek	63.5	52	12	17	4	4	5	0	6	4	0	19.2	17.3	63.5	33	10	9
10	DPRA, h-CLAT, Derek	82.4	34	9	15	4	1	3	0	1	1	0	5.9	11.8	82.4	28	2	4
10	DPRA, GARDskin, Derek	76.5	34	9	13	4	2	3	0	1	2	0	8.8	14.7	76.5	26	3	5
11	DPRA, h-CLAT, Derek	68.4	57	13	22	4	3	5	0	6	4	0	17.5	14	68.4	39	10	8
11	DPRA, U-SENS, Derek	66.7	57	11	23	4	3	5	0	8	3	0	19.3	14	66.7	38	11	8
	LLNA vs human	**61.4**	**44**	**9**	**17**	**1**	**3**	**5**	**0**	**6**	**3**	**0**	**20.5**	**18.2**	**61.4**	**27**	**9**	**8**

Acc = Accuracy = (Correct predictions) / Total
Correct = predictions correctly predicted by the DA as 1A/1B/Not Classified (NC)
Over = predictions misclassified by the DA, NC predicted to be 1B or 1A, 1B predicted to be 1A
Under = predictions misclassified by the DA, 1A predicted to be 1B or NC, 1B predicted to be NC
Red text = For comparison - results as reported in GL 497 in 2023 for the classic ITS. Iteration = repeating instruction (loop) from algorithm used to calculate each permutation in turn. Allows for easy comparison of permutation against 'classic' dataset.

Maximal Common Dataset

As a reminder, for the maximal common dataset, if there lacked a data point for the method under evaluation, these chemicals were not filtered out as was done for the KE-anchored datasets. This was to avoid creating datasets that were too small for adequate analyses to be conducted. This resulted in slight differences in numbers of chemicals for similar permutations in the KE-anchored datasets, as a decision can be made in the DAs using partial information sources (under specific circumstances detailed in GL 497).

Overall BA (for hazard) and accuracy (for potency) was similar across the alternate DAs, with generally no more than 6% between the classic and the permutation of interest. Frequently where the larger differences were noted, the permutation performed better than the classic ITS DA. When considering hazard, generally, the similar methods also had similar sensitivity and specificity when compared to both the LLNA and human reference datasets, and all performed better than the LLNA against the human reference data. For potency, the percentage of under- and over-predictions was consistent between the alternate DAs and the classic DAs, with the exception of several DA permutations which are described more in the ITS underpredictions section.

LLNA Hazard

Take home message: All alternate DAs evaluated here performed similarly to the classic ITS (provided as a comparison for each permutation; Table 9) when evaluated using a common dataset. The percentage difference in performance between the alternate DA and the classic ITS against the LLNA reference dataset was generally within 5% for all performance metrics, with many permutations performing slightly better than the classic ITS.

Table 9. Performance for all DA permutations compared to the ITSv1 when evaluated against the LLNA reference dataset, for maximal available data.

Iteration	ITS permutation	BA (%)	Acc (%)	Sens (%)	Spec (%)	Total	TP	FP	TN	FN
-	Classic' ITSv1 (all data from GL 497 2023)	80.7	87.4	91.5	70	159	118	9	21	11
-	Classic' ITSv2 (all data from GL 497 2023)	79.8	87.8	92.9	66.7	156	117	10	20	9
5	DPRA, h-CLAT, Derek	80.7	87.4	91.5	70	159	118	9	21	11
5	ADRA, h-CLAT, Derek	82.8	88.7	92.2	73.3	159	119	8	22	10
10	DPRA, h-CLAT, Derek	81.9	92	94.6	69.2	125	106	4	9	6
10	DPRA, GARDskin, Derek	81.5	91.2	93.8	69.2	125	105	4	9	7
11	DPRA, h-CLAT, Derek	81.9	88	91.5	72.4	158	118	8	21	11
11	DPRA, U-SENS, Derek	79.8	86.7	90.7	69	158	117	9	20	12
9	DPRA, h-CLAT, Derek	83.5	89	92.1	75	154	116	7	21	10
9	ADRA, h-CLAT, OECD.TB	82.1	89	92.9	71.4	154	117	8	20	9

Table continues to the next page

Iteration	ITS permutation	BA (%)	Acc (%)	Sens (%)	Spec (%)	Total	TP	FP	TN	FN
18	DPRA, h-CLAT, Derek	88.2	93.4	94.6	81.8	122	105	2	9	6
18	DPRA, GARDskin, OECD.TB	87.8	92.6	93.7	81.8	122	104	2	9	7
19	DPRA, h-CLAT, Derek	82.2	88.3	91.4	73.1	154	117	7	19	11
19	DPRA, U-SENS, OECD.TB	78	86.4	90.6	65.4	154	116	9	17	12
20	DPRA, h-CLAT, Derek	82.6	91.3	93.8	71.4	127	106	4	10	7
20	ADRA, GARDskin, Derek	82.2	90.6	92.9	71.4	127	105	4	10	8
21	DPRA, h-CLAT, Derek	83.2	88.5	91.5	75	157	118	7	21	11
21	ADRA, U-SENS, Derek	81.5	87.9	91.5	71.4	157	118	8	20	11
28	DPRA, h-CLAT, Derek	85.3	91.9	93.7	76.9	124	104	3	10	7
28	ADRA, GARDskin, OECD.TB	84.9	91.1	92.8	76.9	124	103	3	10	8
29	DPRA, h-CLAT, Derek	87	90.1	91.4	82.6	151	117	4	19	11
29	ADRA, U-SENS, OECD.TB	82.3	88.1	90.6	73.9	151	116	6	17	12

BA = balanced accuracy = (Sens + Spec) / 2
Acc = Accuracy = (TP + TN) / Total
Sens = Sensitivity = TP / ((TP + FN)
Spec = Specificity = TN / (TN + FP)
TP = True Positives, FP = False Positives, TN = True Negatives, FN = False Negatives
Red text = For comparison - results as reported in GL 497 in 2023 for the classic ITS. Iteration = repeating instruction (loop) from algorithm used to calculate each permutation in turn. Allows for easy comparison of permutation against 'classic' dataset.

Human Hazard

Take home message: All alternate DAs_evaluated here performed similarly to the classic ITS (provided as a comparison for each permutation; Table 10) when evaluated using a common dataset. The percentage difference in performance between the permutations and the classic against the human reference dataset was generally within 5% for all performance metrics, with many permutations performing better than the classic ITS. The comparison of the LLNA to the human reference data (maximal number of chemicals) is also provided. **It should be noted that the balanced accuracy and specificity of the LLNA is significantly lower than any of the evaluated permutations, even with low specificity of all the permutations, including the classic ITS.**

Table 10. Performance for all DA permutations compared to the ITSv1 when evaluated against the human reference dataset, for maximal available data. LLNA vs human is provided as a reference.

Iteration	ITS permutation	BA (%)	Acc (%)	Sens (%)	Spec (%)	Total	TP	FP	TN	FN
-	Classic ITSv1 (all data from GL 497 2023)	68.6	85.9	92.7	44.4	64	51	5	4	4
-	Classic ITSv2 (all data from GL 497 2023)	69.4	87.1	94.3	44.4	62	50	5	4	3
5	DPRA, h-CLAT, Derek	68.6	85.9	92.7	44.4	64	51	5	4	4
5	ADRA, h-CLAT, Derek	67.7	84.4	90.9	44.4	64	50	5	4	5
10	DPRA, h-CLAT, Derek	68.5	85.7	92.6	44.4	63	50	5	4	4
10	DPRA, GARDskin, Derek	66.7	82.5	88.9	44.4	63	48	5	4	6
11	DPRA, h-CLAT, Derek	68.6	85.9	92.7	44.4	64	51	5	4	4
11	DPRA, U-SENS, Derek	71.4	82.8	87.3	55.6	64	48	4	5	7
9	DPRA, h-CLAT, Derek	68.5	85.7	92.6	44.4	63	50	5	4	4
9	ADRA, h-CLAT, OECD.TB	69.4	87.3	94.4	44.4	63	51	5	4	3
18	DPRA, h-CLAT, Derek	77.5	92.5	97.8	57.1	53	45	3	4	1
18	DPRA, GARDskin, OECD.TB	76.4	90.6	95.7	57.1	53	44	3	4	2
19	DPRA, h-CLAT, Derek	68.4	85.2	92.3	44.4	61	48	5	4	4
19	DPRA, U-SENS, OECD.TB	68.4	85.2	92.3	44.4	61	48	5	4	4
20	DPRA, h-CLAT, Derek	73.9	90.7	97.8	50	54	45	4	4	1
20	ADRA, GARDskin, Derek	70.7	85.2	91.3	50	54	42	4	4	4
21	DPRA, h-CLAT, Derek	68.5	85.7	92.6	44.4	63	50	5	4	4
21	ADRA, U-SENS, Derek	67.6	84.1	90.7	44.4	63	49	5	4	5
28	DPRA, h-CLAT, Derek	73.9	90.4	97.7	50	52	43	4	4	1
28	ADRA, GARDskin, OECD.TB	70.5	84.6	90.9	50	52	40	4	4	4
29	DPRA, h-CLAT, Derek	72	87.9	94	50	58	47	4	4	3
29	ADRA, U-SENS, OECD.TB	70	84.5	90	50	58	45	4	4	5
	LLNA vs human	**57.9**	**82.1**	**93.6**	**22.2**	**56**	**44**	**7**	**2**	**3**

BA = balanced accuracy = (Sens + Spec) / 2
Acc = Accuracy = (TP + TN) / Total
Sens = Sensitivity = TP / ((TP + FN)
Spec = Specificity = TN / (TN + FP)
TP = True Positives, FP = False Positives, TN = True Negatives, FN = False Negatives
LLNA vs human added at bottom row for comparison against DAs
Red text = For comparison - results as reported in GL 497 in 2023 for the classic ITS. Iteration = repeating instruction (loop) from algorithm used to calculate each permutation in turn. Allows for easy comparison of permutation against 'classic' dataset.

LLNA Potency

Take home message: All alternate DAs evaluated here performed similarly to the classic ITS (provided as a comparison for each permutation; Table 11) when evaluated using a common dataset. The percentage difference in performance between the maximal and the classic ITS against the LLNA reference dataset was generally within +/- 2-3% or accuracy, with many permutations performing slightly better than the classic ITS. The permutations were consistently greater than 70% for accuracy, and over/under-predictions similar between permutations and the classic comparison. Overall, performance with a similar method was equivalent if not better than the classic ITS across the permutations.

Table 11. Performance for potency estimates for all DA permutations compared to the ITSv1 when evaluated against the LLNA reference dataset, for maximal available data.

It_	ITS permutation	Acc	Total	Cor. 1A	Cor. 1B	Cor. NC	Over (1B as 1A)	Over (NC as 1B)	Over (NC as 1A)	Und.(1A as 1B)	Und (1B as NC)	Und (1A as NC)	Under %	Over %	Correct %	Correctn	Und.n	Over n
-	Classic' ITSv1 (all data from GL 497 2023)	70.2	141	28	50	21	12	9	0	10	11	0	14.9	14.9	70.2	99	21	21
-	Classic' ITSv2 (all data from GL 497 2023)	70.6	136	26	49	21	12	9	0	10	9	0	14	15.4	70.6	96	19	21
5	DPRA, h-CLAT, Derek	70.2	141	28	50	21	12	9	0	10	11	0	14.9	14.9	70.2	99	21	21
5	ADRA, h-CLAT, Derek	73	141	28	53	22	11	8	0	10	9	0	13.5	13.5	73	103	19	19
10	DPRA, h-CLAT, Derek	78.7	75	19	31	9	3	2	0	5	6	0	14.7	6.7	78.7	59	11	5
10	DPRA, GARDskin, Derek	78.7	75	21	29	9	4	2	0	3	7	0	13.3	8	78.7	59	10	6
11	DPRA, h-CLAT, Derek	70.5	139	27	50	21	12	8	0	10	11	0	15.1	14.4	70.5	98	21	20
11	DPRA, U-SENS, Derek	67.6	139	26	48	20	14	9	0	11	11	0	15.8	16.5	67.6	94	22	23
9	DPRA, h-CLAT, Derek	71.3	136	27	49	21	12	7	0	10	10	0	14.7	14	71.3	97	20	19

It_	ITS permutation	Acc	Total	Cor. 1A	Cor. 1B	Cor. NC	Over (1B as 1A)	Over (NC as 1B)	Over (NC as 1A)	Und.(1A as 1B)	Und (1B as NC)	Und (1A as NC)	Und. %	Over %	Cor. %	Cor. n	Und.n	Over n
9	ADRA, h-CLAT, OECD.TB	72.8	136	27	52	20	11	8	0	10	8	0	13.2	14	72.8	99	18	19
18	DPRA, h-CLAT, Derek	79.7	74	18	32	9	3	1	0	5	6	0	14.9	5.4	79.7	59	11	4
18	DPRA, GARDskin, OECD.TB	79.7	74	20	30	9	4	1	0	3	7	0	13.5	6.8	79.7	59	10	5
19	DPRA, h-CLAT, Derek	69.9	133	25	49	19	12	7	0	10	11	0	15.8	14.3	69.9	93	21	19
19	DPRA, U-SENS, OECD.TB	66.2	133	24	47	17	14	9	0	11	11	0	16.5	17.3	66.2	88	22	23
20	DPRA, h-CLAT, Derek	77.3	75	19	29	10	3	2	0	5	7	0	16	6.7	77.3	58	12	5
20	ADRA, GARDskin, Derek	77.3	75	20	28	10	3	2	0	4	8	0	16	6.7	77.3	58	12	5
21	DPRA, h-CLAT, Derek	70.4	135	27	47	21	12	7	0	10	11	0	15.6	14.1	70.4	95	21	19
21	ADRA, U-SENS, Derek	70.4	135	24	51	20	9	8	0	13	10	0	17	12.6	70.4	95	23	17
28	DPRA, h-CLAT, Derek	77.8	72	19	28	9	3	1	0	5	7	0	16.7	5.6	77.8	56	12	4
28	ADRA, GARDskin, OECD.TB	79.2	72	20	27	10	3	0	0	4	8	0	16.7	4.2	79.2	57	12	3
29	DPRA, h-CLAT, Derek	71.1	128	26	46	19	12	4	0	10	11	0	16.4	12.5	71.1	91	21	16
29	ADRA, U-SENS, OECD.TB	70.3	128	23	50	17	9	6	0	13	10	0	18	11.7	70.3	90	23	15

It_ = Iteration
Acc = Accuracy = (Correct predictions) / Total
Cor. = predictions correctly predicted by the DA as 1A/1B/Not Classified (NC)
Over = predictions misclassified by the DA, NC predicted to be 1B or 1A, 1B predicted to be 1A
Und. = predictions misclassified by the DA, 1A predicted to be 1B or NC, 1B predicted to be NC
Red text = For comparison - results as reported in GL 497 in 2023 for the classic ITS. Iteration = repeating instruction (loop) from algorithm used to calculate each permutation in turn. Allows for easy comparison of permutation against 'classic' dataset.

Human Potency

Take home message: All alternate DAs evaluated here performed similarly to the classic ITS (provided as a comparison for each permutation; Table 12) when evaluated using a common dataset. Where larger performance gaps between the permutation and the classic are noted, the permutation performed better in most cases. It should be remembered that one or two misclassifications can also result in large percentage differences. The permutations were consistently greater than 65% for accuracy, and over/under-predictions generally similar between permutations and the classic comparison. Overall, performance with a similar method was equivalent if not better across the permutations. The comparison of the LLNA to the human reference data (maximal number of chemicals) is also provided. It should be noted that the accuracy of the LLNA is lower than most of the evaluated permutations. For those permutations with an accuracy lower than that of the LLNA vs human, these are discussed more below.

Table 12. Performance for potency estimates for all DA permutations compared to the ITSv1 when evaluated against the human reference dataset, for maximal available data. LLNA against human is provided as a reference.

Iteration	ITS permutation	Acc	Total	Correct 1A	Correct 1B	Correct NC	Over (1B pred as 1A)	Over (NC pred as 1B)	Over (NC pred as 1A)	Under (1A pred as 1B)	Under (1B pred as NC)	Under (1A pred as NC)	Under. %	Over %	Correct %	Correct n	Under n	Over n
-	Classic' ITSv1 (all data from GL 497 2023)	72.7	44	9	20	3	5	1	0	3	3	0	13.6	13.6	72.7	32	6	6
-	Classic' ITSv2 (all data from GL 497 2023)	71.4	42	8	19	3	5	1	0	3	3	0	14.3	14.3	71.4	30	6	6
5	DPRA, h-CLAT, Derek	67.8	59	13	23	4	3	5	0	7	4	0	18.6	13.6	67.8	40	11	8
5	ADRA, h-CLAT, Derek	62.7	59	12	21	4	4	5	0	8	5	0	22	15.3	62.7	37	13	9
10	DPRA, h-CLAT, Derek	80	40	9	19	4	1	3	0	3	1	0	10	10	80	32	4	4
10	DPRA, GARDskin, Derek	75	40	9	17	4	2	3	0	3	2	0	12.5	12.5	75	30	5	5
11	DPRA, h-CLAT, Derek	67.8	59	13	23	4	3	5	0	7	4	0	18.6	13.6	67.8	40	11	8
11	DPRA, U-SENS, Derek	66.1	59	11	24	4	3	5	0	9	3	0	20.3	13.6	66.1	39	12	8

9	DPRA, h-CLAT, Derek	70.2	57	13	23	4	3	4	0	6	4	0	17.5	12.3	70.2	40	10	7
9	ADRA, h-CLAT, OECD.TB	64.9	57	12	21	4	5	4	0	7	4	0	19.3	15.8	64.9	37	11	9
18	DPRA, h-CLAT, Derek	78.9	38	8	18	4	1	3	0	3	1	0	10.5	10.5	78.9	30	4	4
18	DPRA, GARDskin, OECD.TB	73.7	38	8	16	4	2	3	0	3	2	0	13.2	13.2	73.7	28	5	5
19	DPRA, h-CLAT, Derek	67.3	55	11	22	4	3	5	0	6	4	0	18.2	14.5	67.3	37	10	8
19	DPRA, U-SENS, OECD.TB	63.6	55	9	22	4	3	5	0	8	4	0	21.8	14.5	63.6	35	12	8
20	DPRA, h-CLAT, Derek	73	37	9	14	4	1	4	0	4	1	0	13.5	13.5	73	27	5	5
20	ADRA, GARDskin, Derek	62.2	37	9	10	4	2	4	0	4	4	0	21.6	16.2	62.2	23	8	6
21	DPRA, h-CLAT, Derek	69	58	13	23	4	3	5	0	6	4	0	17.2	13.8	69	40	10	8
Iteration	**ITS permutation**	**Acc**	**Total**	**Correct 1A**	**Correct 1B**	**Correct NC**	**Over (1B pred as 1A)**	**Over (NC pred as 1B)**	**Over (NC pred as 1A)**	**Under (1A pred as 1B)**	**Under (1B pred as NC)**	**Under (1A pred as NC)**	**Under. %**	**Over %**	**Correct %**	**Correct n**	**Under n**	**Over n**
21	ADRA, U-SENS, Derek	55.2	58	7	21	4	4	5	0	12	5	0	29.3	15.5	55.2	32	17	9
28	DPRA, h-CLAT, Derek	73.7	38	9	15	4	1	4	0	4	1	0	13.2	13.2	73.7	28	5	5
28	ADRA, GARDskin, OECD.TB	60.5	38	9	10	4	3	4	0	4	4	0	21.1	18.4	60.5	23	8	7
29	DPRA, h-CLAT, Derek	71.2	52	12	21	4	3	4	0	5	3	0	15.4	13.5	71.2	37	8	7
29	ADRA, U-SENS, OECD.TB	53.8	52	6	18	4	4	4	0	11	5	0	30.8	15.4	53.8	28	16	8
	LLNA vs human	**61.4**	**44**	**9**	**17**	**1**	**3**	**5**	**0**	**6**	**3**	**0**	**20.5**	**18.2**	**61.4**	**27**	**9**	**8**

Acc = Accuracy = (Correct predictions) / Total
Correct = predictions correctly predicted by the DA as 1A/1B/Not Classified (NC)
Over = predictions misclassified by the DA, NC predicted to be 1B or 1A, 1B predicted to be 1A
Under = predictions misclassified by the DA, 1A predicted to be 1B or NC, 1B predicted to be NC

Red text = For comparison - results as reported in GL 497 in 2023 for the classic ITS. Iteration = repeating instruction (loop) from algorithm used to calculate each permutation in turn. Allows for easy comparison of permutation against 'classic' dataset.

ITS underpredictions

Some combinations of assays in the ITS had a higher number of underpredictions (1A human sensitisers predicted as 1B) than the classic. This is mainly observed in the ITS permutations containing both ADRA and U-SENS. These were investigated further, and the impact of replacing a classic method with an alternate calculated. The human data was also considered, using Annex 4 of GL 497, to understand which substances were likely to have been classified using human data where the predicted sensitising dose was close to the threshold between GHS 1A and GHS 1B sub-classifications (termed GHS 1A-) .The DASS dataset was filtered to only keep 1A human sensitisers. Then, any of the ADRA and U-SENS permutations with an underprediction of 1B was taken forward. This was then filtered against those which the classic ITS correctly predicted (7 of the 13 1A sensitisers).

Of these underpredictions, 3 (glutaraldehyde, trans-hex-2-enal and methylisothiazolinone) were classified as human 1A sensitisers and the classic ITS had a total score of 7, compared to 5 and 4, respectively for the permutations (*reminder: score of 2-5 = 1B, score of 6-7 = 1A in the ITS)*. For two of these (glutaraldehyde and trans-hex-2-enal), human-based sub-classifications were derived from human data where the predicted sensitising dose was close to the threshold between GHS 1A and GHS 1B sub-classifications (termed GHS 1A- in **Annex 4** of the *Supporting document to the OECD Guideline 497 on Defined Approaches for Skin Sensitisation* for more information on how this was derived) suggesting that these are less reliable classifications.

All other underpredictions are based on a difference of 1 point e.g. ITS permutation score = 5, classic ITS = 6) and in most of these, it is only one of the assays (ADRA or U-SENS) having a different score to the classic, as opposed to both assays. This could be a result of the relative sensitivities of each assay, as the h-CLAT. in the classic ITS is known to be slightly more sensitive than other KE3 assays. Alternatively, it could be a natural consequence of the scoring system within the ITS, it applies a broad score to assay data and if one or more of the assays is close to the border then these are more likely to be underpredicted.

To summarise, for potency, 10/10 (100%) ITS permutations perform well against the LLNA (Table 11**) when compared to the classic ITS DA. 27/32 (84%) Most ITS permutations have equivalent, or better, performance than the LLNA vs human (61% accuracy) (**Table 12**). Two ITS permutations (Iterations 21 and 29 –** Table 12**) had lower performance when compared to the LLNA vs human accuracy for the reasons above (**Table 13**). This lower performance is specific to these ITS DA permutations only and should be taken into consideration when selecting specific DAs to use. At this time selection of those specific DA permutations may need to be justified.**

Table 13. Comparison of the 1A human sensitisers underpredicted by the ITS permutations using ADRA and U-SENS.

										ITS scores (total)					
		Human data		**KE1 score**		KE1 score difference	**KE3 score**		KE3 score difference	**Classic**	**ADRA, U-SENS permutation**				
Name	CAS	HU. GHS. SUB	HU.GHS. BORDER (Annex 4)	DPRA score	ADRA score	KE1 score difference	h-CLAT score	U-SENS score	KE3 score difference	Classic	DX				OECD TB
Benzylidene acetone	122-57-6	1A		3	2	**1**	2	2	0	6	5				5
Formaldehyde	50-00-0	1A		2	2	0	3	2	**1**	6	5				5
Glutaraldehyde	111-30-8	1A	Borderline	3	2	**1**	3	1	**2**	7	4				4
trans-Hex-2-enal	6728-26-3	1A	Borderline	3	2	**1**	3	2	**1**	7	5				5
Methyl oct-2-ynoate	111-12-6	1A		3	3	0	2	1	**1**	6	5				5
Methyliso thiazolinone	2682-20-4	1A		3	2	**1**	3	2	**1**	7	5				5
Safranal	116-26-7	1A		3	2	**1**	2	2	0	6	5				5

4 Conclusion

As the performance metrics from each permutation of the 2o3 and ITS slightly differ depending on which dataset and reference set is used, it is challenging to pinpoint a specific threshold for allowing or disallowing a single permutation to be used for regulatory purposes. Further, the vast majority of alternate DAs outperformed the LLNA against human reference data for both hazard and potency. **Therefore, the project leads recommend that all permutations remain in the guideline, but lower performance of specific DAs should be taken into consideration when selecting which DAs to use. At this time selection of those specific DA permutations may need to be justified. Users should also justify their permutation selection based on existing data, substance applicability domain as per individual test method guidelines (TG 442C-E) and/or *in silico* supporting documentation (GL 497 Appendices I-II), and regulatory need, as per agreement with the DASS EG December 2024.**

5 Instructions to generate DA datasets for GL 497

2o3

Table 14. How to generate KE-anchored dataset for 2o3 hazard

Step	Instruction	Column options (from Annex 2)
1	Select KE-anchored assay/model (e.g. ADRA.CALL.BR(correct)) and filter out missing values or NA, keeping only scores of 0 or 1	DPRA.Call_VSBR_MR ADRA.CALL.BR(correct) KS.Call_IR_VSBR LuSens.Call.BR EpiSensA.CALL.BR hCLAT.Call_IR_VSBR hCLAT.Call_IR_VSBR_logKow GARDskin Call BR moIL-8.Luc.assay.call.BR(correct) U-SENS Call.BR
2	Filter out NA from experimental data, keeping only 0 or 1 (e.g. LLNA.Call)	LLNA.Call HDSG.Call
3	Filter out NA or inconclusives from KE-anchored DA output, keeping only 0 or 1 (e.g. DA.2o3_[ADRA, KeratinoSens, h-CLAT])	DA.2o3_[DPRA, KeratinoSens, GARDskin] DA.2o3_[DPRA, KeratinoSens, U-SENS] DA.2o3_[ADRA, KeratinoSens, h-CLAT] DA.2o3_[DPRA, LuSens, h-CLAT] DA.2o3_[DPRA, EpiSensA, h-CLAT] DA.2o3_[DPRA, KeratinoSens, IL-8]
4	Filter out NA or inconclusives from ‘classic’ DA output, keeping only 0 or 1	DA.2o3_[DPRA, KeratinoSens, h-CLAT]
5	Apply hazard calculation by comparing the ‘classic’ and the KE-anchored DA to the experimental data	

Table 15. How to generate maximal dataset for 2o3 hazard

Step	Instruction	Column options (from Annex 2)
1	Filter out NA from experimental data, keeping only 0 or 1	LLNA.Call HDSG.Call
2	Filter out NA or inconclusives from DA output , keeping only 0 or 1	DA.2o3_[DPRA, KeratinoSens, GARDskin] DA.2o3_[DPRA, KeratinoSens, U-SENS] DA.2o3_[ADRA, KeratinoSens, h-CLAT] DA.2o3_[DPRA, LuSens, h-CLAT] DA.2o3_[DPRA, EpiSensA, h-CLAT] DA.2o3_[DPRA, KeratinoSens, IL-8] DA.2o3_[ADRA, EpiSensA, GARDskin] DA.2o3_[ADRA, EpiSensA, U-SENS] DA.2o3_[ADRA, EpiSensA, h-CLAT] DA.2o3_[ADRA, EpiSensA, IL-8] DA.2o3_[ADRA, KeratinoSens, GARDskin] DA.2o3_[ADRA, KeratinoSens, U-SENS] DA.2o3_[ADRA, KeratinoSens, IL-8] DA.2o3_[ADRA, LuSens, GARDskin] DA.2o3_[ADRA, LuSens, U-SENS] DA.2o3_[ADRA, LuSens, IL-8] DA.2o3_[DPRA, EpiSensA, GARDskin] DA.2o3_[DPRA, EpiSensA, U-SENS] DA.2o3_[DPRA, EpiSensA, IL-8] DA.2o3_[DPRA, LuSens, GARDskin] DA.2o3_[DPRA, LuSens, U-SENS] DA.2o3_[DPRA, LuSens, IL-8] DA.2o3_[ADRA, LuSens, h-CLAT]
3	Filter out NA or inconclusives from ‘classic’ DA output, keeping only 0 or 1	DA.2o3_[DPRA, KeratinoSens, h-CLAT]
4	Apply hazard calculation by comparing the ‘classic’ and the KE-anchored DA to the experimental data	

ITS

Table 16. How to generate KE-anchored dataset for ITS hazard

Step	Instruction	Column options (from Annex 2)
1	Select KE-anchored assay/model (e.g. DA.ITS.Score.Derek) and filter out missing values or NA, keeping only scores between 0-3	DA.ITS.Score.ADRA DA.ITS.Score.DPRA DA.ITS.Score.GARDskin DA.ITS.Score.h-CLAT DA.ITS.Score.U-SENS DA.ITS.Score.Derek DA.ITS.Score.OECD.TB
2	Filter out NA from experimental data, keeping only 0 or 1	LLNA.Call HDSG.Call
3	Filter out NA or inconclusives from KE-anchored DA output, keeping only 0 or 1 (e.g. DA.ITS.Call.Conf.[APRA, h-CLAT, Derek)	DA.ITS.Call.Conf.[DPRA, h-CLAT, OECD.TB] DA.ITS.Call.Conf.[DPRA, U-SENS, Derek] DA.ITS.Call.Conf.[DPRA, GARDskin, Derek] DA.ITS.Call.Conf.[ADRA, h-CLAT, Derek]
4	Filter out NA or inconclusives from ‘classic’ DA output, keeping only 0 or 1	DA.ITS.Call.Conf.[DPRA, h-CLAT, Derek]
5	Apply hazard calculation by comparing the ‘classic’ and the KE-anchored DA to the experimental data	

Table 17. How to generate KE-anchored dataset for ITS potency

Step	Instruction	Column options (from Annex 2)
1	Select KE-anchored assay/model and filter out missing values or NA, keeping only scores between 0-3	DA.ITS.Score.ADRA DA.ITS.Score.DPRA DA.ITS.Score.GARDskin DA.ITS.Score.h-CLAT DA.ITS.Score.U-SENS DA.ITS.Score.Derek DA.ITS.Score.OECD.TB
2	Filter out NA from experimental data, keeping only 1A, 1B, NC	LLNA.GHS.SUB HU.GHS.SUB
3	Filter out NA or inconclusives from KE-anchored DA output, keeping only 1A, 1B, NC	DA.ITS.Pot.Conf.[DPRA, h-CLAT, OECD.TB] DA.ITS.Pot.Conf.[DPRA, U-SENS, Derek] DA.ITS.Pot.Conf.[DPRA, GARDskin, Derek] DA.ITS.Pot.Conf.[ADRA, h-CLAT, Derek]
4	Filter out NA or inconclusives from ‘classic’ DA output, keeping only 1A, 1B, NC	DA.ITS.Pot.Conf.[DPRA, h-CLAT, Derek]
5	Apply potency calculation by comparing the ‘classic’ and the KE-anchored DA to the experimental data	

Table 18. How to generate maximal dataset for ITS hazard

Step	Instruction	Column options (from Annex 2)
1	Filter out NA from experimental data, keeping only 0 or 1	LLNA.Call HDSG.Call
2	Filter out NA or inconclusives from 'new' DA output, keeping only 0 or 1	DA.ITS.Call.Conf.[DPRA, h-CLAT, OECD.TB] DA.ITS.Call.Conf.[DPRA, U-SENS, OECD.TB] DA.ITS.Call.Conf.[DPRA, U-SENS, Derek] DA.ITS.Call.Conf.[DPRA, GARDskin, OECD.TB] DA.ITS.Call.Conf.[DPRA, GARDskin, Derek] DA.ITS.Call.Conf.[ADRA, h-CLAT, OECD.TB] DA.ITS.Call.Conf.[ADRA, h-CLAT, Derek] DA.ITS.Call.Conf.[ADRA, U-SENS, OECD.TB] DA.ITS.Call.Conf.[ADRA, U-SENS, Derek] DA.ITS.Call.Conf.[ADRA, GARDskin, OECD.TB] DA.ITS.Call.Conf.[ADRA, GARDskin, Derek]
3	Filter out NA or inconclusives from 'classic' DA output, keeping only 0 or 1	DA.ITS.Call.Conf.[DPRA, h-CLAT, Derek]
4	Apply hazard calculation by comparing the 'classic' and the KE-anchored DA to the experimental data	

Table 19. How to generate maximal dataset for ITS potency

Step	Instruction	Column options (from Annex 2)
1	Filter out NA from experimental data, keeping only 1A, 1B, NC	LLNA.GHS.SUB HU.GHS.SUB
2	Filter out NA or inconclusives from 'new' DA output, keeping only 1A, 1B, NC	DA.ITS.Pot.Conf.[DPRA, h-CLAT, OECD.TB] DA.ITS.Pot.Conf.[DPRA, U-SENS, OECD.TB] DA.ITS.Pot.Conf.[DPRA, U-SENS, Derek] DA.ITS.Pot.Conf.[DPRA, GARDskin, OECD.TB] DA.ITS.Pot.Conf.[DPRA, GARDskin, Derek] DA.ITS.Pot.Conf.[ADRA, h-CLAT, OECD.TB] DA.ITS.Pot.Conf.[ADRA, h-CLAT, Derek] DA.ITS.Pot.Conf.[ADRA, U-SENS, OECD.TB]

		DA.ITS.Pot.Conf.[ADRA, U-SENS, Derek] DA.ITS.Pot.Conf.[ADRA, GARDskin, OECD.TB] DA.ITS.Pot.Conf.[ADRA, GARDskin, Derek]
3	Filter out NA or inconclusives from 'classic' DA output, keeping only 1A, 1B, NC	DA.ITS.Pot.Conf.[DPRA, h-CLAT, Derek]
4	Apply potency calculation by comparing the 'classic' and the KE-anchored DA to the experimental data	