Advancing systemic toxicity risk assessment: Evaluation of a NAM-based toolbox approach

SEAC | **Unilever**

Sophie Cable, Maria Teresa Baltazar, Fazila Bunglawala, Paul L. Carmichael, Leonardo Contreas, Matthew Philip Dent, Jade Houghton, Predrag Kukic, Sophie Malcomber, Beate Nicol, Katarzyna R Przybylak, Ans Punt, Georgia Reynolds, Joe Reynolds, Sharon Scott, Dawei Tang, Alistair M Middleton
Safety and Environmental Assurance Centra, Unilever, Colworth, MK44 1lQ

For many years, a method that allowed systemic toxicity safety assessments to be conducted without generating new animal test data seemed out of reach. However, several different research groups and regulatory authorities are beginning to use a variety of *in silico, in chemico* and *in vitro* techniques to inform safety decisions. To manage this transition to animal-free safety assessments responsibly, it is important to ensure that the level of protection offered by a safety assessment based on new approach methodologies (NAMs), is at least as high as that provided by a safety assessment based on traditional animal studies. To this end, we have developed an evaluation strategy to assess both the level of protection and the utility offered by a NAM-based systemic safety 'toolbox'. We have previously proposed a NAM-based toolbox for integration into a risk assessment framework for the evaluation of systemic toxicity (Middleton et al, 2022; Cable et al 2024, submitted) and Fig.1 shows a tiered approach to NGRA following the ICCR principles, and as utilised through various case studies (Dent et al 2018; Baltazar et al, 2020; Rajagopal et al, 2022; Wood et al, 2024)

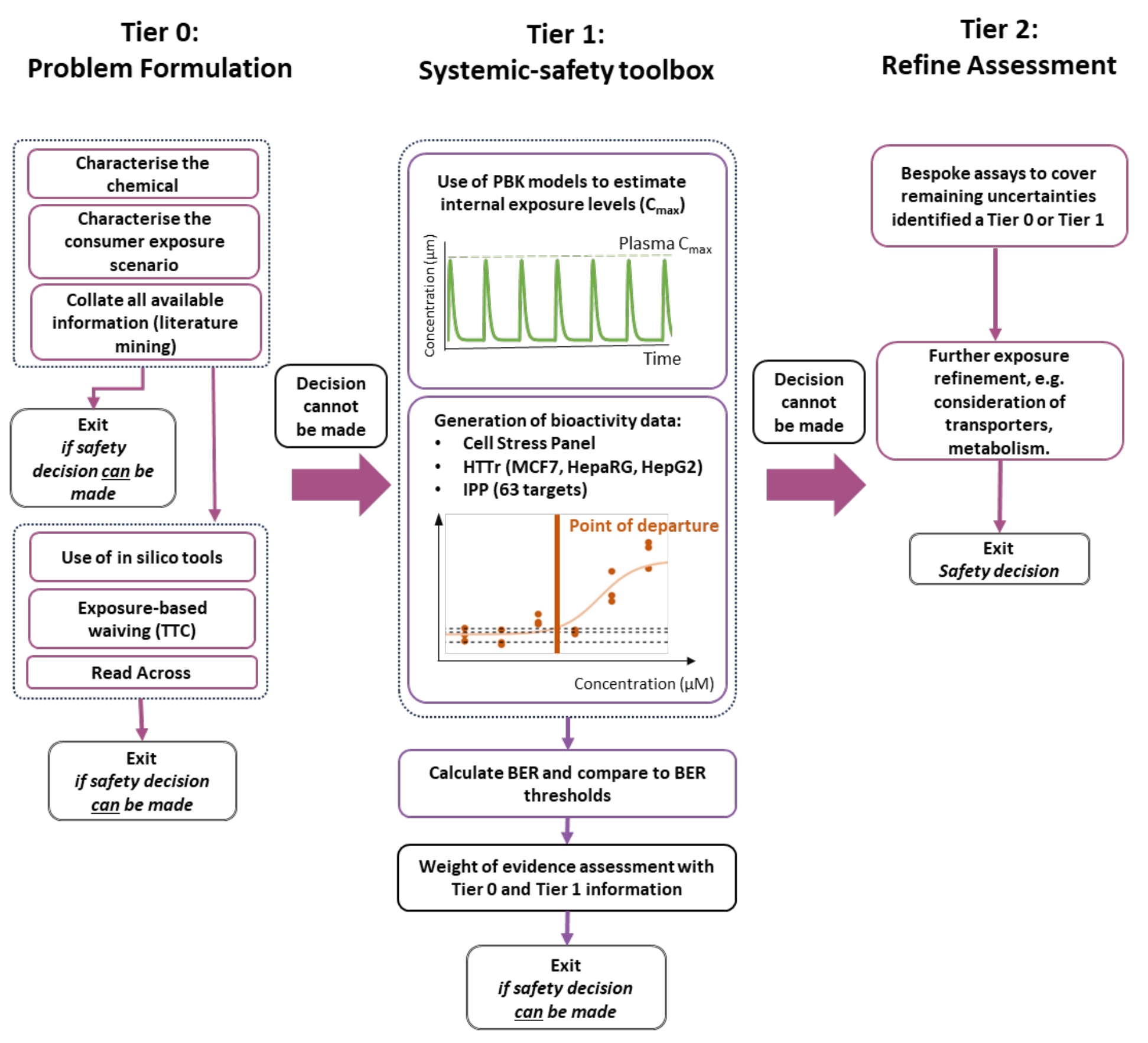
Tier 0:
Problem Formulation
Characterise the chemical
Characterise the consumer exposure scenario
Collate all available information (literature mining)
Exit
if safety decision can be made
Use of in silico tools
Exposure-based waiving (TTC)
Read Across
Exit
if safety decision can be made
Decision cannot be made
Tier 1:
Systemic-safety toolbox
Use of PBK models to estimate internal exposure levels (C_{max})
Plasma C_{max}
Concentration (µm)
Time
Generation of bioactivity data:
Cell Stress Panel
HTTr (MCF7, HepaRG, HepG2)
IPP (63 targets)
Point of departure
Concentration (µM)
Calculate BER and compare to BER thresholds
Weight of evidence assessment with Tier 0 and Tier 1 information
Exit
if safety decision can be made
Decision cannot be made
Tier 2:
Refine Assessment
Bespoke assays to cover remaining uncertainties identified a Tier 0 or Tier 1
Further exposure refinement, e.g. consideration of transporters, metabolism.
Exit
Safety decision

Fig 1 – A risk assessment framework inspired by those of the Seurat-1 project, the EPA blueprint (Thomas et al, 2019) and previous NGRA case studies (Baltazar et al, 2020; OECD IATA 2021a) showing where this systemic-safety toolbox could sit in an early tier data generation phase following collation and appraisal of all existing information at Tier 0. At each potential 'exit' the outcome can be a safety decision of low risk or uncertain risk. Cases where the risk is uncertain can progress to higher tier testing if this can address any remaining uncertainties identified at earlier tiers, e.g. mechanism of action-based testing. The use of the systemic-safety toolbox at Tier 1 is intended to address cases where there are data gaps at Tier 0 specifically regarding systemic toxicity.

STEP 1: DEFINE TOOLBOX COMPONENTS AND PERFORM PROOF OF PRINCIPLE STUDY

The core components of a toolbox and workflow were decided to be:

Estimation of internal exposure using different levels of input parameters to build the physiologically-based kinetic (PBK) models. Plasma Cmax values are estimated for every chemical-exposure scenario using either *in silico* only parameter estimates (L1), *in vitro* parameters from experimental data where available (L2), or calibrated model estimates using human clinical data (L3).

Estimation of a bioactivity point of departure (PoD) was done across 3 different assays consisting of the investigation of 63 specific protein targets (GPCRs, ion channels, enzymes etc.) as well as cellular stress mechanisms and effects on the transcriptome of 3 cell lines (HepG2, HepaRG, MCF7). Bayesian statistical models were built to analyse the cellular stress and transcriptomics data in a concentration-response manner and establish the most likely concentration at which an effect begins, thus determining a bioactivity platform PoD.

Calculation of a Bioactivity Exposure Ratio (BER) combines inputs from the exposure and bioactivity assay modules, calculating the ratio between the plasma Cmax estimates and the lowest platform PoD.

Conceptually a BER > 1 indicates a low risk of adverse effects in consumers if the following assumptions are true:

1. The *in vitro* measures of bioactivity provide appropriate biological coverage
2. There is confidence that the test systems are at least as sensitive to perturbation as human cells *in vivo*
3. The exposure estimate is conservative for the exposed population

Table 1

PBK Level	BER Threshold
L1	110
L2	11
L3	2.5

However there has been limited work up to this point to evaluate if this concept holds true in real cases. A pilot study using 10 chemicals and 24 benchmark chemical exposure scenarios was performed. The results of this pilot study were used to define a threshold for benchmark chemicals at which all exposure scenarios with a greater BER would be considered low risk. The results are shown in Fig 2 and Table 1.

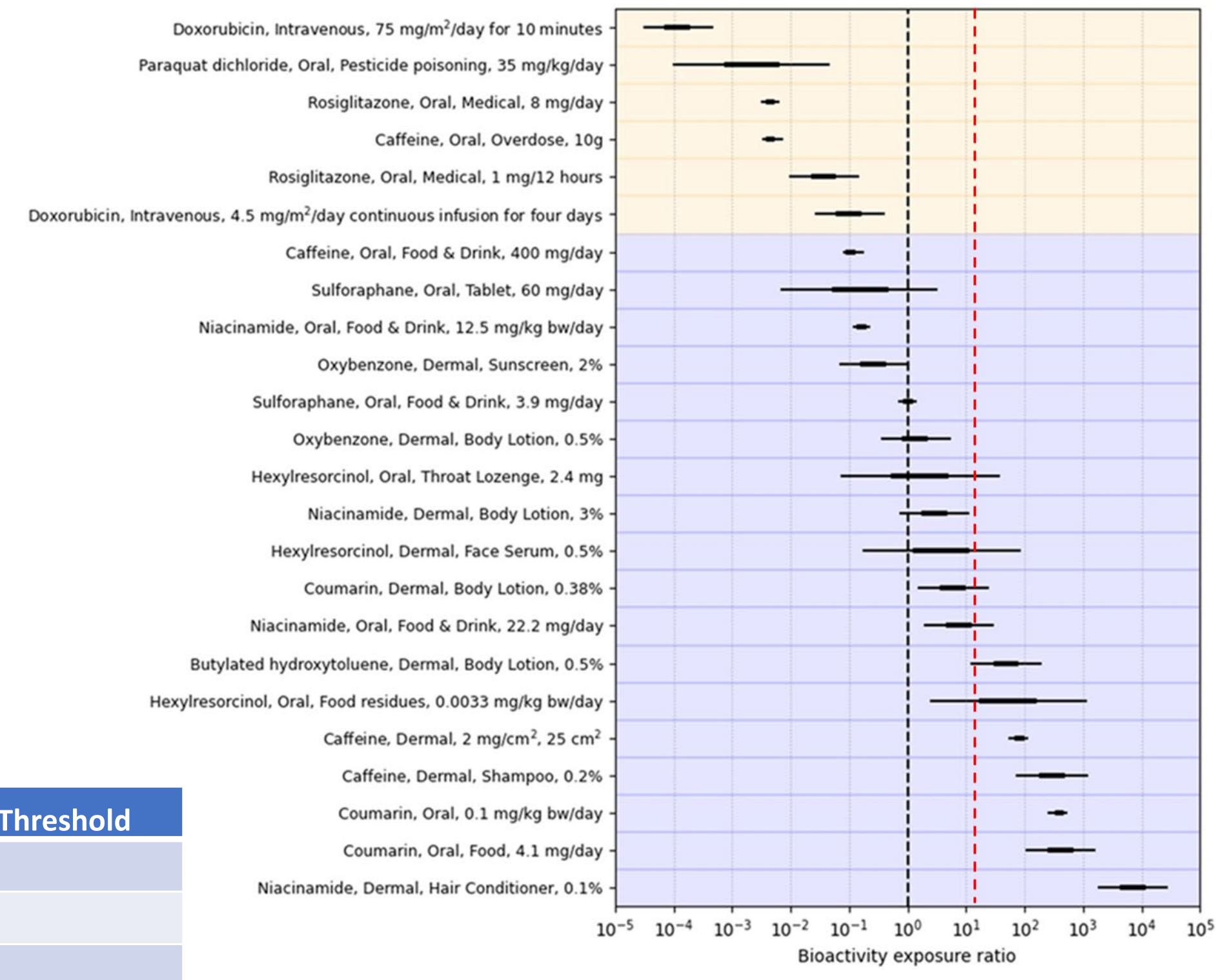

Fig.2 - Calculated BER values for 24 chemical exposure scenarios as determined in Middleton et al, 2022. High risk chemical exposure scenarios are shown in yellow, low risk chemical exposure scenarios are shown in blue. The bars represent the 95% confidence interval of the calculated BER when considering uncertainty in the exposure estimate. The dashed lines represent BER = 1 (black) and the L2 BER threshold (red).

STEP 2: SELECT TEST CHEMICALS AND SET PERFORMANCE CRITERIA

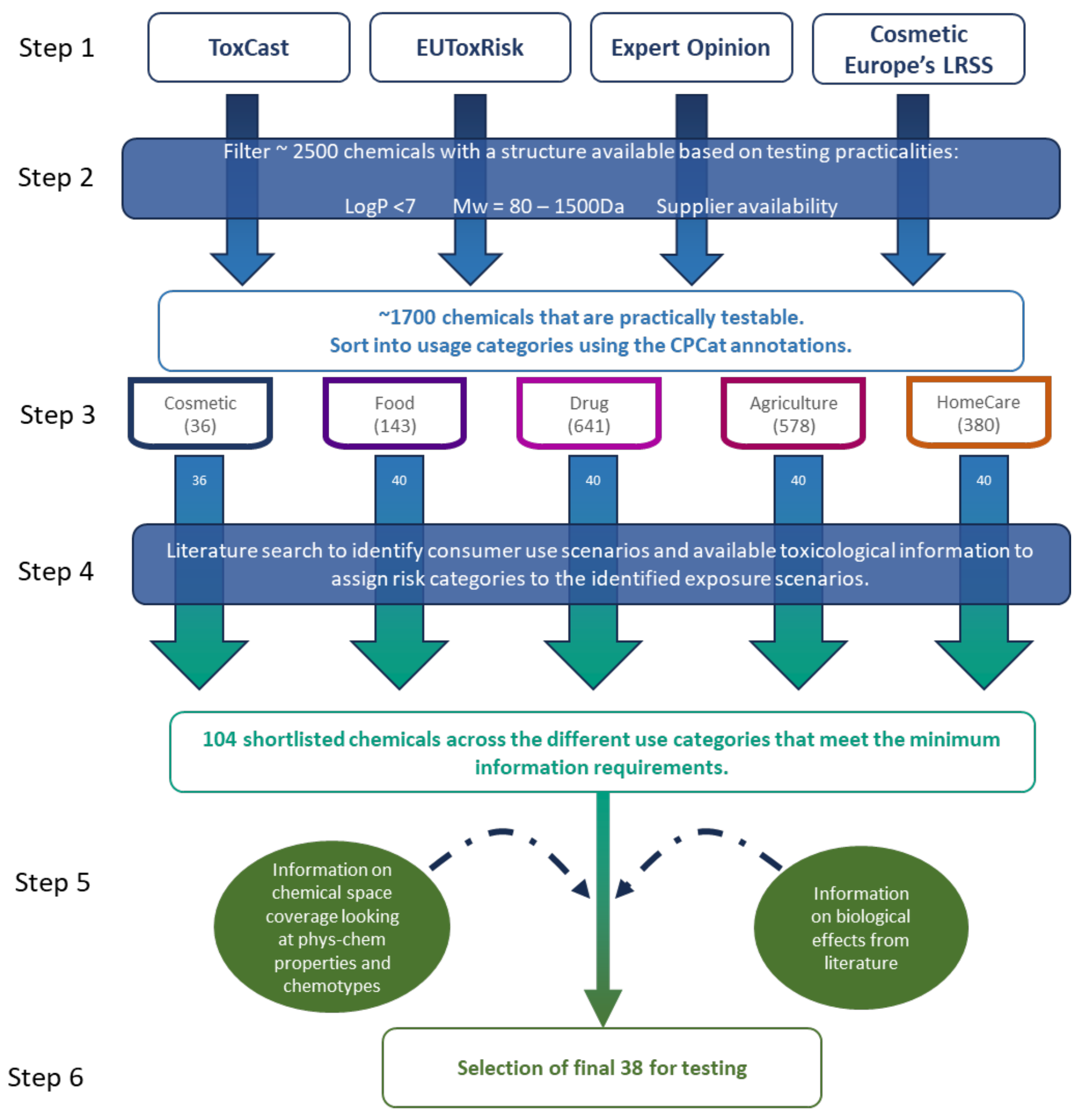
Step 1
ToxCast
EUToxRisk
Expert Opinion
Cosmetic Europe's LRSS
Step 2
Filter ~ 2500 chemicals with a structure available based on testing practicalities:
LogP <7
Mw = 80 – 1500Da
Supplier availability
~1700 chemicals that are practically testable.
Sort into usage categories using the CPCat annotations.
Step 3
Cosmetic (36)
Food (143)
Drug (641)
Agriculture (578)
HomeCare (380)
36
40
40
40
40
Step 4
Literature search to identify consumer use scenarios and available toxicological information to assign risk categories to the identified exposure scenarios.
104 shortlisted chemicals across the different use categories that meet the minimum information requirements.
Step 5
Information on chemical space coverage looking at phys-chem properties and chemotypes
Information on biological effects from literature
Step 6
Selection of final 38 for testing

Fig 3 – Overview and flow of the chemical selection process resulting in 38 chemicals to progress to data generation for evaluation of the NAM-based systemic toolbox and workflow.

Aims:

- Avoid biasing the evaluation through selection of only 'extreme' cases, e.g. fatally toxic chemicals and biologically inert chemicals
- Select chemicals covering a broad range of chemistries and biology
- Select chemicals with exposure scenarios for which a risk classification could be assigned using the available literature.

Fig.3. shows an overview of the chemical selection process, including several filtering steps to remove any chemicals that would be incompatible with the nature of the testing being conducted or for which there wasn't sufficient information available to define an exposure scenario with a defined risk classification.

The final selection of chemicals that met the criteria included 9 chemicals primarily associated with cosmetic use, 21 primarily associated with medicinal use, 3 associated with food exposures, 5 agricultural chemicals and 1 primarily associated with occupational use.

PERFORMANCE CRITERIA

PROTECTIVENESS:

Does the toolbox workflow identify all high-risk exposure scenarios as uncertain risk, i.e. BER < 11 at L2?

Potential reasons for lack of protectiveness:

- The chemical has a specific mode of action not picked up in our test systems:
  - Specific receptors, enzymes or assay endpoints may be missing from the toolbox and might be necessary to enable a protective decision where very little other bioactivity is observed.
- The chemical requires metabolic activation
  - The metabolic competency of the cells used within these assays is known to be less than in vivo. Therefore, the toolbox output might not pick up specific activity driven by the output metabolite and not parent compound.
- The Cmax estimate calculated at L2 is an underestimate of the in vivo exposure
  - The current L2 definition does not specify which parameters need to be derived experimentally, key parameters could be *in silico* and this might not be reflected in the error calculated under the assumption of an L2 prediction.
  - The chemical might rely on active transport to enter cells, which isn't reflected in the PBK model without specific information. This is the case for Digoxin where the L2 prediction underestimates the L3 value by more than 50 times due a lack of consideration of transporters.

UTILITY:

Does the toolbox workflow identify all low-risk exposure scenarios as low risk, i.e. BER > 11 at L2?

Potential reasons for lack of utility:

- The exposure estimate is a significant overestimate of the likely *in vivo* exposure and more data would be needed to refine this.
  - E.g. not all dermal exposure scenarios have good quality dermal penetration data available and so a default of 100% is assumed.
- The concentration-response analysis method used is overly sensitive and does not correct for all false positives
  - This is likely to be the case for examples where the BER is being driven by a small number of genes with low level responses.
- The test systems measure bioactivity and not adversity, so do not differentiate situations where observed activity does not lead to adversity, i.e. is adaptive. Integration of this data in to a weight of evidence framework for safety decision-making will allow for detailed interpretation of the results.

STEP 3: EVALUATE TOOLBOX

This toolbox and workflow is intended for use in quantitative early-tier risk assessment, where the primary goal is protectiveness: i.e. no classification of high-risk chemical exposure scenarios as low risk. It does this for over 90% of the benchmark chemical exposure scenarios

- ❑ There are a total of 8 different PoD types generated by the systemic-safety toolbox: one associated with receptor profiling (IPP), one with cellular stress (CSP) and two for each of the three HTTr cell lines that were tested (one based on gene level changes and one on pathway level changes). Across the different chemicals tested in this work, IPP gave the lowest PoD for 11 chemicals, CSP gave the lowest PoD for 5 chemicals and HTTr (gene level) gave the lowest PoD for 25 chemicals (8 in HepaRG, 6 in HepG2, 11 in MCF-7).
- ❑ BERs were calculated using the lowest PoD across all bioactivity platforms tested and dividing them by the plasma Cmax estimates for each chemical exposure scenario. Fig.4 shows the resulting BER plot when L2 PBK estimates are used and compared to the previously determined threshold of 11, giving a protectiveness and utility of 93% and 27% respectively.
  - ❑ This is comparable to the performance of using traditional in vivo toxicology data for the risk assessment, as demonstrated in Fig.5. Where the NAM PoDs are more conservative than the *in vivo* PoDs in 22/24 cases (in vivo data were not found for all chemicals tested).
- ❑ Only the therapeutic doses of warfarin and occupational exposure to Trimellitic anhydride are misclassified as low risk using this toolbox alone. However, the intended use is within a tiered and iterative framework encompassing all lines of evidence.
  - ❑ Trimellitic anhydride is a known sensitiser, and it is likely that in a risk assessment framework the risk posed by sensitisation via the inhalation route would limit the exposure below that which poses a systemic risk.
  - ❑ In vitro data available for the activity of Warfarin at its target, VKORC1, would change the risk assessment conclusion with a measured IC50 giving a BER<<1.
- ❑ It can reasonably be envisaged that PBK models parameterised with in vitro data are the most likely future scenario for a novel risk assessment, although the performance metrics improve as PBK models can be calibrated against human clinical data. Table 1 shows the resulting protectiveness and utility scores for the different PBK levels.

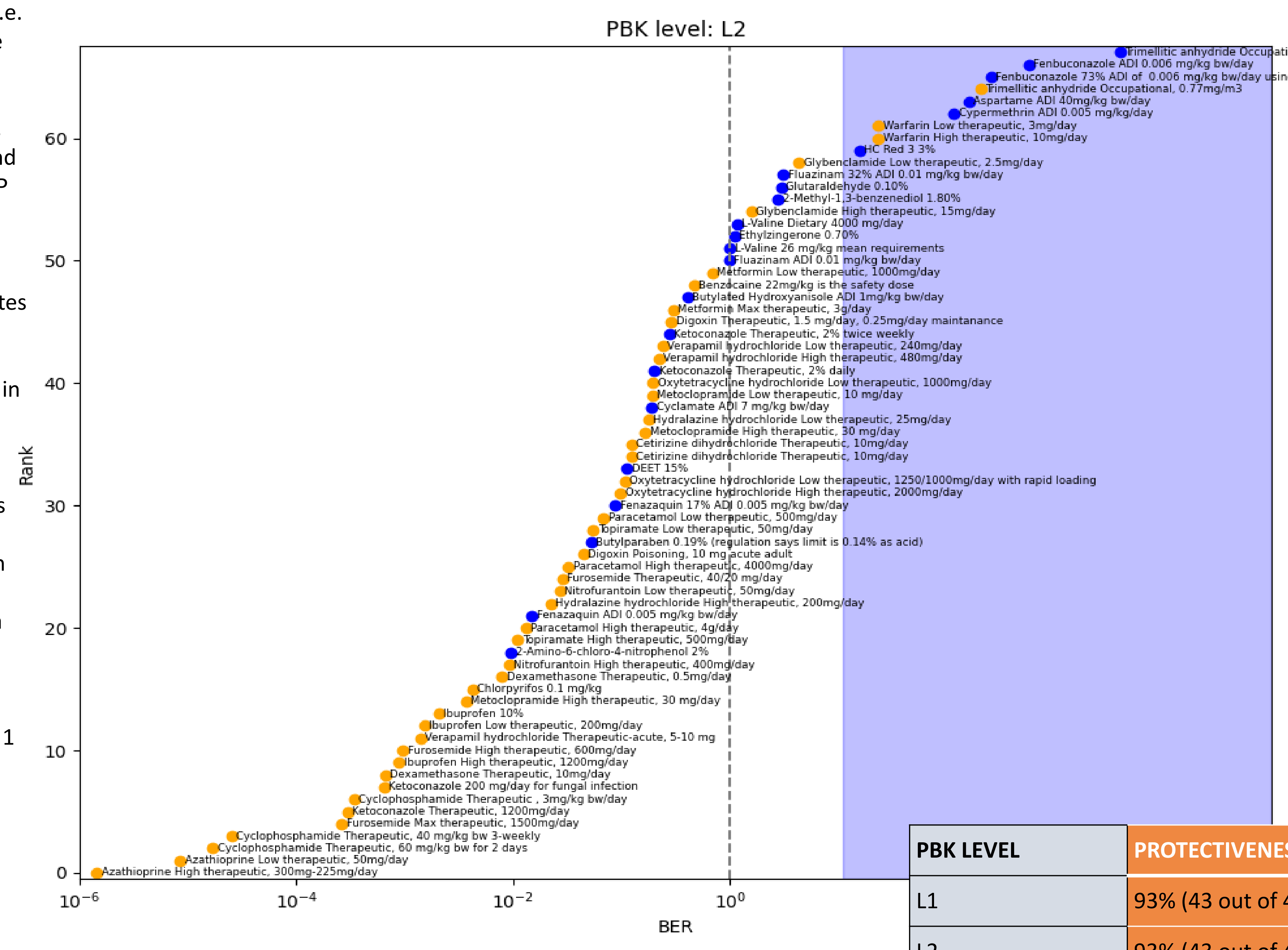

Fig.4 Plot showing BER values for 68 chemical exposure scenarios where L2 exposure estimates were available. Yellow dots represent chemical exposure scenarios classified as high risk, blue dots represent chemical exposure scenarios classified as low risk. The blue shaded region represents a BER of 11 or above, with a BER > 11 being the previously determined threshold for a low risk decision at L2. The black dashed line represents BER = 1.

PBK LEVEL	PROTECTIVENESS	UTILITY
L1	93% (43 out of 46)	8% (2 out of 24)
L2	93% (43 out of 46)	27% (6 out of 22)
L3	98% (40 out of 41)	0% (0 out of 3)
Highest	96% (44 out of 46)	29% (7 out of 24)

Table 2. Protectiveness and Utility statistics for the toolbox varies depending on the parameterisation of the PBK model used. The availability of in vitro and clinical data means that the number of benchmarks at each PBK level changes. The 'Highest' available PBK level statistics consider all available models and compare to appropriate benchmarks to derive the performance statistics, i.e. where L1 C_{max} available the BER threshold of 110 is used, but for L2 and L3 the thresholds of 11 and 2.5 respectively are used.

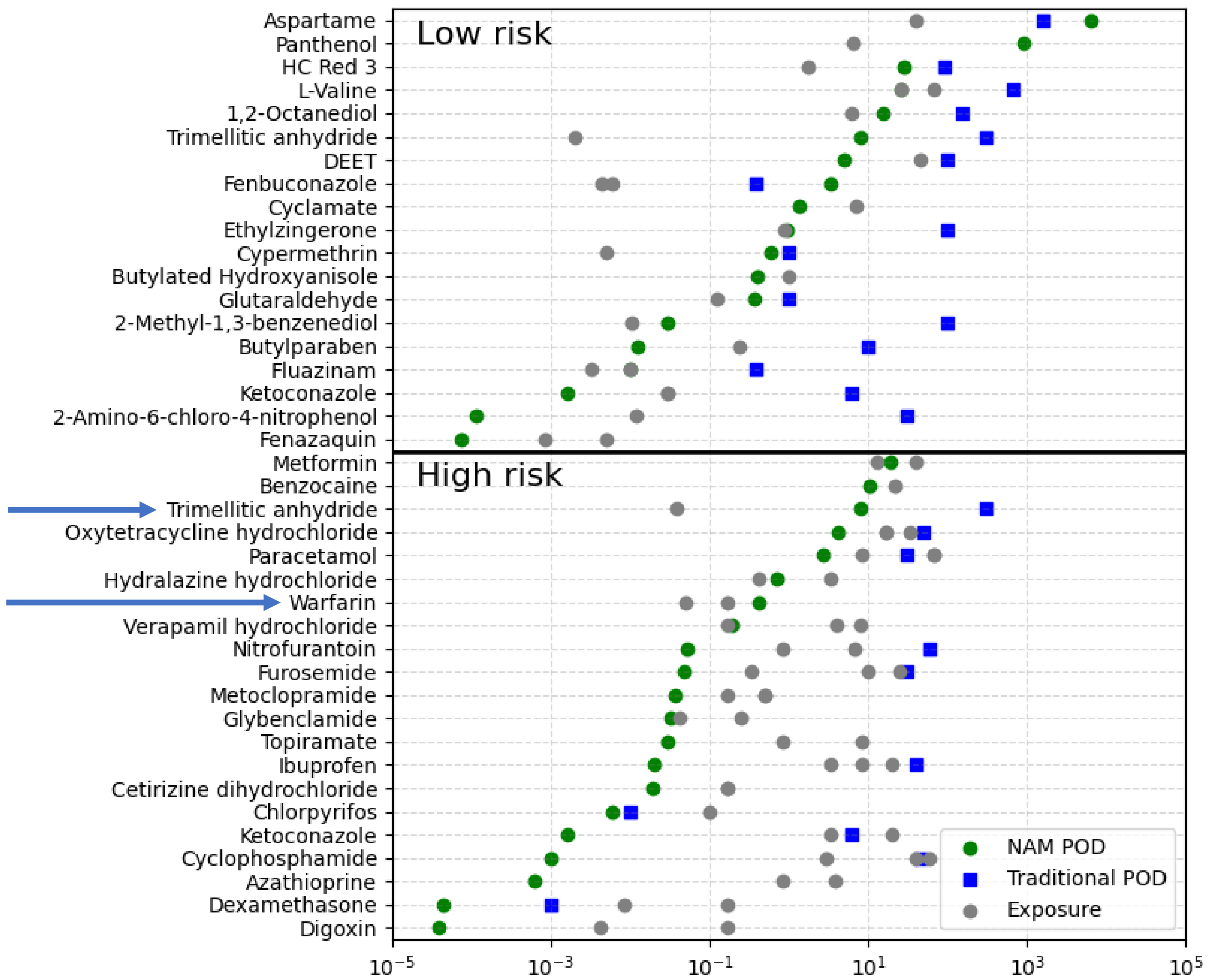

Fig. 5 Summary plot of the external exposure estimates with the converted minimum NAM PoDs and traditional PoDs, separated by the risk classification of the corresponding exposure scenarios. Traditional PoDs are only reported for the 25 chemicals where data were available. Blue arrows highlight the two examples where the Cmax calculated for a high risk chemical exposure scenario is below the NAM PoD

CONCLUSIONS

A NAM-based toolbox can be used to make decisions that **are protective of human health in at least 93% of cases**, despite not predicting the mode of action.

The current proposed toolbox is intended to sit **within a tiered risk assessment framework** and does not differentiate bioactivity from adversity at this stage. The observed low utility could be addressed by the incorporation of further testing or more detailed interpretation of the Tier 0 and Tier 1 results.

More chemicals should be tested to build the reference database from 38 chemicals and 70 benchmark exposure scenarios to **increase confidence** in the applicability of this approach.

seac.unilever.com

References:

Dent et al, 2018: Principles underpinning the use of new methodologies in the risk assessment of cosmetic ingredients. Computational Toxicology. 7: 20-26
Middleton et al 2022: Are Non-animal Systemic Safety Assessments Protective? A Toolbox and Workflow. Toxicological Sciences. 189(1):124-147
Wood et al 2024: Next generation risk assessment for occupational chemical safety – A real world example with sodium-2-hydroxyethane sulfonate. Toxicology. 506:153835
Baltazar et al 2020: A Next-Generation Risk Assessment Case Study for Coumarin in Cosmetic Products. Toxicological Sciences. 178(1):236-252
Rajagopal et al 2022: Beyond AOPs: A Mechanistic Evaluation of NAMs in DART Testing. Frontiers in Toxicology. 4: 838466.
Thomas et al 2019: The Next Generation Blueprint of Computational Toxicology at the U.S. Environmental Protection Agency. Toxicological Sciences. 1;169(2):317-332