The long way from raw data to NAM-based information: overview on data layers and processing steps

Blum, J.; Brüll, M.; Hengstler, J.G.; Dietrich, D.R.; Gruber, A.J.; Dipalo, M.; ... ; Leist, M.

If you can't describe what you are doing as a process, you don't know what you're doing.
W. Edwards Deming

Bench Marks!

The Long Way from Raw Data to NAM-Based Information: Overview on Data Layers and Processing Steps

Jonathan Blum[1], Markus Brüll[1], Jan G. Hengstler[2], Daniel R. Dietrich[3], Andreas J. Gruber[4], Michele Dipalo[5], Udo Kraushaar[6], Iris Mangas[7], Andrea Terron[7], Ellen Fritsche[8,9], Philip Marx-Stoelting[10], Barry Hardy[11], Andreas Schepky[12], Sylvia E. Escher[13], Thomas Hartung[14,20], Robert Landsiedel[15], Alex Odermatt[8,16], Magdalini Sachana[17], Katharina Koch[18,19], Arif Dönmez[18,19], Stefan Masjosthusmann[21], Kathrin Bothe[21], Stefan Schildknecht[22], Mario Beilmann[23], Joost B. Beltman[24], Suzanne Fitzpatrick[25], Aswin Mangerich[26], Markus Rehm[27], Silvia Tangianu[1,20], Franziska M. Zickgraf[15], Hennicke Kamp[28], Gerhard Burger[24], Bob van de Water[24], Nicole Kleinstreuer[29], Andrew White[30] and Marcel Leist[1,20]

[1]In vitro Toxicology and Biomedicine, University of Konstanz, Konstanz, Germany; [2]Leibniz Research Centre for Working Environment and Human Factors (IfADo), Technical University of Dortmund, Dortmund, Germany; [3]Human and Environmental Toxicology, University of Konstanz, Konstanz, Germany; [4]Department of Biology, University of Konstanz, Konstanz, Germany; [5]IIT Istituto Italiano di Tecnologia, Genoa, Italy; [6]NMI Natural and Medical Sciences Institute at the University of Tuebingen, Reutlingen, Germany; [7]EFSA – European Food Safety Authority, PREV Unit, Parma, Italy; [8] SCAHT – Swiss Centre for Applied Human Toxicology, Basel, Switzerland; [9]Department of Pharmaceutical Sciences, University of Basel, Basel, Switzerland; [10]German Federal Institute for Risk Assessment (BfR), Berlin, Germany; [11]Edelweiss Connect GmbH, Basel, Switzerland; [12]Beiersdorf AG, Global Toxicology, Hamburg, Germany; [13]In silico Toxicology, Fraunhofer Institute for Toxicology and Experimental Medicine, Hannover, Germany; [14]Doerenkamp-Zbinden-Chair for Evidence-based Toxicology, Center for Alternatives to Animal Testing (CAAT), Johns Hopkins University, Bloomberg School of Public Health and Whiting School of Engineering, Baltimore, MD, USA; [15]BASF SE, Experimental Toxicology and Ecology, Ludwigshafen am Rhein, Germany; [16]Division of Molecular and Systems Toxicology, Department of Pharmaceutical Sciences, University of Basel, Basel, Switzerland; [17]Organisation for Economic Co-operation and Development (OECD), Environment Health and Safety Division, Paris, France; [18]IUF – Leibniz Research Institute for Environmental Medicine, Duesseldorf, Germany; [19]DNTOX GmbH, Duesseldorf, Germany; [20]CAAT-Europe, University of Konstanz, Konstanz, Germany; [21]Regulatory Toxicology, Research and Development, CropScience, Bayer AG, Monheim am Rhein, Germany; [22]Albstadt-Sigmaringen University, Faculty of Life Sciences, Sigmaringen, Germany; [23]Boehringer Ingelheim Pharma GmbH & Co. KG, Global Nonclinical Safety & DMPK, Biberach an der Riss, Germany; [24]Division of Cell Systems and Drug Safety, Leiden Academic Centre for Drug Research, Leiden University, Leiden, The Netherlands; [25]US Food and Drug Administration, College Park, MD, USA; [26]Nutritional Toxicology, Institute of Nutritional Science, University of Potsdam, Nuthetal, Germany; [27]Institute of Cell Biology and Immunology, University of Stuttgart, Stuttgart, Germany; [28]BASF Metabolome Solutions GmbH, Berlin, Germany; [29]National Toxicology Program Interagency Center for the Evaluation of Alternative Toxicological Methods, NIEHS, RTP, NC, USA; [30]Unilever Safety and Environmental Assurance Centre, Bedfordshire, UK

Received December 17, 2024;

Correspondence:
Jonathan Blum, PhD and
Marcel Leist, PhD,
In vitro Toxicology and Biomedicine,
University of Konstanz,
Universitaetsstrasse 10,
D-78464 Konstanz, Germany
(jonathan.blum@uni-konstanz.de and
marcel.leist@uni-konstanz.de)

ALTEX 42(1), 167-180.
doi:10.14573/altex.2412171

Abstract

Toxicological test methods generate raw data and provide instructions on how to use these to determine a final outcome such as a classification of test compounds as hits or non-hits. The data processing pipeline provided in the test method description is often highly complex. Usually, multiple layers of data, ranging from a machine-generated output to the final hit definition, are considered. Transition between each of these layers often requires several data processing steps. As changes in any of these processing steps can impact the final output of new approach methods (NAMs), the processing pipeline is an essential part of a NAM description and should be included in reporting templates such as the ToxTemp. The same raw data, processed in different ways, may result in different final outcomes that may affect the readiness status and regulatory acceptance of the NAM, as an altered output can affect robustness, performance, and relevance. Data management, processing, and interpretation are therefore important elements of a comprehensive NAM definition. We aim to give an overview of the most important data levels to be considered during the development and application of a NAM. In addition, we illustrate data processing and evaluation steps

between these data levels. As NAMs are increasingly standard components of the spectrum of toxicological test methods used for risk assessment, awareness of the significance of data processing steps in NAMs is crucial for building trust, ensuring acceptance, and fostering the reproducibility of NAM outcomes.

Plain language summary

Toxicological test methods initially generate raw data. These need to be further processed to determine a final outcome, such as the classification of test compounds as hits or non-hits. The processing of the raw data is often highly complex and proceeds stepwise. This process generates many layers of data connected by several processing steps. Any change to these processing steps can impact the final output of new approach methods (NAMs). This means that the same raw data, processed in different ways, may result in different final outcomes. Data management, processing and interpretation are therefore considered important elements of a comprehensive NAM definition. We illustrate data processing and evaluation steps that play an important role. Awareness of the significance of data processing steps in NAMs is crucial for building trust, ensuring acceptance, and fostering the reproducibility of NAM outcomes.

1 Setting the scene

The 21st century is widely acknowledged as the "century of data". Various economists have labelled data as the world's new "petroleum"[1], due to its immense value. Yet, as Hal Varian, a former chief economist at Google, pointed out in an interview in 2009, it's not just about data itself but about how we interact with it. He remarked, "*The ability to take data – to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it – will be a hugely important skill in the next decades*." In the same year, *Nature* published an article titled "*Toxicology for the twenty-first century*", which highlighted the critical role of new approach methodologies (NAMs) in revolutionizing toxicology. The paper emphasized that the field must embrace data-driven methods and utilize NAMs to address the high demand for data on chemicals (Hartung, 2009).

But what is data? The complexity of data acquisition, data structures, and their many processing steps is easily underestimated. Data can be "raw" or "processed", often with several intermediate layers. Examples for data types are quantitative continuous numeric values (with associated uncertainty) such as test compound potency in molarity units, and categorical outputs such as binary (positive/negative) classifications or semi-quantitative groupings (low, medium, high).

In life sciences such as toxicology, experimental data is essential for accepting or refuting hypotheses, for generating knowledge, and for taking data-based decisions. This is especially important when deriving regulatory decisions on approval or classification and labelling of a compound. Therefore, it is important to carefully consider the different types of data being referenced and how specific types of data at various levels are processed.

Despite the importance of the topic, stand-alone publications on general principles of data processing workflows (here also called data processing pipelines) and their typical architecture are hard to find in the literature. However, many related aspects are well covered: (i) There is coverage of data structure in the context of dedicated data processing pipelines. Such pipelines are often used in the context of automated (algorithmic), standardized, and traceable data processing. They may contain some decision points and input by users, as well as quality assurance steps found in dashboards such as the National Interagency Center for Alternative Toxiological Methods' (NICEATM) Integrated Chemical Environment and the U.S. Environmental Protection Agency's (EPA) CompTox Chemicals Dashboard (Bell et al., 2020; Daniel et al., 2022; Feshuk et al., 2023).

Here, we use the term "pipeline" in a general sense, i.e., not suggesting or discussing specific operations and types of decisions, but rather outlining common elements of data processing relevant to many types of NAMs. (ii) One field for which data processing has been described repeatedly and extensively is computational toxicology (Knudsen et al., 2013; Lynch et al., 2024). However, this has largely been done with a technical focus on the data pipelining process or on data integration from various assays. With the advent of artificial intelligence (AI) and machine learning (ML) applied to large data sets, the challenges and opportunities of data analysis and interpretation have grown (Hartung, 2023a,b; Kleinstreuer and Hartung, 2024). (iii) Specific initiatives to document the various steps of data analysis also covered omics technologies and chemical structure curation. Initial efforts focused on the Omics Data Analysis Framework (R-ODAF) (Verheijen et al., 2020) and the National Toxicology Program's (NTP) approach to genomic dose-response modeling (NTP, 2018) to address regulatory concerns with respect to omics data. These efforts parallel international initiatives to cover all steps in reporting omics data analysis via the OECD Omics Reporting Framework (OORF) (Harrill et al., 2021; OECD, 2023), which aims to foster

[1] Expression coined by British mathematician Clive Humby in 2006.

reproducibility and transparency for regulatory toxicological application by providing a framework that ensures all the required data, together with associated metadata and data analytical processes, are reported for review by the data evaluators. Also under the auspices of the OECD, the (Q)SAR community developed a (Q)SAR Result Reporting Format (QRRF) to complement the (Q)SAR Assessment Framework (QAF) for similar purposes (OECD, 2024), and substantial work has been done by the US EPA and NICEATM to develop chemical structure standardization and curation pipelines (e.g., Mansouri et al., 2024).

Many users of NAMs or of their data would benefit from an open, pipeline-independent display and discussion of the steps that may be relevant to process raw data into final NAM test outcomes. The implications have been captured by Kessel et al. (2023), using example data of NAMs for developmental neurotoxicity (DNT), to clearly highlight that different approaches to management of raw data and different uses of biostatistical methods can affect the outcome of NAMs, e.g., a hit definition (Zhu et al., 2013) or a point of departure (PoD) (Sturla, 2018). Kessel et al. (2023) gave five examples for biostatistical procedures that can affect the final test results. Another example comes from the field of mixture evaluations, where five mathematical models led to different outcomes of the same input dataset (Lasch et al., 2020).

Here, we provide a generalized, broadly applicable framework to position such key procedures and other steps that affect a NAM outcome. We explain typical levels of NAM data processing, illustrate how the data might be processed, and exemplify some pitfalls that should be avoided in this context. The intention is to give an overview of choices to be considered for NAM data processing and to indicate what needs to be covered by a complete NAM description. It should be noted that the majority of validated NAMs address hazard identification and characterization. However, some also produce data to parametrize toxicokinetic models (physiology-based kinetic modelling, PBK) (Tsaioun et al., 2016) or to derive PoDs for quantitative risk assessment (e.g., Reinke et al., 2025), and most considerations also apply to such NAMs. Moreover, the issue did not arise from the increased use of NAMs. In classical toxicology, complex processing of raw data (e.g., pathological scoring of tissue sections) to final outputs (e.g., tumorigenicity) is widely applied, and the impact of data processing pipelines on outcomes tends to be underestimated there also.

2 Delimitation to related topics

The topic of data processing to generate a final NAM output is sometimes considered of minor importance compared to other, more obvious issues. Moreover, not all stakeholders realize the definition and requirements for data processing. Finally, data processing is often confused with other topics, such as data storage or data interpretation.

We do not address issues of data structure and data storage in detail here. However, important aspects related to this topic are crucial: data of public concern or generated with the help of public funds should be findable and accessible. If they are accessible, they should be documented to the extent that the process used to get from the raw data to the final output is sufficiently detailed to be independently reproduced. Also, documentation should allow for assessment of the quality of the reported outcome and the relevance of the assay. Findability and accessibility are key features of the FAIR principles (next to interoperability and reusability of data) (Wilkinson et al., 2016). Consideration and standardization of metadata and definition of protocols for the data processing pipeline are key elements of FAIR data[2] and thus overlap with the topic of our article. We hope to raise awareness here that metadata and processing protocols are not just background details of data storage but are prerequisites for NAM development, validation and use, and especially for the interpretation of NAM data. Moreover, quality control is a formal requirement for Good Cell Culture Practice (GCCP) (Pamies et al., 2022), Good In Vitro Method Practice (GIVIMP) (OECD, 2018), and within the extensive Good Laboratory Practice (GLP) framework[3] for data used in regulatory decision making.

Some important background information has already been covered elsewhere: Extensive guidance already exists on, e.g., criteria for test method readiness (Bal-Price et al., 2018; Crouzet et al., 2023), detailed NAM descriptions (OECD, 2017b; Krebs et al., 2019), integration of NAM data in a regulatory context (OECD, 2017a; Schmeisser et al., 2023), or components of test method papers (Leist and Hengstler, 2018; Collen et al., 2024). Here, we also do not cover the experimental part of data generation with NAMs. Moreover, discussions on data context or suggestions on how to proceed with hits or non-hits, e.g., in the developmental neurotoxicity *in vitro* battery (DNT-IVB) (Blum et al., 2023), are covered elsewhere (Pallocca and Leist, 2022; Hartung et al., 2024; Smirnova et al., 2024). Our presentation of a generic set of elements to be considered for a data processing pipeline should also not be confused with dedicated and implemented pipelines (Bell et al., 2020; Daniel et al., 2022; Feshuk et al., 2023) or with issues of data integration, e.g., by combining NAM data in a defined approach for skin sensitization (Kleinstreuer et al., 2018; OECD, 2021; Strickland et al., 2022).

3 Multi-step data processing

For the generic data processing pipeline (procedure/pathway) described here, we focus on the "life stages" of data, covering the period from when they are born (data acquisition, method/machine output) to when they reach adulthood (final test method outcome). It is, for most NAMs, an oversimplification to assume that test methods generate raw data that are converted in one step to the final output. In reality, six or more levels of data may be defined (Fig. 1; Tab. 1). To understand the complexity of data processing it is useful to first look at the different data layers, i.e., the types of

[2] Also part of the readiness criteria considered in the ongoing revision of OECD GD 34 (OECD, 2005) on method validation.

[3] https://www.oecd.org/en/topics/sub-issues/testing-of-chemicals/good-laboratory-practice-and-compliance-monitoring.html

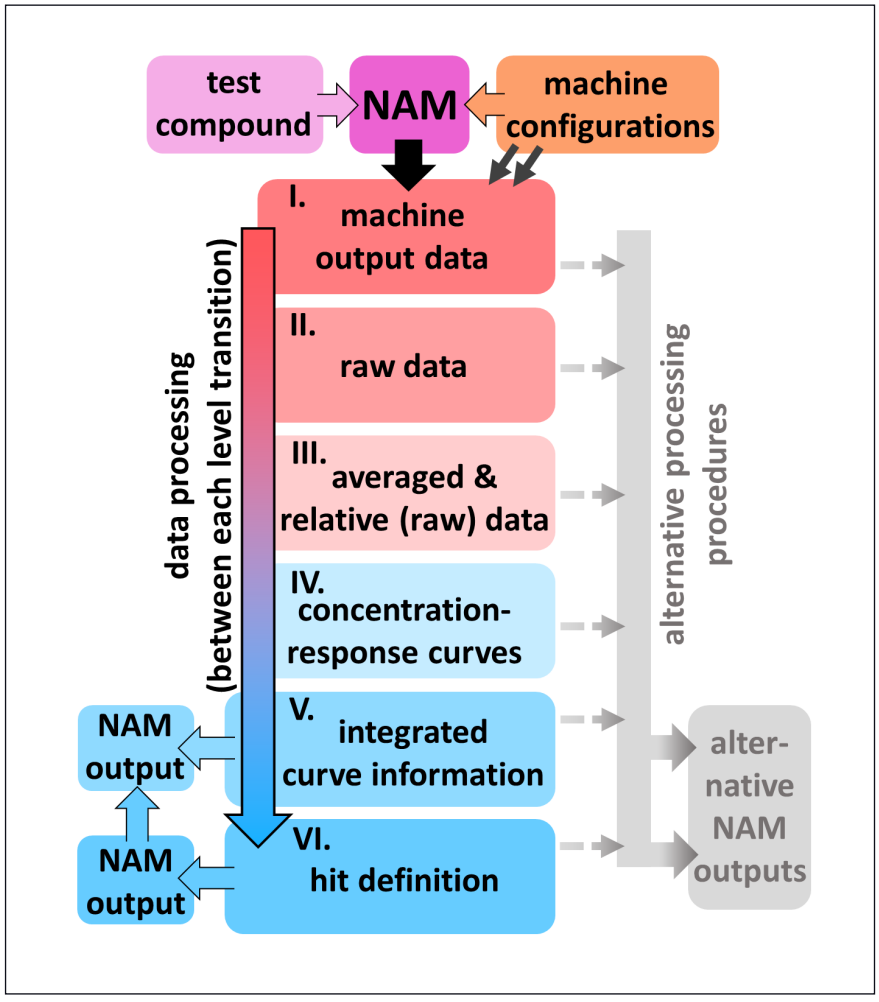

Fig. 1: Overview and concept of data levels in NAMs
Compound testing in NAMs generates initial machine output data. The raw data are processed in various steps (exemplified by six levels) towards the final NAM output in a data processing pipeline. The characteristics of every level on this path are important features for the design and validation of a NAM, as well as for the use of data generated by NAMs. Alternative processing procedures can alter the NAM outputs. Level I data (machine output) not only depends on the test method and the test item (chemical), but also on the machine used to measure the analytical endpoint of the NAM, and on specific machine configurations (which may be considered level zero). While this generic overview applies to many NAMs, there may be cases where one of the levels is missing, or where additional levels are added.

data generated at different steps of data processing (this section), and then more closely examine how data transition from one level to the next, i.e., what types of procedures and inputs are relevant (Section 4). Staying within the metaphor of data "life stages", there is also a fetal stage (Level zero data) and an adult life stage (Level IATA). These are briefly discussed before and after Level I-VI to provide an overall frame; their detailed discussion is outside the scope of this review.

3.1 Level zero: Machine configurations

We use level zero here to indicate a data domain not generally considered in data processing pipelines yet important for complete test method descriptions. One could also cover this level zero as processing input, i.e., as part of the data processing section (Section 4). We prefer to mention it here, as in the processing section we focus on steps between typical data levels, but not on the steps before the first level. Notably, the term "level zero" has an entirely different meaning for ToxCast data (where it is used to describe the usual data entry format).

Many "machine parameters", i.e., properties and physical settings/configurations of the instruments used to assess the test method analytical endpoint (microscopic imaging systems, fluorimeters/luminometers, mass spectrometers, electrophysiological recording setups, polymerase chain reaction machines, fluorescence-activated cell sorters/scanners, etc.), can affect level I data. Modern laboratory instruments often only have an on/off button, a start button, and some sample input. Researchers may not be allowed or able to change hardware parameters or associated software settings. An example of a simple "machine" is a fluorescence microscope. Usually, the lenses, light source, filters and camera are a fixed setup. However, microscope types may differ between laboratories (and be equipped with different lenses, light sources, filters, and cameras), thus resulting in different results. For many reasons, the "machine" cannot be defined in every detail in a NAM description. Moreover, suppliers, procedures, technology, etc. continuously change, so that the analytical device even within one given lab can change. Hence, it may happen that variability is observed among results although the test system, the exposure scheme, and many other NAM elements defined in the test method description remain unchanged.

Some classic examples of machine specifications are: (i) optical filters that are narrow or wide; photo multiplier or camera

Tab. 1: Exemplification of data levels
More detailed description of data levels that are depicted in Fig. 1. The six main data levels I-VI are displayed between bold lines. The data level in the grey boxes is not covered in this manuscript.

Data level	Description	Examples	Typical steps leading to this level
***"Level zero"*:** Machine configurations	Defines the hardware conditions for data acquisition; settings and technical parameters that influence measurements	Fluorescence filter settings; camera sensitivity settings	Strictly speaking, this is rather a *processing step* than a level. It is included here, as it is *before* the first level
Level I: Machine output data	Initial, machine-generated data, often in raw form	Voltage changes; light intensity; pixel values in image data	Data acquisition
Level II: Raw data	Data sets with measurable values (e.g., numbers, areas) that are directly accessible	Cell count in an image; fluorescence intensity	Conversion of machine data to human-readable; numbers, values, etc.; quality control; filtering of erroneous data
Level III: Averaged and relative (raw) data	Normalized or relative values, often in relation to control data	Ratios/percentages relative to the control	Data aggregation; data normalization; check of AC
Level IV: Concentration-response curves	Curves representing responses to different concentrations	Sigmoidal toxicity curves	Curve fitting and modeling (data aggregation)
Level V: Integrated curve information	Reducing curve data to single numerical values, such as threshold concentrations	BMC; PoD (may also be expressed as probability function instead of a single value)	Threshold calculation; uncertainty analysis of PoD
Level VI: Hit definition	Final decision or classification	"Hit" definition; toxicological classification; activity call	(Calibration to positive control items); application of prediction models (PM, DIP) for classification
"Level IATA" *(not covered here)*	Integration of outputs of various NAMs in an IATA or computational model	Use of DA, such as OECD GD 497 or TG 467 to result in a GHS classification as output	Data may be integrated at, e.g., level III or VI (with different outcomes)

AC, acceptance criteria; BMC, benchmark concentration; DA, defined approach; DIP, data interpretation procedure; GHS, Globally Harmonized System of Classification and Labelling of Chemicals; IATA, Integrated Approaches to Testing and Assessment; OECD GD, OECD guideline; PM, prediction model; PoD, point of departure

(CMOS, CCD) settings for high/low sensitivity or gain; (iii) electrode impedance/amplifier characteristics for microelectrode array (MEA) devices; (iv) ambient control (temperature, CO_2) for many devices such as automated microscopes or MEA devices. How can one ensure that data level I (machine output data) remains as consistent as possible based on the inputs from level zero data? A common approach is to define performance standards, i.e., to define a set of test conditions (e.g., various compounds at certain concentrations used in the NAM) and to require results to remain within a certain range. This is commonly done based on test outcome data (level V or VI), and it may require considerable evaluation efforts (Petersen et al., 2023). Sometimes, a definition based on lower data levels can be more stringent and more precise. The most immediate approach to this is sometimes the "calibration" of the machine or the analytical endpoint assessment procedure.

Guidance for NAMs should consider exact settings described in standard operating protocols (SOP) and in detailed method descriptions such as the ToxTemp (OECD, 2017b; Krebs et al., 2019). Good In Vitro Reporting Standards (GIVReSt) aim for this (Samuel et al., 2016; Hartung et al., 2019) but are specific to particular categories of NAMs. Beyond this, awareness of machine settings/characteristics on the outcome of level I data is important, especially if NAMs are transferred across laboratories. Typically, machine configurations are described in SOP sections on the analytical endpoint; in addition, some may feature as metadata.

3.2 Level I: Machine output data

All NAMs use some form of machine-based measurement as their analytical endpoint. This can be considered the initial level of data generated by the NAM (Fig. 1). This data level (level I) represents the "rawest data" produced by a NAM, often with a very complex structure and hard to read or interpret by the experimenter. We prefer the term "machine data", as we reserve the term "raw data" for the next data level. However, there is no fixed rule on how "raw data" are defined. It depends on the analytical endpoint used, and there are differences amongst NAMs. It is more important to explain the terminology behind the levels, and to be transparent in what is done at which level.

For several NAMs, level I data undergo some processing before they can be considered "classic raw data" (level II). After the processing, data are more readily human-readable. For this reason,

the first data level that is stored in repositories and/or used for further processing is sometimes not level I, but level II.

A few examples may help to clarify this: Machine data can be current or voltage changes, often combined with some time information (within the machine). Some of this is further processed within the machine (e.g., in a spectrophotometer, the photomultiplier current of the samples and a reference light beam are combined and converted to an absorbance signal). This is done by proprietary software linked to the hardware. Users of NAMs must have sufficient trust in the reliability of the hardware's data output.

For many NAMs, level I data generated by an analytical device are accessible (and storable). However, in many cases, they are processed further (for a number of technical and practical reasons) before storage and subsequent use in the NAM data processing pipeline. Examples are (i) the generation of time-dependent spike sequences (e.g., in electrophysiological readouts from initial voltage recordings); (ii) generation of base sequences from initial electrical or optical readings in DNA sequencing runs; and (iii) fragment/molecule-specific displays for mass spectrometric data. Sometimes, positional information is added to primary data (e.g., from a photomultiplier), and data are then stored as 2D arrays, i.e., pixels of a raw image. In many such cases, the machine output data cannot be interpreted without substantial context and additional processing. Such data are often processed to the next level, greatly reducing the data volume (and storage space requirements).

As some of the processes done by dedicated software within (or associated with) analytical devices (machines) may not be described within the method description, the hardware used for the data acquisition and initial processing should be specified in conjunction with the latest software version running the hardware. For these reasons, it is useful to define "level zero" data that takes the possible effects of different machine types and configurations into account (see Section 3.1). These aspects are also important details to consider when assessing potential sources of variability within NAM experiments (Petersen et al., 2023).

3.3 Level II: Raw data

At the level II stage, the data mostly takes the format of numbers, linked to a unit of measurement or a quantitative scale. Examples include (i) object area/size/count, if the test endpoint is based on microscopic images; (ii) abundancy counts or cycle numbers in transcriptomic/PCR-based measurements; (iii) nucleotide sequences for sequencing data; (iv) absorbance or fluorescence values. Out of these examples, the first point most clearly demonstrates the enormous data volume reduction (several orders of magnitude) needed, e.g., from a series of images to the number of cells in these images (a single figure). At level I also several primary analytical endpoints (e.g., fluorescence channels) may be combined to generate the actual raw data. One example is the combination of information from various fluorescence channels in image-based data to identify biological features (e.g., size and number of cells or cell organelles).

At this stage of data processing, all single sample replicates (often referred to as technical replicates) of an independent assay run (often referred to as biological replicates) are still fully separated. This data tier is often considered the most suitable "raw data" input format for databases. It provides a high level of detail (granularity). Thus, it leaves many choices open (e.g., types of data integration, types of regression analysis, and uncertainty calculations) for alternative analysis pipelines. An important consideration at level II is that some data may already have been flagged (considered invalid) and therefore excluded from further processing. Curation of data at the transition of level I to level II is based on technical quality control measures. For instance, cells may have been contaminated, or an image may be all black, or a current spike may have resulted in an obvious artifact, or errors may have occurred during the experimental study part (no cells plated in one well, results affected by construction work, etc.). Removal of such data (flagging for exclusion) should not be considered as loss, but as necessary curation required for subsequent analysis that should be traceable.

3.4 Level III: Averaged and relative (raw) data

Many types of data on level II cannot be interpreted if they are not presented relative to control data. Such data comprise, e.g., the object area of images, electric signals in MEA, counts of nucleic acid species, and fluorescence/absorbance values. In level III, the necessary context is incorporated. As a first step, multiple data points such as technical or biological replicates must be summarized. Depending on the data summary method, i.e., forming a mean, median, or minimum activity concentration, results might change (Kessel et al., 2023). At level III, data are normalized to values obtained from reference items. Moreover, level III is distinct from level II in that samples generated under identical conditions within one experiment (technical replicates) are averaged. In some data processing pipelines, data generated in different experiments may also be averaged at level III. Some test conditions, specified as negative controls[4], serve to define a baseline or reference level of the test endpoint. The level III data of test chemicals are often given as fractions of this reference level. As data are given "relative" to "no disturbance" (a negative control), level III is the first stage where the data explicitly indicates an "effect" that can be easily interpreted by humans (notably, the effect is already present in the data before), often expressed as a percentage or a "fold change". An example may help to clarify the difference between levels II and III: If alcohol is consumed, the performance in an attention test may be reduced by 27 points (level II data). This data cannot be interpreted if one does not know how many points full performance represents. The relative reduction in performance could be 1% or 80% compared to control. The average performance of, e.g., a sober control group would need to serve as a reference, similar to negative control samples in a NAM. Thus, only data from level III onward provide information on the effect size.

Two special cases are mentioned here for completeness: (i) for NAMs with multiparametric readouts (e.g., cell painting or transcriptomics), new composite endpoints may be generated for further

[4] The topic of defining controls falls outside the scope of our overview and is covered elsewhere (e.g., GIVIMP (OECD, 2018), GD 34 (OECD, 2005) revision).

processing through levels IV-VI. A typical example would be the defined combination of a group of transcripts in a pathway, a gene ontology term group or a weighted gene correlated network analysis module (Callegaro et al., 2021). In all such cases, the combined group readout would be a single average or summary response parameter; (ii) sometimes level III data can be adjusted for response dynamics by the use of positive controls. This type of normalization differs from normalization to negative controls (baseline).

3.5 Level IV: Concentration-response curves

Level IV data differ from level III data in that they give information on how the test endpoint changes when the concentration of test compounds is increased. This level is not always mandatory, but it is necessary to later derive toxicological information on effect/hazard thresholds, e.g., a PoD (Box 1). The data format at level IV is a "curve", which is usually a mathematical function with concentrations plotted on the abscissa (x-axis) and effect size as a dependent variable on the ordinate (y-axis). Such curves often have certain features, such as monotonicity. "Toxicological curves" also tend to reach an asymptote so that further increases in concentration do not change the outcome. On a semi-log axis such curves appear sigmoidal.

The data processing procedure by regression analysis (colloquially termed "curve fitting") may have rigid rules (e.g., only allow certain curve shapes) or be very open; for instance, it may include non-monotonic curve shapes and non-parametric approaches (Kappenberg et al., 2021; Wheeler, 2023). The requirements for the underlying data structure increase with the complexity of the considered curve shape. The requirements for a straight line are low (few data points may be sufficient), while fitting a meaningful higher polynomial curve requires data from a larger number of concentrations. Typical sigmoidal curves, such as 4-parameter logistic regressions, may be constructed from 6 to 8 "meaningful" data points. This means that not all data points can be in the saturation part and that there should be a minimum number of points in the part with the steepest slope. Another very important aspect is that sufficient data points define the baseline (start level) of the curve (Krebs et al., 2018; Kappenberg et al., 2020).

Level IV data are defined to a large degree via the curve fit selection criteria applied. Decisions on the type of curve fitted (regression model used) may be based on the correlation coefficient of the data points and on considerations like the avoidance of overfitting (e.g., Akaike information-criteria (AIC) or residual analysis) (Krebs et al., 2020a). It has been exemplified previously, e.g., in the DNT-IVB, that variations in curve fitting models can yield different results and impact subsequent level V data (Kessel et al., 2023).

Although averaging of independent experimental runs, i.e., "biological replicates", seems like a standard procedure, it can sometimes be a remarkably complex issue. It also may be addressed in different ways in otherwise quite similar data processing pipelines. This depends, in part, on how an "independent experiment" is defined (this topic will be addressed in a follow-up article). Here, we discuss the need for combining data from several experiments in general, as the transition from level III to level IV may include a data averaging step when experiments have been run several times. One approach that considers all data in one step would be to apply mixed effect models (mathematically complex). Three alternative approaches are (i) averaging of the data points of each test condition (mathematically easier) before curve fitting (this would be an additional step leading from level II to level III); (ii) fitting the curves through all available data points without consideration of independent runs; (iii) calculating the integrated curve information (e.g., the PoD) for each experiment and averaging the values at level V (mathematically easiest). The procedure chosen for integrating data from several experimental runs is usually defined in a NAM's SOP. Each approach has some experimental, statistical, practical, and conceptual considerations, but also disadvantages. Not covered here are Bayesian approaches, which result in probability distributions over the concentration range and can potentially better quantify propagated uncertainty throughout data modelling to results (Reinke et al., 2025). The type of procedure chosen can affect the test outcome and therefore needs to be clearly and transparently defined in the method description and/or in the output data repository.

Box 1: Nomenclature on assay "hits" important for level VI

The field of alternative methods recognized early on that NAMs need a prediction model (PM), i.e., something converting NAM test data (level V) into a prediction of a toxicological outcome (Leist et al., 2010; Crofton et al., 2011; Ferrario et al., 2014; Collen et al., 2024). Initially, it was believed that a single NAM would predict an apical adverse outcome (e.g., eye irritation). This concept has been mostly abandoned, but the concept of the PM remains. It is considered now to interpret the data produced by the NAM in terms of a classification of test items as having or not having "an effect" in the respective NAM. PM should not be mistaken to be only an issue of data science, i.e., a specific type of algorithm to transform data. The role is much broader and of high importance to modern toxicology: a PM is the essential link from data (generated by NAMs) to a biological/toxicological interpretation. The exact definitions of the term have been changing with time and among organisations. The below set of definitions gives a general overview:

Prediction model: An "*algorithm that converts each test result into a prediction of the (toxic) effect of interest*" (OECD, 2005). The PM procedure is used to convert the results from a test method into a prediction of the toxic effect of interest. A PM contains four elements: a definition of the specific purpose(s) for which the test is to be used, a definition of all possible results that may be obtained, an algorithm that converts each test result into a prediction of the toxic effect of interest, and an indication of the accuracy of the prediction[5].

[5] https://ntp.niehs.nih.gov/sites/default/files/iccvam/docs/about_docs/validate.pdf

Data analysis procedure (DAP): A procedure according to which raw data of a NAM are converted into a result (e.g., the IC_{50}) by a specified algorithm (pipeline). The last step may involve the PM.

Data interpretation procedure (DIP): An interpretation procedure used to determine how well the results from the test predict or model the biological effect of interest. In many cases it converts test data (e.g., IC_{50} values for various test endpoints) into a classification (e.g., non-toxic, borderline, toxic). Without a DIP, classical validation procedures (according to OECD, 2005) cannot determine predictivity, and under such circumstances an important element of validation would fail (Leist et al., 2012; Collen et al., 2024). For most practical purposes, a DIP is an alternative term for a PM.

Screening hit: Originally defined as a compound that exhibits the desired biological activity towards a drug target in a pharmacological screening. It has been suggested that the activity must be confirmed by re-testing to call an "assay positive" a real screening hit (Smirnova et al., 2024; Magel et al., 2024). The definition was transferred to toxicology as: a compound that exhibits an effect in a NAM with effect size boundaries predefined by the PM. Notably, some pipelines (e.g., tcpl, used in ToxCast) use continuous hit calling (instead of a categorical call). The output is then a potency value (in concentration units), which indicates that the test item is a "hit at concentration x µM".

Point of departure (PoD): In toxicology, this relates to the lowest concentration (or dose) at which a biological response is first observed. In a narrow sense, the PoD refers to responses considered to be adverse. The PoD is the starting point for calculations/extrapolations needed for risk assessment. In next generation risk assessment, PoDs are derived from NAMs. Often they are considered as starting points for *in vitro* to *in vivo* extrapolations. In a general sense, the PoD is the link between experimental toxicology and risk assessment processes that affect regulatory policies and risk management.

3.6 Level V: Integrated curve information

In contrast to the data format of level IV (which usually consists of concentration-response curves), data in level V are individual numbers. These numbers capture information from the level IV curves. They reduce a two-dimensional information matrix (concentration vs effect) to a single number that is representative of the intended measurement. Classical examples of such information are a PoD, a lowest observed (adverse) effect level (LO(A)EL), or a benchmark concentration (BMC) for individual endpoints (Krebs et al., 2019, 2020b; Delp et al., 2021; Blum et al., 2023; Zobl et al., 2024) or for omics data (OECD, 2023). It is important to understand that this "magic" data reduction is only possible based on certain assumptions and pre-defined processes. For instance, a BMC can only be derived if a benchmark response (BMR) is set (Krebs et al., 2020a), i.e., the critical concentration (BMC) depends entirely on a pre-determined critical effect (BMR). Such definitions (parameter settings) are essential elements of NAM descriptions. Changing such parameters, e.g., in an alternative analysis pipeline, will likely influence some NAM outputs.

For some purposes, level V data may serve as the NAM output (Fig. 1), e.g., where it provides information on the potency of a test item concerning one (or more) test endpoints. Some NAMs, or some problem formulations, do not require a classification or hit definition. The NAMs originally validated by the OECD were intended to make direct toxicological predictions for classification according to the Globally Harmonized System of Classification and Labelling of Chemicals (GHS) system. All such NAMs in OECD test guidelines (TG) have a classification endpoint (level VI), e.g., categorizing compounds to be mutagens, eye irritants, or phototoxicants. For mechanistic NAMs, such a classification is not always given, and in such cases, level V may be considered a final data level and the exit of the processing pipeline. At this stage, measures of uncertainty around the output data (e.g., a 95% confidence interval of a BMC or PoD) are nearly as important as the data values themselves. Conversely, the level V output loses some value if it is not associated with uncertainty measures. The procedure to assess uncertainty (e.g., algorithms used to determine the coefficient of variation or a lower bound of a BMC confidence interval) is part of good test method descriptions (Krebs et al., 2019) and data integration (OECD, 2017a; Watt and Judson, 2018; Krishna et al., 2021). The uncertainty of a PoD may be calculated and indicated in different ways that can be communicated via level V outputs (Leontaridou et al., 2017, 2019; Gabbert et al., 2022; Kessel et al., 2023).

3.7 Level VI: NAM output/hit definition

For many NAMs, level VI data results in a hit identification or a "statement of bioactivity". In some databases, e.g., Tox21/ToxCast, this is also termed an "activity call". The data are neither a curve nor a number, but a classification or decision outcome that is dependent on a data interpretation procedure (DIP) or prediction model (PM) (Box 1) and uses level V data as input to an algorithm or a formula (Schmidt et al., 2017; Collen et al., 2024). This critical aspect of NAM development requires time and considerable resources, as large data sets need to be generated and analyzed. It is usually the basis for validation of a method and for defining performance metrics (Leist et al., 2010; Crofton et al., 2011; Aschner et al., 2017).

Level VI data are typical NAM outputs used to support regulatory (and research) decision making. However, the output of some NAMs can be perceived as being derived collectively from level V and VI. Consider the example of a neurite outgrowth assay to investigate DNT (Krug et al., 2013; Blum et al., 2023; Suciu et al., 2023). A compound that affects neurite outgrowth specifically would be classified as positive/a hit/an activity at level VI, and the lowest concentration at which it affects neurites is X µM (compound potency/PoD, determined at level V). Therefore, a typical "combined output" is "hit at concentration X". An alternative outcome is "no hit until concentration X" or "test negative until highest investigated concentration".

Instead of interpreting levels V and VI separately, some pipelines (e.g., tcpl, the ToxCast Data Analysis Pipeline) use a continuous hit call (instead of a categorical call). In simplified terms, the combined output of levels V/VI is then a potency value, which indicates that the test item is a "hit at concentration X µM". This procedure works well for assays with a single endpoint or with multiple endpoints that are independent in their biological implication.

In cases where endpoints strongly affect one another (usually in assays with multiple functional endpoints plus cytotoxicity), a categorical hit call has some advantages (elimination of false hit calls of the functional parameter under conditions of cytotoxicity). It allows also a sub-specification of non-hits into "no effect" compounds and "cytotoxic" compounds. One may also include borderline categories whose definition is sometimes based on uncertainty measures that were derived at the previous data level (Leontaridou et al., 2017, 2019; Delp et al., 2018; Gabbert et al., 2022). Moreover, statistical tools and measures may allow data interpretation procedures based on probabilistic assessment (Leist et al., 2014; Maertens et al., 2022).

Notably, a feedback connection may be required between level VI and level V data when a NAM has more than one endpoint (Fig. 1). In such cases, a hit definition may require integration of test endpoints. For instance, a cell function endpoint could be compared to a cytotoxicity endpoint, considering that dead cells cannot grow neurites or have mitochondrial function. This means that the data on reduced neurite outgrowth or mitochondrial function are likely to be wrong (i.e., meaningless in a biological sense) because the preconditions for the measurement (i.e., having similar numbers of live cells in each culture compartment) have not been fulfilled. Therefore, many NAM method descriptions specify that there must be a data integration step (at level VI) that flags data on specific cell function effects as valid or non-valid, based on the measurement under consideration. This step determines whether data on an individual endpoint, such as neurite outgrowth determined at level V, may be used, and in turn explains why level V data on specific cell functions may become invalid (not usable for hazard evaluation) if level VI has certain outcomes (identification of potent cytotoxicity). Re-analysis by alternative pipelines (e.g., not using such a feedback connection) may not consider such rules in the same way and thus lead to an altered overall output.

3.8 Level IATA

As mentioned for level zero, level IATA is outside the core data processing pipeline and therefore not discussed extensively here. However, it is an essential subsequent step for the regulatory use of data (Tab. 1). As the output of a single NAM is usually not sufficient for risk assessment, data from many NAMs may be integrated to come to a regulatory weight-of-evidence conclusion on specific forms of activity, such as developmental neurotoxicity (DNT), non-genotoxic carcinogenicity, endocrine disruption or various other forms of systemic or topical toxicity (Kleinstreuer et al., 2017; Casey et al., 2018; Ebmeyer et al., 2024; Jochum et al., 2024; Najjar et al., 2024a,b). This integration procedure can consider various regulations and may have different objectives and outcomes, ranging from setting health-based guidance values to classifying toxicities (according to GHS or the CLP regulations) to deriving PoDs for quantitative risk assessment. A large portfolio of case studies on the use of IATA is available as the OECD Case Studies Project[6]. Data types used may not be limited to hazard but may also address exposure and toxicokinetics (Chang et al., 2022). Some forms of data integration use only expert judgement, while others aspire to be fully computational; many combinations and tiered approaches are currently being explored. One of the recent approaches to establish IATA level data on hazard is to develop quantitative adverse outcome pathways (qAOPs), in which the relations between molecular initiating events (MIEs), key events (KEs), and adverse outcomes (AOs) are described mathematically (Perkins et al., 2019; Di Tillio and Beltman, 2024). Here, level V and VI NAM data are used as representatives of MIEs, KEs, and/or AOs. An extension of this concept is the combination of qAOP with toxicokinetic models in quantitative systems toxicology approaches (Polak et al., 2019; Beattie et al., 2024) as the highest level of computational data integration to assess human risks of chemicals.

4 General overview of data processing steps between the data levels

As demonstrated above, similar raw data can result in different NAM outputs when processing steps are altered. For this reason, we will give a general overview on what "data processing" may imply.

Since data processing occurs between all levels, one data level may always be considered the upstream data level and the one resulting from processing the downstream data level (Fig. 2). The term processing is used here in its widest sense, including (i) the mathematical transformation of data point numeric values; (ii) the integration of data points; (iii) the amending of data points with additional information; (iv) the relation of data points to one another or (v) the flagging of data points (sometimes leading to exclusion from further processing). At each transition from one level to the next, data may be "processed" in three fundamental ways (category 1-3). Categories 1 and 2 have subcategories (1A, 1B, 2A, 2B). A general overview of types of data processing is given below. An important notion is that data processing implies the use of additional "information" or rules. This means that the steps between the levels are not just determined by the raw data. Or, seen from a different perspective, the raw data are not sufficient to determine the final outcome at level V or VI. Here, a general overview is given, while a more extensive description and exemplification by application to various case studies is planned as a follow-up publication, and descriptions of some exemplary pipelines (e.g., tpcl[7] or CRSTATS) can be found elsewhere (Filer et al., 2017; Daniel et al., 2022; Kessel et al., 2023).

6 https://www.oecd.org/en/topics/sub-issues/assessment-of-chemicals/integrated-approaches-to-testing-and-assessment.html

7 https://www.epa.gov/comptox-tools/exploring-toxcast-data

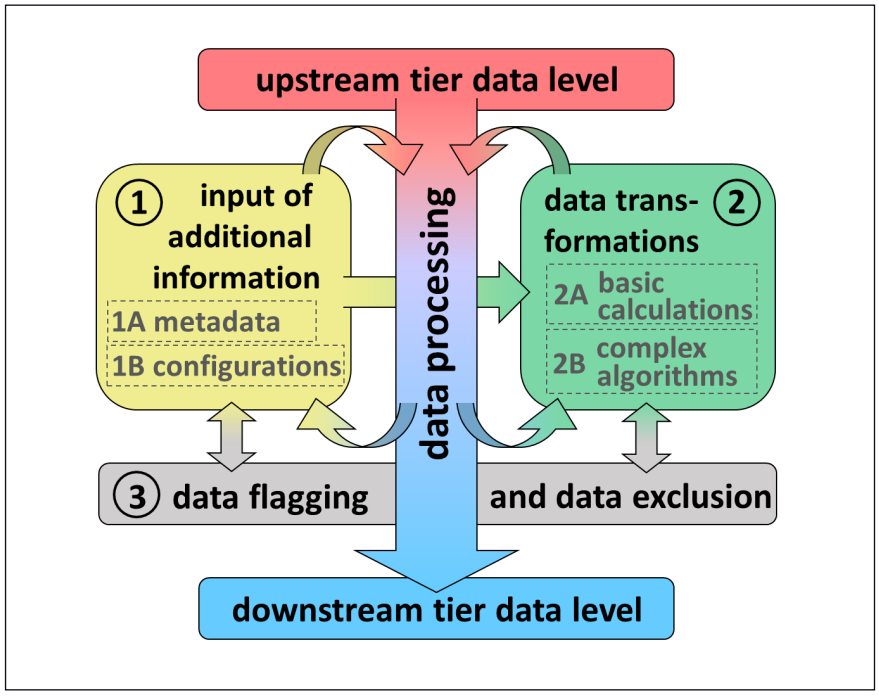

Fig. 2: Processing categories between any two data levels
Data processing occurs between all data levels. This means that one level can always be considered upstream, and the next level as its downstream data level. Data processing between the levels is affected by three types of input/actions. These are depicted here as different process categories. Category 1 involves the input of additional information to the data. Category 2 involves the transformation of data. Category 3 includes flagging and exclusion of data. There are feedbacks and interactions between these categories. For instance, category 1 input often affects data transformations (category 2). Moreover, there can be feedback from the data stream into the processing categories.

4.1 Data processing category 1: Input of additional information

The additional information that is added to the test data consists of an "information package". This may be a number (e.g., calibration curve), spatial information (e.g., plate layout) or an experimental variable (e.g., analysis time). It therefore differs from category 2, where the input is an action (e.g., a calculation rule). The two main types of category 1 input are metadata (1A) and configurations of processes (1B) (Fig. 2). Input information from category 1 may be used to parametrize category 2 actions.

Metadata (1A input)

They may apply to any transition from one data level to another to support data processing. Metadata must be reported where they apply, as they include information required for data transformations. The usual procedure is to define them as fixed values/parameters in detailed SOPs. An example is a plate map (with sample positioning), information on reference values or information on variable experimental parameters. Such metadata information is, e.g., required at level II (raw data). To process the data to level III (averaged and relative (raw) data), metadata is needed to indicate which samples are treatment conditions, controls or replicates, and how they have to be calculated. Another example is the specification of a BMR so that, e.g., a BMC25 or a BMC50 may be calculated (level IV to V processing).

Configurations (1B input)

They differ from 1A input in that the 1B input is not necessarily in the form of fixed values. The input information for process configurations may be partially derived from the experimental data stream. This means that 1B input may differ from experiment to experiment within the same NAM. Prominent examples are noise-levels/noise-thresholds or background corrections of machine data. In the latter case, the background level is determined from the data, then the data are processed using this information in category 2 (e.g., subtraction of a local or global background for image data). Rules on how to handle process configurations are part of SOPs or the method description of a NAM. For some level transitions setting of configurations may be fully automated and performed by a data processing pipeline. In other cases, expert knowledge is required to optimally parametrize the processing procedure. A typical automated configuration of a process/algorithm (e.g., at the transition of level III to level IV) is the generation of concentration-response curve data. Here, the data structure may determine in which way control data are used (Kappenberg et al., 2020; La et al., 2023) and which curve-fit model is chosen (Krebs et al., 2020a; Daniel et al., 2022; Kessel et al., 2023). A typical operator-dependent process is the setting of background and contrast in Western blots. The level zero data (see Section 3.1) may be considered a special type of configuration (1B input) in the sense that it occurs even before data level I.

4.2 Data processing category 2: Data transformations

This category involves actions, such as executable steps or processing rules. It includes both basic calculations (2A) and complex algorithms (2B) for data processing and integration.

Basic calculations (2A)

Fixed rules and standard mathematical operations are applied in the sense of non-iterative processes and the application of fundamental arithmetic functions (e.g., division or addition). For instance, a background value may be subtracted from data. The subtracted

value may be fixed or derived from category 1. Examples in NAMs could be the subtraction of background signals in a fluorescence/absorbance measurement or subtraction of baseline currents in electric signaling test endpoints. In a similar way, basic additions or averaging may occur at this level, like taking into account signals of multiple images or electrodes as summary data for one set of samples exposed to the same test compound concentration. At the transition from level II to level III, an example of a 2A transformation may be the computation of the mean (arithmetic, geometric or median) of several replicate samples and the normalization of data to control samples. Often, metadata (1A) and configurations (1B) of category 1 are required for such basic calculations (2A).

Complex algorithms (2B)
Often, iterative complex procedures (including machine learning) cannot be described with one of the basic arithmetic operations alone. Category 2B may require input from other categories (1 and 3) and can also be affected by the data processing stream. For instance, image processing algorithms are applied to raw image data, e.g., for object recognition and classification. Algorithms may also be used for the normalization process of multi-dimensional datasets (e.g., transcriptomics, proteomics and metabolomics data) or for classification (e.g., by using support vector machines or random forests). While automatic configurations (category 1B) may determine curve-fit decisions between level III and IV data, the curve fitting process itself (and its quality control) may be considered a complex algorithm (category 2B). Deriving BMC/PoD values or their uncertainty measures may also require category 2B processing.

4.3 Data processing category 3: Data flagging and data exclusion

In category 3, data may be flagged and/or excluded. The processing of data in categories 1 and 2 can affect actions in category 3 that alter the structure of the data and therefore may impact data configurations and transformations (Fig. 2). Data flagging and exclusion can happen at any data transition step. Machine output data (level I) may require the exclusion of data due to machine errors in measurements or operator mistakes. Some data are flagged (and may be excluded from the data processing stream) at level II (raw data) or level III (averaged and relative (raw) data). They may be detected and labelled as "outliers", e.g., because of bacterial contamination of a single sample, a cell seeding error, or other technology or operator problems. Sometimes not just single data points but whole cell preparations or assay plates may be affected. Data outlier exclusion at this stage is a necessary part of data processing. This means that aberrant NAM outputs would be generated if data error correction was neglected or omitted. It is an important aspect of databases and data pipelines to flag such outliers when data is deposited, to ensure exclusion from downstream data processing (Bell et al., 2020; Feshuk et al., 2023).

Another issue is the flagging of whole "experimental runs" based on acceptance criteria (AC) (Holzer et al., 2023). Entire assay plates or sets of plates within an experiment ("run") may be excluded from the downstream analysis based on AC. AC may be set for an acceptable degree of variation of baseline data and/or of data of positive controls, and AC are commonly based on prespecified data ranges of positive controls. Several sets of AC can be defined within the method description of a NAM, but at least one set for positive and negative control items is mandatory.

5 Conclusions and outlook

In summary, this paper discusses the complex procedure of processing generated data to a final NAM output (e.g., a hit call or information supporting a regulatory decision). As this processing pipeline may decisively impact the results, it should be considered a core element of a NAM description. To support this process, the paper illustrates the data levels that may be considered (Fig. 1), and an overview of input and transition steps from one data level to another (Fig. 2). This is meant to provide general knowledge and awareness of the complexity of data processing. Examples of detailed data processing pipelines and their application to toxicology and drug development may be found elsewhere (Krebs et al., 2020b; Bell et al., 2020; Feshuk et al., 2023).

For some NAMs, a dozen or more processing steps can be required. The set of steps (the "pipeline") is adequate when adapted to specific requirements and features of a given NAM. Conversely, this means that data processing pipelines may differ among NAMs, and understanding the influence of varying steps and identifying consensus results supports generating optimal outcomes. The use of different data pipelines for a set of different NAMs is not problematic as the processing pipeline is part of each NAM definition, laid down in the prediction model (PM), and is thus part of the test method description and validation (Box 1). This does not exclude that some NAMs use the same pipeline, or that several pipeline elements are standardized (when appropriate) across a larger panel or battery of NAMs.

In the context of the global use of data, harmonization of data processing may be desirable. Also, it is important to recognize that changing the pipelines of already established NAMs can have serious consequences. In the worst case, the NAM may need to be re-validated, as the alteration of data processing may have changed the output. A typical measure of NAM performance is the comparison of NAM output with a gold standard or a ground truth (ideally human toxicological data). If the output of a NAM changes (because of the use of an alternative data processing pipeline), then the apparent performance of the NAM could change. With NAMs becoming a standard component for risk assessment, such considerations on the impact of data processing on final NAM outcomes will be critical to secure their acceptance. Moreover, insight into data processing is essential for stakeholders (e.g., regulators) working with heterogeneous sets of NAM data (from different laboratories, from different time periods, etc.). NAM results may differ under such conditions, even with relatively similar raw data. Over time, guidance documents may be developed on parameters that need to be kept constant or that need to be transparently disclosed. Developing similar reporting templates for data processing, like those developed for omics technologies and (Q)SARs, can provide the necessary information required to understand, interpret, and reproduce NAM-derived results and ensure robust, trustworthy data to support improved public health protection.

References

Aschner, M., Ceccatelli, S., Daneshian, M. et al. (2017). Reference compounds for alternative test methods to indicate developmental neurotoxicity (DNT) potential of chemicals: Example lists and criteria for their selection and use. *ALTEX 34*, 49-74. doi:10.14573/altex.1604201

Bal-Price, A., Hogberg, H. T., Crofton, K. M. et al. (2018). Recommendation on test readiness criteria for new approach methods in toxicology: Exemplified for developmental neurotoxicity. *ALTEX 35*, 306-352. doi:10.14573/altex.1712081

Beattie, K. A., Verma, M., Brennan, R. J. et al. (2024). Quantitative systems toxicology modeling in pharmaceutical research and development: An industry-wide survey and selected case study examples. *CPT Pharmacometrics Syst Pharmacol 13*, 2036-2051. doi:10.1002/psp4.13227

Bell, S., Abedini, J., Ceger, P. et al. (2020). An integrated chemical environment with tools for chemical safety testing. *Toxicol In Vitro 67*, 104916. doi:10.1016/j.tiv.2020.104916

Blum, J., Masjosthusmann, S., Bartmann, K. et al. (2023). Establishment of a human cell-based in vitro battery to assess developmental neurotoxicity hazard of chemicals. *Chemosphere 311*, 137035. doi:10.1016/j.chemosphere.2022.137035

Callegaro, G., Kunnen, S. J., Trairatphisan, P. et al. (2021). The human hepatocyte TXG-MAPr: Gene co-expression network modules to support mechanism-based risk assessment. *Arch Toxicol 95*, 3745-3775. doi:10.1007/s00204-021-03141-w

Casey, W. M., Chang, X., Allen, D. G. et al. (2018). Evaluation and optimization of pharmacokinetic models for in vitro to in vivo extrapolation of estrogenic activity for environmental chemicals. *Environ Health Perspect 126*, 97001. doi:10.1289/ehp1655

Chang, X., Tan, Y. M., Allen, D. G. et al. (2022). IVIVE: Facilitating the use of in vitro toxicity data in risk assessment and decision making. *Toxics 10*, 232. doi:10.3390/toxics10050232

Collen, E., Tanaskov, Y., Holzer, A. K. et al. (2024). Elements and development processes for test methods in toxicology and human health-relevant life science research. *ALTEX 41*, 142-148. doi:10.14573/altex.2401041

Crofton, K. M., Mundy, W. R., Lein, P. J. et al. (2011). Developmental neurotoxicity testing: Recommendations for developing alternative methods for the screening and prioritization of chemicals. *ALTEX 28*, 9-15. doi:10.14573/altex.2011.1.009

Crouzet, T., Grignard, E., Brion, F. et al. (2023). ReadEDTest: A tool to assess the readiness of in vitro test methods under development for identifying endocrine disruptors. *Environ Int 174*, 107910. doi:10.1016/j.envint.2023.107910

Daniel, A. B., Choksi, N., Abedini, J. et al. (2022). Data curation to support toxicity assessments using the integrated chemical environment. *Front Toxicol 4*, 987848. doi:10.3389/ftox.2022.987848

Delp, J., Gutbier, S., Cerff, M. et al. (2018). Stage-specific metabolic features of differentiating neurons: Implications for toxicant sensitivity. *Toxicol Appl Pharmacol 354*, 64-80. doi:10.1016/j.taap.2017.12.013

Delp, J., Cediel-Ulloa, A., Suciu, I. et al. (2021). Neurotoxicity and underlying cellular changes of 21 mitochondrial respiratory chain inhibitors. *Arch Toxicol 95*, 591-615. doi:10.1007/s00204-020-02970-5

Di Tillio, F. and Beltman, J. B. (2024). Developing quantitative adverse outcome pathways: An ordinary differential equation-based computational framework. *Comput Toxicol 32*, 100330. doi:10.1016/j.comtox.2024.100330

Ebmeyer, J., Najjar, A., Lange, D. et al. (2024). Next generation risk assessment: An ab initio case study to assess the systemic safety of the cosmetic ingredient, benzyl salicylate, after dermal exposure. *Front Pharmacol 15*, 1345992. doi:10.3389/fphar.2024.1345992

Ferrario, D., Brustio, R. and Hartung, T. (2014). Glossary of reference terms for alternative test methods and their validation. *ALTEX 31*, 319-335. doi:10.14573/altex.1403311

Feshuk, M., Kolaczkowski, L., Dunham, K. et al. (2023). The ToxCast pipeline: Updates to curve-fitting approaches and database structure. *Front Toxicol 5*, 1275980. doi:10.3389/ftox.2023.1275980

Filer, D. L., Kothiya, P., Setzer, R. W. et al. (2017). tcpl: The ToxCast pipeline for high-throughput screening data. *Bioinformatics 33*, 618-620. doi:10.1093/bioinformatics/btw680

Gabbert, S., Mathea, M., Kolle, S. N. et al. (2022). Accounting for precision uncertainty of toxicity testing: Methods to define borderline ranges and implications for hazard assessment of chemicals. *Risk Anal 42*, 224-238. doi:10.1111/risa.13648

Harrill, J. A., Viant, M. R., Yauk, C. L. et al. (2021). Progress towards an OECD reporting framework for transcriptomics and metabolomics in regulatory toxicology. *Regul Toxicol Pharmacol 125*, 105020. doi:10.1016/j.yrtph.2021.105020

Hartung, T. (2009). Toxicology for the twenty-first century. *Nature 460*, 208-212. doi:10.1038/460208a

Hartung, T., De Vries, R., Hoffmann, S. et al. (2019). Toward good in vitro reporting standards. *ALTEX 36*, 3-17. doi:10.14573/altex.1812191

Hartung, T. (2023a). ToxAIcology – The evolving role of artificial intelligence in advancing toxicology and modernizing regulatory science. *ALTEX 40*, 559-570. doi:10.14573/altex.2309191

Hartung, T. (2023b). Artificial intelligence as the new frontier in chemical risk assessment. *Front Artif Intell 6*, 1269932. doi:10.3389/frai.2023.1269932

Hartung, T., King, N. M. P., Kleinstreuer, N. et al. (2024). Leveraging biomarkers and translational medicine for preclinical safety – Lessons for advancing the validation of alternatives to animal testing. *ALTEX 41*, 545-566. doi:10.14573/altex.2410011

Holzer, A. K., Dreser, N., Pallocca, G. et al. (2023). Acceptance criteria for new approach methods in toxicology and human health-relevant life science research – Part I. *ALTEX 40*, 706-712. doi:10.14573/altex.2310021

Jochum, K., Miccoli, A., Sommersdorf, C. et al. (2024). Comparative case study on NAMs: Towards enhancing specific target organ toxicity analysis. *Arch Toxicol 98*, 3641-3658. doi:10.1007/s00204-024-03839-7

Kappenberg, F., Brecklinghaus, T., Albrecht, W. et al. (2020). Handling deviating control values in concentration-response curves. *Arch Toxicol 94*, 3787-3798. doi:10.1007/s00204-020-02913-0

Kappenberg, F., Grinberg, M., Jiang, X. et al. (2021). Comparison of observation-based and model-based identification of alert con-

centrations from concentration-expression data. *Bioinformatics 37*, 1990-1996. doi:10.1093/bioinformatics/btab043

Kessel, H. E., Masjosthusmann, S., Bartmann, K. et al. (2023). The impact of biostatistics on hazard characterization using in vitro developmental neurotoxicity assays. *ALTEX 40*, 619-634. doi:10.14573/altex.2210171

Kleinstreuer, N. C., Ceger, P., Watt, E. D. et al. (2017). Development and validation of a computational model for androgen receptor activity. *Chem Res Toxicol 30*, 946-964. doi:10.1021/acs.chemrestox.6b00347

Kleinstreuer, N. C., Hoffmann, S., Alépée, N. et al. (2018). Non-animal methods to predict skin sensitization (II): An assessment of defined approaches. *Crit Rev Toxicol 48*, 359-374. doi:10.1080/10408444.2018.1429386

Kleinstreuer, N. and Hartung, T. (2024). Artificial intelligence (AI) – It's the end of the tox as we know it (and I feel fine). *Arch Toxicol 98*, 735-754. doi:10.1007/s00204-023-03666-2

Knudsen, T., Martin, M., Chandler, K. et al. (2013). Predictive models and computational toxicology. *Methods Mol Biol 947*, 343-374. doi:10.1007/978-1-62703-131-8_26

Krebs, A., Nyffeler, J., Rahnenfuhrer, J. et al. (2018). Normalization of data for viability and relative cell function curves. *ALTEX 35*, 268-271. doi:10.14573/1803231

Krebs, A., Waldmann, T., Wilks, M. F. et al. (2019). Template for the description of cell-based toxicological test methods to allow evaluation and regulatory use of the data. *ALTEX 36*, 682-699. doi:10.14573/altex.1909271

Krebs, A., Nyffeler, J., Karreman, C. et al. (2020a). Determination of benchmark concentrations and their statistical uncertainty for cytotoxicity test data and functional in vitro assays. *ALTEX 37*, 155-163. doi:10.14573/altex.1912021

Krebs, A., van Vugt-Lussenburg, B. M. A., Waldmann, T. et al. (2020b). The EU-ToxRisk method documentation, data processing and chemical testing pipeline for the regulatory use of new approach methods. *Arch Toxicol 94*, 2435-2461. doi:10.1007/s00204-020-02802-6

Krishna, S., Berridge, B. and Kleinstreuer, N. (2021). High-throughput screening to identify chemical cardiotoxic potential. *Chem Res Toxicol 34*, 566-583. doi:10.1021/acs.chemrestox.0c00382

Krug, A. K., Balmer, N. V., Matt, F. et al. (2013). Evaluation of a human neurite growth assay as specific screen for developmental neurotoxicants. *Arch Toxicol 87*, 2215-2231. doi:10.1007/s00204-013-1072-y

La, V. N. T., Nicholson, S., Haneef, A. et al. (2023). Inclusion of control data in fits to concentration-response curves improves estimates of half-maximal concentrations. *J Med Chem 66*, 12751-12761. doi:10.1021/acs.jmedchem.3c00107

Lasch, A., Lichtenstein, D., Marx-Stoelting, P. et al. (2020). Mixture effects of chemicals: The difficulty to choose appropriate mathematical models for appropriate conclusions. *Environ Pollut 260*, 113953. doi:10.1016/j.envpol.2020.113953

Leist, M., Efremova, L. and Karreman, C. (2010). Food for thought ... Considerations and guidelines for basic test method descriptions in toxicology. *ALTEX 27*, 309-317. doi:10.14573/altex.2010.4.309

Leist, M., Hasiwa, N., Daneshian, M. et al. (2012). Validation and quality control of replacement alternatives – Current status and future challenges. *Toxicol Res 1*, 8-22. doi:10.1039/c2tx20011b

Leist, M., Hasiwa, N., Rovida, C. et al. (2014). Consensus report on the future of animal-free systemic toxicity testing. *ALTEX 31*, 341-356. doi:10.14573/altex.1406091

Leist, M. and Hengstler, J. G. (2018). Essential components of methods papers. *ALTEX 35*, 429-432. doi:10.14573/altex.1807031

Leontaridou, M., Urbisch, D., Kolle, S. N. et al. (2017). The borderline range of toxicological methods: Quantification and implications for evaluating precision. *ALTEX 34*, 525-538. doi:10.14573/altex.1606271

Leontaridou, M., Gabbert, S. and Landsiedel, R. (2019). The impact of precision uncertainty on predictive accuracy metrics of non-animal testing methods. *ALTEX 36*, 435-446. doi:10.14573/altex.1810111

Lynch, C., Sakamuru, S., Ooka, M. et al. (2024). High-throughput screening to advance in vitro toxicology: Accomplishments, challenges, and future directions. *Annu Rev Pharmacol Toxicol 64*, 191-209. doi:10.1146/annurev-pharmtox-112122-104310

Maertens, A., Golden, E., Luechtefeld, T. H. et al. (2022). Probabilistic risk assessment – The keystone for the future of toxicology. *ALTEX 39*, 3-29. doi:10.14573/altex.2201081

Magel, V., Blum, J., Dolde, X. et al. (2024). Inhibition of neural crest cell migration by strobilurin fungicides and other mitochondrial toxicants. *Cells 13*, 2057. doi:10.3390/cells13242057

Mansouri, K., Moreira-Filho, J. T., Lowe, C. N. et al. (2024). Free and open-source QSAR-ready workflow for automated standardization of chemical structures in support of QSAR modeling. *J Cheminform 16*, 19. doi:10.1186/s13321-024-00814-3

Najjar, A., Kuhnl, J., Lange, D. et al. (2024a). Next-generation risk assessment read-across case study: Application of a 10-step framework to derive a safe concentration of daidzein in a body lotion. *Front Pharmacol 15*, 1421601. doi:10.3389/fphar.2024.1421601

Najjar, A., Wilm, A., Meinhardt, J. et al. (2024b). Evaluation of new alternative methods for the identification of estrogenic, androgenic and steroidogenic effects: A comparative in vitro/in silico study. *Arch Toxicol 98*, 251-266. doi:10.1007/s00204-023-03616-y

NTP (2018). NTP research report on National Toxicology Program approach to genomic dose-response modeling: NTP RR 5. Durham, NC. *National Toxicology Program (5)*, 1-44. doi:10.22427/ntp-rr-5

OECD (2005). Guidance Document on the Validation and International Acceptance of New or Updated Test Methods for Hazard Assessment. *OECD Series on Testing and Assessment, No. 34*. OECD Publishing, Paris. doi:10.1787/e1f1244b-en

OECD (2017a). Guidance Document for the Use of Adverse Outcome Pathways in Developing Integrated Approaches to Testing and Assessment (IATA). *OECD Series on Testing and Assessment, No. 260*. OECD Publishing, Paris. doi:10.1787/44bb06c1-en

OECD (2017b). Guidance Document for Describing Non-Guideline In Vitro Test Methods. *OECD Series on Testing and Assessment, No. 211*. OECD Publishing, Paris. doi:10.1787/9789264274730-en

OECD (2018). Guidance Document on Good In Vitro Method Practices (GIVIMP). *OECD Series on Testing and Assessment, No. 286*. OECD Publishing, Paris. doi:10.1787/9789264304796-en

OECD (2021). Guideline No. 497: Defined Approaches on Skin Sensitisation. *OECD Guidelines fot the Testing of Chemicals, Section 4*. OECD Publishing Paris. doi:10.1787/b92879a4-en

OECD (2023). OECD Omics Reporting Framework (OORF): Guidance on reporting elements for the regulatory use of omics data from laboratory-based toxicology studies. *OECD Series on Testing and Assessment, No. 390*. OECD Publishing, Paris. doi:10.1787/6bb2e6ce-en

OECD (2024). (Q)SAR assessment framework: Guidance for the Regulatory Assessment of (Quantitative) Structure Activity Relationship Models and Predictions, Second Edition. *OECD Series on Testing and Assessment, No. 405*. doi:10.1787/bbdac345-en

Pallocca, G. and Leist, M. (2022). On the usefulness of animals as a model system (part II): Considering benefits within distinct use domains. *ALTEX 39*, 531-539. doi:10.14573/altex.2207111

Pamies, D., Leist, M., Coecke, S. et al. (2022). Guidance document on good cell and tissue culture practice 2.0 (GCCP 2.0). *ALTEX 39*, 30-70. doi:10.14573/altex.2111011

Perkins, E. J., Ashauer, R., Burgoon, L. et al. (2019). Building and applying quantitative adverse outcome pathway models for chemical hazard and risk assessment. *Environ Toxicol Chem 38*, 1850-1865. doi:10.1002/etc.4505

Petersen, E. J., Elliott, J. T., Gordon, J. et al. (2023). Technical framework for enabling high quality measurements in new approach methodologies (NAMs). *ALTEX 40*, 174-186. doi:10.14573/altex.2205081

Polak, S., Tylutki, Z., Holbrook, M. et al. (2019). Better prediction of the local concentration-effect relationship: The role of physiologically based pharmacokinetics and quantitative systems pharmacology and toxicology in the evolution of model-informed drug discovery and development. *Drug Discov Today 24*, 1344-1354. doi:10.1016/j.drudis.2019.05.016

Reinke, E. N., Reynolds, J., Gilmour, N. et al. (2025). The skin allergy risk assessment-integrated chemical environment (saraice) defined approach to derive points of departure for skin sensitization. *Curr Res Toxicol 8*, 100205. doi:10.1016/j.crtox.2024.100205

Samuel, G. O., Hoffmann, S., Wright, R. A. et al. (2016). Guidance on assessing the methodological and reporting quality of toxicologically relevant studies: A scoping review. *Environ Int 92-93*, 630-646. doi:10.1016/j.envint.2016.03.010

Schmeisser, S., Miccoli, A., von Bergen, M. et al. (2023). New approach methodologies in human regulatory toxicology – Not if, but how and when! *Environ Int 178*, 108082. doi:10.1016/j.envint.2023.108082

Schmidt, B. Z., Lehmann, M., Gutbier, S. et al. (2017). In vitro acute and developmental neurotoxicity screening: An overview of cellular platforms and high-throughput technical possibilities. *Arch Toxicol 91*, 1-33. doi:10.1007/s00204-016-1805-9

Smirnova, L., Hogberg, H. T., Leist, M. et al. (2024). Revolutionizing developmental neurotoxicity testing – A journey from animal models to advanced in vitro systems. *ALTEX 41*, 152-178. doi:10.14573/altex.2403281

Strickland, J., Truax, J., Corvaro, M. et al. (2022). Application of defined approaches for skin sensitization to agrochemical products. *Front Toxicol 4*, 852856. doi:10.3389/ftox.2022.852856

Sturla, S. J. (2018). Point of departure. *Chem Res Toxicol 31*, 2-3. doi:10.1021/acs.chemrestox.7b00341

Suciu, I., Delp, J., Gutbier, S. et al. (2023). Definition of the neurotoxicity-associated metabolic signature triggered by berberine and other respiratory chain inhibitors. *Antioxidants (Basel) 13*, 49. doi:10.3390/antiox13010049

Tsaioun, K., Blaauboer, B. J. and Hartung, T. (2016). Evidence-based absorption, distribution, metabolism, excretion (ADME) and its interplay with alternative toxicity methods. *ALTEX 33*, 343-358. doi:10.14573/altex.1610101

Verheijen, M., Tong, W., Shi, L. et al. (2020). Towards the development of an omics data analysis framework. *Regul Toxicol Pharmacol 112*, 104621. doi:10.1016/j.yrtph.2020.104621

Watt, E. D. and Judson, R. S. (2018). Uncertainty quantification in ToxCast high throughput screening. *PLoS One 13*, e0196963. doi:10.1371/journal.pone.0196963

Wheeler, M. W. (2023). An investigation of non-informative priors for Bayesian dose-response modeling. *Regul Toxicol Pharmacol 141*, 105389. doi:10.1016/j.yrtph.2023.105389

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J. et al. (2016). The FAIR guiding principles for scientific data management and stewardship. *Sci Data 3*, 160018. doi:10.1038/sdata.2016.18

Zhu, T., Cao, S., Su, P. C. et al. (2013). Hit identification and optimization in virtual screening: Practical recommendations based on a critical literature analysis. *J Med Chem 56*, 6560-6572. doi:10.1021/jm301916b

Zobl, W., Bitsch, A., Blum, J. et al. (2024). Protectiveness of NAM-based hazard assessment – Which testing scope is required? *ALTEX 41*, 302-319. doi:10.14573/altex.2309081

Disclaimer

The views expressed in this article are those of the authors and do not necessarily represent the views or policies of the institutions of employment of the authors.

Conflict of interest

No conflicts of interest have been identified.

Acknowledgments

We gratefully acknowledge funding by the Deutsche Forschungsgemeinschaft (DFG) (grant TRR353, project number 471011418). The work was facilitated by grants from the BMBF (161L0243B, 016LW0146K), ZonMW (114027005), EFSA, the Swiss Centre for Applied Human Toxicology (SCAHT-GL-21-06), and by the IHI2 project Vict3R. Funding was also received from the European Union's Horizon 2020 research and innovation program under grant agreements No. 964537 (RISK-HUNT3R), No. 964518 (ToxFree), No. 101057014 (PARC) and No. 963845 (ONTOX). This work was supported in part by the NIH Intramural Research Program.