P32-11

Predicting CLP classification for skin sensitisation using the CLASS model

Gaspard Levet[1], Emel Ay-Albrecht[1], Marie Darracq-Ghitalla-Ciock[1], Franklin J. Bauer[1], Pablo Vidal[1], Zlatomir Todorov[1], Paul C. Thomas[1], Carole Charmeau-Genevois[1]

[1]KREATiS, 38080 L'Isle d'Abeau – France

Contact us : www.kreatis.eu / contact@kreatis.eu

Introduction

The new version, v1.6, of the **iSafeRat® CLASS (Classification & Labelling Assessment for Skin Sensitisation) model**, presented here, now includes category and EC3 value predictions. It has been trained on an extended dataset of 838 LLNA data curated and validated according to previously described methodologies, selecting only high-quality data following OECD Guideline No. 429, No. 442a and No. 442b requirements. The applicability domain was refined, model statistics were improved and compliance with the QSAR Assessment Framework (QAF)[1] requirements was verified. In this newest version, a recently updated predictive method has been developed in iSAFERAT® to assign substance **skin sensitisation** (**SS**) CLP categories with confidence. This poster describes the main changes introduced in this version.

Methods

Data: LLNA data obtained from NiceATM, CosUE, ECHA dossiers and proprietary data.

Curation and evaluation: high-quality dataset of 1,075 substances; excluding inorganics or multi-constituents and duplicates, with canonical SMILES available, experimental reliability of the studies was validated according to the OECD 429, 442a and 442b Guidelines and worst-case selection when several EC values available.

Splitting: training set (697 substances), Internal Validation set (141 substances), External Test set (237 substances) (**Table 1**)

Model for potential prediction (**Figure 1**): decision tree based on three sub-models: SkinAbs v2.0, Hapt v1.3 and Autox v2.0.

Applicability domain (AD): according to QAF, the iSAFERAT® CLASS model displays all required information for each chemical space. In addition to the descriptors specific to each model, the AD is based on molecular weight, molecular structure, and mechanistic information derived from the training and internal validation sets, which are combined into a single overall training set.

Analogue search: the model was trained with various types of molecular fingerprints (PubChem, MACCS Key, and an internal fingerprint), numbers of analogues (from 2 to 4), similarity methods, thresholds (from 30% to 70%). From the twenty configurations tested, a single configuration was chosen for use by default: (k= 2), based on two similarity calculations, maximum common substructure similarity and MACCS Key fingerprint, with more weight on MACCS Key (95%).

Category prediction: (**Figure 2**): category prediction based on a weighted average of analogue experimental results. Three category are possible: Category 1A, Category 1B, Not classified according to CLP (NC). In the event of conflicting results, it selects the most severe category.

Reliability indices: three indices (similarity index, accuracy index and concordance index) are calculated. A final index called Reliability Index of the Prediction (RIP) combines these three indices. A prediction is considered reliable when RIP ≥ 70%, a fixed threshold defined by the model developers as an indication of the confidence in the prediction.

Presented performances: statistics are presented for the external test set of the model: **Table 2, 3, 4 and 5**. **Case 1**) Substances falling within AD. **Case 2**) Substances falling within AD, AD is based only on the substance in the training set, analogue search by default method, RIP threshold is fixed at 70%, analogue search based only on training. **Case 3**) Substance falling within AD, analogue search by default method on overall training (train + validation), RIP threshold is fixed at 70%. **Case 4**) substance falling within AD, manual selection of analogues, RIP threshold is fixed at 70%, analogue search based on overall training set.

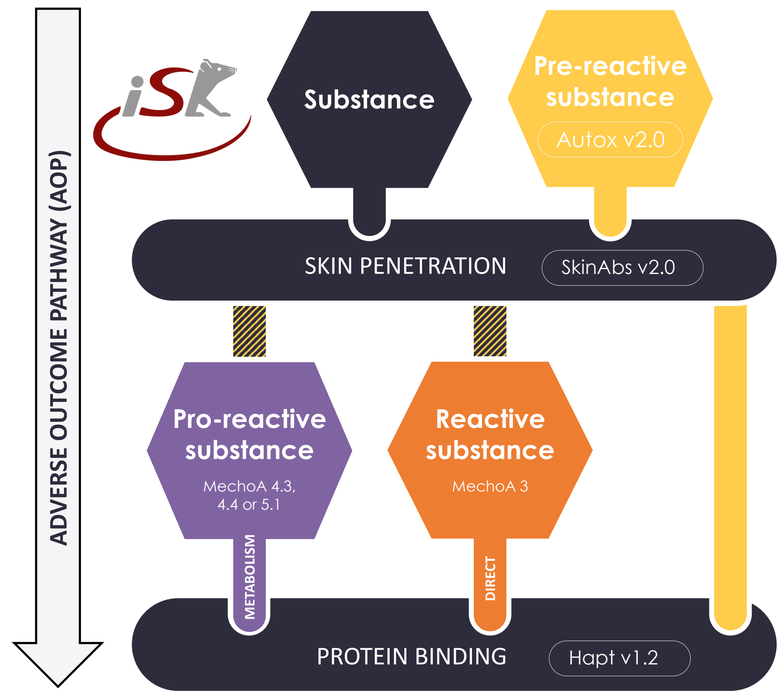

Figure 1. *Principles of CLASS model: potential prediction based on 3 sub-models (Autox v2.0, Skin Abs v2.0, Hapt v1.2). Decision tree is based on AOP of skin sensitisation (induction phase).*

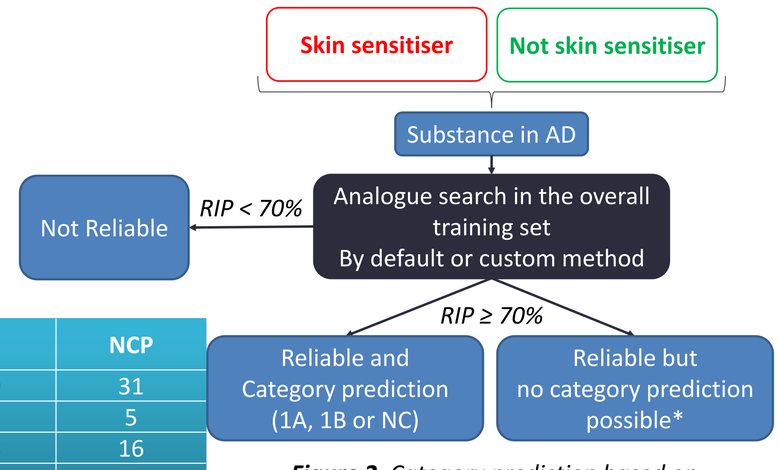

Figure 2. *Category prediction based on analogue found*
**Analogues found are NCP*

Results & discussion

Retraining the model on the extended dataset improved the predictive performance for skin sensitising potential for non-sensitisers. It also provided a more comprehensive AD description, which was incorporated in the software. External validation demonstrated robust performance for predictions of SS potential (**Table 2**, **Case 1**).

With the "by-default" setting, it was possible to test the predictivity of the decision tree for predictions with a RIP ≥ 70%. Improvements were noted when we based the RIP on analogues coming from the training set only (**Table 3**, **Case 2**) or from the overall training set (**Table 4**, **Case 3**). Additionally, statistics on category predictions could be obtained from **Cases 2** and **3**, (respectively **Table 3** and **4**). Considering the RIP index, a boost in the predictivity of the model of about 10%, was observed for each statistical parameter. Although a decrease in coverage of around 45% was observed, the confidence in the predictions of SS makes this model a good choice when considering regulatory submissions (e.g. for REACH). However, when only one positive analogue was identified, the predicted category should be interpreted with caution, even if the RIP index exceeds the threshold.

Among the category predictions (NC, 1A or 1B) 75% were correctly predicted for **Case 3, Table 4**. Categories 1A and 1B contain relatively few substances and exhibit partially overlapping chemical characteristics, making their statistical discrimination more challenging and limiting the number of clearly relevant analogues.

The main difference between **Cases 3** and **4** lies in analogue selection. In **Case 4**, analogues are no longer selected by default but are manually selected by an expert based on their relevance to the test substance. This expert selection leads to an overall improvement in performance, particularly for categories 1A and 1B, for which the F1-scores increase from 56% to 61% and from 60% to 74%, respectively. The NC category also shows an improvement, with the F1-score increasing from 88% to 93%. Thus, the manual selection of relevant analogues by an expert can improve the predictive performance across the different categories, particularly when discrimination between categories relies on structural or toxicological differences that may be difficult to identify automatically.

Dataset	1A	1B	NC	NCP
Training set (n=697)	63	194	409	31
Internal validation set (n=141)	18	46	72	5
External test set (n=237)	20	66	135	16
Total (n=1075)	101	306	616	52

Table 1. *Splitting and number of substance with experimental data on categories (NCP=No classification was possible based on experimental collected data)*

External test	Coverage	Accuracy (%)	Sensitivity (%)	Precision (positive) (%)	False negative rate (%)	Specificity (%)	Precision (negative) (%)	False positive rate (%)	Balanced accuracy (%)	Percentage No analogues (%)	Percentage RIP<70% (%)
V1.4	100% (123 molecules)	75,6	85,7	77,6	14,3	58,7	71,1	41,3	72,2	/	/
V1.6 Case 1	100% (237 molécules)	77,6	72,5	74,7	27,5	81,5	79,7	18,5	77,0	/	/
V1.6 Case 2	57% (134 molecules)	78,4	75,4	79,0	24,6	81,2	77,8	18,8	78,3	21,5	21,9
V1.6 Case 3	55% (131 molecules)	87,0	83,1	87,5	16,9	90,3	86,7	9,7	86,7	17,7	27,0
V1.6 Case 4	59% (139 molecules)	92,1	90,7	89,1	9,3	92,9	94	7,1	91,8	8,0	25,3

Table 2. *Prediction of potential (skin sensitiser=positive, not skin sensitiser=negative).*
Statistics of iSAFERAT® CLASS v1.6, AD based on overall training set before and after reliability index use (by-default and custom methods).

Category	precision	recall	f1-score	Percentage no analogues
1a	43%	60%	51%	0%
1b	67%	55%	61%	21%
NC	88%	89%	88%	25%

Table 3. *Prediction of category based on Case 2*

Category	precision	recall	f1-score	Percentage no analogues
1a	41%	71%	56%	0%
1b	70%	51%	60%	18%
NC	88%	88%	88%	21%

Table 4. *Prediction of category based on Case 3*

Category	precision	recall	f1-score	Percentage no analogues
1a	45%	77%	61%	0%
1b	82%	66%	74%	8%
NC	94%	93%	93%	10%

Table 5. *Prediction of category based on Case 4 (custom analogue search)*

Conclusion

The updated CLASS model v1.6 has:

1) enhanced applicability domain assessment,
2) enabled CLP skin sensitisation category predictions, and
3) provides regulatory compliant reliability information in line with QAF and ECHA guidance.

The model can be used as a standalone method to predict SS potential and potency based on category prediction.

In the future, it is planed :

- to increase the AD using additional information on other available assays,
- to further adjust the by-default parameter of the k-NN, for example by automatically selecting the best parameters based on a genetic algorithm,
- and to test other methodologies (e.g. SVM).

Scan QR code to download the poster
Available on https://kreatis.eu

[1] OECD. *(Q)SAR assessment framework: Guidance for the regulatory assessment of (quantitative) structure activity relationship models and predictions*. Paris: OECD Publishing. 2023;386.
[2] Bauer FJ *et al. . A new classification algorithm based on mechanisms of action.* Comput Toxicol. 2018;5:8–15.