

***In silico* generation, evaluation and application of
developmental neurotoxicity data derived from high
throughput screening assays**

Dissertation to obtain the degree
Doctor Rerum Naturalium (Dr. rer. nat.)
at the Heinrich-Heine-University Düsseldorf

Submitted by

Hagen Eike Keßel

from Bochum

Düsseldorf, April 2023

***In silico* Generierung, Evaluation und Anwendung von
aus Hochdurchsatzaufnahmen stammenden
entwicklungsneurotoxischen Daten**

Inaugural-Dissertation

Zur Erlangung des Doktorgrades der Mathematisch-Naturwissenschaftlichen
Fakultät der Heinrich-Heine-Universität Düsseldorf

vorgelegt von

Hagen Eike Keßel

aus Bochum

Düsseldorf, April 2023

Angefertigt am Leibniz Institut für umweltmedizinische Forschung (IUF) an der Heinrich-Heine Universität Düsseldorf.

Gedruckt mit der Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Heinrich-Heine-Universität Düsseldorf.

Referentin: Prof. Dr. Ellen Fritsche

Korreferent: Prof. Dr. Vlada B. Urlacher

Tag der mündlichen Prüfung: 26.01.2024

Table of contents

1	Introduction	6
1.1	Paradigm shift in toxicology	6
1.2	Developmental neurotoxicity testing	8
1.3	Computational Bioinformatics for <i>in vitro</i> DNT testing.....	10
1.3.1	Data generation	10
1.3.2	Biostatistical data evaluation	13
1.3.3	Data application	14
1.4	Objectives of the thesis	15
2	Manuscripts	16
2.1	Reliable identification and quantification of neural cells in microscopic images of neurospheres	18
2.2	Biostatistics and its impact on hazard characterization using in vitro developmental neurotoxicity assays	32
2.3	Establishment of a human cell-based in vitro battery to assess developmental neurotoxicity hazard of chemicals	79
2.4	Neurodevelopmental toxicity assessment of flame retardants using a human DNT in vitro testing battery	107
3	Discussion.....	136
3.1	Data generation.....	137
3.2	Data evaluation	140
3.3	Data application	144
3.4	Connecting the dots: How data generation affects the evaluation and what it means for the application in hazard assessment	146
3.5	Conclusion	150
4	Summary.....	152
5	Zusammenfassung.....	153
	List of abbreviations	154
	References	155

1 Introduction

1.1 Paradigm shift in toxicology

In human toxicology, risk assessment is used to assess the risk of a chemical for human health. For this purpose, toxicological hazard and human exposure need to be characterized. For the last decades, hazard characterization of chemicals was primarily done by *in vivo* testing. However, the use of animals for toxicity testing is a very resource intensive procedure, comes with ethical concerns and possible species differences are not considered (Crofton *et al.*, 2012; Krewski *et al.* 2020). To improve human risk assessment and address these issues, the national research council (NRC) proposed a new strategy for toxicity testing in the 21st century, which is based on a paradigm shift from conventional *in vivo* toxicity testing to high throughput, mechanistic *in vitro* assays by development of so called “new approach methods” (NAM) (NRC, 2007; Collins *et al.*, 2008) (Fig.1). NAMs are defined as any technology, methodology, approach, or their combination that can provide information on chemical hazard and risk assessment to avoid the use of animal testing (USEPA 2021).

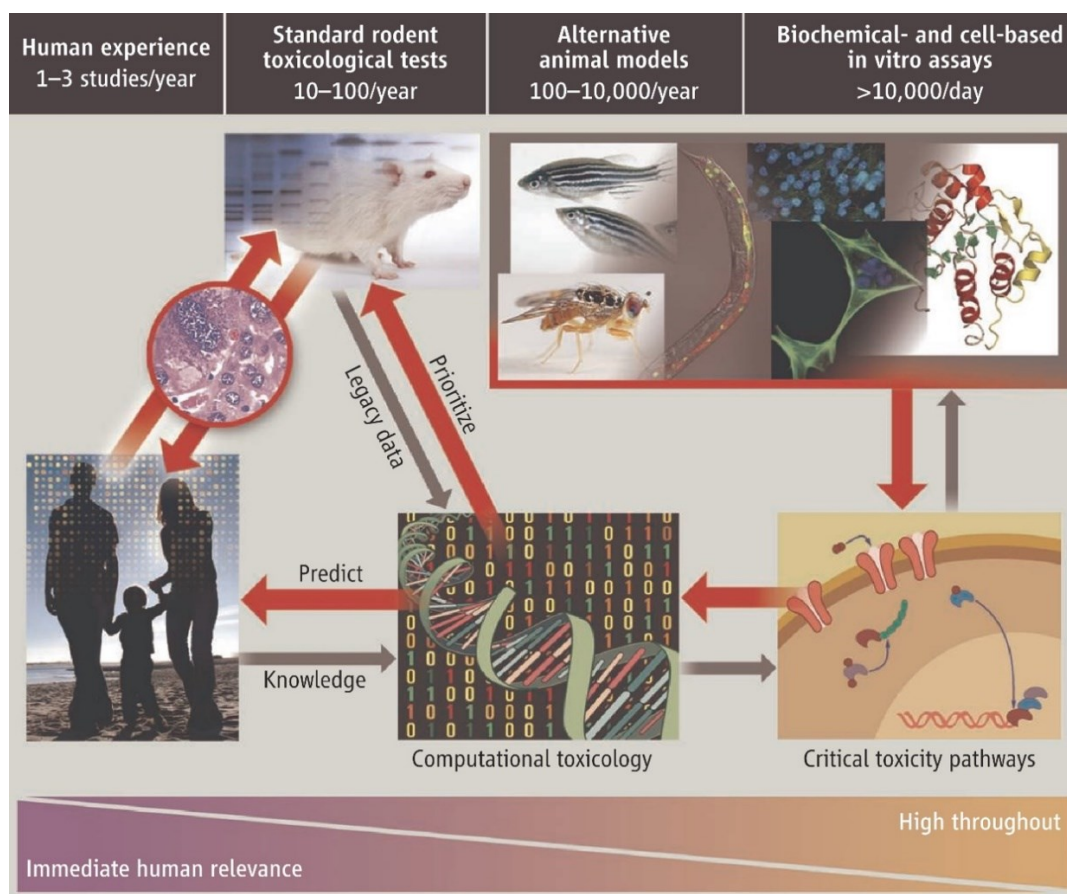


Figure 1: Paradigm shift in toxicology

The US-National Research Council (NRC) proposed a new strategy for toxicity testing in the 21st century, Tox21, which is based on a shift from conventional *in vivo* toxicity testing in rodents to high content, mechanism-based *in vitro* assays. The goal is to increase mechanistic understanding and sample throughput, overcome species differences and reduce animal testing in line with the 3R principle. The shift proposes the use of high content and high throughput data in combination with computational toxicology (Collins *et al.* 2008).

NAMs can be used for different regulatory scenarios, i.e. hazard characterization as well as screening and prioritization. To integrate data from *in vitro* assays into systemic predictions about health-related consequences on the level of whole organs or individuals, the adverse outcome pathway (AOP) framework was established. With this framework, available data on molecular initiating events (MIE) are linked through key event relationships (KER) to biological key events (KE), which result in an adverse outcome (AO; the human health effect) (OECD (Organization for Economic Co-Operation and Development), 2013; Villeneuve *et al.*, 2014; Carusi *et al.*, 2018). Within this framework all available toxicological data (e.g. molecular and cellular data from *in vitro* assays, animal studies or epidemiological studies) can be combined to enable an understanding of the mode of action (MOA) of a compound's toxicity. Furthermore, 'Integrated Approaches to Testing and Assessment' (IATA) frameworks are developed for hazard characterization by relying on integrated analysis of existing information in combination with legacy data. The goal of IATA frameworks is to answer defined questions in a regulatory context and provide sufficient information for confident regulatory decision making. One IATA approach is to link existing data from *in silico* methods with experimental data, enabling informed regulatory decision making on the basis of experimental and *in silico* data (Bal-Price *et al.*, 2015b) (Fig.2). For hazard characterization, the future approach incorporates NAMs (e.g. *in vitro*, *in silico*, omics, physiologically based pharmacokinetic (PBPK) modelling, AOPs) into the IATA framework (Escher *et al.* 2022).

Prioritization of compounds is determined by the use of systems biology (screening for critical pathways or cell biological processes in *in vitro* assays) and subsequent computational toxicology (evaluation of data obtained by systems biology). This strategy also follows the 3R's principle (Reduce, Refine, Replace), as it was introduced by Russell and Burch (Russell and Burch, 1959). For a successful shift, several requirements need to be met. First, the novel test assays need to be able to mimic relevant cell, tissue or organ functions. They furthermore need to be capable of generating data in a medium to high throughput set up to drastically reduce the time and resource intensity of toxicity testing. As for the generation and evaluation of the data deriving from these assays, biostatistical/-informatical tools that can handle the broad amount of data and extract relevant information for toxicological interpretation are necessary. Thus, in the last two decades efforts were made to develop such assays and tools, marking several milestones in the shift from *in vivo* to *in vitro* toxicity testing (Wheeler *et al.*, 2015; Villeneuve *et al.*, 2019). These advances include development and establishment of high throughput screening (HTS) assays, as well as biostatistic and bioinformatic tools for data generation, management and evaluation, subsequently allowing compound prioritization or supporting regulatory decisions (Leist *et al.*, 2014; Villeneuve *et al.*, 2019; Villeneuve *et al.*, 2019).

However, there are still challenges to overcome within the use of HTS methods, such as lacking documentation, expensive licensing, input formats, scalability, operating systems, and reproducibility (Frommolt and Thomas, 2008; Fourches *et al.*, 2014). Furthermore, there is a disconnect between the biological events measured by HTS assays (e.g. gene expression, changes in cell morphology) and the concerns from a risk-management perspective (human health, e.g. IQ rates) (Villeneuve *et al.*, 2019).

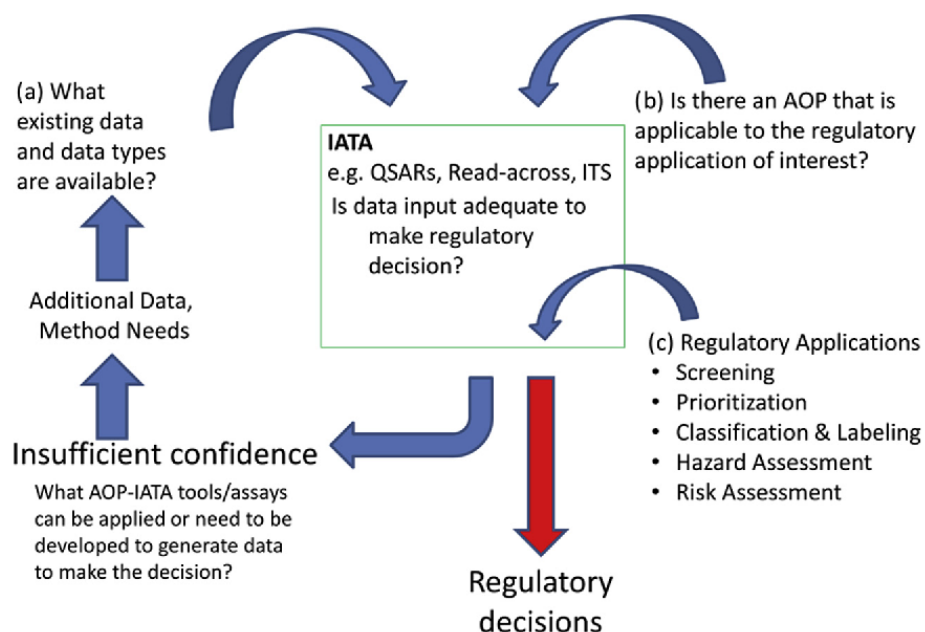


Figure 2: The AOP-informed IATA framework

As a first step, the IATA framework regards to three questions: (a) What data is available? (b) Is there an AOP available? (c) Are regulatory applications involved? With this information, available data is evaluated by non-experimental (e.g. QSARs, Read-across) and experimental approaches (e.g. ITS). The outcome of these evaluations is used to decide, if there is enough confidence in the available data to make regulatory decisions based on it. If that is not the case, further data needs to be collected and/or methods need to be developed to enable confident decision making.

1.2 Developmental neurotoxicity testing

The use of alternative methods within an IATA strategy is of increasing importance for toxicity testing, as there is a desire to implement faster, yet less cost intensive and more human-relevant test methods. Within the field toxicity testing, developmental neurotoxicity (DNT) describes the effect of chemical exposure on the morphology and functionality of the developing nervous system. It has been shown that the pathological changes resulting from exposure of hazardous chemicals can lead a broad variety of cognitive impairments or disfunctions such as decrease of the intelligent quotient (IQ), attention disorders or learning deficits (Grandjean, Landrigan, 2006; Grandjean, Landrigan, 2014; Bennett *et al.*, 2016). Despite the socioeconomic threat that DNT poses, there is a huge knowledge gap on the DNT potential of chemicals (Sachana *et al.*, 2019) with only 110 to 140 chemicals tested to date in one of the international DNT guideline studies (Crofton *et al.*, 2020; Makris *et al.*, 2009; OECD 2008). The

Introduction

cause of this gap is seen in the unpractical nature of *in vivo* DNT testing: Conventional *in vivo* testing strategies following the EPA (environmental protection agency) and OECD testing guidelines (USEPA 1998; OECD 2007) rely on behavioral endpoints with poor statistical power, show high variability (Paparella *et al.*, 2020), raise ethical concerns and furthermore are demanding in terms of time and resources (e.g. animal test subjects, material for animal housing, money, test compounds; Sachana *et al.* 2021). In the course of the shift from *in vivo* to *in vitro*, human based high content *in vitro* assays were developed, which mimic major processes of brain development (Bal-Price *et al.*, 2018). These methods with appropriate readiness (Bal-Price *et al.*, 2018) were combined in a DNT *in vitro* battery (DNT-IVB) to identify the potential of chemicals to trigger DNT by interfering with such key neurodevelopmental processes (Masjosthusmann *et al.*, 2020). To advance the use of such a battery, a pilot study was commissioned by the European Food Safety Authority (EFSA) to test 120 chemicals in a battery of ten *in vitro* methods (Masjosthusmann *et al.*, 2020; Blum *et al.* 2022). This DNT-IVB contains methods that mimic the key neurodevelopmental processes neural progenitor cell (NPC) proliferation, migration of neural crest and radial glia cells, neurons and oligodendrocytes, neuronal differentiation, neurite outgrowth of the peripheral and central nervous system, as well as oligodendrocyte differentiation (Fig. 3). The dataset and the interpretation of the data serves as the basis for an OECD guidance document on the use and interpretation of the DNT-IVB in a regulatory context (Crofton and Mundy 2021), which will be released in 2023.

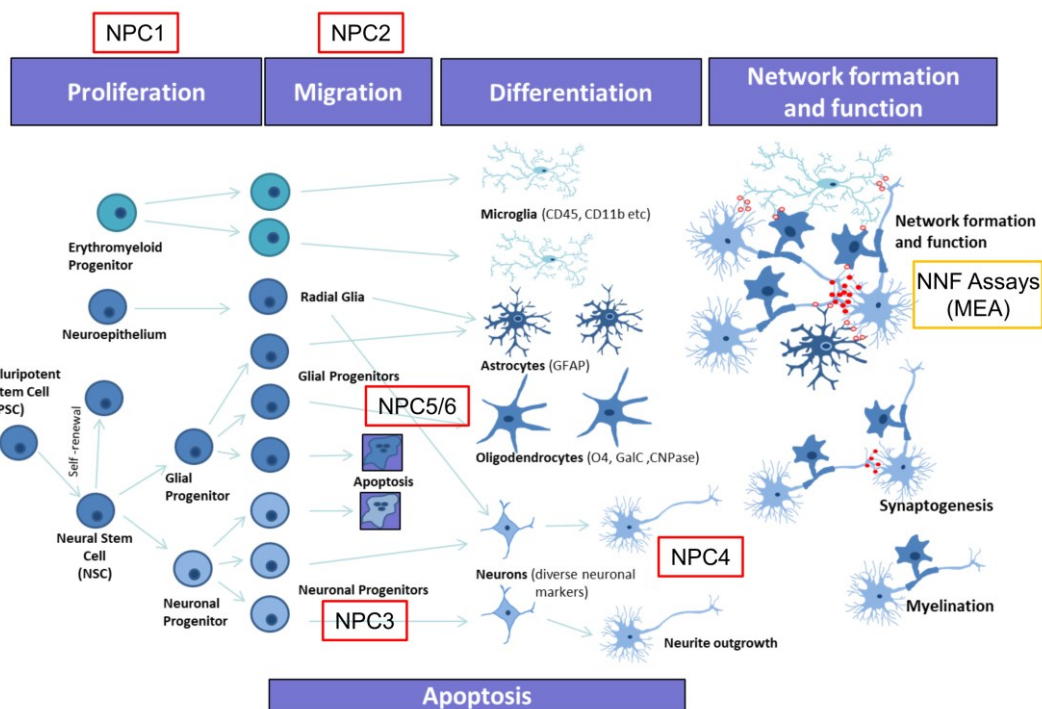


Figure 3: Key neurodevelopmental processes

Primary human neural progenitor cells (NPC) mimic several key neurodevelopmental processes (KNDP, marked in red), which are essential for brain development. The figure above gives an overview over the process of cell development from organ stem cells (left) to neural networks (right) (Bal-Price *et al.*, 2018).

1.3 Computational Bioinformatics for *in vitro* DNT testing

One prerequisite for regulatory acceptance of HTS *in vitro* data is robustness of the data including the test systems themselves, but also a robust and standardized computational workflow that manages and evaluates the broad amounts of data generated by different HTS assays. The processes from generating HTS data to application of this data for regulatory purposes can be divided into three major steps: 1) data generation from cell system readouts, 2) evaluation of the generated data by biostatistical tools and 3) application of the evaluated data for regulatory purposes (Fig. 4).

Major steps	DNT data...	Thesis manuscripts
<ul style="list-style-type: none"> • Image acquisition • Extraction of endpoint data from images 	<div>Generation</div>	<ul style="list-style-type: none"> • Reliable Identification and Quantification of Neural Cells in Microscopic Images of Neurospheres (Förster <i>et al.</i>, 2021)
<ul style="list-style-type: none"> • Estimation of benchmark concentrations and confidence intervals • Compound classification 	<div>Evaluation</div>	<ul style="list-style-type: none"> • Biostatistics and its impact on hazard characterization using <i>in vitro</i> developmental neurotoxicity assays (Keßel <i>et al.</i>, 2022)
<ul style="list-style-type: none"> • Potential hazard identification • Compound prioritization • IATA (QSAR, read-across, etc.) 	<div>Application</div>	<ul style="list-style-type: none"> • Neurodevelopmental toxicity assessment of flame retardants using a human DNT <i>in vitro</i> testing battery (Klose <i>et al.</i>, 2021) • Establishment of a human cell-based <i>in vitro</i> battery (IVB) to assess developmental neurotoxicity hazard of chemicals (Blum <i>et al.</i> 2022)

Figure 4: Major steps of the next generation risk assessment process using new approach methods (NAM) data and applying biostatistics and bioinformatics

HTS data can be generated by different NAMs. Here we suggest data generation by high content imaging (HCI) followed by image analysis as an example (manuscript 2.1). Endpoint data is evaluated by biostatistical methods to gather benchmark responses, corresponding confidence limits and resulting compound classifications (manuscript 2.2). These are finally applied to identify potential hazards (manuscript 2.3 and 2.4).

1.3.1 Data generation

Over the last two decades, major advances were made for automated screening of biological samples (Villeneuve *et al.*, 2019). By today, there is a variety of HTS assays, producing image data with rapid pace in high quantities. To keep up with the abundance of screening data, automated image analysis algorithms are needed. High content image analyses (HCIA) tools have been developed and now find application in a broad variety of applications (cell differentiation, apoptosis, tumor biology, neurodegenerative disorders or arterial hypertension, to name a few areas of application; Villeneuve *et al.*, 2019). With HCIA, high levels of information can be extracted from images of biological samples

and allow the assessment of different biological endpoints. The diversity of testing assays brings along a variety of different systems for identification of single cells and analysis of cellular responses. There is an abundance of different approaches for cell identification, with skeletonization, vectorization, super-ellipsoids (Shariff *et al.*, 2010) and overlap algorithms (Schmuck *et al.*, 2017) being popular choices. In the skeletonization method, images are segmented (often by the use of thresholds such as brightness or contrast thresholds). The segments are then further processed by systematically removing pixels (usually by considering the surrounding neighborhood). In the end, a skeleton structure of the segment is achieved, which then can be used to assess the features of interest (e.g. morphological endpoints; Bai, Latecki and Liu, 2007). During vectorization, defined sections of the image are analyzed stepwise. For this purpose, a “starting point” must be identified either manually or automatically. Starting from this defined point, vectorization algorithms recursively explore the regions of interest for feature extraction (Al-Kofahi, Lasek and Szarowski, 2002). Super-ellipsoids algorithms rely on cylinders with an elliptical cross section as a special model for the regions of interest. With this approach, structures can be represented with dense special information, which allows fast feature extraction from the area of interest (Tyrrell *et al.*, 2007). Overlap algorithms compare the pixel-overlap of two different stainings (usually one for nuclei to identify cells and one specific for a cell type) to identify certain cell types. A cell is then identified as a certain type, if the type-specific staining showed enough overlap with an underlying nucleus staining (i.e. if an overlap threshold was reached).

The usage of machine learning (ML) algorithms (often implemented as convolutional neuronal network models) has become the staple of image analysis in recent years (Shariff *et al.*, 2010). ML algorithms are able to learn feature differentiation and to extract relevant features from provided training data, by comparing their own evaluation (e.g. cells detected as neurons) to a ground truth that was set up by a human experimenter (e.g. cells that were marked as neurons by an experimenter). This approach is known as “supervised learning”. Alternative approaches are “semi-supervised learning” and “unsupervised learning”, in which the algorithm is only partially or not guided by a ground truth. ML approaches offer several advantages over the aforementioned traditional methods of image analysis: While conventional methods usually rely heavily on the image acquisition tools to have consistent image properties (e.g. images need to be recorded by the same camera to maintain the same image brightness and focus) and allow only limited flexibility (usually achieved by hardcoded parameters that can be adjusted to certain extends by a user), ML models can be trained to extract the relevant information from a broad variety of images with varying properties (e.g. images acquired by different cameras). In other words, ML models are able to mimic human evaluation far more accurate than conventional methods and are able to learn from existing datasets for further improvement of performance by supervised learning.

Introduction

HCIA approaches for identification of neurons in *in vitro* cultures have been numerous, while other neural cell types like radial glia, oligodendrocytes or astrocytes have so far been mostly neglected. In the field of DNT, a DNT in vitro battery (IVB) was set up that allows identification of a large number of key neurodevelopmental processes such as migration and differentiation of several cell types as well as neurite outgrowth by novel HCIA approaches (Masjosthusmann *et al.*, 2020). One of such HCIA tools is Omnisphero (Schmuck *et al.*, 2017; manuscript 2.1 - Förster *et al.*, 2021), a software tool able to extract information about neural progenitor cell migration and differentiation, as well as information about morphological aspects such as neurite length or number of branching points in a HCIA manner (Fig. 5) by analyzing data from the neurosphere model.

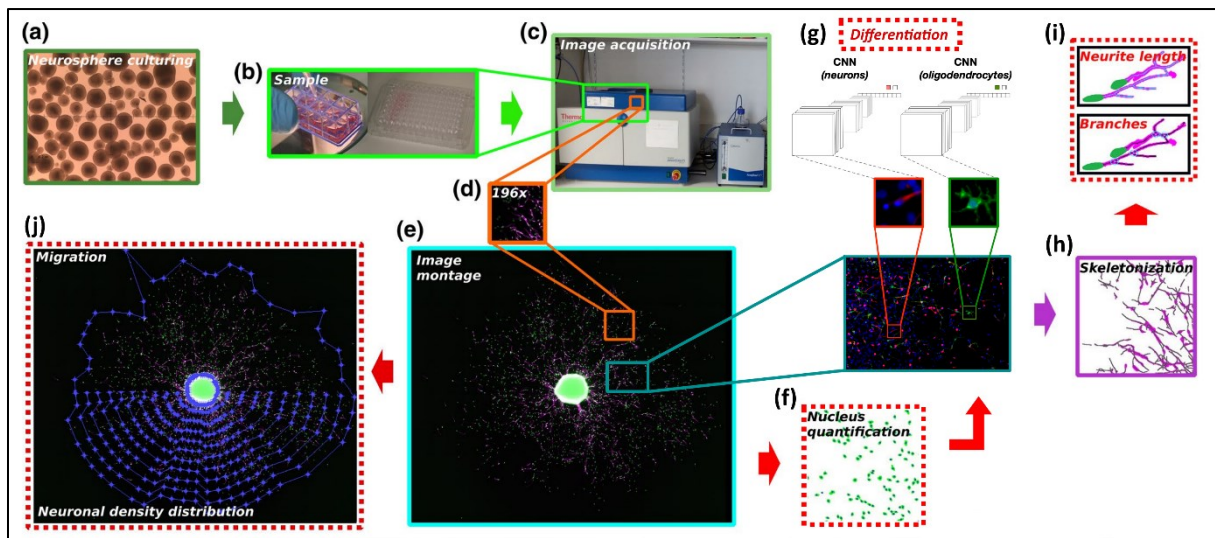


Figure 5: Omnisphero workflow

(a) Neurospheres are plated into 96-well plates and (b) exposed to chemical treatment during incubation. (c) After a 5 day differentiation time, cells are fixed and stained with specific antibodies, i.e. β -III-tubulin for neurons, O4 for oligodendrocytes and Hoechst for nuclei, and scanned with the ArrayScan V⁷ HCS Reader (ThermoFisher Scientific) to (d) get fragmented images for each staining channel and nuclei coordinates within the images. The fragmented images are then joined together to (e) get a completed neurosphere image for each staining channel. (f) Nuclei are located within the jointed image and quantified. Based on the nuclei coordinates, (g) trained ML models identify neurons or oligodendrocytes. (h) Identified corresponding cell types are then skeletonized and further analyzed for (i) their neurite length and number of branches. Based on the nuclei coordinates within the completed image and the identified cell types, (f) the migration distance of different corresponding cell types can be measured.

For this purpose, fluorescence images are acquired and nuclei located by the ArrayScan V⁷ HCS Reader (ThermoFisher Scientific) and vHSC Scan Software. The images are then imported into Omnisphero, where the nuclei locations are used as reference to identify neurons and oligodendrocytes within the corresponding staining (commonly β -III-tubulin for neurons and O4 for oligodendrocytes). Originally, Omnisphero relied on overlap-algorithms to identify neurons and oligodendrocytes (Schmuck *et al.*, 2015). However, ML algorithms were implemented and trained by supervised learning to vastly improve the performance (manuscript 2.1 – Förster *et al.*, 2021). With different cell types identified,

cell type-specific endpoints such as number of differentiated cells, their migration distance and density of identified cells can be measured. Identified cells are skeletonized to further analyze morphological endpoints such as neurite length or branching points (Schmuck *et al.*, 2015).

1.3.2 Biostatistical data evaluation

With the rising usage of HTS and HCIA tools, biostatistical data evaluation tools also emerged and became publicly available (Villeneuve *et al.*, 2019). Due to the broad variety of *in vitro* assays, there is an abundance of different approaches for data evaluation with these tools. Despite many scientific publications describing biostatistical methods, as well as guidelines for general concentration response data evaluation, e.g. published by the EFSA Scientific Committee (2016), there is no clear consensus on the use of biostatistical methods for *in vitro* toxicity data (Wheeler *et al.*, 2015; Sand *et al.*, 2017). Furthermore, different assays come with differences in concentration-response behavior (e.g. factors like variability within and between experiments or possible response levels may vary between assays). This leads to the challenge of finding one data evaluation protocol which appropriately evaluates all data deriving from different assays in one automated evaluation pipeline. In previous large-scale studies examining *in vitro* data for regulatory purposes, it was already shown that careful statistical evaluation is important to optimize test systems (Prieto *et al.*, 2013; Kropp-Schneider *et al.*, 2013) and that differences in statistical approaches can alter the outcome (Jensen *et al.*, 2020; Fischer *et al.*, 2020).

In order to quantify biological effects, a point-of-departure is estimated for concentration-response relationships. In recent years, the BMC (benchmark concentration) *method* introduced by Crump (1995) became the standard approach of effect readout and is now seen as a superior alternative to the no-observed-adverse-effect-level (Bokkers and Slob, 2005; Davis *et al.*, 2011). The BMC is defined as a concentration resulting in an effect at a predefined limit (benchmark response; BMR) below expected control treatment noise and is thus similar to Effective Concentration (EC) estimations or Lethal Concentration (LC) estimations (Jensen *et al.*, 2020). The uncertainty of a BMC is estimated by a confidence interval (CI), which is defined as the range between lower and upper limits (BMCL and BMCU respectively) and most commonly the 5% quantiles are used as limits. With the BMC and CI readout of single endpoints, the effect of chemical exposure on the cell system behavior is quantified.

This quantification then allows the assessment of potential hazard by classification of the biological specificity of exposure effects. For this purpose, classification models are applied. Classification models utilize decision trees and consider the BMC and its uncertainty to characterize potential hazards for compounds. This is usually done by applying classification categories such as “hit” or “not hit”. These

or similar categorizations then serve as orientation for which compounds might exert hazard and should thus e.g. be prioritized for further testing. As to identify DNT-specific effects, reference to general cell health is required. If a toxic effect is observed in an endpoint specific for developmental neurogenesis (“specific endpoint”; e.g. neuronal differentiation) and the effect is clearly distinct from an effect or the absence of such on general cell health (“unspecific endpoint”; e.g. viability or cytotoxicity), it can be considered as a specific DNT effect. In scenarios where there is uncertainty, if a hit is specific or unspecific, a borderline classification is recommended, to respect estimations with high uncertainty instead of separating them into either specific or unspecific hit categories (Leontaridou *et al*, 2017). In scenarios where there is high uncertainty in the data required for classification (e.g. large CI width or missing unspecific endpoint data), a flagging for subsequent expert judgement can be applied.

1.3.3 Data application

The evaluated data gives insight into the hazard potential of tested compounds and this information can subsequently be applied for regulatory purposes. Compound classifications give an overview over the general DNT potential of a compound, while BMCs and uncertainties can be used to narrow down concentrations causing disturbances of key neurodevelopmental processes (KNDP) and thus may cause DNT in an *in vitro* system. Compounds triggering several specific DNT effects or trigger a DNT effect at relatively low concentration can be prioritized. These metrics can be employed into IATA approaches and used as point-of-departure for subsequent steps, such as physiology-based kinetic modelling followed by *in vitro*-to-*in vivo* extrapolations (IVIVE) to convert the BMCs to estimated adverse doses. On the scale of industrial application, support for the approval process of pesticides or registration of a chemical would be one example of application. For some applications, also non-hits (i.e. compounds without any observed toxic effects) play an important role as well, e.g. the status of the safety of food constituents or contaminants. Furthermore, classifications can be used as reference for follow-up testing with orthogonal assays (assays which tackle the same biological phenomenon). If no or not sufficient *in vivo* DNT data is available for a regulatory question, *in vitro* data can be used to support the *in vivo* data and allow a regulatory decision. If available *in vivo* DNT data is inconclusive, *in vitro* testing can be used to inform the assessment based on Weight-of-Evidence (OECD 2019) for DNT (Crofton and Mundy, 2021).

The traditional method to evaluate DNT hazard potential is based on animal studies following the OECD test guideline TG426 or the EPA protocol oppts 970.6300. Due to their high demands on time and resources, only about 140 compounds have DNT guideline data available, revealing a vast gap in

compound DNT knowledge. In line with the shift from *in vivo* to *in vitro* test methods, data derived from NAMs can be applied to close this gap. Several steps need to be taken as prerequisites, before *in vitro* data can be used for regulatory purposes. First, established *in vitro* assays need to cover the most relevant biological processes, i.e. most crucial KNDP which may lead to DNT in humans. Second, used assays need to be validated as sufficiently robust and reliable in terms of DNT predictivity, which requires sound bioinformatical analysis and biostatistical evaluation.

1.4 Objectives of the thesis

The use of animals for toxicity testing is a very resource- and time-intensive procedure. To improve human risk assessment and attend these issues, the national research council proposed a new strategy for toxicity testing in the 21st century, in which a shift from conventional *in vivo* toxicity testing to high throughput but physiologically relevant *in vitro* assays is proposed. To make this shift possible, new procedures with novel technologies require novel data acquisition, management and evaluation software algorithms. Here, plated neurospheres, a secondary 3D highly complex *in vitro* model, was used as the basis. Neurospheres contribute to a DNT-IVB that is planned to be implemented into e.g. pesticide regulation. Due to the primary nature of this cell system, higher variabilities are observed than e.g. with immortalized tumor cell lines. Hence, a bioinformatics/biostatistics workflow is needed that accounts for such variabilities. To tackle this challenge, the overall aim of this thesis was to establish a bioinformatic workflow capable of

- 1) generating image data from plated neurospheres and analyzing the images for endpoint data,
- 2) evaluating the endpoint data with adequate statistical methodology and thus
- 3) enabling informed application of the evaluated data for hazard identification.

2 Manuscripts

The first manuscript (2.1) 'Reliable identification and quantification of neural cells in microscopic images of neurospheres' (Förster *et al.*, 2021) describes how ML approaches are utilized for HTS data generation. By analyzing neurosphere images, Omnisphero enables automatic generation of endpoint data reflecting neurological processes of brain development. CNN models that were developed for identification and quantification of either neurons and oligodendrocytes are described and validated for their performance in the context of application for toxicology screening.

The second manuscript (2.2) 'Biostatistics and its impact on hazard characterization using in vitro developmental neurotoxicity assays' explores the different biostatistical approaches and monitors them for their impact on hazard identification. Five key aspects of biostatistical DNT data evaluation were identified and monitored: 1) Experiment summary by either median or mean, 2) normalization by re-normalization or sole control normalization, 3) application of a best-fit algorithm for model fitting or enforcement of only one fit model, 4) different BMC and CI estimation approaches such as inverse regression, delta method, bootstrapping and model averaging and 5) measurement of different BMRs. The basis for this study is a compound screening project performed on behalf of an EFSA procurement during the years 2017-2020 (OC/EFSA/PRAS/2017/01). The DNT-IVB described in 1.2 was exposed with 148 compounds from different compound classes including expected positive and negative control compounds (Masjosthusmann *et al.*, 2020). These controls were used to assess the performance of the monitored biostatistical approaches to identify hazardous compounds with accuracy.

The third manuscript (2.3) 'Establishment of a human cell-based *in vitro* battery to assess developmental neurotoxicity hazard of chemicals' explores the feasibility of DNT hazard assessment based on NAMs. For this purpose, ten NAMs were combined into one DNT-IVB which covers relevant KNDPs such as proliferation of neuronal progenitor cells (NPCs), migration of several brain cell types, differentiation of neurons and oligodendrocytes, as well as neurite outgrowth. Additionally, several cell viability assays (often measuring as cytotoxicity) are included. A set of 120 compounds was analyzed and evaluated by the bioinformatics workflow presented in this thesis. To validate the accuracy of the DNT-IVB, pre-defined control compounds that are known to either induce toxic effects (positive controls) or to have no toxic effects (negative controls) were used as reference. A sensitivity of 82% and specificity of 100% was reached (manuscript 2.3 – Blum *et al.*, 2022), indicating the applicability of the DNT-IVB for regulatory purposes.

The fourth manuscript (2.4) 'Neurodevelopmental toxicity assessment of flame retardants (FRs) using a human DNT *in vitro* testing battery' utilized the bioinformatics workflow presented in this thesis to

identify potential DNT hazards deriving from flame retardants (including phased-out polybrominated FRs and organophosphorus FRs). For this purpose, the compounds were tested in the DNT-IVB, resulting in BMCs and classifications for each flame retardant, enabling informed assessment of potentially DNT-specifically hazardous retardants.

2.1 Reliable identification and quantification of neural cells in microscopic images of neurospheres

Nils Förster, Joshua Butke, **Hagen Eike Keßel**, Farina Bendt, Melanie Pahl, Lu Li, Xiaohui Fan, Ping-chung Leung, Jödis Klose, Stefan Masjosthumann, Ellen Fritsche, Axel Mosig

Cytometry part A

Aus primären humanen neuronalen Stamm-/Vorläuferzellen (hNPC) bestehende Neurosphären werden verwendet, um *in vitro* durch Substanzen induzierte Effekte auf frühe entwicklungsneurologische Prozesse zu untersuchen. Sobald auf geeigneter extrazellulärer Matrix ausplattiert, migrieren und differenzieren hNPCs zu Radialgliazellen, Neuronen, Astrozyten und Oligodendrozyten, und modellieren somit Prozesse der frühen neuronalen Entwicklung. Um Änderungen der Entwicklung von hNPCs zu charakterisieren, ist es notwendig den Zelltyp jeder Zelle innerhalb der Migrationsfläche zu identifizieren. Zu diesem Zweck präsentieren und validieren wir ein Ansatz des maschinellen Lernens zur Identifizierung und Quantifizierung von Zelltypen in mikroskopischen Bildaufnahmen von differenzierten hNPCs. Wie hier demonstriert, identifiziert unser Ansatz mit hoher Präzision und ist robust gegenüber typischen potentiellen Störfaktoren. Wir zeigen, dass unser Ansatz des maschinellen Lernens die Konzentrationswirkung von gut etablierten entwicklungsneurotoxischen Substanzen und Kontrollen reproduziert, was sein Potential für den Einsatz in Medium- bis Hochdurchsatz *in vitro* Screening Studien nachweist. Unser Ansatz kann somit für die Untersuchung von Substanzeffekten auf neurale Differenzierungsprozesse in einem automatisierten und unvoreingenommenen Prozess verwendet werden.

Reliable identification and quantification of neural cells in microscopic images of neurospheres

Nils Förster¹  | Joshua Butke²  | Hagen Eike Keßel³  | Farina Bendt³  |
 Melanie Pahl³  | Lu Li^{4,5,6}  | Xiaohui Fan⁴  | Ping-chung Leung⁶  |
 Jödis Klose³  | Stefan Masjosthusmann³  | Ellen Fritsche³  | Axel Mosig¹ 

¹Department of Bioinformatics, Center for Protein Diagnostics, Ruhr-University Bochum, Gesundheitscampus 4, Bochum, Germany

²Bioinformatics, Faculty of Biology and Biotechnology, Ruhr-University Bochum, Universitätsstr 150, Bochum, Germany

³IUF - Leibniz Research Institute for Environmental Medicine, Düsseldorf, North Rhine-Westphalia, Germany

⁴College of Pharmaceutical Sciences, Zhejiang University, Hangzhou, China

⁵Department of Obstetrics and Gynaecology, The Chinese University of Hong Kong, Shatin New Town, Hong Kong

⁶Institute of Chinese Medicine, The Chinese University of Hong Kong, Shatin New Town, Hong Kong

Correspondence

Axel Mosig, Department of Bioinformatics, Center for Protein Diagnostics, Ruhr-University Bochum (RUB), Gesundheitscampus 4, 44801 Bochum, NRW, Germany.
Email: axel.mosig@ruhr-uni-bochum.de

Funding information

Center for Alternatives to Animal Testing (CERST), Grant/Award Number: 233-1.08.03.03-121972/131-1.08.03.03; European Food Safety Authority (EFSA), Grant/Award Number: OC/EFSA/PRAS/2017/01

Abstract

Neurosphere cultures consisting of primary human neural stem/progenitor cells (hNPC) are used for studying the effects of substances on early neurodevelopmental processes in vitro. Differentiating hNPCs migrate and differentiate into radial glia, neurons, astrocytes, and oligodendrocytes upon plating on a suitable extracellular matrix and thus model processes of early neural development. In order to characterize alterations in hNPC development, it is thus an essential task to reliably identify the cell type of each migrated cell in the migration area of a neurosphere. To this end, we introduce and validate a deep learning approach for identifying and quantifying cell types in microscopic images of differentiated hNPC. As we demonstrate, our approach performs with high accuracy and is robust against typical potential confounders. We demonstrate that our deep learning approach reproduces the dose responses of well-established developmental neurotoxic compounds and controls, indicating its potential in medium or high throughput in vitro screening studies. Hence, our approach can be used for studying compound effects on neural differentiation processes in an automated and unbiased process.

KEYWORDS

deep learning, high content image analysis, neurospheres, neurotoxicology

1 | INTRODUCTION

Neurospheres are spherical cell aggregates consisting of (NPCs) [1]. While proliferating in 3D floating cultures, upon removal of growth factors and contact to an extracellular matrix, NPCs radially migrate out of the sphere and thereby differentiate into the major cell types of the brain, namely

radial glia, astrocytes, neurons, and oligodendrocytes [2–4] (Figure 1). Due to the neurospheres' ability to mimic such early neurodevelopmental processes, they are by now an established in vitro model for studying early neural development. In a broader sense, neurospheres as a spheroid cell culture system have become a popular and versatile tool for cancer drug research when consisting of tumor cells due to their ability to model

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial-NoDerivs](https://creativecommons.org/licenses/by-nc-nd/4.0/) License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.
© 2021 The Authors. *Cytometry Part A* published by Wiley Periodicals LLC on behalf of International Society for Advancement of Cytometry.

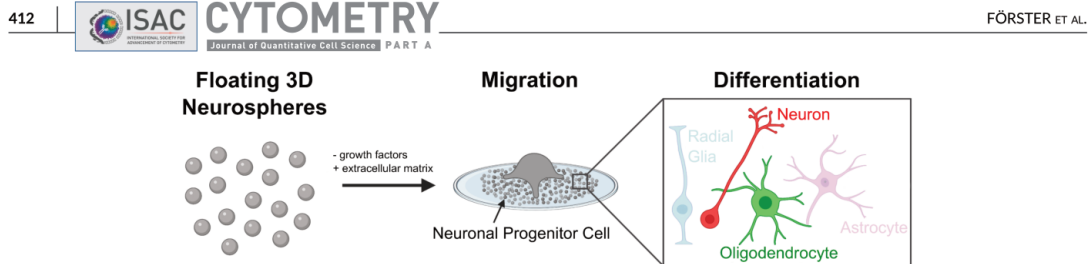


FIGURE 1 Schematic illustration of the neurosphere cell culture system. Free floating 3D neurospheres begin to migrate upon removal of growth factors and by plating on an extracellular matrix consisting of poly-D-lysine and laminin (see Section 2.1). During migration, they differentiate into radial glia, astrocytes, neurons, and oligodendrocytes. Following the protocol outlined in that section, this work aims to automate the quantification of cells differentiated from neurospheres. We aim to automatically identify differentiated neurons and oligodendrocytes using machine learning (see Section 3) based on morphological features (see Section 2.2) and a large manually annotated data set (Table 1). Figure created with [BioRender.com](https://www.biorender.com) [Color figure can be viewed at [wileyonlinelibrary.com](https://onlinelibrary.wiley.com)]

tumor architecture more accurately than conventional 2D cell cultures [5–7]. Furthermore, there are multiple applications for neurospheres in biomedical research, for example for inter-species comparison in evolutionary developmental biology, [8] for disease modeling, [9–12] as well as for investigating developmental neurotoxic effects of chemical compounds [13–16]. Thus, the opportunity to model neurodevelopmental processes in vitro, especially neural migration and differentiation, offers an elegant possibility to study adverse effects and modes-of-action of exogenous noxae like environmental chemicals in the developing nervous system without the use of animals [13, 17–19]. Evaluating (DNT) in vitro is not only of academic and ethical interest, but also caught attention of regulatory agencies like the European Food Safety Authority and the United States Environmental Protection Agency. [20–23]

Thus, capturing a microscopic image using suitable fluorescent markers (see Section 2.1) facilitates the quantification of differentiation status of the migrated cells as well as phenotypic features. Therefore, a core computational challenge addressed in our current contribution is the development of image processing and machine learning techniques to identify, classify and characterize the migrated cells in fluorescence microscopic images of differentiated neurospheres. While systematic approaches to quantitative image analysis have not yet been exploited for microscopic images of spheroids, that is, grown in secondary 3D, [24] our present contribution introduces such approaches for the neurosphere assay, implemented in a (HCA) fashion.

Image analysis procedures for microscopic imagery of neurospheres have been investigated previously, [25, 26] yet not by implementing machine learning. For example in Reference [26], a number of toxicological endpoints, such as the relative number of neurons, the integral neurite outgrowth and the migration pattern of cells, are obtained from microscopic images of differentiated neurospheres with fluorescently labeled nuclei and neurites. As demonstrated in this work and subsequent studies, [2, 13, 16] adversity in these endpoints reflects developmental neurotoxic effects of analyzed chemicals. Hence, automated image analyses can be considered as the computational core component of a promising in vitro approach for testing larger numbers of compounds. Yet, there is a remarkable lack of systematic studies of image processing approaches for the analysis of neurosphere microscopic imagery. Despite recently proposed methods to identify differentiation

patterns in neurospheres, [27] there is a lack in methods that are validated towards their robustness when using neurospheres as a screening assay, in particular in the context of toxicology.

For conventional neuronal 2D cell culture systems, numerous approaches have been developed for computationally quantifying or classifying phenotypic features of neurons. Besides elementary approaches for segmenting nuclei, tracing neurites has attracted major and systematic attention [28, 29]. Such facilitated the use of in vitro screening for chemical effects based on microscopic images of neuronal 2D cell cultures, [30–33] thus paving the way for high-content image analysis for neurotoxicity testing [34–36]. While HCA approaches for neurons cultured in vitro have been numerous, other neural cell types like radial glia, astrocytes or oligodendrocytes have so far been neglected.

While a plethora of well-established approaches is available for analyzing microscopic images of conventional neuronal cell cultures, especially those using only one cell type and adjustable cell densities, these approaches do not answer the computational questions arising from neurosphere image data. Beside the common first step of segmenting nuclei, most downstream analysis steps tend to be highly neurosphere-specific. Most notably, nuclei need to be assigned to either of the cell types that are observable in the mixed cell population of neurosphere assay. Experimentally, this is commonly accomplished by immunocytochemical stainings for specific structures, for example, neurons, in the sample. Computationally, this necessitates dedicated and well-validated approaches for cell type classification. Classification models need to be robust against the particularities of neurospheres. Specifically, compared to conventional neuronal cell culture, cell density cannot be controlled experimentally as it inherently varies from very dense areas around the sphere core to sparse cell concentrations in the periphery of the migration area. In addition, migrated cells differentiate into a lawn of mixed cell types [2, 26] (Figure 1), reflecting a rather complex situation for automated image analysis.

1.1 | Contribution and approach outline

The main goal of this study is to identify minimum standards required to train robust machine learning procedures that are required to utilize

differentiated spheroid cell culture models in combination with HCA in pharmacological or toxicological screening applications involving larger numbers of samples. Such minimum standards are essential, since training deep neural networks in biomedical settings involves careful considerations: the quality of deep neural networks relies essentially on the quality of the training data, which does not only depend on the amount of training data, but also whether all relevant sources of variance, bias and confounders are sufficiently covered. For complex spheroid models, obtaining training data comes at a particular price, since extremely labor intensive annotations are required. Thus, for bringing differentiated spheroid models into screening practice, annotation requirements must be well understood, motivating our effort to identify minimum standards for the underlying training data sets.

Our approach is outlined in Figure 2. Given an image of a complete neurosphere as an input, we identify nuclei as described in Section 2.1. A tile covering an area of roughly $56 \times 56 \mu\text{m}$ each around every nucleus is extracted for each tile. Each tile is then classified as a neuron, an oligodendrocyte or other cell type using a convolutional neural network.

Since cell types in the neurosphere assay occur in unbalanced ratios, one of the main challenges to overcome is class imbalance when training and validating our neural networks. Class imbalance is a notorious problem when training and validating classification models and has thus been investigated in much detail, in particular in the context of training convolutional neural networks [37, 38]. This imbalance also requires particular attention during validation, as has been observed by Reference [39].

2 | DATA AND METHODS

2.1 | Sample preparation and data acquisition

A detailed description of the experimental procedure can be found in Reference [3, 13]. Briefly, neurospheres (hNPCs) isolated from whole brains at GW16-19 (Lonza® Group, Verviers, Belgium) [4]) are grown as free-floating 3D aggregates under proliferative culture conditions for up to 7 weeks. Between weeks 3 and 7, the spheres are passed every week by mechanical dissociation using a tissue chopper. After two to 3 days, spheres with a defined size of 0.3 mm are plated for hNPC migration analyses onto poly-D-lysine/laminin-coated 96-well plates. Per well, one sphere was plated in 100 μl of differentiation medium. Spheres settle down and NPCs migrate radially out of the sphere core and differentiate for 5 days in presence and absence of compounds. Respective treatment conditions are summarized in Table S1. As seen in Table 1, plate preparation including compound dilutions and cell feeding was performed manually (11 plates) or by using an automated liquid handling system (8 plates).

As outlined in Figure 1, after a 5-day (120 h) differentiation period, cells were fixed with 4% paraformaldehyde (PFA) for 30 min at 37°C and labeled by performing immunocytochemical stainings. Therefore, cells were stained overnight at 4°C with IgM oligodendrocyte O4 antibody solution (1:400 in Phosphate-buffered saline [PBS] with 10% Goat Serum (GS) and 1% bovine serum albumin [BSA]), followed by secondary antibody solution (1:400 Alexa Fluor 488 anti-mouse IgM) which was added for 30 min at 37°C. For neuronal

FIGURE 2 Approach overview. Human neural progenitor cell were treated as described in Section 2.1. Pseudo-colored composite images were created by overlapping the Hoechst33258 (nuclei, blue), TUBB3 (neurites, red) and O4 (oligodendrocytes, green) antibody stainings. For every nucleus within a ROI (see Section 2.2) contained in the neurosphere image (A) contained within the migration area, a 64×64 pixel tile is created with the corresponding nucleus at the center. Every such tile covers an area of $\sim 56 \times 56 \mu\text{m}$. (A) Originally has an image resolution of 5520×5520 pixel, representing $\sim 4858 \times 4858 \mu\text{m}$. (B) Two examples, representing neurites (left) and oligodendrocytes (right). Every tile (C) has the unused color channel removed and a manual color adjustment is applied to each. (D) Both tiles being loaded into the corresponding CNN. The CNNs were trained based on annotated subregions throughout a large number of wells, as exemplified in (E), where also four annotated cells are highlighted [Color figure can be viewed at wileyonlinelibrary.com]

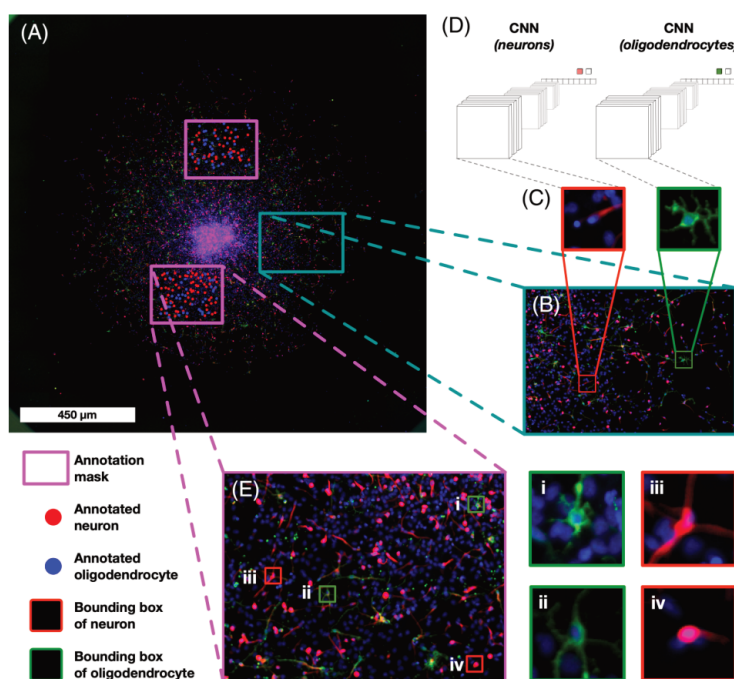


TABLE 1 Data set composition. Neurospheres from 19 different 96-well plates derived from four different individuals (column Individ.) were stained with antibodies detecting neurons and oligodendrocytes column plating indicates whether plates were prepared manually or by using an automated liquid handling system. The two columns under developmental neurotoxicity indicate chemical compounds and if the compound affects differentiation of the respective cell type, as further detailed in Table S1. The compound index in that table matches the index in the column compound. In general, 40 wells with one neurosphere each, were developed on each plate. The column staining shows the CNN(s) the plate was used for and the last three columns indicate the respective well distribution sets for the neural network(s)

Index		Stainings used		Plating	Developmental neurotoxicity		Number of wells distributed		
Plate	Individ.	Neurons	Oligos	Automated	Compound	Effects observed	Train	Val	Test
1	I1	✓			3	Both	36	4	0
2	I1	✓			11		36	4	0
3	I1	✓			11		36	4	0
4	I1	✓	✓		3	Both	36	4	0
5	I1	✓	✓		13	Oligodendrocytes	36	3	0
6	I1	✓	✓	✓	4		36	4	0
7	I1	✓	✓		1	Both	36	3	0
8	I2		✓		1	Both	36	3	0
9	I3		✓		9	Both	36	4	0
10	I1		✓		13	Oligodendrocytes	36	4	0
11	I1		✓		6	Both	36	3	0
12	I2		✓		6	Both	36	3	0
13	I2		✓	✓	12	Oligodendrocytes	36	4	0
14	I4		✓	✓	2	Oligodendrocytes	36	4	0
15	I4		✓	✓	5	Both	36	4	0
16	I1	✓	✓	✓	8	Neurons	35	4	1
17	I2	✓	✓	✓	7	Neurons	35	4	1
18	I1	✓	✓	✓	10	Oligodendrocytes	35	4	1
19	I4	✓	✓	✓	11		35	4	1

staining, cells were incubated for 1 h at 37°C with a conjugated rabbit TUBB3 antibody (1:400 in PBS with 10% Rabbit Serum [RS], 1% BSA). In parallel nuclei were stained with 5% Hoechst33258 [13].

Image acquisition was performed using the high content fluorescence imaging microscope Thermo Scientific ArrayScan® VTI (Thermo Fisher Scientific Inc.). Thereby one sphere and the migration area within on well were imaged in a 200-fold magnification in 64–100 individual images with an image resolution of 520 × 520 pixel each. These images were merged using [26] and nucleus identification was performed by the spot detector BioApplication of the ArrayScan® VTI scan software (Version 6.6.0; Thermo Fisher Scientific Inc.). Each well thus yields an image of roughly 5520 × 5520 pixels in size, representing an area of roughly 4.9 × 4.9 mm for each well.

2.2 | Data set and ground truth annotations

Altogether, RGB image data from nineteen 96-well plates were used, where each plate contained roughly 40 wells with one differentiated and stained neurosphere each. As summarized in Table 1, this data set represents different sources of variability as they occur in the application of the neurosphere assay. First, the neurospheres were obtained from four different human individuals labeled as I1–I4 throughout the manuscript, representing inter-individual differences. Furthermore,

plate preparation was performed manually or by using an automated liquid handling system representing differences in the plate preparation procedure. Finally, different plates contain different chemical compounds representing differences in the effect on the cell differentiation, including compound-dependent morphological differences.

In the resulting image, the sphere core was masked out and nuclei were detected in the remaining migration area as described previously [26]. For each well, two regions of interest (ROIs), each 800 × 600 pixel (~704 × 528 μm) in size, were randomly selected for annotation. Among nuclei in randomly selected sub-regions, neurons were manually annotated for 11 out of the 19 plates (referred to as neuron plates). Correspondingly, oligodendrocytes were manually annotated in 16 out of the 19 plates (referred to as oligo plates). For eight plates, both neurons and oligodendrocytes were annotated. Manual annotations of cell types in the data set were performed by four annotators and each annotation was controlled by a second annotator. For training and validation in the neuron data set, 32, 414 out of 211, 478 nuclei in the ROIs were annotated as neurons. In the oligo data set, 21, 929 out of 298, 808 nuclei were annotated as oligodendrocytes.

As indicated in Table 1, we followed a strict separation of our data set into fixed subsets for training, validation and testing. One well from the last four plates shown in Table 1 (indices 16–18) was reserved as the test set. In these wells six to eight ROIs were selected. This resulted in a

total number of 10, 945 nuclei in the neuron set. Of those, 1, 114 were annotated as neurons. In the oligodendrocyte set, 718 out of 10, 945 nuclei were annotated correspondingly as oligodendrocytes (Table 2).

2.3 | Preprocessing

For each nucleus within each ROI, a 64×64 pixel 8-bit RGB image tile with the nucleus centered within the tile is generated each (Figure 2A). The nuclei were identified and segmented using Reference [26]. Each well contains up to two ROIs.

To adjust for global, plate dependent intensity variance among fluorescence microscopic images of different plates, intensities and contrast were adjusted manually for each individual plate. Subsequent to manual adjustment of whole plates, we subsequently conduct normalization at the level of wells and tiles. At this level, we implemented and compared different procedures to normalize the integer-valued image intensities to a floating point value between 0.0 and 1.0:

- *flat normalization*: Each value is divided by 255.
- *per tile, separate channels*: Each channel in each tile is min-max normalized independently.
- *per tile, across channels*: Min-max is determined across all color channels with a tile, so that all three color channels of the tile are normalized using the same min and max values.
- *per well, across channels*: Min-max is determined across all color channels within a complete well, and all tiles in the well are normalized using the same min and max values across the well.

2.4 | Convolutional neural networks

We established convolutional neural networks to classify 64×64 pixel tiles into different cells types as follows: we trained two separate networks, one for distinguishing neurons from non-neurons, and a

TABLE 2 Nuclei based cell counts in data sets for training, validation and test. As indicated, roughly 15% of nuclei have been annotated in the neuron data set; in the oligodendrocyte data set, roughly 7% of nuclei have been annotated. The annotation quantifies the imbalance of cell types with less than one out of five cells being a neuron and less than one out of 11 cells an oligodendrocyte. Numbers refer to nuclei counts within the ROIs of neurospheres

		Neuron set	Oligo set
	Experiments	11	16
	Total nuclei	222, 423	309, 753
Label annotated	Training	28, 548	19, 486
	Validation	3, 866	2, 443
	Test	1, 114	718
Label not annotated	Training	158, 932	248, 908
	Validation	20, 132	27, 971
	Test	9, 831	10, 227

second network that was trained to distinguish oligodendrocytes from non-oligodendrocytes. While both networks technically operate on images with three channels, the putatively uninformative channel for each network was masked as plain background, that is, in the neuron model, the O4-stained oligodendrocyte channel was blackened for all training, validation, and test data, and the same was done for the TUBB3-stained neurite channel in the oligodendrocyte neural network processes. This procedure was motivated by the largely non-overlapping annotations in the training data (Table 1) and further legitimized through a preliminary network trained on the set of plates with both neuron and oligodendrocytes, which did not achieve the performance of separately trained models (see Section 3). This process also mirrors the manual annotation process in, [26] where plates with only one label had the unused color channel missing or disabled.

As a topology for both CNNs, we employed a slightly modified version of the VGG topology [40] summarized in Table 3. The cornerstone of the employed topology is the repetition of the core convolutional feature extraction block. Each of these blocks configured in exactly the same manner, except for the number of kernels the convolutional 2D operation outputs as its feature maps. Typically, this number increases in modern CNNs applications before the final feature maps are flattened into a feature vector for subsequent classification by a fully connected network. Following common practice, we used ReLU as an activation function, except for a the last output neuron which uses a sigmoidal activation function in combination with binary cross-entropy as loss function.

All layers are initialized according to Reference [41]. To prevent the network from over-fitting, we followed common practice and introduced dropout [42] within the fully connected layer with a dropout rate of 50%. For training each of the resulting models, different optimizer strategies and settings were tested.

Training was conducted with a batch size of 100 over 5000 epochs. During training the learn rate (initially 0.001) was halved automatically, whenever 100 epochs passed with no improvement to the validation loss. If the validation loss did not improve further after two such reductions, the training would stop early.

All deep learning computations were implemented in Tensorflow 1.13 with Keras 2.2 [43] running on Python 3.8.2 and performed on a GPU server running Ubuntu 18.04 LTS with four Nvidia® GeForce® GTX 1080 Ti graphics cards. All other processing and data analysis tasks were

TABLE 3 Overview of the neural network topology used for both neuron and oligodendrocyte classification

Layer	Type
1	Conv2D(3,1,0)-32 + ReLU + BatchNorm
2	Conv2D(3,1,0)-32 + ReLU + BatchNorm
3	MaxPooling2D(2,2)
4-10	Two repetitions of layers 1-3 as blocks with increasing kernel size 32-64-128
11	FC256 + ReLU + Dropout(0.5)
12	FC1 + Sigmoid classification

carried out with an extension of the previously published OmniSphero software, [26] running on MATLAB® 2014b (The MathWorks®, Inc.).

2.5 | Validation and dealing with class imbalance

As indicated in Table 2, cell identification in the neurosphere model constitutes a machine learning problem with substantial class imbalance: Only one out of five nuclei is a neuronal nucleus, and only one out of 12 nuclei belongs to an oligodendrocyte. This imbalance in our data set, which is in line with figures reported in Reference [13], needs to be taken into account for both training and validation of machine learning models.

While under-sampling is prohibitive in our given setting, we assessed two over-sampling approaches in order to increase the number of samples in the under-represented classes in the training data set by introducing randomly rotated and mirrored tiles as further data points. The second over-sampling approach we assessed is SMOTE [44] which depends on the k -nearest neighbors principle to find (in this case) similar tiles and synthesize more, where k is an essential parameter that influences the characteristics of the over-sampled data set. As a baseline to test the effectiveness of SMOTE, both data sets were trained using $k = 5$. Then, another model was trained and tested, determining k to be 4% of the size of the minority class, avoiding over-fitting and resulting in $k_{\text{Neuron}} = 1142$ and $k_{\text{Oligodendrocyte}} = 780$. Lastly, both methods were tried with the original training-data being over-sampled first, before being augmented. The respective (PR) curves can be seen in Figure 3.

In order to properly validate the trained CNNs on our class-imbalanced data sets, we relied on (PR) curves [39] which are not susceptible to misleading interpretation in the presence of class-imbalance as (ROC) curves are. We followed [39] and used the (AUC) of PR curves as the main indicator of classification strength.

3 | RESULTS

3.1 | Validation and comparison of over-sampling and normalization approaches

Panels A and B of Figure 3 display PR curves that compare different over-sampling approaches. While the classification of oligodendrocytes is hardly affected by different forms of over-sampling, the classification of neurons varies greatly. Over-sampling using SMOTE in either variant is neutral at best or leads to a small decrease in the AUC. Following this observation, data augmentation was used for training of any subsequent models. This augmentation was realized in randomly rotating and mirroring the tiles during training.

A similar pattern emerges for the comparison of normalization methods (Figure 3C,D). In general, only the classification of neurons, but not of oligodendrocytes, is affected by normalization techniques. Per-well normalization performs better than per-tile normalization. Likely, tile-based normalization is unfavorable in regions highly

overpopulated or highly underpopulated with nuclei, so that the heterogeneous structure or exposure within a neurosphere well will be grasped better by a well-based normalization procedure. This is further supported by our observation that there is a significant illumination bias for each individual slide, as shown and detailed in Figure S1.

3.2 | Robustness under data heterogeneity

To assess the heterogeneity in our data, we specifically investigated heterogeneity between neurospheres derived from different human individual donors. We trained a neuron classifier on neurons from individual I1 and compared performance on an independent test set containing neurons from one of the remaining individuals. As a baseline comparison, we predicted on a data set containing all three individuals. The resulting PR-curves displayed in panel A of Figure 4 show a strong generalization to individual I4, but a much weaker generalization to I2. For the analogous validation of oligodendrocyte generalization performance (Panel B), internal validation on I1 yields a very strong AUC, which drops to levels comparable to neuron identification on external validation.

This indicates a significant amount of heterogeneity among different individuals for both neurons and oligodendrocytes. In other words, a network trained on spheres derived from one (or more generally too few) individuals will not generalize well on test data derived from other individuals. A network trained on data from more than one individual matches the performance of internal validation with a single individual, illustrating that the network exhibits robust generalization across individuals when the heterogeneity is represented in the training data.

3.3 | Assessment of training data composition

In order to assess how the ratio between training and validation data size affects classification, we varied our default ratio of roughly 10:1 to roughly 4:1 by transferring nuclei from the training to the validation set. As shown in panel D of Figure 4, this negatively affects classification performance, indicating that a small validation set is sufficient for training robust neural networks.

3.4 | Intersections between the models

Since not all training and validation data sets provide stainings of both neurons and oligodendrocytes (Table 1), the identification and differentiation of the two cell types cannot be realized in one CNN. To address this situation, we trained one CNN to distinguish neurons from non-neurons and a second CNN to distinguish oligodendrocytes from non-oligodendrocytes.

In practice, developing neurospheres can partly result in dense nuclei clusters (as seen in Panel E of Figure 2). These clusters can yield the risk of overlapping classification of the same nuclei by both

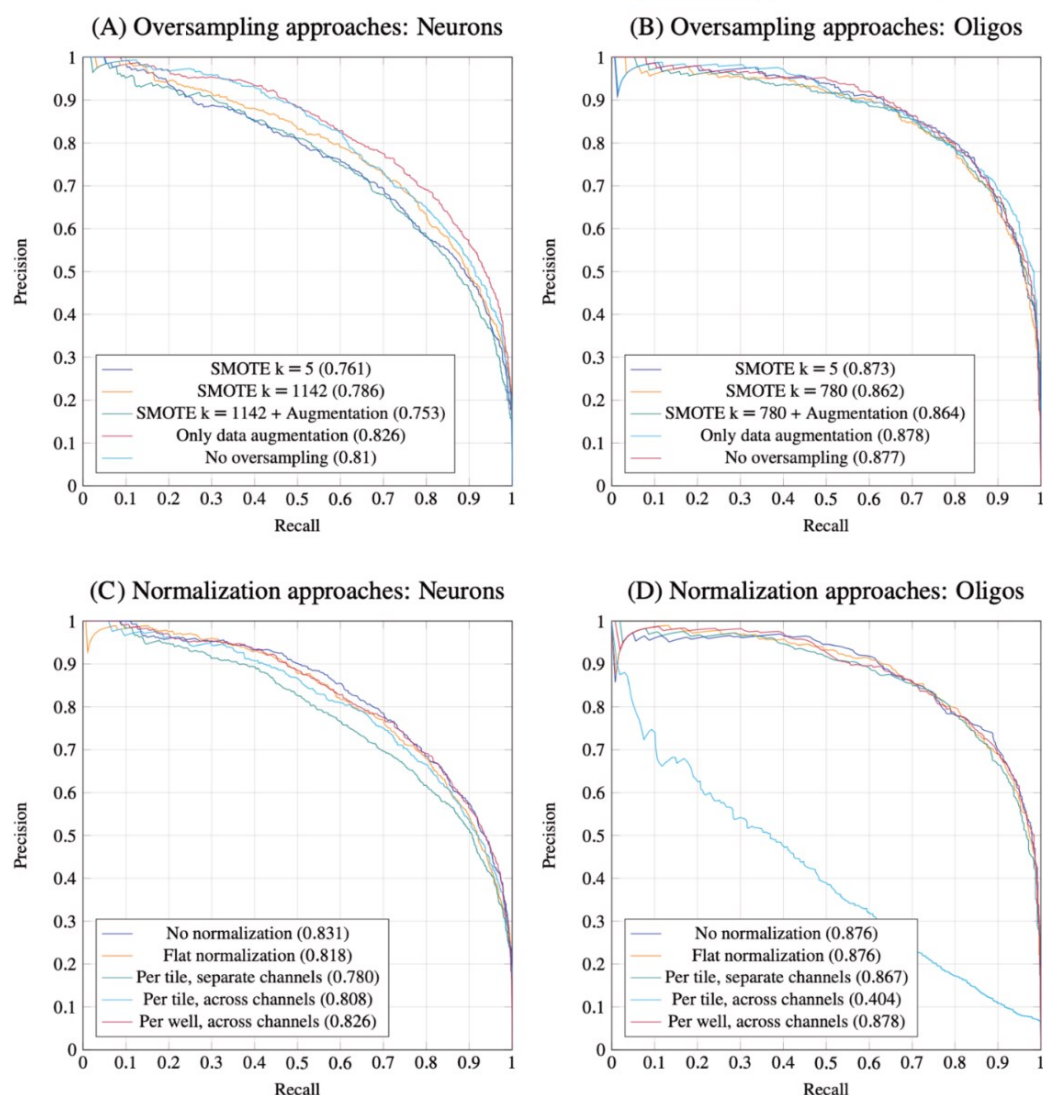


FIGURE 3 Precision-recall (PR)-curves for comparison of computational approaches and training sets. (A,B) PR-curves for different upsampling approaches. Correspondingly, (C) and (D) compare different normalization approaches. Legend entry values within brackets show the respective area under curves [Color figure can be viewed at wileyonlinelibrary.com]

models. As a simple policy, we assign nuclei with overlapping classification as oligodendrocytes. To assess the effects of intersecting classifications, we predicted the validation data set (Table 1) with both models. The resulting PR curves of the neuron model can be seen in panel C of Figure 4. Afterwards, we removed intersecting predictions to determine the second PR curve within the same panel. The resulting AUCs differ only marginally. This shows that intersecting predictions of our models do not lead to significant misclassifications.

3.5 | Uncovering concentration-response relations

In order to assess whether the precision and recall achieved by the neural networks is sufficient for compound screening applications, we used relative frequencies of neuron and oligodendrocyte counts along a concentration series of chemicals, yielding the concentration-response curves displayed in Figure 5. As previously outlined in, [13, 19, 26, 45] the compounds methylmercury(II)chloride and 3,3',5,5'

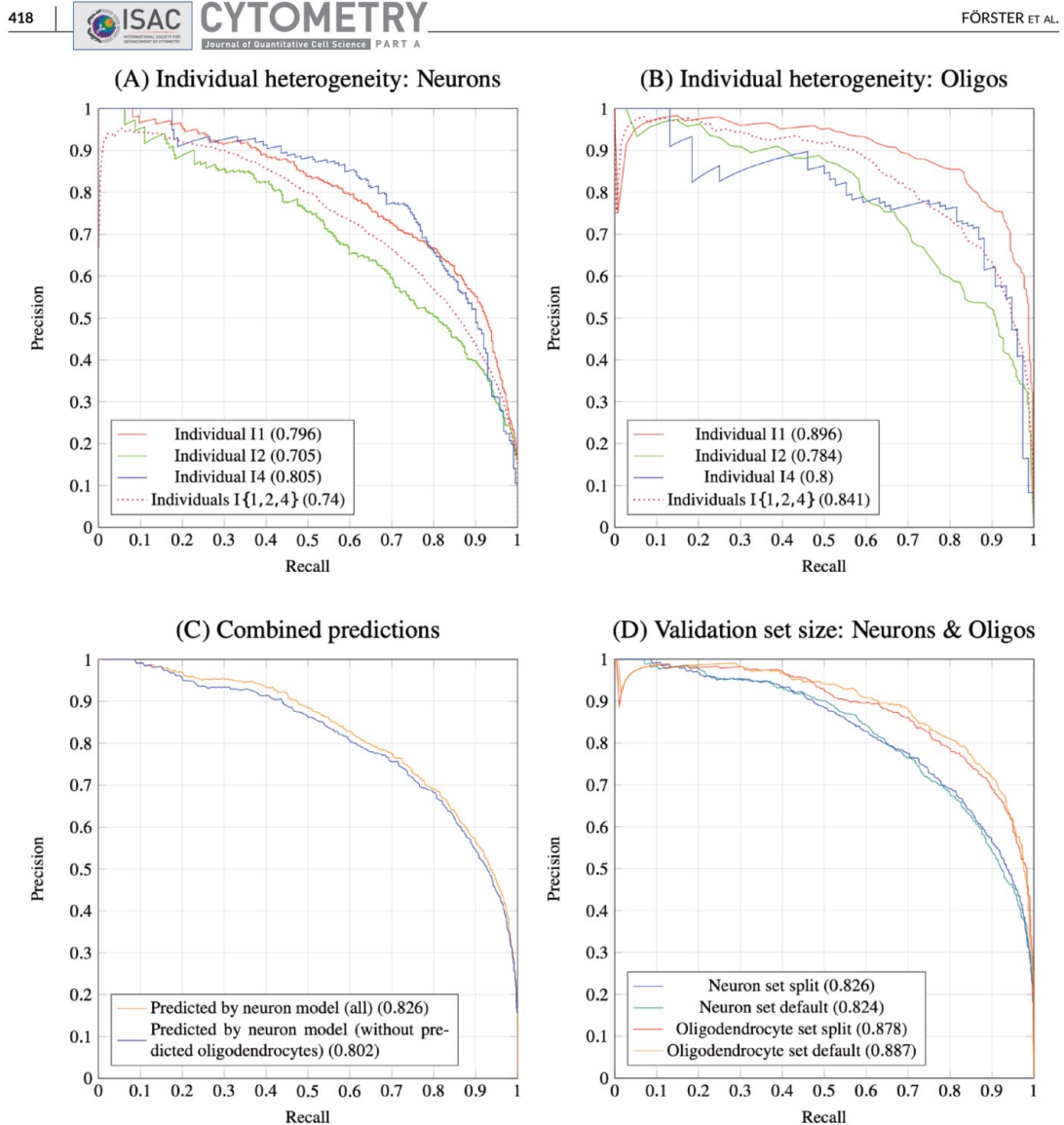


FIGURE 4 Precision-recall-curves for comparison of neurosphere specific conditions. (A,B,D) Effects due to training data heterogeneity on classification performance: In (A) and (B), the heterogeneity among individual human stem cell donors can be seen. (C) Compares the performance of the neuron model when removing prediction-intersections with the oligodendrocyte model. (D) The effect of increasing the size of the validation set at the cost of decreasing the size of the training data. Legend entry values within brackets show the respective area under curves [Color figure can be viewed at wileyonlinelibrary.com]

Tetrabromobisphenol A are established as known human DNT compounds. Moreover, Ibuprofen was used as a negative control, as it provided no DNT effects in hNPCs [46]. Both compounds (see Table S1) and the respective underlying experiments were not used during CNN training. By performing the neurosphere assay as described in Section 2.1 and by replicating these baseline results on the toxicology spectrum, we validate the performance and robustness of the presented approach.

4 | DISCUSSION

The use of spheroid cell culture systems in biomedical research has matured over the past two decades, paving the way for spheroid-based in vitro screening. Microscopic image analysis for capturing and quantifying phenotypic features is at the core of corresponding screening assays, and our present systematic validation study demonstrates how to approach the challenges arising from the relatively

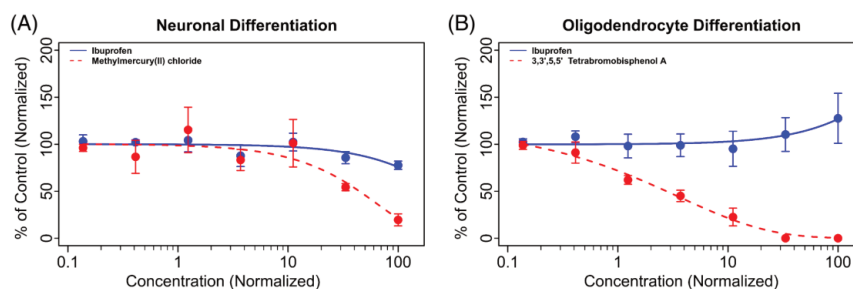


FIGURE 5 Concentration–response curves for validation. We followed the protocol outlined in Section 2.1 and treated human neural progenitor cells (hNPCs) with the known developmental neurotoxicants methylmercury(II)chloride and 3,3',5,5' tetrabromobisphenol A and the negative compound ibuprofen (see Table S1). Quantification of neurons (A) and oligodendrocytes (B) in the migration area was performed using the CNNs presented in this work. The neuronal and oligodendrocyte differentiation was determined as percent of neurons/oligodendrocytes in the migration area (acquired via Reference [26]) for each concentration. The concentrations were normalized to the respective solvent controls for each compound (20 μ M for ibuprofen and 3,3',5,5' tetrabromobisphenol A and 2.22 μ M for methylmercury(II)chloride). Each curve was generated based on three independent experiments (with five replicates for each condition) using hNPC from three different human individuals. Error bars represent the standard error of the mean [Color figure can be viewed at wileyonlinelibrary.com]

complex and heterogeneous image data obtained from neurospheres. While our specific *in silico* classifiers resulting from this study are specific for the neurosphere assay, we expect that many of our approaches and observations can be carried from primary human neurospheres to other multicellular models, in particular tumor spheres or induced pluripotent stem cell-based models, [47] which have been recognized as a promising basis for screening assays [48]. Demonstrating the robustness towards toxicological screening also distinguishes our approach from the recent work by Zhu et al. [27].

Annotation is a major obstacle if not the main bottleneck for training deep learning models that classify image data from models. Training robust deep learning models requires not just sufficient amounts of training data, but, as we show, they also need to cover the variances and heterogeneity encountered when the model is productively used in larger scale screening studies. A large share of the robustness of our presented model is certainly due to the effort of annotating substantial parts of 680 neurospheres covering different compounds and all ranges of concentrations. Our deep learning models are naturally limited to classifying cell types in the neurosphere model, and spheroid systems other than the neurosphere system will require the training of new deep learning models. Yet, our study also allows formulating the following specific guidance and minimum standards for validation in future studies of similar or even more complex cell culture or organoid systems and their respective assays:

1. *Perform external validation:* Over-fitting is an inherent and obvious danger of CNNs, which can potentially lead to display false positive or false negative dose–response relationships in compound screening. The ideal gold standard would be a fully independent test, that is, a validation on data from a different laboratories using different devices and a fully independent sample preparation. This gold standard, however, is impractical in most settings. In larger scale studies, our validation across neurospheres grown from

different individuals exemplifies this: Networks trained on more than one single individual are robust across more individuals. It is thus an essential part of CNNs in screening applications to identify sources of bias and confounding, and validate against heterogeneity under such confounders. For some factors such as fluctuations in fluorescence intensity, there may be a trade-off between normalizing, that is, aiming to eliminate variance, versus reflecting variance in the training data. As a matter of good validation practice, aspects of class-imbalance will be commonplace in complex cell cultures and should be addressed by proper validation measures such as using PR-curves rather than ROC-curves.

2. *Validate end-to-end:* Endpoints such as counting cells of certain types are only intermediate in the sense that they do not immediately display the effect of a substance or other intervention in a single neurosphere, but only in relation to other neurospheres. It is thus essential to validate dose–response relationships for substances with well-established dose–response in comparison to control-substances with well-established neutral effect, and assess whether the expected response can be called from dose–response curves.

Annotation has been recognized as a major bottleneck in several other contexts in bioimage analysis. To address the often prohibitive costs of annotation, some researchers have established crowd-sourcing resources [49] which facilitate to obtain large amounts of annotations through volunteers. To deal with the inherent problem of reliability of such non-expert annotations, the histopathology community has developed a hierarchical panel approach that combines large-scale annotations by volunteers with a systematic review by experts [50]. Such approaches are certainly conceivable for spheroid models as well.

Precision and recall of our models reach AUC values around 0.8 for oligodendrocytes and slightly less for neurons, which is clearly

sufficient for neurotoxicity screening. A limiting factor may be label noise in the annotations which is an inherent problem with human annotations. Whether obtained from crowdsourcing or from experts, large-scale annotations can hardly ever claim 100% accuracy. This problem has been addressed through computational approaches, and it is conceivable that approaches based on either weakly supervised learning [51] or one-shot-learning [52, 53] can be utilized to alleviate the problem of label noise. Since this will require the integration of sophisticated machine learning algorithms into interactive annotation systems, there is an implementation hurdle. In addition, introducing further semi-automated support into the annotation procedure is in danger of introducing an annotation bias which would need to be examined carefully in resulting classifiers. Finally, it can be projected that the analysis of spheroid microscopic image data will become easier with pre-trained models, which can be transferred to new tasks or new cell culture models with comparatively small amounts of training data.

In summary, carefully validated deep learning are promising approaches to further advance the use of spheroid models in screening applications. We suggest that identifying guidelines and minimum standards in this work is an important contribution for such deep learning methods to gain acceptance.

ACKNOWLEDGMENTS

This research received support by the project CERST (Center for Alternatives to Animal Testing) of the Ministry for culture and science of the State of North-Rhine Westphalia, Germany (File No. 233-1.08.03.03-121972/131-1.08.03.03-121972) and EFSA (European Food Safety Authority) [OC/EFSA/PRAS/2017/01].

AUTHOR CONTRIBUTIONS

Nils Förster: Conceptualization (equal); data curation (supporting); formal analysis (lead); methodology (lead); software (lead); validation (supporting); visualization (lead); writing – original draft (equal); writing – review and editing (equal). **Joshua Butke:** Conceptualization (supporting); data curation (supporting); methodology (supporting); software (supporting); validation (supporting); writing – original draft (supporting); writing – review and editing (supporting). **Hagen Eike Keßel:** Validation (supporting); visualization (supporting). **Farina Bendt:** Investigation (equal); validation (supporting). **Melanie Pahl:** Investigation (equal); resources (equal); validation (supporting). **Lu Li:** Resources (supporting). **Xiaohui Fan:** Resources (supporting). **Ping-chung Leung:** Resources (supporting). **Jördis Klose:** Investigation (equal); resources (equal); validation (supporting); visualization (supporting); writing – review and editing (supporting). **Stefan Masjosthusmann:** Data curation (supporting); funding acquisition (supporting); investigation (equal); resources (equal); supervision (supporting); validation (supporting); writing – review and editing (equal). **Ellen Fritsche:** Data curation (supporting); funding acquisition (equal); project administration (supporting); resources (lead); supervision (supporting); validation (supporting); writing – review and editing (equal). **Axel Mosig:** Conceptualization (equal); funding acquisition (equal); project administration (lead); resources (supporting); software

(supporting); supervision (equal); validation (equal); writing – original draft (equal); writing – review and editing (equal).





CONFLICT OF INTEREST

The authors declare no conflict of interest.

PEER REVIEW

The peer review history for this article is available at <https://publons.com/publon/10.1002/cyto.a.24514>.

ORCID

Nils Förster  <https://orcid.org/0000-0003-4226-754X>
Joshua Butke  <https://orcid.org/0000-0002-8395-3469>
Hagen Eike Keßel  <https://orcid.org/0000-0003-3023-1830>
Farina Bendt  <https://orcid.org/0000-0002-1653-1151>
Melanie Pahl  <https://orcid.org/0000-0003-4002-0738>
Lu Li  <https://orcid.org/0000-0002-9746-050X>
Xiaohui Fan  <https://orcid.org/0000-0002-6336-3007>
Ping-chung Leung  <https://orcid.org/0000-0002-0195-4688>
Jördis Klose  <https://orcid.org/0000-0002-2924-555X>
Stefan Masjosthusmann  <https://orcid.org/0000-0003-1493-7980>
Ellen Fritsche  <https://orcid.org/0000-0002-7454-679X>
Axel Mosig  <https://orcid.org/0000-0001-7266-8323>

REFERENCES

1. Reynolds BA, Weiss S. Generation of neurons and astrocytes from isolated cells of the adult mammalian central nervous system. *Science*. 1992;255(5052):1707–10.
2. Masjosthusmann S, Becker D, Petzuch B, Klose J, Siebert C, Deenen R, et al. A transcriptome comparison of time-matched developing human, mouse and rat neural progenitor cells reveals human uniqueness. *Toxicol Appl Pharmacol*. 2018;354:40–55.
3. Nimtz L, Klose J, Masjosthusmann S, Barenys M, Fritsche E. The neurosphere assay as an in vitro method for developmental neurotoxicity (DNT) evaluation. In: Aschner M, Costa L (eds). *Cell Culture Techniques. Neuromethods*, (Vol. 145, pp.141-168). Humana, New York, NY: Springer.
4. Moors M, Rockel TD, Abel J, Cline JE, Gassmann K, Schreiber T, et al. Human neurospheres as three-dimensional cellular systems for developmental neurotoxicity testing. *Environ Health Perspect*. 2009; 117(7):1131–8.
5. Kunz-Schughart LA, Freyer JP, Hofstaedter F, Ebner R. The use of 3-D cultures for high-throughput screening: the multicellular spheroid model. *J Biomol Screen*. 2004;9(4):273–85.
6. Sant S, Johnston PA. The production of 3D tumor spheroids for cancer drug discovery. *Drug Discov Today Technol*. 2017;23:27–36.
7. David P, Zurich M-G, Hartung T. Organotypic models to study human glioblastoma—studying the beast in its ecosystem. *Iscience*. 2020; 23(10):101633.
8. Kitajima R, Nakai R, Imamura T, Kameda T, Kozuka D, Hirai H, et al. Modeling of early neural development in vitro by direct neurosphere formation culture of chimpanzee induced pluripotent stem cells. *Stem Cell Res*. 2020;44:101749.
9. Barenys M, Illa M, Hofrichter M, Loreiro C, Pla L, Klose J, et al. Rabbit neurospheres as a novel in vitro tool for studying neurodevelopmental effects induced by intrauterine growth restriction. *Stem Cells Transl Med*. 2021;10(2):209–21.
10. Tang H, Hammack C, Ogden SC, Wen Z, Qian X, Li Y, et al. Zika virus infects human cortical neural progenitors and attenuates their growth. *Cell Stem Cell*. 2016;18(5):587–90.

11. Lee Y-K, Hwang S-K, Lee S-K, Yang J-E, Kwak J-H, Seo H, et al. Cohen syndrome patient iPSC-derived neurospheres and forebrain-like glutamatergic neurons reveal reduced proliferation of neural progenitor cells and altered expression of synapse genes. *J Clin Med*. 2020;9(6):1886.
12. da SG Pedrosa C, Goto-Silva L, Temerozo JR, Gomes IC, Souza LR, Vitória G, et al. Non-permissive SARS-CoV-2 infection in human neurospheres. *Stem Cell Research*. 2021;54:102436.
13. Masjosthusmann S, Blum J, Bartmann K, Dolde X, Holzer A-K, Stürzl L-C, et al. Establishment of an a priori protocol for the implementation and interpretation of an in-vitro testing battery for the assessment of developmental neurotoxicity. *EFSA Support Publ*. 2020;17(10):1938E.
14. Kelava I, Lancaster MA. Dishing out mini-brains: current progress and future prospects in brain organoid research. *Dev Biol*. 2016;420(2):199–209.
15. Fritsche E, Barenys M, Klose J, Masjosthusmann S, Nimtz L, Schmuck M, et al. Current availability of stem cell-based in vitro methods for developmental neurotoxicity (DNT) testing. *Toxicol Sci*. 2018;165(1):21–30.
16. Klose J, Pahl M, Bartmann K, Bendt F, Blum J, Dolde X, et al. Neurodevelopmental toxicity assessment of flame retardants using a human dnt in vitro testing battery. *Cell Biol Toxicol*. 2021;1–27.
17. Dach K, Bendt F, Huebenthal U, Giersiefer S, Lein PJ, Heuer H, et al. BDE-99 impairs differentiation of human and mouse npcs into the oligodendroglial lineage by species-specific modes of action. *Sci Rep*. 2017;7:44861.
18. Bal-Price A, Hogberg HT, Crofton KM, Daneshian M, FitzGerald RE, Fritsche E, et al. Recommendation on test readiness criteria for new approach methods (nam) in toxicology: exemplified for developmental neurotoxicity (dnt). *ALTEX*. 2018;35(3):306–52.
19. Klose J, Tigges J, Masjosthusmann S, Schmuck K, Bendt F, Huebenthal U, et al. TBBPA targets converging key events of human oligodendrocyte development resulting in two novel AOPs. *ALTEX*. 2020;38(2):215–34.
20. Fritsche E, Crofton KM, Hernandez AF, Hougaard Bennekou S, Leist M, Bal-Price A, et al. OECD/EFSA workshop on developmental neurotoxicity (DNT): the use of non-animal test methods for regulatory purposes. *ALTEX*. 2017;34(2):311–5.
21. Sachana M, Bal-Price A, Crofton KM, Bennekou SH, Shafer TJ, Behl M, et al. International regulatory and scientific effort for improved developmental neurotoxicity testing. *Toxicol Sci*. 2019;167(1):45–57.
22. Fritsche E, Grandjean P, Crofton KM, Aschner M, Goldberg A, Heinonen T, et al. Consensus statement on the need for innovation, transition and implementation of developmental neurotoxicity (DNT) testing for regulatory purposes. *Toxicol Appl Pharmacol*. 2018;354:3–6.
23. Sachana M, Shafer TJ, Terron A. Toward a better testing paradigm for developmental neurotoxicity: OECD efforts and regulatory considerations. *Biology*. 2021;10(2):86.
24. Alépée N, Bahinski A, Daneshian M, De Wever B, Fritsche E, Goldberg A, et al. t4 workshop report: state-of-the-art of 3D cultures (organs-on-a-chip) in safety testing and pathophysiology. *ALTEX*. 2014;31(4):441–77.
25. Poli D, Magliaro C, Ahluwalia A. Experimental and computational methods for the study of cerebral organoids: a review. *Front Neurosci*. 2019;13:162.
26. Schmuck MR, Temme T, Dach K, de Boer D, Barenys M, Bendt F, et al. Omnisphero: a high-content image analysis (HCA) approach for phenotypic developmental neurotoxicity (DNT) screenings of organoid neurosphere cultures in vitro. *Arch Toxicol*. 2017;91(4):2017–2028.
27. Zhu Y, Huang R, Wu Z, Song S, Cheng L, Zhu R. Deep learning-based predictive identification of neural stem cell differentiation. *Nat Commun*. 2021;12(1):1–13.
28. Meijering E, Jacob M, Sarria J-C, Steiner P, Hirling H, Unser M. Design and validation of a tool for neurite tracing and analysis in fluorescence microscopy images. *Cytometry A*. 2004;58(2):167–76.
29. Pool M, Thiemann J, Bar-Or A, Fournier AE. Neuritracer: a novel imagej plugin for automated quantification of neurite outgrowth. *J Neurosci Methods*. 2008;168(1):134–9.
30. Radio NM, Mundy WR. Developmental neurotoxicity testing in vitro: models for assessing chemical effects on neurite outgrowth. *Neurotoxicology*. 2008;29(3):361–76.
31. Harrill JA, Robinette BL, Freudenrich T, Mundy WR. Use of high content image analyses to detect chemical-mediated effects on neurite sub-populations in primary rat cortical neurons. *Neurotoxicology*. 2013;34:61–73.
32. Harrill JA, Freudenrich TM, Robinette BL, Mundy WR. Comparative sensitivity of human and rat neural cultures to chemical-induced inhibition of neurite outgrowth. *Toxicol Appl Pharmacol*. 2011;256(3):268–80. <http://www.sciencedirect.com/science/article/pii/S0041008X11000676>
33. Harrill JA, Robinette BL, Mundy WR. Use of high content image analysis to detect chemical-induced changes in synaptogenesis in vitro. *Toxicol In Vitro*. 2011;25(1):368–87. <https://www.sciencedirect.com/science/article/pii/S0887233310002699>
34. Radio NM, Breier JM, Shafer TJ, Mundy WR. Assessment of chemical effects on neurite outgrowth in pc12 cells using high content screening. *Toxicol Sci*. 2008;105(1):106–18.
35. Li S, Xia M. Review of high-content screening applications in toxicology. *Arch Toxicol*. 2019;93:3387–96.
36. Van Vliet E, Daneshian M, Beilmann M, Davies A, Fava E, Fleck R, et al. Current approaches and future role of high content imaging in safety sciences and drug discovery. *ALTEX*. 2014;31(4):479–93.
37. Buda M, Maki A, Mazurowski MA. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Netw*. 2018;106:249–59.
38. Johnson JM, Khoshgoftaar TM. Survey on deep learning with class imbalance. *J Big Data*. 2019;6(1):27.
39. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One*. 2015;10(3):e0118432.
40. K. Simonyan and A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556, 2014.
41. K. He, X. Zhang, S. Ren, and J. Sun, Delving deep into rectifiers: surpassing human-level performance on imagenet classification, in *Proceedings of the IEEE international conference on computer vision (ICCV)*, December 2015.
42. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. *J Mach Learn Res*. 2014;15(1):1929–58.
43. M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard et al., “Tensorflow: a system for large-scale machine learning,” in *12th {USENIX} symposium on operating systems design and implementation ({OSDI} 16)*, 2016, pp. 265–283.
44. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. Smote: synthetic minority over-sampling technique. *J Artif Intell Res*. 2002;16:321–57.
45. Baumann J, Gassmann K, Masjosthusmann S, DeBoer D, Bendt F, Giersiefer S, et al. Comparative human and rat neurospheres reveal species differences in chemical effects on neurodevelopmental key events. *Arch Toxicol*. 2016;90(6):1415–27.
46. Aschner M, Ceccatelli S, Daneshian M, Fritsche E, Hasiwa N, Hartung T, et al. Reference compounds for alternative test methods to indicate developmental neurotoxicity (DNT) potential of chemicals: example lists and criteria for their selection and use. *ALTEX*. 2017;34(1):49–74.

47. Fritsche E, Haarmann-Stemmann T, Kapr J, Galanjuk S, Hartmann J, Mertens PR, et al. Stem cells for next level toxicity testing in the 21st century. *Small*. 2020;17(15):2006252.
48. Benning L, Peintner A, Finkenzeller G, Peintner L. Automated spheroid generation, drug application and efficacy screening using a deep learning classification: a feasibility study. *Sci Rep*. 2020;10(1):1–11.
49. dos Reis FJC, Lynn S, Ali HR, Eccles D, Hanby A, Provenzano E, et al. Crowdsourcing the general public for large scale molecular pathology studies in cancer. *EBioMedicine*. 2015;2(7):681–9.
50. Amgad M, Elfandy H, Hussein H, Attaya LA, Elsebaie MA, Abo Elnasr LS, et al. Structured crowdsourcing enables convolutional segmentation of histology images. *Bioinformatics*. 2019;35(18):3461–7.
51. Northcutt CG, Jiang L, Chuang IL. Confident learning: estimating uncertainty in dataset labels. *J Artif Intell Res*. 2021;70: 1373–411.
52. Fei-Fei L, Fergus R, Perona P. One-shot learning of object categories. *IEEE Trans Pattern Anal Mach Intell*. 2006;28(4):594–611.
53. G. Koch, R. Zemel, and R. Salakhutdinov, “Siamese neural networks for one-shot image recognition,” in *ICML Deep Learning Workshop*, vol. 2. Lille, 2015.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Förster N, Butke J, Keßel HE, Bendt F, Pahl M, Li L, et al. Reliable identification and quantification of neural cells in microscopic images of neurospheres. *Cytometry*. 2022;101:411–22. <https://doi.org/10.1002/cyto.a.24514>

Reliable identification and quantification of neural cells in microscopic images of neurospheres

Nils Förster, Joshua Butke, **Hagen Eike Keßel**, Farina Bendt, Melanie Pahl, Lu Li, Xiaohui Fan, Ping-chung Leung, Jödis Klose, Stefan Masjosthumann, Ellen Fritsche, Axel Mosig

Journal: Cytometry part A

Impact factor: 4.335 (2022-2023)

Contribution to the publication: 15%

Preparatory work, image analysis software programming,
consultation on biostatistics, writing of the manuscript

Type of authorship: Co-authorship

Status of publication: Published 7th Nov 2021

2.2 Biostatistics and its impact on hazard characterization using in vitro developmental neurotoxicity assays

Hagen Eike Keßel, Stefan Masjosthusmann, Kristina Bartmann, Jonathan Blum, Arif Dönmez, Nils Förster, Jördis Klose, Axel Mosig, Melanie Pahl, Marcel Leist, Martin Scholze, Ellen Fritsche

ALTEX

Im Forschungsfeld der Gefährdungsbeurteilung von Substanzen sind sogenannte „Benchmark concentrations“ (BMC) und deren Unsicherheit von besonderem Interesse für regulatorische Entscheidungen. Zur Ermittlung eines BMCs müssen mehrere statistische Entscheidungen getroffen werden, welche stark von Faktoren wie etwa dem experimentellen Design und Eigenschaften der erhobenen Endpunkte abhängen. In der aktuell gängigen Praxis ist für die Datenauswertung oft der Experimentator verantwortlich, welcher dementsprechend auf statistische Software angewiesen ist. Dabei besteht oft die Gefahr, dass der Experimentator sich nicht über die gegebenen Standardeinstellungen der Software und deren Konsequenz für die Datenauswertung bewusst ist. Um ein besseres Verständnis dafür zu schaffen, wie sich statistische Entscheidungen auf die Datenauswertung auswirken, haben wir Fallstudien auf einen großen Datensatz angewandt, welcher durch Entwicklungsneurotoxizität-Testbatterien produziert wurde. Wir betrachten dabei auf die Ermittlung von BMCs, deren Unsicherheit, sowie resultierende Gefährdungsklassifizierungen. Hier konnten wir fünf kritische statistische Entscheidungen identifizieren, mit welchen sich der Experimentator während der Datenauswertung auseinander setzen muss: i) Wahl der Mittelung von biologischen Replikaten, ii) Datennormalisierung, iii) Regressionsmodellen, iv) Methode der BMC-Ermittlung, sowie v) die Wahl sogenannter „Benchmark response levels“ (BMR). Eine besondere Stärke unserer Datenauswertungssoftware ist dabei die Integration von Endpunkt-spezifischen Gefährdungsklassifikationen, einschließlich eines Warnsystems für unsichere Fälle, was bisher keine andere vergleichbare Software mitbringt. Die in dieser Studie gewonnen Einsichten demonstrieren, wie wichtig geeignete, aufeinander abgestimmte und regulatorisch akzeptierbare Methoden der Datenauswertung für objektive Gefährdungsbeurteilung sind.

Biostatistics and its impact on hazard characterization using in vitro developmental neurotoxicity assays

Hagen Eike Keßel¹, Stefan Masjosthusmann¹, Kristina Bartmann¹, Jonathan Blum², Arif Dönmez¹, Nils Förster⁴, Jödis Klose¹, Axel Mosig⁴, Melanie Pahl¹, Marcel Leist², Martin Scholze^{5}, Ellen Fritsche^{1,3*}*

¹IUF - Leibniz Research Institute for Environmental Medicine, 40225 Düsseldorf, Germany

²In vitro Toxicology and Biomedicine, Dept inaugurated by the Doerenkamp-Zbinden foundation, University of Konstanz, 78457 Konstanz, Germany

³Medical Faculty, Heinrich-Heine-University, Düsseldorf, Germany

⁴Bioinformatics Group, Ruhr University Bochum, 44801 Bochum, Germany

⁵Brunell University, London, UK

*authors contributed equally

Abstract

In the field of hazard assessment, Benchmark concentrations (BMC) and their associated uncertainty are of particular interest for regulatory decision making. The BMC estimation consists of various statistical decisions to be made, which depend largely on factors such as experimental design and assay endpoint features. In current data practice, the experimenter is often responsible for the data analysis and therefore relies on statistical software without being aware about the software default settings and how they can impact the outputs of data analysis. To provide more insight into how statistical decision making can influence the outcomes of data analysis and interpretation, we have used case studies on a large dataset produced by a developmental neurotoxicity (DNT) in vitro battery (DNT IVB). Here we focused on the BMC and its confidence interval (CI) estimation, as well as on the final hazard classification. We identified five crucial statistical decisions experimenter have to face during data analysis: choice of replicate averaging, response data normalization, regression modelling, BMC and CI estimation, as well as choice of benchmark response levels. In addition, the strength of our data evaluation platform is the integration of endpoint-specific hazard classifications, including flagging systems for uncertain cases, which none of the so far existing statistical data analysis platforms provide. The insights gained in this study demonstrate how important fit-for-purpose, internationally harmonized and accepted data evaluation and analysis procedures are for an objective hazard classification.

Keywords

biostatistics, benchmark concentration, confidence interval, hazard characterization, DNT

Corresponding Author

Hagen Eike Keßel, IUF - Leibniz Research Institute for Environmental Medicine, Düsseldorf, GERMANY

Suggested Reviewers

Katie Paul-Friedman (US-EPA; Paul-Friedman.Katie@epa.gov)

Timothy Shafer (US-EPA; Shafer.Tim@epa.gov)

Philip DiSalvo (Carle Foundation Hospital, Urbana, IL; disalvo@gmail.com)

1 Introduction

In 2007, the National Research Council (NRC) of the United States proposed a new strategy for toxicity testing in the 21st century centering around a shift from *in vivo* experiments in animals to mechanism-based *in vitro* testing (NRC, 2007). Since then, major advances in the field of in vitro toxicology have been made, including development and establishment of medium and high throughput screening (HTS) assays, as well as bioinformatics tools for data generation, management and analysis (Leist et al., 2014; Wheeler et al., 2015; Villeneuve et al., 2019). These efforts are contributing to next generation risk assessment (NGRA), which aims at using new approach methods (NAMs) for exposure-based, hypothesis-driven risk assessment without the generation of new animal data (Li et al. 2021; Dent et al. 2021; Palloca et al. 2022).

Typically, an in vitro HTS test system produces hazard data for a relatively large number of test concentrations and thus makes it most suitable for concentration-response regression modelling. This statistical approach allows the interpolative estimation of a concentration value at a given effect level (effect or inhibitory concentration), and of particular regulatory interest is hereby the benchmark concentration (BMC) and its associated uncertainty, expressed as lower limit of a one-sided 95% confidence interval (BLL). A BMC is considered as lowest concentration of the test compound that produces a pre-defined small “relevant” change to the control reference’s response level, and as consequence, the benchmark response (BMR) value should be as “close as possible” to the control response.

In vitro test systems represent a huge variety of different types of assays, from cell-free, cell and tissue-based methods up to multi-response organoid systems, and as consequence, concentration-response data between these systems vary enormously with respect to their test-specific experimental designs, data variability, dynamic ranges and concentration-response pattern. Unique to HTS systems is also

that assay outputs are produced in microplate multi-well readers, with concentration-response data from the same concentration and experiment are considered to reflect technical (intra-replicate) variation and data from repeated experiments more indicative for “biological” (between-study) variation. These hierarchical data are usually simplified by using an average response value per test concentration and experiment (replicate average) as statistical unit for the concentration-response analysis, with the argument that the BMC and BLL estimation should reflect mainly biological and between-study variability.

The BMC estimation consists of various statistical decisions to be made in the concentration-response analysis, which dependent largely on the experimental design, the concentration-response data and assay endpoint features, and which require statistical knowledge that is usually only warranted by experienced biostatisticians. In current data practice, the experimenter is often responsible for the data analysis and therefore relies on statistical software without being aware about the software default settings and how they can impact the outputs of data analysis (Jensen *et al.*, 2020). Existing guidelines for concentration response data analysis are often too general (OECD, 2006; EFSA, 2016), and no clear consensus on a common and standardized biostatistical method for *in vitro* toxicity data have been achieved (Wheeler *et al.*, 2015; Sand *et al.*, 2017).

To provide more insight into how statistical decision making can influence the outcomes of data analysis and interpretation, we have used case studies on a large dataset produced by a developmental neurotoxicity (DNT) *in vitro* battery (DNT IVB; Masjosthusmann *et al.* 2020, Crofton and Mundy 2021). In this DNT IVB, 148 compounds were tested across up to ten test methods representing the neurodevelopmental key events (KE) of neural progenitor cell (NPC) proliferation, migration of neural crest and radial glia cells, neurons and oligodendrocytes, neuronal differentiation, neurite outgrowth of peripheral and central nervous system neurons, as well as oligodendrocyte differentiation, and accomplished by various endpoints measuring cell viability and cytotoxicity (Masjosthusmann *et al.* 2020). Some of the DNT-specific endpoints are derived from primary and organotypic cultures, and thus more prone to a data variability typically observed in animal studies. Here we focused on the BMC and its confidence interval (CI) estimation, as well as the final hazard classification. For this purpose, we identified five crucial statistical decisions the experimenter have to face during the data analysis (Figure 1):

- (i) Statistical unit: shall the median of all replicate responses of an experiment be used, which makes no assumption to the data and thus reduces the negative impact of potential data outlier, or the replicate mean, which has a higher certainty but assumes a symmetric distribution of the replicate responses, and if violated, can lead to a biased estimation of the replicate mean?

- (ii) Response data normalization: shall the responses of an experiment always be normalized to the control's response even if the exposure responses provide clear evidence against the use of control data, or shall in that case the "control reference" be estimated directly from the responses of the exposures ("re-normalization", Krebs et al., 2018)?
- (iii) Regression model: shall the concentration-response data always be described by the same and supposedly flexible mathematical model, or is it better to use several models and either subsequently select the best model by means of goodness-of-fit criteria ("best-fit method", Scholze *et al.*, 2001) or estimate an average of all model fits ("model averaging", Claeskens et al., 2008)?
- (iv) Uncertainty of a BMC estimation: shall the confidence level of a BMC be calculated by a simple and commonly used statistical approximation technique ("Delta method", Cox, 1990) which is known to be inaccurate (Moerbeek, Piersma and Slob, 2004), or by alternative approaches such as bootstrapping or inverse regression (Jensen, Kluxen and Ritz, 2019)?
- (v) Benchmark response (BMR): shall a response level most close to the control reference be selected, which might not always be applicable for the statistical concentration-response analysis and thus might fail to provide a reliable BMC estimation, or a higher BMR, which guarantees a statistically more robust BMC estimation but might fail for compounds that has produced weak responses below the intended BMR?

We designed a standard data evaluation protocol ("standard protocol") which we used as reference to alternative statistical methods, so that their BMCs and confidence intervals (CIs) estimated to the same DNT IVB data could be compared. The statistical methods to be changed were chosen along the questions outlined in (i) to (iv). This was supplemented by measuring their impact on hazard alerts derived from hit classifications, which separate cytotoxic concentration ranges from the respective BMC of the specific DNT endpoint, and by measuring their impact on the DNT IVB's capability on predicting DNT adversity in terms of specificity, sensitivity and accuracy.

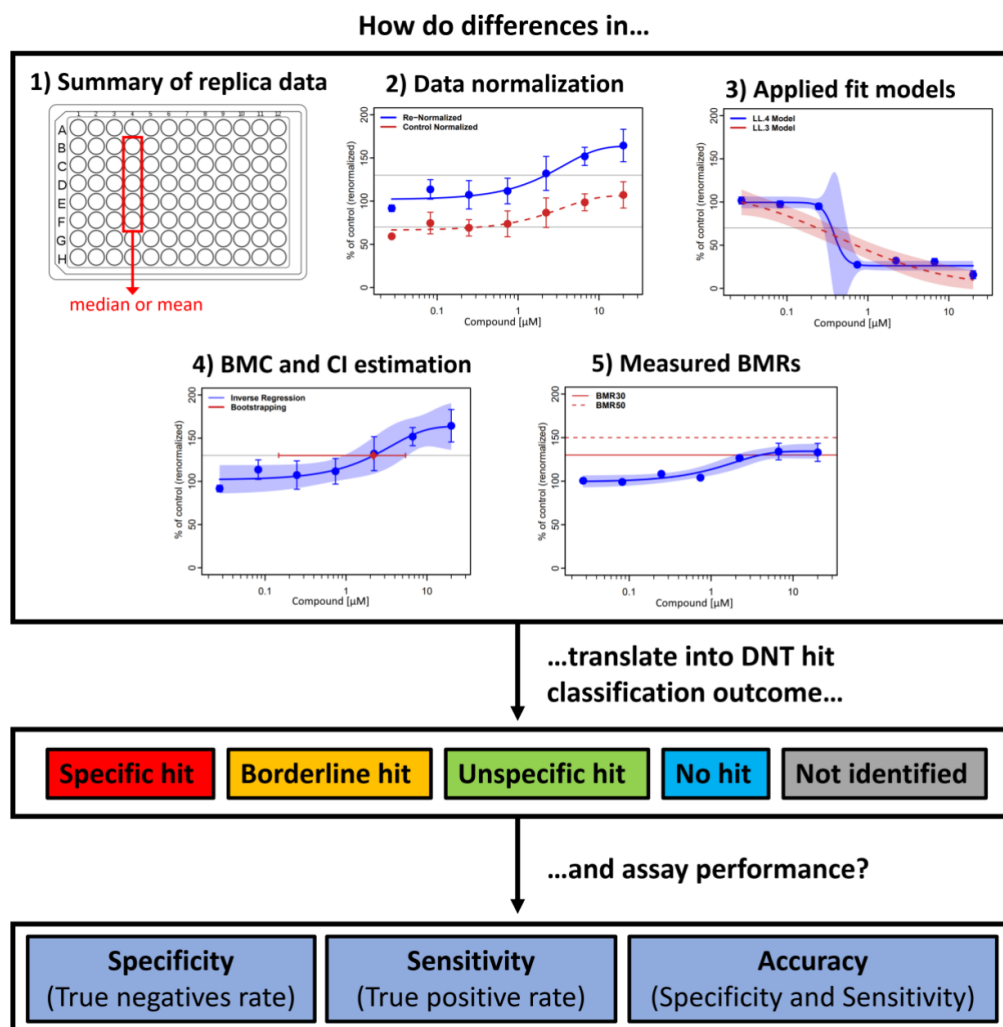


Figure 1: Study Overview

Several biostatistical data analysis and evaluation steps were analysed for their impact on a BMC estimation and subsequent hazard characterization from developmental neurotoxicity (DNT) data: i) how to average replicate responses from an experiment, ii) how to normalize concentration-response data, iii) how to describe concentration-response data by regression modelling, iv) how to estimate a benchmark concentration (BMC) and its uncertainty, and v) which benchmark response (BMR) level to select. Changes between statistical methods were recorded for 148 compounds tested on up to 22 assay endpoints, and their impact translated into the compound's DNT hit classification and the predictivity performance of the overall assay battery.

2 Methods

2.1 DNT data

All concentration response data used in this study are from a DNT *in vitro* battery of 8 assays with 22 endpoints, in which a total of 148 compounds were tested. 120 compounds were tested across all assays, while 28 compounds were tested in at least 2 assays. Fourteen assay endpoints represent major key neurodevelopmental processes, and 8 endpoints measure general cell viability and cytotoxicity (Table 1). This DNT *in vitro* battery was developed in collaboration with EFSA with the aim to advance the application of *in vitro* DNT testing for regulatory purposes. The term “BMC” was used equally for data from DNT-specific, cytotoxicity and viability endpoints.

Depending on the assay, fluorescent readouts using a multiplate reader or fluorescence and brightfield imaging with subsequent artificial intelligence-based image analysis (Schmuck et al., 2015; Förster et al., 2021) was performed as endpoint assessment. Each compound was tested in at least three independent experiments and eight concentrations per experiment, with 5-6 controls and 5-6 replicates per experiment. An overview of the assays, the cell model and the respective endpoints is given in Table 1, and more detailed information about the assay-specific experimental testing procedures and test outcomes is provided in Masjosthusmann *et al.* (2020).

The BMR for each endpoint was derived from the between experimental variability as the coefficient of variation of median plate medians (after normalization) measured at the lowest test concentration and across all independent experiments (Masjosthusmann *et al.*, 2020). To achieve a better comparability across the endpoints, the BMRs were then rounded to the next higher value, resulting into three BMRs: a 10% change was selected for endpoints from the NPC2a and NPC1-5 cytotoxicity assay (BMR10), and a 30% change for endpoints from the NPC1, NPC2a, NPC2b, NPC3-5 and NPC1-5 viability assay (BMR30). For all UKN assays a 25% change was decided (BMR25), and for the viability of the UKN2 a BMR10 was chosen. NPC Assays were conducted with three to five independent experiments and 5 replicates each, UKN Assays with three independent experiments and 6 replicates each (Table 1).

Table 1: Test Assays

An overview over the assays and their key characteristics, including the cell model, assays endpoint and chemical exposure time (in brackets), the BMR that was used for the classification, the number of independent experiments as well as the replicates per experiment. Unspecific endpoints (cell toxicity and viability) are highlighted in *cursive*. Cytotoxicity and cell viability assay endpoints were used as reference for NPC2, NPC3 and NPC5.

Assay	Cell model	Endpoint [exposure time]	BMR [%]	Independent experiments	#Replicate experiment	# of Compounds ²
NPC1	neural progenitor cells	proliferation NPC1 [72h]	30	3 to 5	5	123
		proliferation by area [72h]	30			117
		<i>viability NPC1 [72h]</i>	30			123
		<i>cytotoxicity NPC1 [72h]</i>	10			115
NPC2	neural progenitor cells	migration distance radial glia NPC2a [72h]	10	3 to 5	5	123
		migration distance radial glia NPC2a [120h]	10			
		migration distance neurons NPC2b [120h]	30			
		migration distance oligodendrocytes NPC2c [120h]	30			
NPC3	neural progenitor cells	neuronal differentiation NPC3 [120h]	30	3 to 5	5	123
		neurite length NPC4 [120h]	30			
		neurite area NPC4 [120h]	30			
NPC5	neural progenitor cells	oligodendrocyte differentiation NPC5 [120h]	30	3 to 5	5	123
NPC2-5 ¹	neural progenitor cells	<i>cytotoxicity NPC2-5 [72h]</i>	10	3 to 5	5	123
		<i>cell number [120h]</i>	30			123
		<i>cytotoxicity NPC2-5 [120h]</i>	10			122
		<i>viability NPC2-5 [120h]</i>	30			123
UKN2	hiPSC-derived neural crest cells	Migration UKN2 [24h]	25	3	6	70 ³
		<i>Viability UKN2 [24h]</i>	10			
UKN4	Luhmes cells	Neurite Area UKN4 [24h]	25	3	6	75 ³
		<i>Viability UKN4 [24h]</i>	25			
UKN5	hiPSC-derived sensory neurons	Neurite Area UKN5 [24h]	25	3	6	71 ³
		<i>Viability UKN5 [24h]</i>	25			

¹Cytotoxicity and cell viability assay endpoints were used as classification reference for NPC2, NPC3 and NPC5.

²Shown here is the number of compounds that have concentration response information with at least 5 concentrations

³A total of 140 compounds were tested in these assays. However, some compounds were only tested in a high concentration. If the high concentration was negative in the assay, no concentrations response testing was performed.

2.2 Data Evaluation Platform

For data processing and evaluation, the R package *drc* (R Core Team 2019, Ritz et al. 2019) was extended and optimized for the use of data from multi-well plate experiments. The biostatistical data evaluation software is freely available as open source under the name CRStats (github.com/ArifDoenmez/CRStats), an interactive R Markdown document is available and can freely be assessed for use. All for the comparative study relevant mayor modules are displayed as workflow diagram in Figure 2, starting from minimal data requirements for a BMC estimation (2.2.1) up to the endpoint-specific hazard classification module (2.2.8). The individual modules are explained in more detail below, with the module number in Figure 2 referring to the number of the subsection. From

module 2.2.3 onwards we defined a standard protocol for the evaluation of DNT IVB data, with the following statistical methods chosen: (i) average replicate per experiment estimated by median (2.2.3), (ii) control-normalization followed by re-normalization (2.2.4), (iii) application of several mathematical models to find the ‘best fit’ regression model for a BMC estimation (2.2.6), (iv) CI estimation of the BMC by inverse regression (2.2.7), and (v) selecting endpoint specific BMRs for the hazard classification as outlined in Table 1. This standard setup is shown on the left side of the modules (blue), and all alternative methods that we considered in this study are listed on the right side under “changes” (orange).

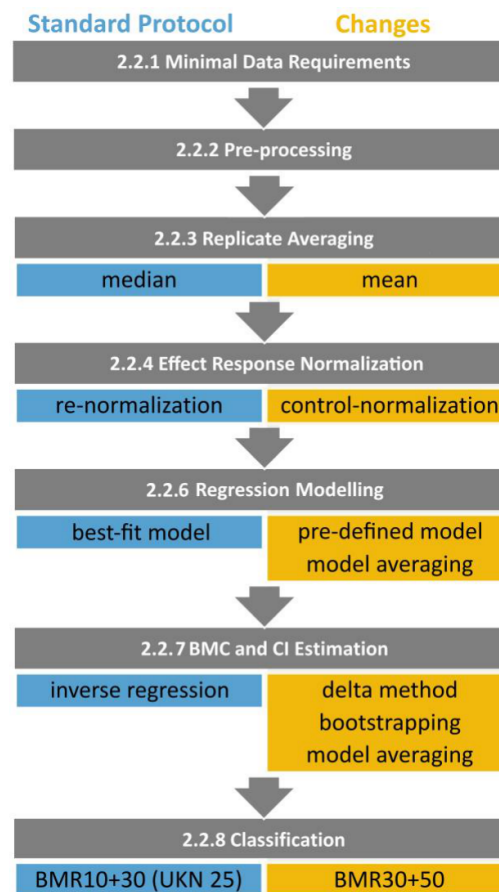


Figure 2: R Evaluation Pipeline Workflow

The R workflow is depicted with subsequent data evaluation steps from top to bottom. Grey boxes indicated the mayor data processing and evaluation steps with reference to the material and methods section in which they are described. Mayor key methods are depicted as coloured boxes below the according step. The blue methods (left) are the ones used for the standard protocol, while changes of one of the standard protocol methods (depicted in orange, right) are used to create the alternative protocols (i.e. for one alternative protocol the statistical unit for replicate averaging was changed from median to mean, while all other key methods remained the same as in the standard protocol, so that only the impact of this particular change can be monitored).

2.2.1 Minimal data requirements

Data were accepted for data analysis only if the following three minimal data requirements were fulfilled: (i) at least two replicas per concentration are available, otherwise all readouts from this concentration were excluded, (ii) at least five concentrations per experiment have provided readouts otherwise the whole experiment was excluded, (iii) at least two control readouts are available otherwise the whole experiment was excluded.

2.2.2 Pre-processing

CRSTATS uses different assay-specific pre-processing steps in order to obtain a single response value for each well. For example, the neuronal differentiation in the NPC3 assay is calculated as the number of neurons divided by the total number of cells with a nucleus:

$$(1) \quad \text{NPC3 neuronal differentiation [120h]} = \frac{\text{NPC3 number neurons [120h]}}{\text{NPC3 number cells [120h]}}$$

All assay specific pre-processing methods that are currently implemented in CRSTATS are listed in Table S1.

2.2.3 Replicate averaging

The average assay response for controls and treatments from the same experiment was either estimated by the arithmetic mean or by the median. The variability between replicates was calculated as standard deviation (SD; for the mean replicate) or as median absolute deviation (MAD; for the median replicate). Outlier detection procedures were not applied and data points from wells where technical problems were known or obvious were excluded from the data analysis.

2.2.4 Effect data normalization

CRSTATS offers different normalization methods which allows the translation of pre-processed effect data into relative values. For this study, we used the following two methods:

- (i) Control normalization: effect responses are normalized to the mean or median of the solvent controls as

$$(2) \quad \frac{\text{replicate response}}{\text{median or mean (solvent control responses)}}$$

- (ii) Control re-normalization: normalized effect responses (Equation 2) are further normalized by a mean value that has been estimated by regression modelling at the lowest test concentration, i.e.

$$(3) \quad \frac{\text{normalised replicate response}}{\text{model estimate of normalised response at lowest test concentration}} \quad (\text{Krebs et al., 2018}).$$

2.2.5 Significance analysis

The presence of at least one exposure concentration that had produced an effect response which differs statistically significantly from the responses of all remaining exposures is a crucial factor in the hazard classification method (2.2.8). To account for that, significant differences between treatment means were identified by using the Tukey Honest Significant Differences test ($\alpha=5\%$, two-sided) (Tukey HSD; Yandell, 1997), with hypothesis testing conducted on normalized replicate averages from at least three independent experiments. As an average control value was always set to 100% (2.2.4), controls were excluded from the significance analysis. Data provided no evidence against the Gaussian assumption.

2.2.6 Concentration-response regression analysis

The R packages *drc* (Ritz *et al.*, 2015) and *bmd* (Jensen *et al.*, 2020) were used for regression analysis and the estimation of a BMC and its associated uncertainty. The *drm* function fits a pre-defined regression model to the concentration-response data, with several options implemented to provide more flexibility for the estimation method. A large number of mathematical nonlinear regression functions was applied to the same data set (Table S2), and the best fitting model then selected on basis of the Akaike's Information Criterion (AIC) ("best fit method", Scholze *et al.*, 2009; Portet, 2020). AIC is commonly used to compare the relative goodness-of-fit among different models and to then choose the model of best predictive power by balancing data support against model complexity. As all effect endpoints in this study are continuous, the estimation method of ordinary least-squares (OLS) was used. OLS relies on two assumptions, i.e. (i) effect data (here replicate average) follow a symmetrical distribution, and (ii) variance homogeneity across all treatment groups. Both assumptions were checked prior to data analysis on basis of pooled endpoint-specific data from all experiments: data variability differed in average by maximally 20% between the treatment groups, with the highest variability often occurring at highest test concentration, and no overall clear evidence was detected that normalized replicate means did not follow a symmetric distribution. These findings were deemed as acceptable for using the unweighted OLS regression analysis.

2.2.7 BMC and its Uncertainty

In the standard protocol the BMC was estimated directly from the best fit model. We also considered model averaging as an alternative option where, similar to the previous best fitting method, a number of suitable concentration-response models were fitted to the same data but in this case all resulting model fits were combined to provide an weighted average BMC estimates (Ritz, Gerhard and Hothorn, 2013). Uncertainty was always expressed as $\alpha=5\%$, i.e. the lower limit (BLL) corresponds to the 2.5% limit and upper limit (BUL) to the 97.5% limit. BLL and BUL were derived by three different methods,

i.e. inverse regression, the delta method and bootstrapping. The estimation of the BMC and its 95% CI by model averaging was always performed in combination with bootstrapping. Inverse regression estimates both BLL and BUL directly from the regression fit around the BMC (Buckley, Piegorsch & West, 2009; Fang, Piegorsch & Barnes, 2015) and therefore puts high emphasis on a successful regression fit in terms of robustness and reliability. The delta method is an asymptotic approach which combines information of the estimated model parameters to derive a Wald-type interval (Jensen et al. 2020). Bootstrapping uses computer-intensive simulation techniques that resamples the original dataset to create a huge number of so-called bootstrap samples, with each sample mirroring the original data set with an identical experimental design but newly simulated effect responses. On each bootstrap sample the same statistical data analysis was performed, resulting into a distribution of resampled BMC values around the original BMC estimation. If the median of this distribution equals the original BMC (unbiased resampling), then the 2.5% and 97.5% quantiles are expected to mirror the BUL and BLL of the original BMC, respectively. For each bootstrap sample, always the same regression model was used as part of the best-fit method, or one model-averaged BMC if model averaging was performed. To simplify the model averaging method, only three regression models were considered (four-parameter loglogistic, four-parameter Weibull and three-parameter exponential model). Bootstrapping was always conducted on 1000 resampled datasets, and due to the small sample sizes, we used always the parametric version (Efron, Bradley and Tibshirani, 1993). All resampling was performed by the function *bmdMA* of the R package *bmd* (Jensen *et al.*, 2020). Bootstrapping can simulate a bootstrap sample which do not allow a BMC estimation or which leads to an unreliable BMC estimation that is well outside the tested concentration range. Therefore, a resampled BMC was excluded from the resampling distribution if it was 1.5-times above the highest test concentration or below the lowest tested concentration.

2.2.8 Hazard Classification

CRSTATS uses a hazard classification approach which judges if data evidence is sufficient to define a compound as active for the specific DNT endpoint and if this can be distinguished from an activity observed in cell health related endpoints (viability and cytotoxicity). Accordingly, the endpoint-specific hazard of a compound is classified into five categories:

- **No hit:** no observed effect on the DNT-specific endpoint or on general cell health.
- **Unspecific hit:** the effect on the DNT-specific endpoint cannot be separated from an effect on the cell health related endpoint.
- **Borderline hit:** the separation between the effects on the DNT-specific endpoint and the effect on cell health related endpoint is statistically not clear (Leontaridou et al, 2017).

- **Specific hit:** the effect on the DNT-specific endpoint is clearly separated from an effect on the cell health related endpoint.
- **Not identified:** data are incomplete und do not allow any classification.

If the automatic classification failed due to a high uncertainty of the BMC or a missing BMC for the cell health related endpoint, the classification was recorded as **expert judgement** and classification into one of these five categories was done by manual inspection on the basis of all data evidence. An overview over all flagging alerts leading to expert judgement are given in Table S4.

The hazard classification approach was operationalized by hazard decision trees which reflect specific assay features and the directionality of the observed concentration response pattern (i.e. either reduction or inhibition). Common to all decisions trees is that they compare the BMC of the DNT-specific endpoints to the respective BMC of the unspecific endpoint (i.e., cytotoxicity or cell viability). For the NPC and UKN assays slightly different versions were developed, with all NPC assay endpoints accounting directly for the statistical uncertainties of both BMC estimations by using their corresponding CIs, and all UKN assay endpoints using pe-defined acceptance ranges instead. The principles of the hazard decision tree for data sets with decreasing concentration-response pattern (reduction) measured in NPC assays (NPC1, NPC2, NPC3 and NPC5, Table 1) are shown in Figure 3, and for increasing concentration-response pattern (induction) in Figure 4. Inductions are handled separately, because the specific and unspecific endpoints do not have the same relationship during an induction, compared to a reduction in the endpoint. A loss in general cell viability for example will likely result in an effect in cell proliferation, while an induction in cell viability does not necessarily increase cell proliferation. If migration (NPC2a) is affected, only cytotoxicity is used as a reference for all specific endpoints of NPC2-5. A reduction in migration also reduces cell viability due to the lower number of cells in the migration area and not necessarily due to cell death. If so, it cannot be used as valid reference to discriminate between a specific and unspecific effect. The same applies to effects in cell viability. In these cases, only cytotoxicity is used as general cell health reference for according specific NPC endpoints. More details can be found in the supplementary material (S1.3) and in table S3, and details about the classification tree applied to data from the UKN assays can be found in Masjosthusmann et al. (2020).

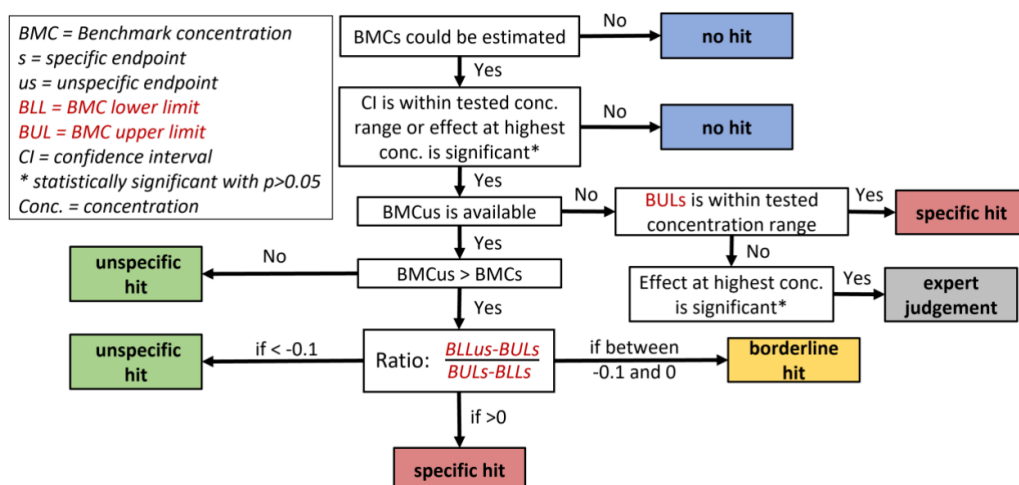


Figure 3: Decision tree for the NPC hazard classification of inhibitory effects

The decision tree shows for NPC1-5 data with decreasing concentration-response pattern how BMC estimations and their uncertainty (expressed as 95% confidence intervals, CI) for data from both specific and unspecific endpoints are used to classify the compound into one of the DNT hit categories (coloured boxes). Hits with the category "expert judgement" (grey box) will be classified into one of the DNT hit categories by manual inspection on the basis of all data evidence.

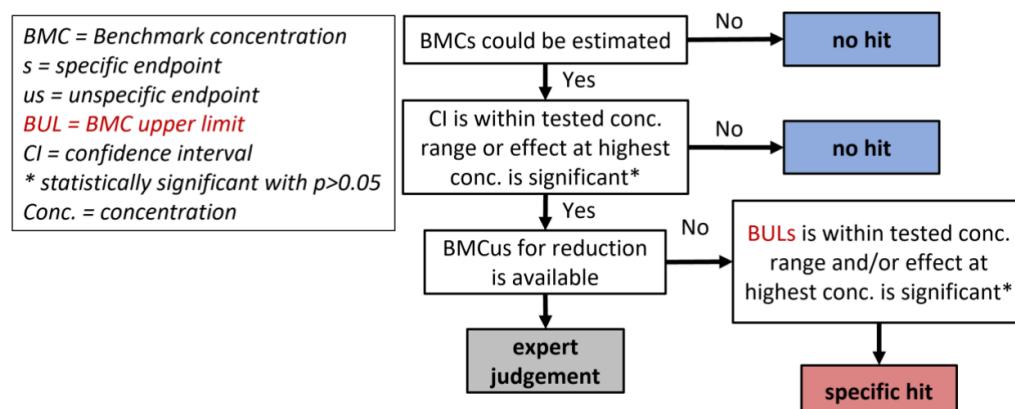


Figure 4: Decision tree for NPC hazard classification of increasing effects

The decision tree shows for NPC1-5 data with increasing concentration-response pattern ("induction") how BMC estimations and their uncertainty (expressed as 95% confidence intervals, CI) for data from both specific and unspecific endpoints are used to classify the compound into a specific or no hit (coloured boxes). The presence of a cytotoxic responses can lead to an artefact in the DNT-specific endpoint and is therefore initially categorized as "expert judgement". These hits will be classified into one of the DNT hit categories by manual inspection on the basis of all data evidence.

2.3 Assay Performance

From the 148 compounds tested in the DNT IVB, a set of 45 reference compounds (17 negative compounds that are known not to cause DNT; 28 positive compounds with proven DNT adversity in humans or mammals) was used for an evaluation of the DNY IVB predictivity. Hit decisions were

derived from the hazard decision trees developed in 2.2.8, and the following performance parameters were used:

$$\text{Specificity} = \frac{\# \text{ true negative hits}}{\# \text{ negative compounds}}$$

$$\text{Sensitivity} = \frac{\# \text{ true positive hits}}{\# \text{ positive compounds}}$$

$$\text{Accuracy} = \frac{\# \text{ true negative hits} + \# \text{ true positive hits}}{\# \text{ negative compounds} + \# \text{ positive compounds}}$$

A negative compound was considered as true negative if it was not classified as specific hit or borderline in any of the assays. A positive compound was considered as true positive, if it was classified as specific hit or borderline in at least one assay.

3 Results

To perform a robust, fast and automated hazard characterization based on high content *in vitro* toxicity testing data, we have set up the R-based data evaluation pipeline CRSTATS (github.com/ArifDoenmez/CRStats), which offers multiple statistical options for the data evaluation of continuous concentration-response data. Based on these options, we have defined a standard data evaluation protocol ("standard protocol") for DNT IVB data (Fig. 2), and by changing statistical methods as part of the protocol we studied their impact on the BMC estimation of the DNT IVB outcomes and the subsequent consequences for the hazard classification and overall DNT IVB performance ("alternative protocol"). The following statistical methods were chosen as alternatives to the standard protocol: 1) average replicate per experiment estimated by the arithmetic mean, 2) control normalization without re-normalization, 3) using a three-parameter log-logistic regression model (LL3rm) for the BMC estimation, 4) using model-averaging for the BMC estimation, 5) CI estimation of the BMC by the delta method, 6) CI estimation of the BMC by bootstrapping, 7) CI estimation of the BMC by model averaging, and 8) increasing the endpoint specific BMRs by 20%. Differences in the BMC estimation, the uncertainty of a BMC (expressed as the width of the central 95% confidence interval of a BMC estimation), the endpoint-specific hazard classification of the compound and the final assay performance were quantified and compared across the various specific assay endpoints.

In total, 148 compounds were tested on up to 14 DNT-specific and 8 cytotoxicity and viability endpoints, of which 2385 data sets fulfilled the minimal data requirements of the data evaluation pipeline. According to the standard protocol, it was possible to perform a regression analysis for 2385 data sets (1953 NPC and 432 UKN) and a hazard hit categorization for 1563 data sets from DNT-specific endpoints (1347 NPC and 216 UKN). In nearly one third of all best-fit model decisions the simplest regression model was chosen, i.e. the exponential function with two model parameters, followed by three-parametric models (55.7%) and by four-parametric models (10.3%). Only in 1.9% of all best-fit model decisions sufficient data were available to support the most complex regression model (5-parameter general log-logistic).

3.1 Impact of different data evaluation methods on the BMC estimation

To allow a better comparison of BMCs from different data scenarios, the BMC was transformed to a relative BMC on a log₁₀ scale by relating the 100-fold BMC estimation to the highest test concentration of the data set:

$$relative\ BMC = \log_{10}\left(\frac{100 \cdot BMC}{highest\ test\ concentration}\right).$$

A relative BMC of 1 corresponds to a BMC that is tenfold below the highest test concentration of the data set, and a relative BMC above 2 corresponds to a BMC that has been extrapolated beyond the highest test concentration. The lower the relative BMC value, the more likely the estimation is supported by effect data from more concentrations.

The relative BMCs from the standard and alternative statistical protocols are shown in Figures 5 A-E for five statistical parameters that were changed, which the BMC of the alternative protocol always referring to the x-axis and the BMC of the standard protocol to the y axis. If a regression analysis could be performed but a BMC not established due to missing data support for the BMR, the BMC was flagged as “BMRnr” (BMR not reached) and included in the plot at the end of the BMR axes, i.e. a BMRnr value on the right side of the plot indicate a BMC estimation which was only possible for the standard protocol, and similarly, a BMC value on the top of the plot area indicate a BMC estimation that could only be established for the alternative protocol. Data sets for which none of the protocols were able to produce a BMC were excluded. Color-coded symbols refer to the 22 bioassay endpoints, and a data point on (or close to) the solid 45-degree line indicates a perfect agreement between the BMCs from both protocols. Three-fold BMC differences are highlighted by a belt around the line of perfect agreement (i.e. values outside of the belt have above three-fold change), and the percentage number of successful regression fits for the alternative protocol are included on top of each plot, with reference to the 1953 data sets for which a successful regression modelling was conducted according to the standard protocol. To identify general deviation patterns, we performed trend regression analyses between the relative BMCs, and the corresponding value of the goodness-of-fit criterion (R^2) is provided in the plot: the higher the coefficient, the more consistent the results between the two protocols. For the trend analysis, we set a relative BMC = 2.47 for a BMRnr, i.e. a 3-fold difference between the highest concentration and a fictional BMC was assumed. Not shown are BMC differences for the bootstrapping and delta method, as both refer to the same BMC and thus would have resulted always into identical BMCs in the plot.

We found the most profound BMC differences between the data re-normalization and control normalization (Fig. 5B), with an R^2 of 0.3. The main reason for the huge number of BMC disagreements is due to huge number of BMRnr's, i.e. regression fits that could establish a reliable BMC for the endpoint-specific BMR in only one of the protocols. Using the mean as replicate average instead of the median (Fig. 5A), using a predefined regression model (LL3rm) instead of the best fit method (Fig. 5C), and using a higher BMR resulted in moderate BMC changes, with R^2 's between 0.59-0.61. The best agreement between relative BMC values was observed for the comparison between the outcomes from model averaging against the best-fit method (Fig. 5D) with an R^2 of 0.85.

The number of datasets for which a regression model could be fitted for the alternative protocol was related to the number of fits for the standard protocol and expressed as relative “fit success rate”. All changes of statistical methods lead to similar success rates, with the exception of the sole application of the three-parameter log-logistic model which led to a noteworthy loss of successful regression fits (68.55% success rate).

To further explore differences between BMC estimates, the number of BMRnr cases that only occurred in the alternative protocol (i.e. the standard protocol did result in a BMC while the alternative protocol did not; Fig. 5F, blue shaded area of bar), the number of BMRnr cases that only turned out in the standard protocol (Fig. 5F, green shaded area of bar) and large differences outside the belt (“outliers”, Fig. 5F red shaded area of bar) were compared to the total number of BMCs that were estimated by the standard protocol. Most protocols that lead to less successful BMCs were caused by the inability of the data to support the regression modelling for the intended BMR level. All alternative protocols together led to less BMCs but more BMRnr cases, with protocol changes to control normalization and higher BMRs resulting into the highest increase towards BMRnr cases (i.e. less BMCs), with an increase of 17.25% and 37.98% of BMRnr cases, respectively. Taking only the cases with huge BMC differences into account (“outliers”), the number of BMCs that were either lost or gained due to the protocol change was further quantified: model averaging led to the smallest number of relevant changes (7.13%), followed by replicate averaging by mean, fixed regression model (LL3rm) and control-normalization with moderate changes (13.43%-28.86%), up to >40% changes were reached if a higher BMR was used.

Differences between the relative BMCs were also expressed as fold-change, and the distribution of all fold changes summarized as median and the interquartile ranges (IQR) (Figure 5G). Here, the BMRnr cases were excluded from the fold change analysis. In alignment to the previous results, the protocol change towards higher BMRs led to the most severe fold-change (median = 1.71, IQR = 1.02). The protocols with mean replicate average, control-normalized data, model averaging and choice of higher BMRs showed moderate median fold-changes of estimated BMCs ranging from 1.04-1.15 (IQR ranging from 0.09-0.3).

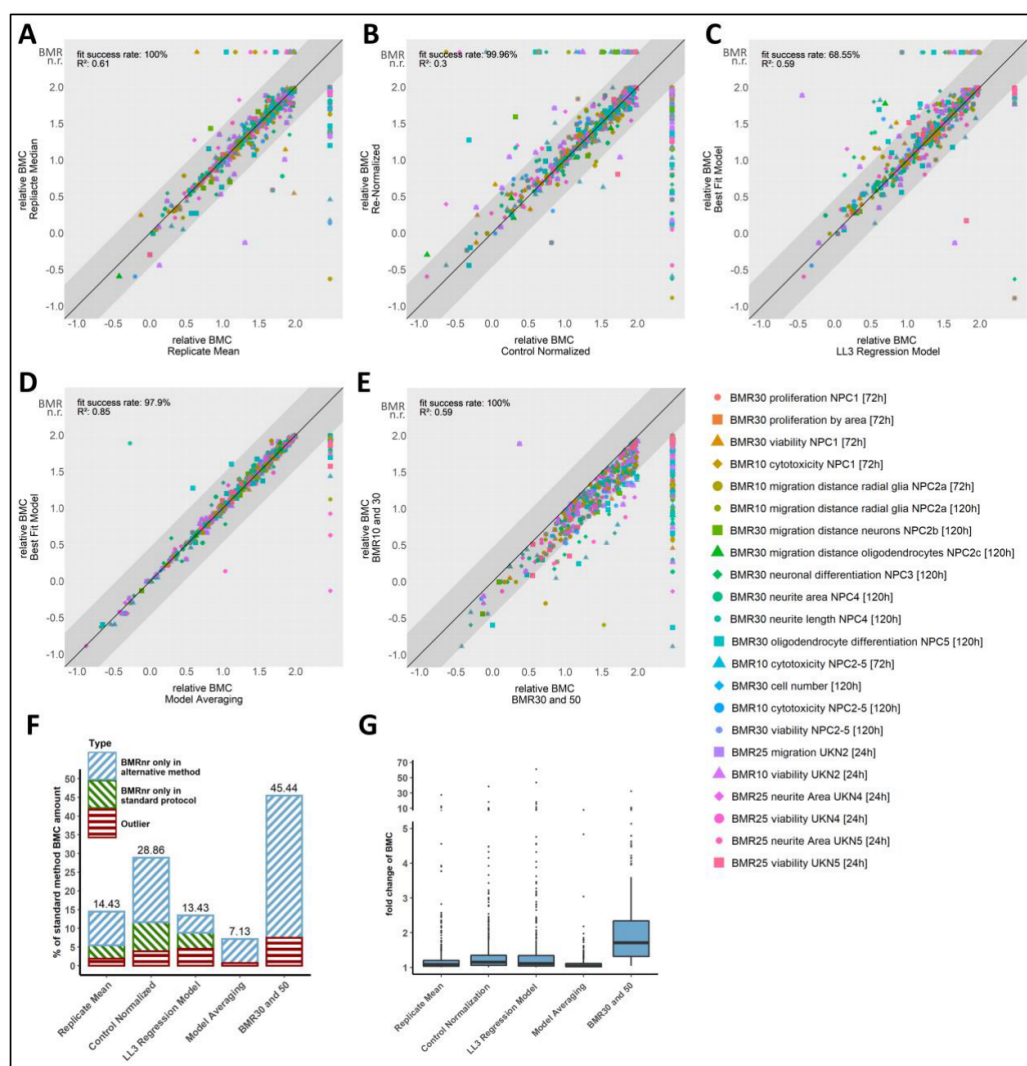


Figure 5: Impact of methodological changes in the data evaluation on the BMC estimation

BMCs for 148 compounds tested on up to 22 endpoints from 8 assays were estimated using the standard protocol and opposing alternative methods. A-E): A relative BMC was expressed as the log₁₀-transformed ratio between the 100-fold BMC and the maximum test concentration, and relative BMCs from all data sets and endpoints but different statistical methods were plotted against each other. The solid black trend line indicates no differences between the relative BMCs, the grey interval around the trend line indicates values within a three-fold range. Values outside this interval are considered as relevantly different between the opposing methods. If a relative BMC could be calculated for only one method, the missing value of the opposing method is plotted as BMRnr area on the right or upper side of the graph. Relative BMCs are colored according to their bioassay endpoint. To indicate the strength of agreement between both data evaluation protocol, the goodness-of-fit coefficient from a trend regression analysis between both relative BMCs is included (R²; top left), and the percentage of successfully applied regression models of the alternative protocol in relation to the standard protocol is shown top left ("fit success rate"). A) Experimental median replicates versus mean replicates (n = 568). B) Re-normalized data versus control-normalized data (n = 630). C) Best fit approach versus a predefined three parameter log-logistic regression model (n = 520). D) Inverse regression versus model averaging (n = 604). E) BMR10+30 (BMR10+25 for UKN) versus BMR30+50 (n = 604). F) Percentage of all data sets for which the protocol change lead to a BMC change in terms of BMRnr (i.e. a BMC could not be determined from the regression fit) or an above three-fold BMC change. G) Distribution of BMC fold-changes in response to statistical method changes from the standard protocol. Box whisker plot shows the median (horizontal line), interquartile range (box), 5% and 95% percentile values (whisker), and extreme values (black dots).

3.2 Impact of different data evaluation methods on the BMC uncertainty

Next, we analyzed how changes to the standard evaluation protocol can influence the overall uncertainty of the BMC estimation. The uncertainty of a BMC is estimated as central 95% confidence interval with the BLL corresponding to the lower 2.5% interval section and the BUL to the upper 97.5% interval section. The width of the interval (i.e. the difference between the BUL and the BLL) is an essential factor in some of the classification models of the hazard characterization. Similar to the analysis of BMC changes, a CI width was transformed to a relative CI width by fixing it to the maximal test concentration of the data set:

$$relative\ CI\ width = \log_{10}\left(\frac{100 \cdot CI\ width}{highest\ test\ concentration}\right),$$

with CI width = BUL – BLL, where BLL and BUL are the 2.5% and 97.5% confidence interval of the BMC estimation. A single relative CI width has no meaningful interpretation, and it was only used in combination with a second value with reference to the same highest test concentration.

Changes to the standard protocol which lead to changes in the relative CI width were visualized in the same way as in the previous section, with the relative CI widths from the standard and alternative protocol shown as endpoint-specific symbols for all data sets in a common scatter plot, with each plot referring to a specific method change (Fig. 6A-G). Values below the line of perfect agreement indicate an increase of the CI width, i.e. the BMC estimation of the alternative protocol is considered as more uncertain, and values above the line indicate a more certain BMC estimation than judged by the standard protocol. Also, a supporting trend analysis was conducted, with the corresponding goodness-of-fit criterion (R^2) provided in the plot, and a belt around the line of perfect agreement between both relative CI widths was included, with larger than three-fold changes outside this belt considered as relevant.

Outcomes of the trend analyses show that protocol changes due to the experimental mean replicate or the sole application of the LL3rm regression model led to the least impact on the CI width, with R^2 s of 0.9 and 0.88, respectively (Fig. 6A and 6C). All remaining protocol changes led to slightly higher changes in the BMC uncertainty, with R^2 between 0.79-0.73 (Fig. 6A-G). The number of increased or decreased confidence intervals around the BMC was balanced across all protocol changes, with the exception of the protocol change towards higher BMRs where the BMC uncertainty was increased for the majority of data sets. The total number of BMC estimations for which a method change led to changes in the CI width that we consider as relevant (i.e. values outside belt around the perfect agreement) was then compared to the total number of BMCs: all method changes directly involved in the estimation of the BMC uncertainty (delta method, bootstrapping and model averaging) changed the CI width of the BMC for ca. 10% of all BMCs, whereat method changes that are expected to impact

the uncertainty estimation of a BMC only indirectly (mean replicate average, control-normalized data, sole application of an LL3rm, higher BMRs) had a minor impact on the determination of the BMC uncertainty (Fig. 6H). The distribution of fold change of CI widths is shown in Figure 6I.

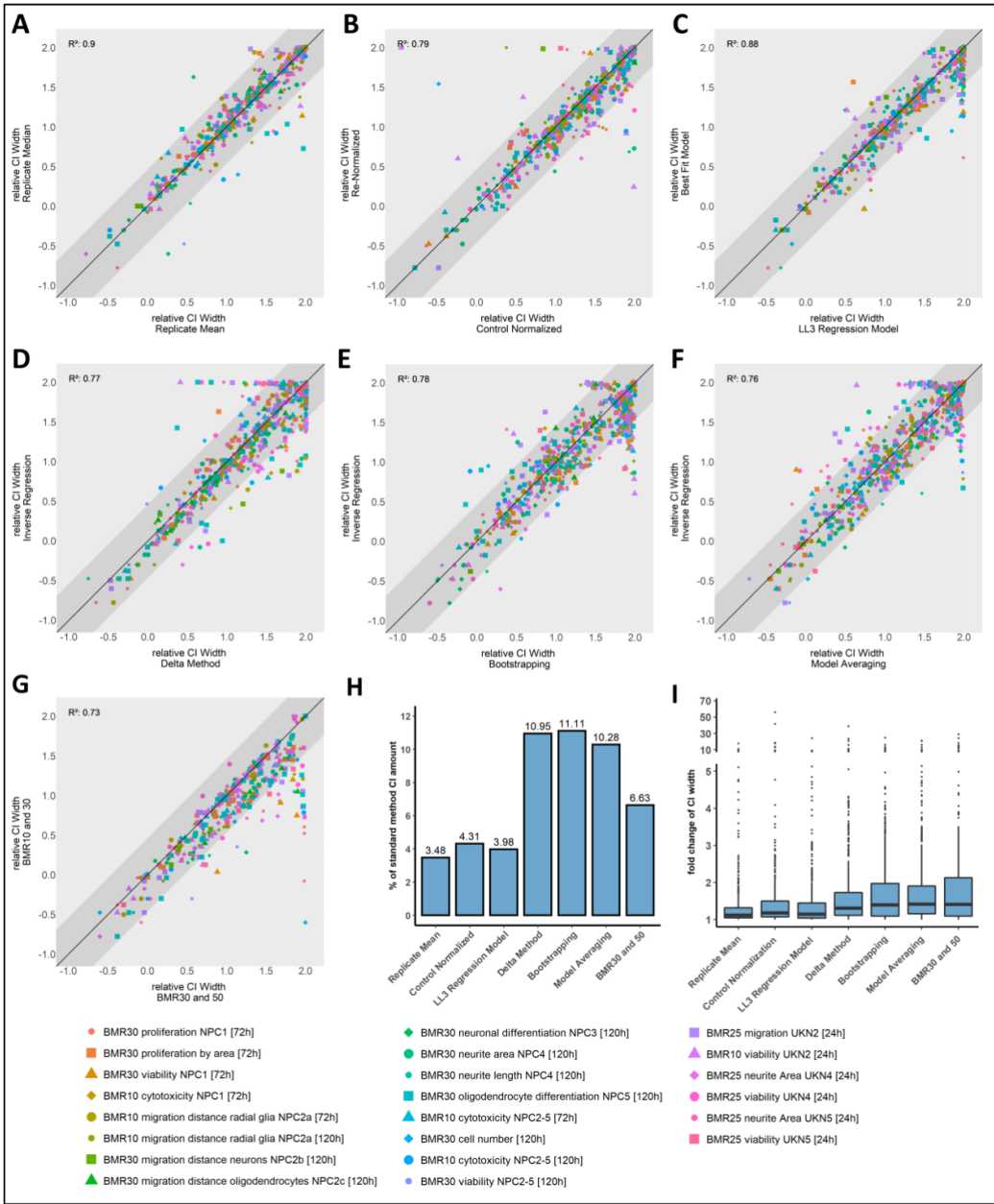


Figure 6: Impact of methodological changes in the data evaluation on the uncertainty of a BMC estimation
The BMC uncertainty was expressed as the width of the central 95% confidence interval (CI) around the BMC estimation, and CI widths for 148 compounds tested on up to 22 endpoints from 8 assays were determined using the standard protocol and opposing alternative methods. A-G): a relative CI width was calculated as the log10-transformed ratio between a 100-fold CI width and the highest test concentration, the relative CI width of the alternative protocol was plotted against the relative CI

width of the standard protocol. The solid black trend line indicates identical CI width's, the grey interval around the trend line indicates values below three-fold change. Values outside of this interval are considered as relevantly different between the opposing methods. Endpoints are indicated by a different color and shape. The goodness-of-fit coefficient from a trend regression analysis between both relative BMCs was calculated (R^2 ; top left), indicating the strength of agreement between both data evaluation protocols. A) Experimental median replicates versus mean replicates ($n = 517$). B) Re-normalized data versus control-normalized data ($n = 502$). C) Best fit approach versus a predefined three parameter log-logistic regression model ($n = 499$). D) Inverse regression versus delta method ($n = 588$). E) Inverse regression versus bootstrapping ($n = 600$). F) Inverse regression versus model averaging ($n = 561$). G) BMR10+30 (BMR10+25 for UKN) versus BMR30+50 ($n = 359$). H) Percentage of all BMC for which the protocol change lead to a threefold change in the relative CI in relation to the total number of BMC estimations. I) Distribution of CI width fold changes in response to statistical method changes from the standard protocol. Box whisker plot shows the median (horizontal line), interquartile range (box), 5% and 95% percentile values (whisker), and extreme values (black dots).

3.3 Examples

In the following we have selected five data examples which demonstrates the impact of methodological changes in the standard data evaluation protocol on a BMC estimation (Figure 7).

(i) Replicate median versus replicate mean (Figure 7A, proliferation by BrdU after 72h exposure): response data were calculated either as the median (blue, standard protocol) or as the arithmetic mean (red, alternative protocol) of the replicate responses and expressed as mean \pm SEM ($n=5$). The corresponding best-fit regression models are shown as solid (standard protocol) and dashed lines (alternative protocol), with the horizontal line corresponding to a BMR30. The combination of a large data variability and presence of individual data outlier led to mean response estimations closer to the control level, which was reflected by distinct regression curve estimates for both protocols. As consequence, a 7-fold higher BMC30 value was estimated for the alternative protocol, with a BMC30 of 2.02 μ M according to the standard protocol and a 14.3 μ M for the alternative protocol.

(ii) Re-normalization versus control normalization (Figure 7B, neuronal differentiation after 120h exposure): response data were either normalized to the average response observed at the lowest test concentration (blue, standard protocol) or to the controls (red, alternative protocol). The four lowest test concentrations produced similar control-normalized responses between 60-70%, with no indications for a trend between them. A regression analysis on all control-normalized responses suggests that only the highest test concentration produced a response not distinguishable from the controls, which we deemed as unrealistic, and, as consequence, the data set was not considered for a reliable BMC estimation. However, data re-normalization led to a more valid induction pattern, and the BMC estimate from the best-fit regression model was accepted.

(iii) Best-fit model selection versus a pre-defined regression model (Figure 7C, cell viability after 72h exposure): re-normalized response data are presented in the same way as in the previous examples, with the blue regression curve corresponding to the best-fit regression model (standard protocol) and the red curve to the three-parameter log-logistic model (alternative protocol). In this example, the

four-parameter log-logistic model (Table S2) was selected as best-fit regression model. In addition, the 95% confidence intervals around the entire curve estimates are included for both regression models. The exposure concentrations produced either no or maximal responses, with no data support for effect responses between. In a strict statistical sense, this data pattern does not allow the estimation of a reliable data curve, which is indicated for the three-parameter model (alternative protocol) by its poor data description and for the four-parameter log-logistic model (standard protocol) by its huge confidence belt for intermediate effect estimates. Nevertheless, the BMC30 (and its uncertainty) derived from the four-parameter log-logistic model (0.347 μM , 95% CI: 0.272-0.572 μM) seems reasonable and it is unlikely that a re-testing of the same compound on a refined concentration range will contradict this BMC30. This example demonstrates that although a BMC estimation might not fulfill all criteria according to “best statistical practice” (and be judged as unreliable by statisticians), it still can provide sufficient information to be assessed by the experimenter as reliable.

(iv) Uncertainty estimation of the BMC by inverse regression versus the delta method (Figure 7D, neuronal differentiation after 120h exposure): the 95% confidence intervals of the BMC30 estimated from the same best-fit regression model and data set are shown either by inverse regression, i.e. the interval along the horizontal 130% response line that intersects with the confidence belt of the regression curve (blue, standard protocol), or by the delta method (red). The confidence belt of the BMC30 by inverse regression provides a reliable expectation about where a BMC30 can be expected if the same experiments would be repeated, whereas the delta method provides a confidence belt which spans the entire test concentration range and therefore provides the misleading conclusion about a non-existing data support.

(iv) Uncertainty estimation of the BMC by inverse regression versus bootstrapping (Figure 7E, neuronal differentiation after 120h exposure): similarly to the previous example, 95% confidence intervals of the BMC30 estimated from the same best-fit regression model and data set are shown for two different statistical methods, and similar to the delta method, bootstrapping provides a large 95% confidence belt of the BMC30 which could be interpreted misleadingly as lacking data support for the regression modelling. The most likely reason for the poor performance of the bootstrap is the combination of a relatively large data variability and the responses observed at the second lowest concentrations which are not well described by the regression model.

Finally, we selected a single data set and analyzed it according to the standard protocol and all the methodological changes we have conducted to estimate a BMC and its 95% CI (cell viability, Figure 7F). It summarizes well how the different statistical methods can change a BMC estimation: a control-normalization of the response data shifted the BMC and its CI to much lower test concentrations, application of the delta method led to a drastic increase of the CI of the BMC30, the sole application

of the LL3rm regression model failed and did not provide any estimation and increasing the BMR30 to a BMR50 led to a non-estimable BMC (BMRnr).

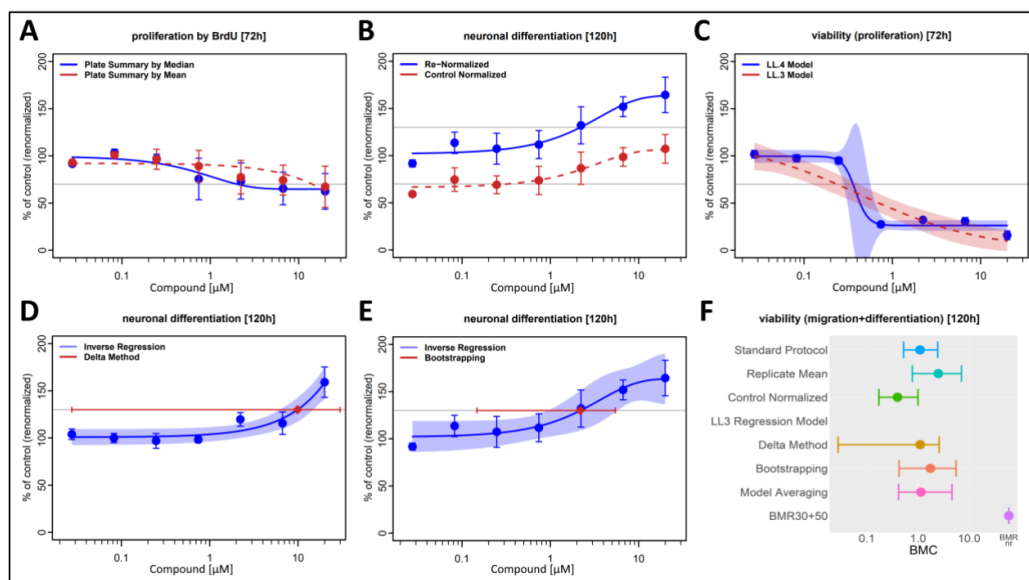


Figure 7: Data set examples and the impact of methodological changes in the data evaluation on the regression modelling and the BMC estimation and its uncertainty

A)-E) For several different steps of the data analysis and evaluation, the data resulting from the standard protocol (blue) is compared to the data deriving from the alternative protocol (red). Error bars show the SEM between summarized experiment data. Horizontal grey lines indicate the BMR. A) Experiment summarization by median and by mean. B) Re-normalized data and control-normalized data. C) Best fit approach and use of only a LL3 regression model. CI is displayed as confidence band around the fit model. Both models are applied to the data shown in blue. D) Inverse regression and delta method. CI of the alternative method is shown as red bar and BMC as red square. E) Inverse regression and bootstrapping. F) All method changes and their resulting BMC (displayed as dots) and CI values (displayed as bars) are shown for one exemplary dataset.

3.4 Method impact on hazard classification

An important application of the BMC estimation is the endpoint-specific hazard classification of the test compound into one of five hit categories, i.e. if the compound produced sufficient data evidence to be judged as a DNT-specific hit, borderline hit, unspecific hit, no hit, or as not identifiable (due to missing data support). Although all decision trees were setup as automatic systems, some data scenario provided insufficient data and were flagged for an expert judgement. The number of data scenarios for which the hazard classification was performed by “expert judgement” are listed in Table 2 for the standard protocol and seven methodological changes, divided according to the main decision trees developed for data from NPC or UKN assays. In total, 1563 classifications were conducted (NPC: 1347, UKN: 216), of which 68 (NPC: 30, UKN: 38) were flagged for an expert judgement according to the standard protocol. All protocol changes led to similar numbers, with the exception of the delta method applied to data outcomes from NPC assay endpoints which required expert input for three-

times more classifications. A marked difference was observed between the decision trees for NPC and UKN assay endpoints, with up to 5 times more classifications flagged for expert judgement for UKN outcomes depending on the statistical method chosen.

Table 2: Number of endpoint-specific DNT hit classifications judged by experts.

The numbers of hit classifications by expert judgement are presented as percentage of all classifications that were supervised by the hazard decision trees.

Method	NPC [%] ¹	UKN [%] ²
Standard Protocol	2.23	17.59
Replicate Mean	2.15	16.67
Control-normalized	3.27	15.74
LL3rm	2.23	12.50
Delta Method	8.09	18.52
Bootstrapping	3.12	17.13
Model Averaging	2.90	15.74
BMR30+50	1.93	12.96

¹NPC = Data outcomes from NPC assays, ²UKN = Data outcomes from UKN assays

Due to the poor performance outcomes of the delta method in judging the uncertainty of a BMC estimation and the consequence of a more likely expert intervention in the automatic hazard classification, we judge this method as too unreliable and have excluded it from all remaining analyses.

Exemplary data sets are shown for three different classification scenarios: (i) a specific DNT hit decision for a significantly inhibited oligodendrocyte differentiation at exposure concentrations above 0.25 μM , but only a marginally reduced cell viability (marker for cytotoxicity) at 20 fold higher concentrations (Figure 8A), (ii) an unspecific hit decision for a significantly inhibited oligodendrocyte differentiation and cytotoxicity observed at same concentration ranges (0.24 to 2.2 μM) (Figure 8B), and (iii) a data scenario which was flagged for an expert judgement because the highest test concentration (20 μM) produced a weak but significant effect reduction for the specific endpoint but the regression analysis estimated a BMC10 (and BLL) that was outside the test concentration range. On closer inspection of the experimental data (Figure 8C, with each color-coded symbol representing the replicate median from an independent experiment) it was decided that responses from both the specific and unspecific endpoint were not distinguishable, and thus the weak response reduction of the specific endpoint was classified as unspecific.

Figure 8D provides an overview about the total number of hit classifications that changed in response to changes of the standard protocol. Expressed as percentages and for each methodological change, the changes of hit classifications are further divided in “gains”, i.e. the percentual increase of hazard hits in relation to the standard protocol, and “losses”, i.e. the percentual decrease of hazard hits in relation to the standard protocol. Here a change toward replicate averaging by mean, control

normalization and bootstrapping caused the lowest number of classification changes (<5%), followed by methodological changes towards model averaging or higher BMR levels which led to almost 7% different hit classifications. Here, model averaging increased the number of “not identified” classifications by 2.56%, mostly at the cost of “no hit” classifications, and the higher BMR levels led to 4.86% more “no hit” classifications (in line with the data example of Figure 7G). The by far most severe changes of hit classifications were observed if only the LL3rm regression model was used to describe the experimental concentration response data (45.87% total difference), which led to 42.03% more “not identified” classifications. The latter is most likely the consequence of unsuccessful regression modelling (and corresponding BMC estimation) due to lack of sufficient data support for this model (see 3.1 and 3.2).

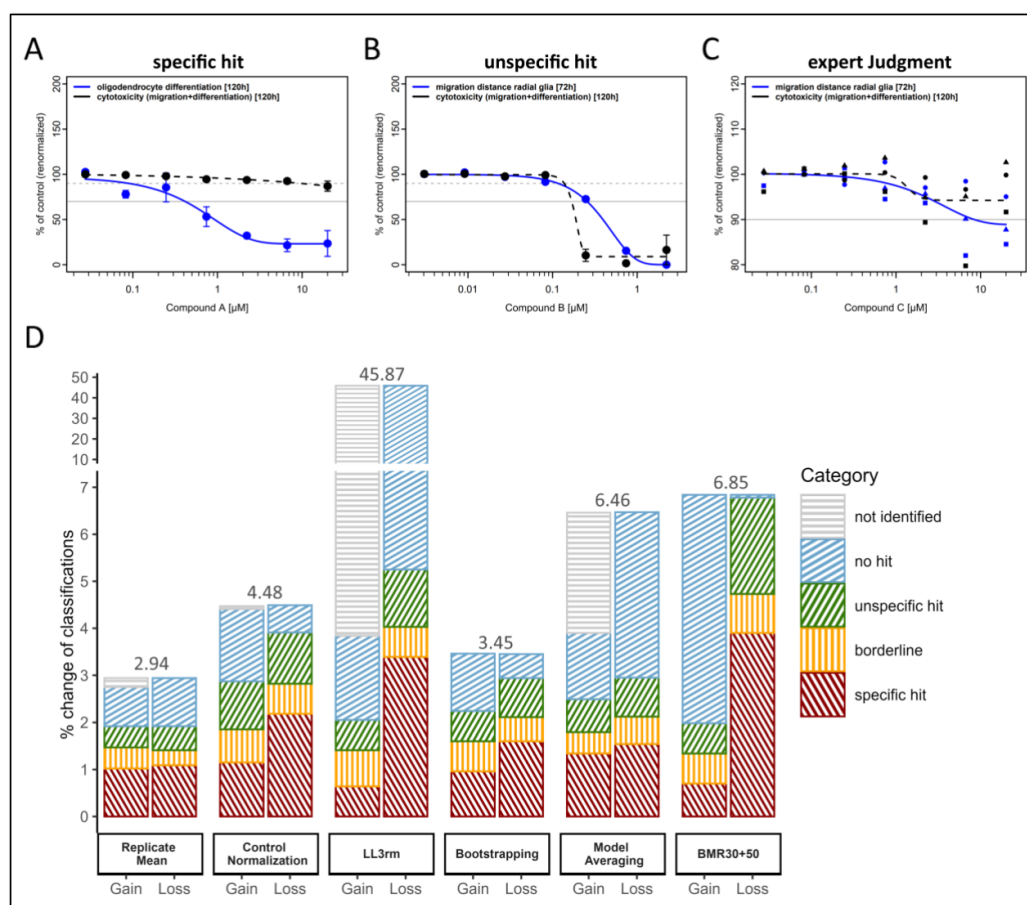


Figure 8: Number of endpoint-specific DNT hit classification changes in response to changes in the standard data evaluation protocol A-C Exemplary data sets for three different classification scenarios: concentration-response data from the specific (blue) and unspecific (black) endpoints are from 5 independent experiments, with effect responses re-normalized to the regression estimate at lowest test concentration and summarized as mean±SEM. Horizontal lines indicate the BMR levels for the BMC estimation, where straight lines indicate the specific endpoint BMR and dotted lines the unspecific endpoint BMR (if they differ). Data were always analyzed according to the standard data evaluation protocol A) Specific hit: the specific endpoint (oligodendrocyte differentiation) is impacted at non-toxic concentrations. B) Unspecific hit: inhibition of

oligodendrocyte differentiation and cell viability are observed at similar concentration ranges. C) Hit classification by expert judgement: an automatic hit classification was prevented by ambiguous data, but judged as “unspecific” by experts. D) For each methodological change to the standard protocol, the number of hit changes is expressed as percentage of the total number of hit classifications, divided into in “gains” (i.e. the percentual increase of hazard hits in relation to the standard protocol) and “losses” (i.e. the percentual decrease of hazard hits in relation to the standard protocol). Different bar segments represent the different classification categories.

3.5 Assay performance

To assess how changes in the data evaluation protocol might impact the evaluation of the DNT IVB’s predictivity, 28 reference chemicals of known DNT and 17 negative control chemicals were selected (Masjosthusmann et al. 2020), with all 45 substances tested in the DNT IVB, and the overall performance of the DNT IVB was quantified by its specificity, sensitivity and accuracy. Outcomes are shown for the standard protocol as well as all relevant changes in Figure 9: (i) Specificity (Fig. 9A): standard protocol and changes of it led always to a specificity between 87.5% and 100%, i.e. a truly DNT negative substances were almost always also judged as negative by the DNT IVB, and the standard protocol seems to be robust against methodological changes in judging false-negatives. (ii) Sensitivity (Fig. 9B): 23 of the 28 DNT substances (82.1 %) were successfully identified by the DNT IVB if the standard protocol was used, but changes to it led always to a lower sensitivity. (iii) Accuracy (Fig. 9C): The best performance was achieved for the standard protocol (88.6%), followed by a methodological change to bootstrapping (86.4%), higher BMR levels (84.1%), mean replicates, control normalization, pre-defined regression model (all 81.8%) and model averaging (77.3%). The latter performed 11.3% below the accuracy value of the standard protocol. A detailed overview over the hit definition of all control compounds is given in supplementary segment 2.1 (Tab. S5-S7).

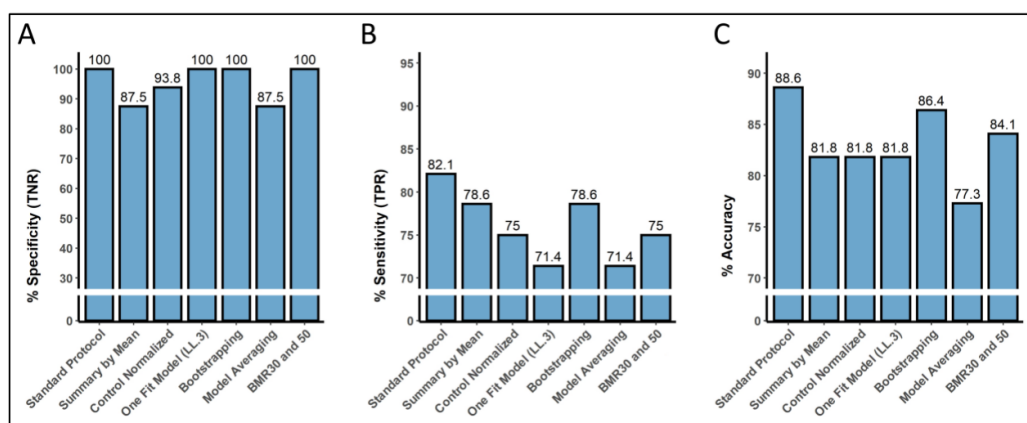


Figure 9: Evaluation of the predictive performance of the DNT IVB based on the standard data evaluation protocol and changes Bar graphs show the results of the predictive capability of the DNT IVB for 28 substances of known DNT and 17 negative control substances in terms of specificity, sensitivity and accuracy.

4 Discussion

The basis for this biostatistical study is a compound screening project performed on behalf of an EFSA procurement during the years 2017-2020 (OC/EFSA/PRAS/2017/01). Twelve DNT test methods with accompanying cytotoxicity and viability assays belonging to an OECD DNT IVB (Crofton and Mundy, 2021) were challenged with 124 compounds from different compound classes including expected negative control compounds (Masjosthusmann et al. 2020). This paper is not about informing on the compounds' effects on specific neurodevelopmental key events, which can be found elsewhere (Masjosthusmann et al. 2020; Blum et al. in revision), but rather analyzes the impact of common biostatistical concentration-response methods on the overall DNT IVB performance. As in vitro methods have been gaining complexity over the last decade, i.e. from simple reporter gene assays towards organotypic cultures, we tested the hypothesis if the selection of a biostatistical method can affect the performance of the DNT IVB. Specifically in the field of developmental toxicity, where in vitro test systems can nowadays assess biologically more complex systems, like changes in key developmental processes over time, such an evaluation seems timely. Hence, a comparative assessment of different biostatistical methods on the BMC estimation, DNT hit classification and DNT IVB performance was performed.

4.1 Experimental mean or median replicate

Instead of the individual experimental readouts, we used always their "average" response per test concentration and experiment (replicate average) as statistical unit in the concentration-response regression analysis. Our main argument for this data reduction was that the BMC and BLL estimation should reflect mainly biological and between-study variability, with the advantage that less complex statistical methods are required, which is a crucial requirement for the robustness of an automatic data evaluation pipeline. An "average" of the replicate responses can be estimated in various ways, with the arithmetic mean calculation the most popular and statistically often best option if the response data follow the rules of a symmetric distribution. However, the presence of an outlier can violate this assumption, and as a consequence it can lead to a biased estimation of an average that does not represent the observed data correctly. To protect the mean against an outlier would not only require outlier detection methods, which are per se problematic for small sample sizes, but also a decision on how to handle these values in further data analyses (e.g., removing, winsorization, trimming). A common alternative to the mean is the median which is more robust against outliers, but also known to be a more uncertain estimate for an average (Maindonald and Brown, 2010). As assay endpoints of the DNT IVB can produce a relatively high data variability within an experiment, we ruled out outlier detection methods and considered the median of the replicate responses as default option for the standard protocol.

Our study shows that the alternative of using the arithmetic mean led in average only to minor changes in BMC estimations and hazard classification outcomes, which might refer to those data sets where either no outliers were present or outliers occurred at concentrations with only little influence on the regression analysis. Nevertheless, for a few data sets a decision towards the median or mean had a strong influence on the best-fit regression analysis such that, at worst case, the subsequent BMC estimation was prevented (“BMRnr”). Although the choice of the average replicate calculation had in comparison to other methodological protocol changes only a minor impact on the hazard classification, it still lowered all performance parameters (specificity, sensitivity, accuracy) that we used to assess the DNT IVB’s predictive power for identifying DNT adversity. Altogether, our study outcomes strengthen the argument for using the median.

4.2 Data normalization to control and re-normalization

Typically, readouts from *in vitro* endpoints can substantially vary between experiments and thus require a normalization to make them comparable across experiments. The classical approach is to “anchor” all values to the average response of a negative control (and, depending on the assay endpoint, positive control). The expectation is that an exposure concentration with no impact on the assay endpoint will produce readouts similar to those from the control reference, which in the concentration response context means that a regression curve is expected to “equal” the control responses at zero and non-effective low exposure concentrations. This expectation does not always hold true in experimental practice, and readouts from concentrations that are expected to show no exposure activity (on basis of mechanistic reasoning or the entire data pattern) can differ from the readouts of the control reference. Although a random explanation is theoretically possible, e.g., the control readouts were not representative and only rare “unlucky” outcomes, the confirmation by independent experiments points to a non-random cause. However, the reasons for this phenomenon are usually unclear, with biological effects and technical issues discussed (Krebs et al., 2018). If ignored, a control normalization can not only suggest false treatment-related effects at non-effective concentration ranges, but also contradict the meaning of a comparable relative effect scale and therefore invalidate a BMR: if non-effective concentrations produced normalized effect responses which are more than 10% different from the negative controls, then a regression modelling cannot establish a BMC10. To overcome the problem of misleading control data it has been suggested to estimate the control reference directly from the responses of the test concentrations, assuming that they can provide sufficient evidence for “non-exposure related endpoint activity”, which is often translated as equal assay responses at the lowest test concentrations over a sufficient large concentration range (Krebs *et al.*, 2018). A refined control reference can then either be estimated directly from these data responses or from response estimated by concentration-response regression

analysis, and then be used to re-normalize the entire data set such that the refined control reference is set to 100%.

The experimental design for the assays from the DNT IVB was chosen such that the lowest test concentrations were expected to produce no treatment-related effect responses. For more than 90% of all data sets the three lowest concentrations provided non-distinguishable effect responses, which we deemed as sufficient for using the control re-normalization on all data, and it was implemented in the automatic data evaluation pipeline as part of the standard protocol. This was mainly motivated by the frequent occurrence of misleading negative control responses observed for some of the assay endpoints.

On this background it is not very surprising that the choice of the normalization method not only led to very different BMC estimations but often completely failed, as documented by the number of BMRnr's (Figure 5B): the intended BMR was not covered by control-normalized responses, and as consequence the regression analysis suggested a best fit model that also did not cover the BMR and therefore could not estimate a BMC. However, a re-normalization of the effect scale guaranteed coverage of the BMR, and accordingly the regression modelling was able to establish a BMC. Figure 7B provides another example for a gross data misinterpretation: normalized effect responses suggest a BMC for inhibition, re-normalized effect data a BMC for induction. Although the majority of data sets did not necessarily require a control-renormalization, a change to the standard control normalization still changed the hit category for approx. 5% of all endpoint-specific DNT hazard classifications (Figure 8D), and impacted all performance parameters about the DNT IVB's predictivity negatively (Figure 9).

A re-normalization should only be applied if sufficient data evidence is provided to do it, otherwise an existing exposure effect can be judged wrongly as technical or biological artifact and misused as zero effect response in the statistical concentration-response analysis. This decision making is only difficult to resolve in an automatized HTS data evaluation and we cannot fully rule out that a re-normalization was wrongly used for some data sets. Therefore, we recommend for the future a list of criteria such that those data scenarios can be identified and flagged for an expert decision. Potential criteria for assuring a successful data re-normalization could be a certain minimum magnitude between the original and re-defined control references based on statistical reasoning, a minimum number of test low concentrations for which the effect responses provide no indications for a positive or negative trend and a minimum concentration range at which no effect can be judged. In case a control normalization can neither be judged by an automatic ruling system nor by an expert we suggest as last solution a repetition of the experiment at lower test concentrations.

4.3 Concentration-response regression model

A BMC is derived from a mathematical function which is fitted as parametric regression model to the experimental concentration-response data. As no unique mechanistic model exists that would allow a 100% accurate representation of every possible shape of a concentration-response pattern, a regression model is considered as empirical and judged as suitable only if it describes the data in the best possible way. Each model is characterized by a limited number of model parameters which provide the flexibility to fit the model as close as possible to the data: the more model parameters are considered the better the model will describe the data, and the more complex the observed data pattern is (e.g. non-monotony) the more model parameters are required to describe the data. However, the more model parameters are considered the more data evidence are required to support a reliable regression fit. If a model with too many parameters is chosen that is not supported by a sufficient amount of data (over-parametrization), the estimation method will result into overinflated model estimates and therefore in an unreliable BMC estimate with an extremely large CI (high difference between BLL and BUL). In the concentration-response plot, the latter is usually indicated by a huge “uneven” confidence belt around the regression fit with extreme peaks at concentrations at which the highest nonlinearity of the concentration-response pattern occurred. Moreover, two different models might describe the concentration-response data similarly well with identical BMC estimates, but the model with the lower number of model parameters will result into smaller confidence interval around the BMC (and which might therefore influence the hazard classification). Generally, the model with the lowest number of parameters is favored over a more complex model as long as it can describe the data almost as accurate as the more complex model (parsimony). The accuracy vs. parsimony trade-off is utilized by the AIC criterion in the model selection process of the best-fit method.

Data and the experimental design decide on which model can be chosen for a BMC estimation, and the minimal data requirements can be assigned directly to the model parameters: each parameter requires a specific data support from effect responses of at least one concentration, i.e. a model with five parameters would require five concentrations with effect responses that are specifically addressing the nature of the parameter. For instance, the model parameter describing the upper control asymptote (d in Table S3) requires that at least one concentration has produced responses that can be used as average control reference, the parameter describing the lower maximal asymptote (d in Table S3) requires that at least one concentration has produced responses that can be used as average control reference, and all other model parameters require data support from concentrations that neither have produced maximal or minimal responses. If these data requirements are not given, a BMC estimation should be considered always as unreliable from a strict statistical point of view. For example, the data example C in Figure 7 provides no effect responses between the minimal and maximal asymptote, and therefore a priori no regression model can be fitted to the data on sound

statistical criteria. Nevertheless, the LL4 model (blue line) provides an BMC estimation including an 95% confidence belt which, in this case, appears to be well in line with the experimental data, and can be considered as sufficiently accurate for regulatory purposes as all three independent experiments produced nearly identical data outcomes.

A concentration-response function commonly used in pharmacology and toxicology is the Hill function, which is a reparametrized form of the 3-parameter log-logistic function (LL3, but without the log10 transformation of the concentration, Table S2). We chose this function as pre-defined model in comparison to the best-fit model approach since it a popular approach for regression of strict monotonic concentration-response patterns in comparable software (*tcpl*, Filer *et al.* 2016), and since it is also recommended by the OECD for continuous data (EFSA Scientific Committee, 2017). In line with theoretical expectations and previously reported simulation studies (Zhu *et al.*, 2007; West *et al.*, 2012; Piegorsch *et al.*, 2013), the best-fit model approach responded more flexible to data sets and therefore resulted often to BMC estimations that differed significantly from those derived by the Hill model (Figure 5C). As a consequence, the sole application of the Hill model occasionally prevented the estimation of a BMC and its uncertainty, and therefore led to less data sets for which a hazard identification could be performed. This strongly suggests the use of several regression models in a best-fit approach, including functions with maximal two model parameters for “data-poor” sets and more complex functions for “data rich” scenarios.

Model averaging is historically motivated by the typically small number of doses in animal studies that can provide meaningful data for the regression modeling, and the subsequent problem that different regression models can describe the observed dose-response data equally well but interpolation in a dose region with little or no data may result into very different response (and BMD) estimates (EFSA, 2017). A statistical argument in favor of model averaging is that uncertainty of the model selection process of the best fitting method is not incorporated in the BMD and associated BMDL estimation (West *et al.*, 2012). Our study shows no big differences between both methods, and we attribute the higher number of failed BMC estimates for model averaging (Figure 5D) due to the fact that the models with the lowest number of model parameters were not included in the pool of candidate models for model averaging: the 2-parameter exponential model (Table S2) was selected as best-fit model in approx. 33% of all model decisions (standard protocol), indicating that the data sets did not allow the selection of a more complex model with 3 model parameters, and as consequence model averaging did fail. It demonstrates that simple regression models are essential for “poor data” scenarios, i.e., data sets where maximally two concentrations responded with significant but often weak assay responses.

4.4 Uncertainty estimation of the BMC

Not only the BMC estimation is crucial for the hazard classification but also a correct derivation of its uncertainty, usually expressed as lower and upper confidence interval (CI). We have used various statistical methods which are implemented in the *drc* and *bmd* R package (delta approximation, inverse regression, resampling methods), and investigated how they can impact the hazard classification. It should be noted that these methods do not change the BMC estimation but try to calculate the uncertainty of the BMC estimation from the estimated regression model and experimental data. All methods have their pros and cons with different requirements to the data and regression models, and none of them can a priori be ruled out as inappropriate for the BMC estimation of a DNT IVB data set. Assuming that the correct regression model was chosen and the estimation method led to only one reliable BMC estimation, we used the inverse regression as CI reference for a best-fit model estimated BMC, and compared it to the ones derived from the delta method and parametric bootstrapping. The advantage of the inverse regression method is that the confidence of a regression curve can easily be assessed by a non-statistician by showing the concentration-response data together with the regression fit and its associated confidence interval in a common plot.

As shown in Figure 6H, different CI methods often resulted in largely different outcomes, with a moderate impact on the hazard classification (Figure 8D) and corresponding DNT IVB performance parameters (Figure 9). The delta method provides a means to estimate the approximate variance and CI of a model function when the function consists of one or more estimated model parameters, and where there is an estimate for the variance of each model parameter, with both derived from the successful fit estimation. The method implemented in the *drc* package (Ritz *et al.*, 2015) is based on a first order approximation for the variance of the BMC, and thus expected to be accurate only for concentration-response pattern that show a minor non-linearity (Zhu, Wang and Jelsovsky, 2007; Moerbeek, Piersma and Slob, 2004). Higher order approximations of the delta method would be progressively more flexible and provide a better description of the BMC uncertainty but are currently not implemented. Therefore, it is not surprising that the delta approximation often failed with an unreliable CI spanning the entire range of test concentrations (Figure 7D), especially for the 2-parameter exponential function with a concentration term that is not log10-transformed. Based on the study outcome we deem this method as unfit for an automatic HTS data evaluation.

Whereas the delta method is entirely based on the outcomes from the regression fit and therefore provides a quick and easy way to calculate the IC for a BMC, resampling methods use only the regression model(s) and BMC estimation and develop the BMC uncertainty entirely from re-doing the regression analysis and BMC derivations on a huge number of concentration response data resampled from the original experimental data sets. This method puts strong emphasis on a “representative” data

set for the resampling, and if violated, it is prone to biased interval estimations (i.e. mode of the resampled BMC distribution differs from the original BMC estimation) or, in worst-case, the simulations lead to an interval that hardly mirrors the observed data variability. Typically for DNT IVB, endpoints often produced responses with a relatively high between-study variability (documented in the corresponding BMRs, Table 1), with only a small sample size for resampling (3-5 experimental replicate medians), and with often only two or less test concentrations which provided significant responses distinguishable from the controls. These data scenarios are not optimal for regression resampling, and therefore it is not surprising that bootstrapping often resulted in very different, too wide confidence belts compared to those from inverse regression, or even completely failed (Figure 7E). To some extent this might also explain the different outcomes for model averaging, which was performed always in combination with bootstrapping.

Until generally applicable decision rules about the minimal data requirements for bootstrapping can be implemented in an automatic data evaluation platform, it is only difficult for the non-expert to make decisions about the usefulness of resampling for a particular data scenario. Therefore, our advice is that the user should have some experience regarding statistical resampling, or, if applicable, use inverse regression or related methods.

4.5 The choice of the BMR on the Hazard identification

The optimal choice of an endpoint-specific BMR level is always a compromise between “as close as possible” to the control reference (i.e. a BMC estimation as low as possible) and the statistical demands for providing a reliable BMC estimation for as many data sets as possible. In principle, each data set has its own optimal BMR, mainly defined by its between-experimental data variability and how well it can support the regression analysis. A BMR which guarantees a BMC for all future data sets would need to be chosen from the data set with the largest observed between-experimental data variability, and by this a relatively large BMR would be favored, with response levels around 50% for some of the assay endpoints (i.e. a BMC would equal an EC50 or IC50). However, a larger BMR leads to a higher BMC, and the consequence for all data sets which a much lower data variability is that their substance responses observed at concentration ranges below the BMC are ignored, and therefore contradict the intended regulatory meaning of a benchmark concentration. But more important, it would also rule out those data sets for a BMC estimation where the observed maximal responses are below the BMR (and thus a BMC cannot be established). The latter was decisive for ruling out many sets for a BMC estimation after we increased the BMRs by 20% in our standard protocol (e.g., changing BMC30 to BMC50, Figure 5E), and as consequence our simulations resulted into 5% different hazard classifications, with a change mainly from “hit” to “no hit” (Figure 8D). Therefore, the use of the most

common descriptor for concentration response data in pharmacology and in vitro toxicology, an IC50 or EC50, cannot be recommended as surrogate for a BMC for endpoints of the DNT IVB.

A critical aspect is how a BMR can be derived: we used the 1.5 sigma rule, with sigma estimated as standard deviation from the between-experimental variation from a large set of historical data sets (Masjosthusmann et al., 2020). For a sample size of 3-5 independent experiments, we expected for the majority of data sets the estimation of a BMC if a true BMC was present in the data, but nevertheless our standard protocol might have failed to identify a hazard because the BMR was selected as too low for a particular data set. Our study outcomes do not provide the exact number how often this might have happened as it would require for each individual data set a statistical power analysis and corresponding estimation of the detection limit, however, assuming that the scatter between the experimental replicate medians always followed the Gaussian distribution we expect this to be the case in less than 1% of all cases.

4.6 Hazard identification and software

A huge number of free software packages for the statistical analysis of dose-response data and dose-response modelling are available, with PROAST (RIVM National Institute for Public Health and the Environment), BMDS (US EPA), ToxCast pipeline (tcpl, Filer et al. 2017) or BMCeasy (Krebs et al., 2019) just to mention a few. Similar to the R packages we use (drc and bmd, Ritz et al., 2015 and Jensen et al., 2020), most of these software packages provide a variety of options in order to respond as flexible as possible to the various data scenarios a user can possibly face, and as consequence, always a minimum of statistical knowledge is demanded from the user. Similar to the tcpl pipeline we became interested in an automated data evaluation platform with no required user intervention and addressing the specific features of DNT data or other data from organotypic cultures. To our experience, the proposed standard protocol is for an automated data evaluation pipeline the best compromise between the various statistical methods without “overcomplicating” the regression analysis and the corresponding BMC estimation. The drawback of an automated analysis is always the danger of not being prepared to deal with an unusual data set, a scenario that most likely can only be avoided by analysing each data set individually by an expert. The strength of our data evaluation platform is the integration of endpoint-specific hazard classifications, including flagging systems for uncertain cases, which none of the software packages mentioned above offer. We consider it crucial for the hazard assessment to differentiate between general cell toxicity and specific DNT hits.

4.7 Conclusion

The comparative study between various statistical methods involved in the estimation of a BMC and its associated uncertainty for a huge number of concentration response data sets from the DNT IVB revealed the following main conclusions:

- 1) The normalization of effect data to the outcomes of test concentrations can be a viable option to safeguard against an ill-defined negative control reference and therefore avoid a biased BMC estimation and incorrect hazard alerts. This re-normalization of response data should be done whenever sufficient data evidence is provided for non-exposure related effect responses at lowest concentrations and which have been confirmed by independent experiments. Optimally it should be decided on a case-by-case basis by the experimenter, and more efforts are required to integrate decisions for a re-normalization in automatic data evaluation routines.
- 2) The pool of candidate models for the parametric regression analysis should include as simple as possible mathematical functions in order to enable a BMC estimation for data sets which provide only little data support for the regression modelling. This can be either the exponential or linear function, with both including maximally only two model parameters.
- 3) Simple common statistical methods such as the delta method do not necessarily guarantee a reliable estimation about the uncertainty of a BMC confidence and depend strongly on the chosen regression model and its non-linearity close to the BMR. In contrast, more sophisticated methods such as resampling require more data support which is often not given by the experiments. Invers regression provided the best way to judge a BMC uncertainty.
- 4) Data sets with only two or less effective concentrations are often borderline to a reliable statistical analysis, but nevertheless provide sufficient data for a “pragmatic” solution. The BMC for the specific DNT endpoints of these data sets is usually at high concentrations and within (or close) to the cytotoxic concentration ranges, and thus are most likely not too be classified as “specific hit”.
- 5) The BMR level should be chosen as close as possible to the control level without compromising the statistical concentration-response analysis. Setting it too high (e.g. 50%) involves the danger of overlooking hazard responses which can lead to erroneous hazard hit classifications.
- 6) An endpoint-driven hazard classification method is essential for a reliable identification of hazard alerts, and DNT-specific endpoints should always take general cell health into account. The automatized data evaluation should include a decision making that pinpoint to data scenarios which require a manual expert judgment for the hazard classification.

Although this study was conducted on concentration response data from only the DNT IVB, we think many of the conclusions can be generalized to data from other specific toxicological endpoints, especially in the rising field of organotypic/stem cell-based cultures. It demonstrates that statistical decisions which seem to be of minor importance can become decisive if it comes to the hazard classification of a test substance. It also demonstrates how important fit-for-purpose, internationally harmonized and accepted data evaluation and analysis procedures are for an objective hazard classification.

5 References

Buckley, B.E., Piegorsch, W.W., West, R.W. Confidence limits on one-stage model parameters in benchmark risk assessment. *Environ Ecol Stat* 16, 53–62 (2009). <https://doi.org/10.1007/s10651-007-0076-2>

Claeskens, G. and Hjort, N. (2008). *Model Selection and Model Averaging*. Cambridge University Press, Cambridge. <http://dx.doi.org/10.1017/CBO9780511790485>

Cox, C. Fieller's Theorem, the Likelihood and the Delta Method. *Biometrics*, vol. 46, no. 3, 1990, pp. 709–18. *JSTOR*, <https://doi.org/10.2307/2532090>. Accessed 13 Aug. 2022.

Crofton, K.M., Mundy, W., R, (2021). External Scientific Report on the Interpretation of Data from the Developmental Neurotoxicity In Vitro Testing Assays for Use in Integrated Approaches for Testing and Assessment. *EFSA supporting publication*; 18(10):EN-6924. 42 pp. doi:10.2903/sp.efsa.2021.EN-6924

Dent, M.P., Vaillancourt, E., Thomas, R.S., Carmichael, P.L., Ouedraogo, G., Kojima, H., Barroso, J., Ansell, J., Barton-Maclaren, T.S., Bennekou, S.H., Boekelheide, K., Ezendam, J., Field, J., Fitzpatrick, S., Hatao, M., Kreiling, R., Lorencini, M., Mahony, C., Montemayor, B., Mazaro-Costa, R., Oliveira, J., Rogiers, V., Smegal, D., Taalman, R., Tokura, Y., Verma, R., Willett, C., Yang, C., Paving the way for application of next generation risk assessment to safety decision-making for cosmetic ingredients, *Regulatory Toxicology and Pharmacology*, Volume 125, 2021, 105026, ISSN 0273-2300, <https://doi.org/10.1016/j.yrtph.2021.105026>

Efron, B. and Tibshirani, R.J. *An introduction to the bootstrap*. CRC press, 1994.

EFSA Scientific Committee, Hardy, A., Benford, D., Halldorsson, T., Jeger, M.J., Knutsen, K.H., More, S., Mortensen, A., Naegeli, H., Noteborn, H., Ockleford, C., Ricci, A., Rychen, G., Silano, V., Solecki, R., Turck, D., Aerts, M., Bodin, L., Davis, A., Edler, L., Gundert-Remy, U., Sand, S., Slob, W., Bottex, B., Abrahantes, J.C., Marques, D.C., Kass, G. and Schlatter, J.R., 2017. Update: Guidance on the use of the benchmark dose approach in risk assessment. *EFSA Journal* 2017;15(1):4658, 41 pp. <https://doi.org/10.2903/j.efsa.2017.4658>

Fang, Q., Piegorsch, W. W., and Barnes, K. Y. (2015) Bayesian benchmark dose analysis. *Environmetrics*, 26: 373– 382. doi: 10.1002/env.2339.

Filer, D.L., Kothiyi, P., Setzer, R.W., Judson, R.S., Martin, M.T., tcpl: the ToxCast pipeline for high-throughput screening data, *Bioinformatics*, Volume 33, Issue 4, 15 February 2017, Pages 618–620, <https://doi.org/10.1093/bioinformatics/btw680>

Förster, N., Butke, J., Keßel, H.E., Bendt, F., Pahl, M., Li, L., et al. Reliable identification and quantification of neural cells in microscopic images of neurospheres. *Cytometry*. 2022; 101: 411– 422 <https://doi.org/10.1002/cyto.a.24514>

Gerrard, P., (2013), J. Maindonald, & W.J. Braun (2010) *Data Analysis and Graphics Using R: An Example Based Approach* (Cambridge Series in Statistical and Probabilistic Mathematics) Third Edition, *Psychometrika*, 78, issue 4, p. 856-857.

Jensen, SM., Kluxen, FM., Ritz, C. A Review of Recent Advances in Benchmark Dose Methodology. *Risk Anal.* 2019 Oct;39(10):2295-2315. doi: <https://doi.org/10.1111/risa.13324> Epub 2019 May 2. PMID: 31046141.

Jensen, SM., Kluxen, FM., Streibig, JC., Cedergreen, N., Ritz, C. (2020). *bmd*: an R package for benchmark dose estimation. *PeerJ* 8:e10557 <https://doi.org/10.7717/peerj.10557>

Krebs, A., Nyffeler, J., Karreman, C., Schmidt, B.Z., Kappenberg, F., Mellert, J., Pallocca, G., Pastor, M., Rahnenführer, J. and Leist, M. (2020) "Determination of benchmark concentrations and their statistical uncertainty for cytotoxicity test data and functional in vitro assays", *ALTEX - Alternatives to animal experimentation*, 37(1), pp. 155–163. doi: <https://doi.org/10.14573/altex.1912021>

Krebs, A., Nyffeler, J., Rahnenführer, J., Leist, M., (2018). *Normalization of data for viability and relative cell function curves*. In: *Alternatives to Animal Experimentation : ALTEX*. 35(2), pp. 268-271. ISSN 1868-596X. eISSN 1868-8551. Available under: doi: <https://dx.doi.org/10.14573/1803231>

Leist, M., Hasiwa, N., Rovida, C., Daneshian, M., Basketter, D., Kimber, I., Clewell, H., Gocht, T., Goldberg, A., Busquet, F., Rossi, A.-M., Schwarz, M., Stephens, M., Taalman, R., Knudsen, T. B., McKim, J., Harris, G., Pamies, D. and Hartung, T. (2014) "Consensus report on the future of animal-free systemic toxicity testing", *ALTEX - Alternatives to animal experimentation*, 31(3), pp. 341–356. doi: <https://doi.org/10.14573/altex.1406091>

Leontaridou, M., Urbisch, D., Kolle, S. N., Ott, K., Mulliner, D. S., Gabbert, S. and Landsiedel, R. (2017) "The borderline range of toxicological methods: Quantification and implications for evaluating precision", *ALTEX - Alternatives to animal experimentation*, 34(4), pp. 525–538. doi: <https://doi.org/10.14573/altex.1606271>.

Li, H., Yuan, H., Middleton, A., Li, J., Nicol, B., Carmichael, P., Guo, J., Peng, S., Zhang, Q., Next generation risk assessment (NGRA): Bridging in vitro points-of-departure to human safety assessment using physiologically-based kinetic (PBK) modelling – A case study of doxorubicin with dose metrics considerations, *Toxicology in Vitro*, Volume 74, 2021, 105171, ISSN 0887-2333, <https://doi.org/10.1016/j.tiv.2021.105171>

Masjosthusmann, S., Blum, J., Bartmann, K., Dolde, X., Holzer, A.-K., Stürzl, L.-C., Hagen, Keßel H. E., Förster, N., Dönmez, A., Klose, J., Pahl, M., Waldmann, T., Bendt, F., Kisitu, J., Suci, I., Hübenthal, U., Mosig, A., Leist, M., Fritsche, E., 2020. Establishment of an a priori protocol for the implementation and interpretation of an in-vitro testing battery for the assessment of developmental neurotoxicity. *EFSA supporting publication* 2020: 17(10): EN-1938. 152 pp. doi: 10.2903/sp.efsa.2020.EN-1938

Moerbeek, M., Piersma, AH., Slob, W. A comparison of three methods for calculating confidence intervals for the benchmark dose. *Risk Anal.* 2004 Feb;24(1):31-40. doi: 10.1111/j.0272-4332.2004.00409.x PMID: 15027998.

OECD (2006), *Current Approaches in the Statistical Analysis of Ecotoxicity Data: A guidance to application (annexes to this publication exist as a separate document)*, OECD Series on Testing and Assessment, No. 54, OECD Publishing, Paris, <https://doi.org/10.1787/9789264085275-en>.

Pallocca, G., Moné, M. J., Kamp, H., Luijten, M., van de Water, B. and Leist, M. (2022) "Next-generation risk assessment of chemicals – Rolling out a human-centric testing strategy to drive 3R implementation: The RISK-HUNT3R project perspective", *ALTEX - Alternatives to animal experimentation*, 39(3), pp. 419–426. doi: <https://doi.org/10.14573/altex.2204051>.

Piegorsch, W.W., (2014) Model Uncertainty in Environmental Dose–Response Risk Analysis. *Statistics and Public Policy* 1:1, pages 78-85. <https://doi.org/10.1080/19466315.2012.757559>

Piegorsch, W.W., An, L., Wickens, A.A., Webster West, R., Peña, E.A. and Wu, W. (2013), Information-theoretic model-averaged benchmark dose analysis in environmental risk assessment. *Environmetrics*, 24: 143-157. <https://doi.org/10.1002/env.2201>

Portet, S., A primer on model selection using the Akaike Information Criterion, *Infectious Disease Modelling*, Volume 5, 2020, Pages 111-128, ISSN 2468-0427, <https://doi.org/10.1016/j.idm.2019.12.010>.

Ritz, C., Baty, F., Streibig, J.C., Gerhard, D. (2015) Dose-Response Analysis Using R. *PLoS ONE* 10(12): e0146021. <https://doi.org/10.1371/journal.pone.0146021>

Sand, S., Parham, F., Portier, C.J., Tice, R.R., Krewski, D. 2017. Comparison of points of departure for health risk assessment based on high-throughput screening data. *Environ Health Perspect* 125:623–633; <http://dx.doi.org/10.1289/EHP408>

Schmuck, M.R., Temme, T., Dach, K., et al. Omnisphero: a high-content image analysis (HCA) approach for phenotypic developmental neurotoxicity (DNT) screenings of organoid neurosphere cultures in vitro. *Arch Toxicol*. 2017;91(4):2017-2028. doi: <https://doi.org/10.1007/s00204-016-1852-2>

Scholze, M., Boedeker, W., Faust, M., Backhaus, T., Altenburger, R. and Grimme, L.H. (2001), A general best-fit method for concentration-response curves and the estimation of low-effect concentrations. *Environmental Toxicology and Chemistry*, 20: 448-457. <https://doi.org/10.1002/etc.5620200228>

Villeneuve, D.L., Coady, K., Escher, B.I., Mihaich, E., Murphy, C.A., Schlekot, T., Garcia-Reyero, N. High-throughput screening and environmental risk assessment: State of the science and emerging applications. *Environ Toxicol Chem*. 2019 Jan;38(1):12-26. doi: doi.org/10.1002/etc.4315. Epub 2018 Dec 20. PMID: 30570782; PMCID: PMC6698360.

West, R.W., Piegorsch, W.W., Peña, E.A., An, L., Wu, W., Wickens, A.A., Xiong, H. and Chen, W. (2012), The impact of model uncertainty on benchmark dose estimation. *Environmetrics*, 23: 706-716. <https://doi.org/10.1002/env.2180>

Wheeler, M.W., Park, R. M., Bailer A.J., Whittaker, C., (2015) Historical Context and Recent Advances in Exposure-Response Estimation for Deriving Occupational Exposure Limits, *Journal of Occupational and Environmental Hygiene*, 12:sup1, S7-S17, DOI: 10.1080/15459624.2015.1076934

Xu S., Chen M., Feng T., Zhan L., Zhou L. and Yu G. (2021) Use *ggbreak* to Effectively Utilize Plotting Space to Deal With Large Datasets and Outliers. *Front. Genet*. 12:774846. doi: 10.3389/fgene.2021.774846

Yandell, B.S. (1997). *Practical Data Analysis for Designed Experiments* (1st ed.). Routledge. <https://doi.org/10.1201/9780203742563>

Zhu, Y., Wang, T. and Jelsovsky, J.Z. (2007), Bootstrap Estimation of Benchmark Doses and Confidence Limits with Clustered Quantal Data. *Risk Analysis*, 27: 447-465. <https://doi.org/10.1111/j.1539-6924.2007.00897.x>

Conflict of interest

Kristina Bartmann, Arif Dönmez, Ellen Fritsche and Axel Mosig are co-founders of the start-up company DNTOX.

Author contribution

All authors read, commented, and approved the manuscript. **Hagen Eike Keßel**: study conception, data analysis, software development, supervision, figure design, writing of article. **Stefan Masjosthusmann**: study conception, figure design, supervision. **Kristina Bartmann**: investigation. **Jonathan Blum**: investigation. **Arif Dönmez**: software development, data analysis. **Nils Förster**: software development, data analysis. **Jördis Klose**: investigation. **Axel Mosig**: software development, supervision. **Melanie Pahl**: investigation. **Marcel Leist**: supervision. **Martin Scholze**: supervision, software development, data analysis, writing of article. **Ellen Fritsche**: study conception, supervision, funding acquisition, project administration, writing of article.

Acknowledgements

The authors are grateful to Katie Paul Friedmann (US EPA) for data integration and transfer to the ToxCast data base. We would like to thank Signe Marie Jensen (University of Copenhagen) for her help with proper application of the *drc* and *bmd* R packages for data analysis. This work was supported by the European Food Safety Authority (EFSA- Q - 2018 – 00308), the Danish Environmental Protection Agency (EPA) under the grant number MST-667-00205 and the project CERST (Center for Alternatives to Animal Testing) of the Ministry for culture and science of the state North-Rhine Westphalia, Germany (file number 233- 1.08.03.03- 121972/131 – 1.08.03.03 – 121972).

Impact of biostatistical data evaluation methods on hazard characterization using the neurosphere model as case study

Supplementary Data

1 Supplementary material and methods

1.1 Pre-processing of endpoints

Table S1: Pre-processing of endpoints

Pre-processed endpoints (left) are calculated with raw endpoints (right).

Pre-processed endpoint	Formular (raw endpoints)
neuronal differentiation [120h]	$\frac{\text{number of neurons}}{\text{number of all cells}}$
oligodendrocyte differentiation [120h]	$\frac{\text{number of oligodendrocytes}}{\text{number of all cells}}$
migration distance neurons [120h]	$\frac{\text{migration distance of neurons}}{\text{migration distance of all cells}}$
migration distance oligodendrocytes [120h]	$\frac{\text{migration distance of oligodendrocytes}}{\text{migration distance of all cells}}$
Viability UKN4	$\frac{\text{number of selected objects}}{\text{number of valid objects}}$
Viability UKN5	$\frac{\text{number of selected objects}}{\text{number of valid objects}}$

1.2 Regression models

Table S2 shows all parametric regression functions that defined the pool of candidate models for the best fit method. All models were applied by the *drm* function of the *drc* package with all key parameters (in the following written in cursive) set as follows: *formular* was given by a list of averaged replicate values and corresponding dose values, *type* was set to “continuous” for continuous data regression, *robust* was set to “mean” for least-square-estimation of continuous data, *fct* was set to one of the regression models listed in the *drc* syntax column of Table S2.

Table S2: Regression models

Model ²⁾	drc syntax	Model equation ¹⁾
general logistic	logistic2()	$f(x) = c + \frac{d - c}{(1 + \exp(b(\log(x) - \log(e))))^f}$
3-parameter log-logistic	LL.3()	$f(x) = 0 + \frac{d - 0}{1 + \exp(b(\log(x) - \log(e)))}$
4-parameter log-logistic	LL.4()	$f(x) = c + \frac{d - c}{1 + \exp(b(\log(x) - \log(e)))}$
2-parameter exponential	EXD.2()	$f(x) = 0 + (d - 0)(\exp\left(-\frac{x}{e}\right))$
3-parameter exponential	EXD.3()	$f(x) = c + (d - c)(\exp\left(-\frac{x}{e}\right))$
3-parameter Weibull	w1.3()	$f(x) = 0 + (d - 0)\exp(-\exp(b(\log(x) - e)))$
4-parameter Weibull	w1.4()	$f(x) = c + (d - 0)\exp(-\exp(b(\log(x) - e)))$

¹⁾ The parameters b, c, d, e of the model equations are estimated by concentration-response analysis: d is always estimated as model asymptote for the negative control response, c describing maximal responses at high concentrations (lower model asymptote), c and e provide flexibility in describing the location and steepness of the concentration-response pattern.

²⁾ Model name and abbreviation from *Analysis of Dose-Response Curves* (Ritz et al. 2016).

1.3 Classification model

Table S3: Specific, unspecific and viability-related endpoint correlations for the classification model

Specific endpoints are shown with their affiliated unspecific and viability-related endpoints. To gain DNT specific classifications, DNT specific endpoints are compared to either one or two unspecific endpoints (measuring general cell health). If an effect was detected in one of the affiliated viability-related endpoints, only cytotoxicity endpoints were used as reference.

Specific Endpoint	Unspecific Endpoint 1	Unspecific Endpoint 2	Viability-related Endpoint 1	Viability-related Endpoint 2
migration distance radial glia [72h]	cytotoxicity (migration) [72h]			
migration distance radial glia [120h]	viability (migration+differentiation) [120h]	cytotoxicity (migration+differentiation) [120h]	migration distance radial glia [120h]	cell number [120h]
cell number [120h]	viability (migration+differentiation) [120h]	cytotoxicity (migration+differentiation) [120h]	migration distance radial glia [120h]	cell number [120h]
proliferation by BrdU [72h]	viability (proliferation) [72h]	cytotoxicity (proliferation) [72h]		
proliferation by area [72h]	viability (proliferation) [72h]	cytotoxicity (proliferation) [72h]		
neuronal differentiation [120h]	viability (migration+differentiation) [120h]	cytotoxicity (migration+differentiation) [120h]	migration distance radial glia [120h]	cell number [120h]
oligodendrocyte differentiation [120h]	viability (migration+differentiation) [120h]	cytotoxicity (migration+differentiation) [120h]	migration distance radial glia [120h]	cell number [120h]
migration distance neurons [120h]	viability (migration+differentiation) [120h]	cytotoxicity (migration+differentiation) [120h]	migration distance radial glia [120h]	cell number [120h]
migration distance oligodendrocytes [120h]	viability (migration+differentiation) [120h]	cytotoxicity (migration+differentiation) [120h]	migration distance radial glia [120h]	cell number [120h]
neurite length [120h]	viability (migration+differentiation) [120h]	cytotoxicity (migration+differentiation) [120h]	migration distance radial glia [120h]	cell number [120h]
neurite area [120h]	viability (migration+differentiation) [120h]	cytotoxicity (migration+differentiation) [120h]	migration distance radial glia [120h]	cell number [120h]
Migration UKN2 [24h]	Viability UKN2 [24h]			
Neurite Area UKN4 [24h]	Viability UKN4 [24h]			
Neurite Area UKN5 [24h]	Viability UKN5 [24h]			

Table S4: Alerts used to flag classification data for manual evaluation

During endpoint classification, the data is checked for uncertainties and an alert is produced, if the resulting classification has high uncertainty. The according alerts (left) with reasoning (right) are shown.

Alert	Explanation
BMCU above concentration testrange	If the upper confidence limit of the BMC is above the tested concentration range, it has a high uncertainty and automated classification cannot be made.
high CI width	A high CI width indicates uncertainty in the BMC estimation and should therefore be checked for the final classification. A CI width was considered as high, if $BMCU/BMCL > (BMC*5)/(BMC/5)$.
no data for significance	Statistical significance could not be calculated (likely due to sample size of $n \leq 2$), but is needed for classification.
Issues with predict CI	Algorithmic failure did not allow the calculations of the BMC upper or lower limit. Therefore, no automated classification can be made.
no data	Necessary data for the classification was missing.

2 Supplementary results

2.1 Assay performance of all control compounds

To assess how changes in the data evaluation protocol might impact the evaluation of the DNT IVB's predictivity, 28 reference chemicals of known DNT and 17 negative control chemicals were selected (Masjosthusmann et al. 2020), with all 45 substances tested in the DNT IVB. A negative compound was considered as true negative (abbreviated as "TN" in the table below) if it was not classified as specific hit or borderline in any of the assays. Else, it was considered as false positive (FP). A positive compound was considered as true positive (TP), if it was classified as specific hit or borderline in at least one assay. Else, it was considered as false negative (FN).

Table S5: Assay performance for negative controls

Negative controls (n=17) used to determine the specificity of different protocols. Negative control compounds are listed in the first column. Remaining columns represent the compound classifications for each protocol used in this study. Specificity is given as true negative rate at bottom row and is calculated as the percentage of true negatives (TN) from all negative controls.

Negative controls	Standard Protocol	Replicate Mean	Control-Normalized	LL3rm	Bootstrapping	Model Averaging	BMR30+5 0
Amoxicillin	TN	TN	TN	TN	TN	TN	TN
Aspirin	TN	TN	FP	TN	TN	FP	TN
Buspirone	TN	TN	TN	TN	TN	TN	TN
Chlorpheniramine maleate	TN	FP	TN	TN	TN	FP	TN
D-Glucitol	TN	TN	TN	TN	TN	TN	TN
D-Mannitol	TN	TN	TN	TN	TN	TN	TN
Diethylene glycol	TN	TN	TN	TN	TN	TN	TN
Doxylamine succinate	TN	TN	TN	TN	TN	TN	TN
Famotidine	TN	TN	TN	TN	TN	TN	TN
Ibuprofen	TN	TN	TN	TN	TN	TN	TN
Metformin	TN	TN	TN	TN	TN	TN	TN
Metoprolol	TN	TN	TN	TN	TN	TN	TN
Penicillin VK	TN	TN	TN	TN	TN	TN	TN
Saccharin	TN	TN	TN	TN	TN	TN	TN
Sodium benzoate	TN	FP	TN	TN	TN	TN	TN
Warfarin	TN	TN	TN	TN	TN	TN	TN
Specificity (True Negative Rate in %)	100.0	87.5	93.8	100.0	100.0	87.5	100.0

Table S6: Assay performance for human positive controls

Human positive controls (n=9) used to determine the sensitivity of different protocols. Human positive control compounds are listed in the first column. Remaining columns represent the compound classifications for each protocol used in this study. Sensitivity is given as true positive rate at bottom row and is calculated as the percentage of true positives (TP) from all human positive controls.

Human positives	Standard Protocol	Replicate Mean	Control-Normalized	LL3rm	Bootstrapping	Model Averaging	BMR30+50
2,2',4,4'-Tetrabromodiphenyl ether	TP	TP	TP	TP	TP	TP	TP
Cadmium chloride	TP	TP	TP	TP	TP	TP	TP
Chlorpyrifos	TP	TP	FN	TP	TP	TP	TP
Dexamethasone	TP	TP	TP	TP	TP	TP	TP
Hexachlorophene	TP	TP	TP	TP	TP	TP	TP
Lead(II) acetate trihydrate	TP	TP	TP	TP	TP	TP	TP
Manganese(II) chloride	TP	TP	FN	TP	TP	TP	FN
Methylmercury(II) chloride	TP	TP	TP	TP	TP	TP	TP
PBDE 99	TP	TP	TP	TP	FN	FN	TP
Sensitivity (True Positive Rate in %)	100.0	100.0	77.8	100.0	88.9	88.9	88.9

Table S7: Assay performance for *in vivo* positive controls

In vivo positive controls (n=19) used to determine the sensitivity of different protocols. *In vivo* positive control compounds are listed in the first column. Remaining columns represent the compound classifications for each protocol used in this study. Sensitivity is given as true positive rate at bottom row and is calculated as the percentage of true positives (TP) from all *In vivo* positive controls.

<i>in vivo</i> positives	Standard Protocol	Replicate Mean	Control-Normalized	LL3rm	Bootstrapping	Model Averaging	BMR30+50
(+)-Ketamine hydrochloride	FN	FN	FN	FN	FN	FN	FN
(-)-Nicotine	FN	FN	FN	FN	FN	FN	FN
5,5-Diphenylhydantoin	FN	FN	FN	FN	FN	FN	FN
Acrylamide	TP	TP	TP	TP	TP	TP	TP
Chlorpromazine hydrochloride	TP	TP	TP	TP	TP	TP	TP
Deltamethrin	TP	TP	TP	TP	TP	TP	TP
Domoic acid	FN	FN	FN	FN	FN	FN	FN
Haloperidol	TP	TP	TP	TP	TP	TP	TP
Heptadecafluorooctanesulfonic acid potassium salt	TP	TP	TP	TP	TP	TP	TP
Maneb	TP	FN	TP	TP	TP	FN	FN
Methylazoxymethanol acetate	TP	TP	TP	TP	TP	TP	TP
Paraquat dichloride hydrate	TP	TP	TP	TP	TP	TP	TP
Perfluorooctanoic acid	FN	FN	TP	FN	FN	FN	FN
Sodium valproate	TP	TP	TP	TP	TP	TP	TP
Tebuconazole	TP	TP	TP	FN	TP	TP	TP
Tributyltin chloride	TP	TP	TP	TP	TP	TP	TP
Trichlorfon	TP	TP	TP	FN	TP	TP	TP
Triethyltin bromide	TP	TP	FN	FN	TP	FN	TP
all-trans-Retinoic acid	TP	TP	TP	TP	TP	TP	TP
Sensitivity (True Positive Rate in %)	73.7	68.4	73.7	57.9	73.7	63.2	68.4

Biostatistics and its impact on hazard characterization using in vitro developmental neurotoxicity assays

Hagen Eike Keßel, Stefan Masjosthusmann, Kristina Bartmann, Jonathan Blum, Arif Dönmez, Nils Förster, Jödis Klose, Axel Mosig, Melanie Pahl, Marcel Leist, Martin Scholze, Ellen Fritsche

Journal:	ALTEX
Impact factor:	6.250 (2021)
Contribution to the publication:	85%
	Study design, biostatistics and image analysis software development, data analysis and evaluation, performance and evaluation, writing of manuscript
Type of authorship:	Authorship
Status of publication:	Submitted 17 th Oct 2022

2.3 Establishment of a human cell-based in vitro battery to assess developmental neurotoxicity hazard of chemicals

Jonathan Blum, Stefan Masjosthusmann, Kristina Bartmann, Farina Bendt, Xenia Dolde, Arif Donmez, Nils Forster, Anna-Katharina Holzer, Ulrike Hübenthal, **Hagen Eike Keßel**, Sadiye Kilic, Jödis Klose, Melanie Pahl, Lynn-Christin Stürzl, Iris Mangas, Andrea Terron, Kevin M. Crofton, Martin Scholze, Axel Mosig, Marcel Leist, Ellen Fritsche

Die Entwicklungsneurotoxizität (DNT) ist ein wesentliches Sicherheitsproblem für alle Chemikalien des menschlichen Exposoms, doch DNT-Daten aus Tierstudien sind nur für wenige dieser Substanzen verfügbar. Daher werden dringend Testmethoden mit einem höheren Durchsatz als im Tierversuch und einer besseren Relevanz für den Menschen benötigt. Wir untersuchten daher die Durchführbarkeit einer DNT-Gefährdungsbeurteilung auf der Grundlage von sogenannten new approach methods (NAM). Eine in vitro-Batterie (IVB) wurde aus einzelnen NAMs zusammengestellt, die in den letzten Jahren entwickelt wurden, um die Auswirkung von Chemikalien auf verschiedene grundlegende Prozesse der Gehirnentwicklung zu untersuchen. Für alle Tests wurden menschliche neurale Zellen in verschiedenen Entwicklungsstadien entweder in 2D, 3D oder sekundärem 3D verwendet. Auf diese Weise konnten Störungen (i) der Proliferation neurale Vorläuferzellen (NPC), (ii) der Migration von Neuralleistenzellen, radialen Gliazellen, Neuronen und Oligodendrozyten, (iii) der Differenzierung von NPCs in Neuronen und Oligodendrozyten und (iv) des Neuritenwachstums peripherer und zentraler Neuronen in Verbindung mit Messungen der Zytotoxizität/Viabilität beurteilt werden. Die Durchführbarkeit eines konzentrationsabhängigen Screenings und einer zuverlässigen biostatistischen Verarbeitung der komplexen multidimensionalen Daten wurde mit einer Reihe von 120 Testsubstanzen untersucht, die eine Auswahl von vordefiniert positiven und negativen DNT-Substanzen enthielten. Die Batterie lieferte Hinweise (Hit oder Borderline) für 24 von 28 bekannten DNT-Toxika (82% Sensitivität), und die Spezifität lag bei >94%. Auf der Grundlage dieser Daten wurden Strategien entwickelt, wie die Daten im Rahmen von Risikobewertungsszenarien unter Verwendung integrierter Ansätze für die Prüfung und Bewertung verwendet werden können.



Contents lists available at ScienceDirect

Chemosphere

journal homepage: www.elsevier.com/locate/chemosphere

Establishment of a human cell-based in vitro battery to assess developmental neurotoxicity hazard of chemicals

Jonathan Blum^{a,1}, Stefan Masjosthusmann^{b,1}, Kristina Bartmann^b, Farina Bendt^b, Xenia Dolde^a, Arif Dönmez^b, Nils Förster^d, Anna-Katharina Holzer^a, Ulrike Hübenthal^b, Hagen Eike Keßel^b, Sadiye Kilic^a, Jödis Klose^b, Melanie Pahl^b, Lynn-Christin Stürzl^b, Iris Mangas^e, Andrea Terron^e, Kevin M. Crofton^f, Martin Scholze^g, Axel Mosig^d, Marcel Leist^{a,*}, Ellen Fritsche^{b,c,*}

^a In Vitro Toxicology and Biomedicine, Dept Inaugurated By the Doerenkamp-Zbinden Foundation, University of Konstanz, 78457, Konstanz, Germany

^b IUF - Leibniz Research Institute for Environmental Medicine, 40225, Düsseldorf, Germany

^c Medical Faculty, Heinrich-Heine University, 40225, Düsseldorf, Germany

^d Bioinformatics Group, Ruhr University Bochum, 44801, Bochum, Germany

^e European Food Safety Authority, PREV Unit, 43126, Parma, Italy

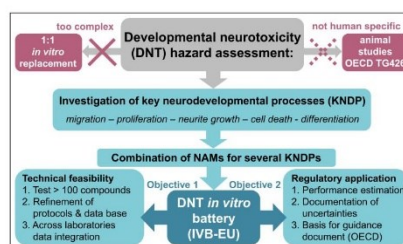
^f R3Fellows LLC, Durham, NC, USA

^g Institute of Environment Health and Societies, Brunel University London, UK

HIGHLIGHTS

- An in vitro testing battery (IVB) that allows screening of chemicals for developmental neurotoxicity (DNT) has been assembled.
- Performance estimates (>80% accuracy) have been obtained for the IVB, based on 45 negative/positive controls.
- Concentration-response data for altogether 120 compounds have been obtained for ten tests covering altogether 21 endpoints.
- Gaps of the IVB have been analyzed, and recommendations for the use of the IVB for regulatory testing have been put forward.

GRAPHICAL ABSTRACT



ARTICLE INFO

Handling Editor: Jian-Ying Hu

Keywords:

Testing battery
Stem cell
Brain development

ABSTRACT

Developmental neurotoxicity (DNT) is a major safety concern for all chemicals of the human exposome. However, DNT data from animal studies are available for only a small percentage of manufactured compounds. Test methods with a higher throughput than current regulatory guideline methods, and with improved human relevance are urgently needed. We therefore explored the feasibility of DNT hazard assessment based on new approach methods (NAMs). An in vitro battery (IVB) was assembled from ten individual NAMs that had been developed during the past years to probe effects of chemicals on various fundamental neurodevelopmental

* Corresponding author.

** Corresponding author. IUF - Leibniz Research Institute for Environmental Medicine, 40225, Düsseldorf, Germany.

E-mail addresses: jonathan.blum@uni-konstanz.de (J. Blum), marcel.leist@uni-konstanz.de (M. Leist), ellen.fritsche@iuf-duesseldorf.de (E. Fritsche).

¹ These authors contributed equally.

<https://doi.org/10.1016/j.chemosphere.2022.137035>

Received 8 July 2022; Received in revised form 20 October 2022; Accepted 24 October 2022

Available online 31 October 2022

0045-6535/© 2022 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

In vitro testing
DNT

processes. All assays used human neural cells at different developmental stages. This allowed us to assess disturbances of: (i) proliferation of neural progenitor cells (NPC); (ii) migration of neural crest cells, radial glia cells, neurons and oligodendrocytes; (iii) differentiation of NPC into neurons and oligodendrocytes; and (iv) neurite outgrowth of peripheral and central neurons. In parallel, cytotoxicity measures were obtained. The feasibility of concentration-dependent screening and of a reliable biostatistical processing of the complex multi-dimensional data was explored with a set of 120 test compounds, containing subsets of pre-defined positive and negative DNT compounds. The battery provided alerts (hit or borderline) for 24 of 28 known toxicants (82% sensitivity), and for none of the 17 negative controls. Based on the results from this screen project, strategies were developed on how IVB data may be used in the context of risk assessment scenarios employing integrated approaches for testing and assessment (IATA).

1. Introduction

Screening of chemicals for a potential neurodevelopmental toxicity (DNT) hazard has been recognized as a pressing need by several large governmental and international organizations concerned with consumer safety. For instance, the US EPA and the European JRC took important roles in the organisation of a conference series (TestSmart) that was devoted to the development of a DNT test strategy useful in a regulatory context (Coecke et al., 2007; Lein et al., 2007; Crofton et al., 2011; Bal-Price et al., 2012). Also EFSA and the OECD embarked on similar efforts (Fritsche et al., 2017). In this context, several experimental programs were launched to probe novel approaches and to accelerate their implementation (Crofton et al., 2012; van Thriel et al., 2012; Krug et al., 2013b; Bal-Price et al., 2015; Baumann et al., 2016; Fritsche et al., 2018; Harrill et al., 2018; Behl et al., 2019; Lupu et al., 2020; Pistollato et al., 2021; Sachana et al., 2021; Vinken et al., 2021; Koch et al., 2022).

DNT is a field of toxicology concerned with effects of chemicals on the developing nervous system. Several experimental and epidemiological studies (on metals, pesticides and drugs) link compound exposure during early life phases (of the embryo, fetus or child) to functional alterations of the nervous system in adolescents or adults (Grandjean and Landrigan, 2014; Smirnova et al., 2014; Bennett et al., 2016). A particular concern is the possible role of DNT in the increased frequency of neurodevelopmental disorders, such as autism-spectrum disorders (Grandjean and Landrigan, 2006, 2014; Bellinger, 2012; Modafferi et al., 2021). The assessment is particularly challenging due to the multitude of potential toxicity manifestations (structural and functional). Moreover, there may be a time offset between toxicant exposure (before or after birth) and manifestation of effects (Grandjean et al., 2019).

The traditional methods to evaluate DNT hazard potential are based on animal studies following the OECD (OECD, 2007) or U.S. EPA (USEPA, 1998) test guidelines. To date only about 180 compounds world-wide have been tested using these guidelines (Crofton and Mundy, 2021). Several factors contribute to the limited availability of such studies: extensive time (e.g. 1–2 years) and resource requirement; limited triggered testing by chemical alerts; the need to reduce animal use; and the limited regulatory requirement for DNT testing as compared to some other test guidelines (e.g., carcinogenicity). The data available suffer from many uncertainties, and they require species extrapolation from rodents to humans. Moreover, they provide limited information on toxicity mechanisms. This can make them difficult to use in human risk assessments (Makris et al., 2009; Tsuji and Crofton, 2012; Tohyama, 2016; Paparella et al., 2020).

The strategic concepts of next generation risk assessment and of “toxicology for the 21st century” (Leist et al., 2008; Thomas et al., 2018; Pallocca et al., 2022a) suggest reductions in use of animal studies and development of new approach methods (NAMs) for toxicity assessment. The non-animal test methods should ideally be based on human-relevant test systems, reduce costs, allow a high throughput of test chemicals, and provide information on the toxicity mechanisms of toxicants. Many recent activities on scientific and regulatory levels have been undertaken to apply this strategy to the field of DNT (Sachana et al., 2019).

The establishment of DNT NAMs followed two major principles

(Bal-Price et al., 2015; Aschner et al., 2017). First, a concept was developed on how complex in vivo events and their disturbances could be modeled by simplified in vitro systems. It was found that the biological process of nervous system development can be broken down to less complex key neurodevelopmental processes (KNDP). Moreover, it was assumed that the disturbance of any KNDP may lead to DNT in humans. On this basis, NAMs were developed for most of the crucial KNDP. The second principle was that the performance and robustness of the NAMs should be at a high level, so that data could be used with high confidence. The concept of test readiness was developed to provide a measure of the NAM validation status (Bal-Price et al., 2018; Krebs et al., 2019, 2020b), and several assays were deemed ready and suitable for use in chemical screening. They include: proliferation, migration and differentiation assays based on neurospheres (NPC1-5 test methods); the neurite growth assays NeuriTox and PeriTox; the neural crest migration assay (cMNC); and an assays for neural network formation and synaptogenesis (Masjosthusmann et al., 2020; Crofton and Mundy, 2021; Carstens et al., 2022). Instead of a formal OECD-type validation (e.g. skin sensitization NAMs (OECD, 2021; Strickland et al., 2022)), the concept of a fit-for-purpose biological validation based on regulatory needs has been suggested (Leist et al., 2012; Hartung et al., 2013; Judson et al., 2013; Cote et al., 2016; Griesinger et al., 2016; Bal-Price et al., 2018; Andersen et al., 2019; Masjosthusmann et al., 2020). Its application to DNT NAM involved: understanding of all technologies related to test systems and endpoint assessment; a comparison of pivotal in vitro signaling pathways to those relevant in vivo; and an assessment of the cellular presence of toxicity targets known to play a role for human DNT (Aschner et al., 2017; Bal-Price et al., 2018; Koch et al., 2022).

No individual NAM covers all key aspects of neurodevelopmental biology. Thus no single test will detect effects on all KNDP. Therefore, a battery of assays is needed, to sufficiently cover all DNT toxicants. In 2016, participants of a meeting jointly organized by the European Food Safety Authority (EFSA) and the organisation for Economic Co-operation and Development (OECD) agreed that “an in vitro testing battery (based on available DNT NAM) could be used immediately to screen and prioritize chemicals” (Fritsche et al., 2017). A test run for such a battery was planned, in order to evaluate the technical feasibility, to identify potential gaps and to provide data and experience for setting up a draft guidance on how to run battery testing, and how to interpret data therefrom (Crofton and Mundy, 2021). The purpose of this manuscript is to describe the first test run of a DNT in vitro test battery based on methods available in European laboratories (IVB-EU). Extensive raw data and method documentations can be found in a report by EFSA (Masjosthusmann et al., 2020), and the experience and learnings from the IVB-EU have led to the preparation of the draft of an OECD guidance document, which is currently (July 2022) under revision in member countries (Crofton and Mundy, 2021). However, the data from 10 assays on 120 compounds (including 28 positive and 17 negative controls) have not been made available to academia and the interested public in a peer-reviewed publication. The same applies to the preliminary performance evaluation of the IVB-EU as a whole and the considerations concerning further use. The purpose of this manuscript is to make this important information available, and to provide a basis for further developments in academia, industry and by regulatory institutions

concerned with NAM-based DNT testing.

2. Materials and methods

2.1. Chemicals

A list of screen compounds ($n = 120$) was assembled by a working group, using the member's experience as members/employees at the US EPA, EFSA or in OECD working groups. Compounds were selected to be chemically and biologically somewhat diverse and to reflect groups of compounds with concern for a potential DNT hazard. For instance, flame retardants and pesticides were included, as some compounds in these groups are known for biological properties of relevance to DNT. One aspect of the selection process was also to allow for diversity of effects on different fundamental neurodevelopmental processes (and respective assays), and it was important to cover the full spectrum from compounds with no or low evidence for DNT liability to compounds with rich background data to allow for a wide spread of screen results. A subset of compounds ($n = 28$) were included as positive controls for DNT hazard, based on human data or robust animal data (Grandjean and Landrigan, 2006, 2014; Mundy et al., 2015; Ryan et al., 2016; Aschner et al., 2017) (Fig. S1). Another subset ($n = 17$) were compounds considered as negative controls. They were selected for their safe use during human pregnancy or because the available extensive data on their toxicity gave no evidence (by observation or mechanism) of any effects related to DNT (at the test concentrations used) (Fig. S2). A description of chemicals, including exact chemical identity and suppliers is found in the suppl. file 2 - sheet 1.

2.2. Test methods

All test methods used for screening were selected based on their high readiness level (Bal-Price et al., 2018), as well as a very comprehensive test description compatible with the OECD Guidance Document GD211 for in vitro test method descriptions. These ToxTemp files (Krebs et al., 2019) are included in suppl. file 1. Below, only brief descriptions are given for a quick overview. Notably, most assays had at least two endpoints, and some assays were run in more than one version, e.g. measurement after 72 and 120 h.

UKN2 Assay (cMNC): The assay, is based on neural crest cells differentiated from hiPSC (Nyffeler et al., 2017). Cells were seeded into 96-well plates around a stopper. The stopper was removed after 24 h to allow migration into the cell free area. Cells were exposed to the test compound for 24 h, and then stained with calcein-AM and Hoechst H-33342. The number of migrated double positive cells was quantified independent of an observer by high content imaging and image analysis (RingAssay software; <http://invitro-tox.uni-konstanz.de>). The cell viability was also determined by an automated imaging algorithm. Concentration-response curves from this test were based on six test compound concentrations (plus solvent control).

UKN4 assay (NeuriTox): The assay is based on LUHMES cells that were cultured and handled as previously described (Lotharius et al., 2005; Scholz et al., 2011; Krug et al., 2013a). It assesses neurite outgrowth in central nervous system neurons (Delp et al., 2018). Cells were pre-differentiated for two days to commit them towards the neuronal fate. They were then re-seeded in 96-well plates and exposed to the chemical for 24 h. Viability and neurite area were determined by high-content imaging after staining with calcein-AM and H-33342. The neurite area was defined by a fully automated algorithm as the area of calcein-positive pixels minus the area of all cell soma (Stiegler et al., 2011). Concentration-response curves from this test were based on ten test compound concentrations (plus solvent control).

UKN5 Assay (PeriTox): The assay is based on immature sensory neurons differentiated from hiPSC as previously described (Hoelting et al., 2016; Holzer et al., 2022). The test measures neurite outgrowth in peripheral neurons. Frozen lots of peripheral neuron precursors were

thawed and seeded into 96-well plates. After 1 h, the cells were exposed to test chemicals for 24 h. Testing and endpoint measurements were exactly as for the UKN4 assay (despite 6 instead of 10 compound concentrations tested).

NPC1-5 Assays: The neurosphere assays (NPC1-5) are based on primary human neural progenitor cells (hNPCs; gestational week 16–19), that are grown as floating 3D neurospheres. Their growth and viability is assessed in the 3D neurospheres (NPC1). Alternatively, spheres can be plated onto a laminin-coated matrix, where the cells start migration and differentiation to form a secondary 3D co-culture. The latter approach allows the simultaneous assessment of radial glia migration (NPC2a), neuronal differentiation (NPC3), neuronal migration (NPC2b) and neurite outgrowth (NPC4) as well as oligodendrocyte differentiation (NPC5) and their migration (NPC2c) by fully automated high content imaging. Data were obtained and analyzed from recorded microscope images by a dedicated image processing software, trained on positive and negative control images, as described earlier in detail (Forster et al., 2022; Koch et al., 2022).

For the NPC1 assay, spheres (0.3 mm) were plated in 96-well plates (U-bottom; 1 sphere/well) and directly exposed to the test compound (in proliferation medium). DNA synthesis was assessed as functional endpoint after 3 days in vitro (DIV), using a luminescence-based bromodeoxyuridine (BrdU) ELISA (Nimtz et al., 2019). Cytotoxicity was assessed as a membrane integrity assay (CytoTox-ONE Assay) measuring the LDH release into the supernatant.

For the NPC2-5 assays, spheres (0.3 mm) were plated in poly-D-lysine/laminin-coated 96-well plates (F-bottom; 1 sphere/well) and directly exposed to the test compounds (in differentiation medium). Under control conditions, NPCs migrate radially out of the attached sphere and differentiate into radial glia, neurons and oligodendrocytes. Data were obtained after 72 h and 120 h. After 72 h (3 DIV), bright field images were taken of live cell cultures, and radial glia migration (NPC2a [72 h]) was assessed using ImageJ software. The medium was partially removed (50%) and used to assess cytotoxicity (CytoTox-ONE Assay). To continue the assay, the medium was replenished and cells were allowed to further differentiate and migrate for 48 h. At 5 DIV, cells were fixed and stained for TUBB3 (neuronal marker), O4 (oligodendrocyte marker) and Hoechst H-33258 (nuclear marker). The endpoint assessment was done by high content imaging followed by different image analysis algorithms. Neuronal and oligodendrocyte differentiation (NPC3 and NPC5) was assessed as the number of all TUBB3-positive and O4-positive cells in percent of the total number of nuclei in the migration area. Neurons and oligodendrocytes were automatically recognized by a machine learning software based on convolutional neural networks (Forster et al., 2022). The high-content image analysis software Omnisphere was used to determine radial glia migration (NPC2a [120 h]), neuronal migration (NPC2b) and oligodendrocyte migration (NPC2c) as well as neuronal morphology (NPC4a: neurite length; NPC4b: neurite area) (Schmuck et al., 2017). Cytotoxicity was assessed from samples of medium removed before the fixation by the CytoTox-ONE LDH Assay. Some additional cell viability data were obtained by using a resazurin reduction assay (CellTiter-Blue Assay). Concentration-response curves from all these tests were based on seven test compound concentrations.

2.3. Screen strategy

Most of the compounds ($n = 75$) were provided by EPA's ToxCast chemical contractor (Evotec, South San Francisco, CA) in v-bottom 96 well plates. Separate plates were provided for different assays, and volumes shipped ranged from 50 to 300 μ l as DMSO stock solutions (always 20 mM). Other compounds were obtained from commercial sources (indicated in the suppl. 2 Excel file). In some of these cases stock solution was higher than 20 mM and compounds were dissolved in water if they were highly water-soluble (e.g. valproic acid). The University of Konstanz robotics platform was used to either produce replicates of the

master plate for different screening runs and different assays (UKN assays) or to directly prepare the compound dilutions (1:3 steps) in the media in 96-well plates (NPC assays). Operators were blinded to the compound identity. For the UKN assays serial dilutions (1:3 steps) were prepared from the cloned master plates for each compound in DMSO on 96-well plates, and each of these stocks was transferred to a pre-dilution plate. On these plates compounds were diluted 1:3 in medium plus 1% DMSO to have constant levels of DMSO among all concentrations. Finally, pre-dilutions were transferred to assay plates with cells (e.g. 20 μ l transfer to 180 μ l cells corresponding to 1:10) in medium to a maximum DMSO level of 0.1% in each assay. Exact volumes and pre-dilutions were assay-dependent and are detailed in ToxTemps; suppl. file 1. Some compounds were tested in an adapted concentration range (e.g. it is known that valproic acid is a human teratogen and DNT toxicant at clinically used concentrations of 0.5–1 mM. Therefore, higher concentrations were also tested, and master stocks were prepared accordingly).

For some assays (e.g. UKN2), a pre-screening step was included, in which only 1–2 (highest) test compound concentrations were run. When they showed no effect, screening was ended. When there was an effect (at least 20% change of endpoint), a full concentration-response was obtained. Pre-screen and full concentration-response screen were performed three times independently for all assays. For the UKN assays this meant the use of different cell lots for each run, for the NPC assays it meant the use of cells from different donors and/or passages for each run. Each screen run contained 2–6 technical replicates (details in ToxTemps; suppl. file 1). In some cases, follow-up tests were run, when e.g. only the highest concentration showed a response. Then new stocks were produced, and the concentration range was extended to 60 or 100 μ M, depending on the solubility of the compound.

2.4. Data analysis

A fully automated data analysis workflow was implemented on the programming platform R (Keßel, 2022). Original code and source files are available on GitHub at (<https://github.com/iuf-duesseldorf/fritsche-lab-CRStats>). It included the following steps and outputs: (1) Pre-processing of data, where required by the definitions of the assay endpoints (see ToxTemps; suppl. file 1). For instance, the background signal was subtracted from all data points for the BrdU fluorescence readings. (2) Normalization of test compound data to the median of solvent controls. (3) Calculation of the median of the replicates for each experimental condition. (4) Concentration response fitting of the data for each compound. The best-fitting model (general logistic, 3-parameter log-logistic, 4-parameter log-logistic, 2-parameter exponential, 3-parameter exponential, 3-parameter Weibull, 4-parameter Weibull) was selected by the AKAIKE information criteria (Ritz et al., 2015; Jensen et al., 2020). (5) Re-normalization of the data, so that the upper asymptote of the selected curve fit was at 100% (Krebs et al., 2018; Kappenberg et al., 2020). (6) Calculation of the mean re-normalized values for each condition across independent test runs. (7) Concentration response fitting of the data for each compound. The best-fitting model (general logistic, 3-parameter log-logistic, 4-parameter log-logistic, 2-parameter exponential, 3-parameter exponential, 3-parameter Weibull, 4-parameter Weibull) was selected by the AKAIKE information criteria. (8) Determination of the benchmark concentration (BMC) as the point of the concentration-response curve that intersected with the benchmark response level (BMR). The BMR was determined and described for each assay (see ToxTemp; suppl. file 1), based on a biological and statistical rationale. It marked the extent of response considered to be statistically significant and toxicologically meaningful. It thus depended on the endpoint and on the base line noise. For most functional endpoints it was set at 75% (= 25% reduced normal function). For some assays it was set at 70% (higher baseline noise). For some viability measures it was set at 90% (a deviation of >10% was considered to potentially influence the functional endpoint). (9) After

determination of the BMC, the upper (BMCU) and lower limit (BMCL) of its 95% confidence interval were calculated (Krebs et al., 2020a).

2.5. Hit definitions and prediction models

The prediction models (Worth and Balls, 2001; Leist et al., 2010; Griesinger et al., 2016; Schmidt et al., 2017; Bal-Price et al., 2018; Krebs et al., 2020b) of the NAM used in the IVB-EU had been defined during the original test setup, as documented in the literature and the ToxTemp files. A key feature of all assays was that they had a specific functional endpoint (related to a KNDP) and an endpoint characterizing compound effects on cell viability. Within each NAM, a compound was considered a specific hit (toxicant), when it affected the functional endpoint at least at one concentration that did not affect viability (Fig. S3). Notably, this does not mean that specific cytotoxicity of a given cell population (e.g. neural crest cells) would not lead to DNT. However, specific toxicity to a subpopulation can only be determined across assays, not within one assay. At present, a procedure for such a cross-IVB interpretation has not been established. Within a given assay, cytotoxicity makes the interpretation of the functional endpoint difficult. Therefore, (i) functional endpoint data were only used for concentrations that were non-cytotoxic, and (ii) specific cytotoxicity to subpopulations was not considered in this first application of the IVB-EU. For the UKN assays, specific effects were determined by the ratio of benchmark concentrations for the functional endpoint (e.g. neurite growth in UKN4) and cytotoxicity (e.g. a 4-fold offset for UKN4). For the NPC assays, specific toxicity was assumed when the 95% confidence intervals of the functional endpoint and the viability endpoint did not overlap. As the separation between “hit” and “non-hit” leads to binary data with high uncertainties at the hit/non-hit boundary (Leontaridou et al., 2017; Delp et al., 2018), we introduced a borderline category for transition compounds (e.g. when confidence intervals in NPC assays overlapped by > 10%). Thus, a given compound was classified in each assay as “no hit”, “unspecific hit”, “specific hit” or “borderline hit” (Fig. S3).

2.6. Performance parameters

A set of 45 reference compounds (28 DNT positives; 17 DNT negatives) was used for a preliminary evaluation of the IVB-EU predictivity (more may be added in the future). Various hit definitions were used (e.g. only specific hits, or specific + borderline hits). If a positive control was a hit, it was considered true positive (TP), if it was not a hit, it was considered a false negative (FN). If a negative control was a hit, it was considered a false positive (FP) and if it was not a hit, it was considered a true negative (TN). Using these four numbers (FP, FN, TP, TN), the following performance parameters were defined:

$$\text{sensitivity} [\%] = \frac{TP}{(TP + FN)} * 100$$

$$\text{specificity} [\%] = \frac{TN}{(TN + FP)} * 100$$

$$\text{accuracy} = \frac{(TP + TN)}{(TP + TN + FP + FN)} * 100$$

$$\text{balanced accuracy} = \frac{\text{sensitivity} + \text{specificity}}{2}$$

$$\text{positive predictive value (PPV)} = \frac{TP}{(TP + FP)} * 100$$

$$F1 \text{ score} = \frac{2}{\frac{1}{\text{sensitivity}} + \frac{1}{\text{PPV}}} = \frac{1}{2} * (\text{sensitivity} + \text{PPV})$$

$$\text{Matthews correlation coefficient (MCC)} = \frac{(TP * TN) - (FP * FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

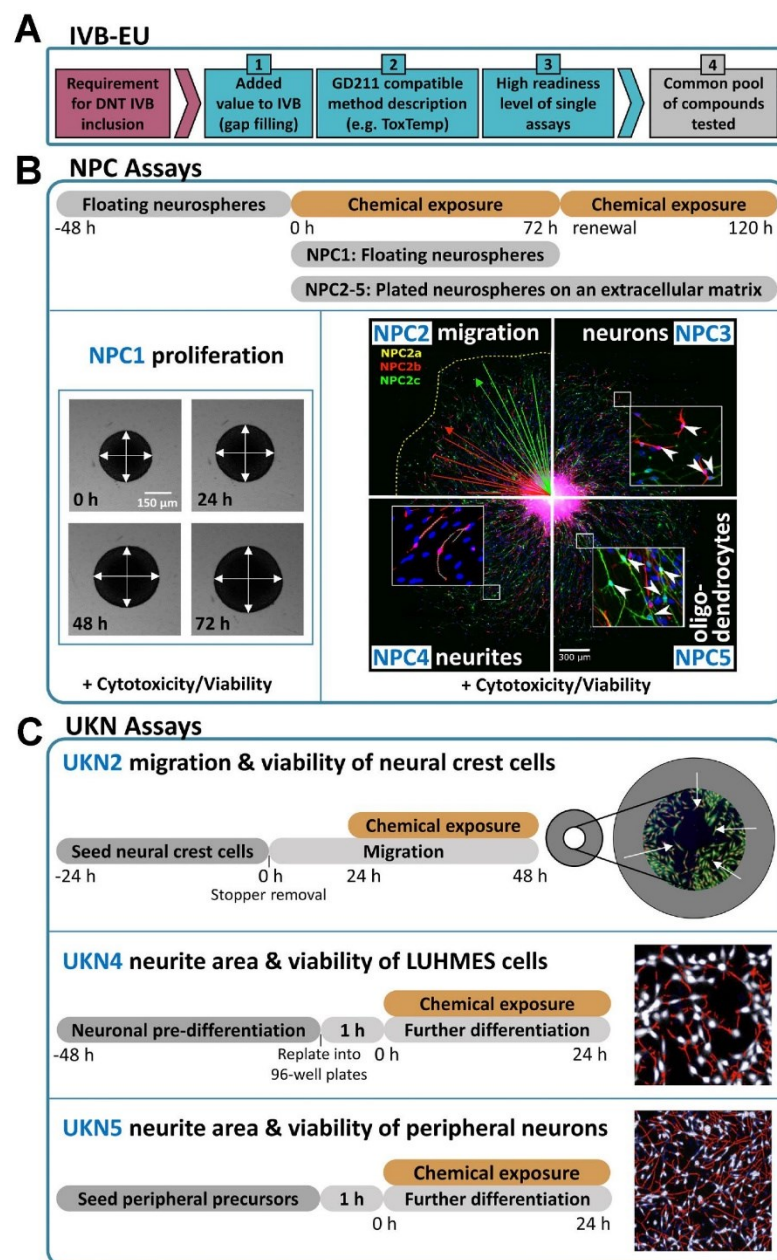


Fig. 1. Requirements and composition of the IVB-EU. (A) Criteria for assays to be included in the DNT test battery designated here IVB-EU. Criteria 1–3 were applied to this study. Criterion 4 was fulfilled in the course of this study and is suggested to be considered for future battery expansion. GD211 = OECD guidance document 211 on documentation of in vitro methods. (B) Schematic representation of the assays based on human neural progenitor cells (NPC) and their progeny. The general test system generation and exposure scheme is indicated on top. For the NPC1 test, floating neurospheres were exposed to toxicants for 72 h, and bromodeoxyuridine (BrdU) incorporation was used as endpoint for proliferation of NPC. For the NPC2-5 assays, neurospheres were plated and allowed to form secondary co-cultures of various cell types. Endpoints related to migration (NPC2), neuronal differentiation (NPC3), neurite growth (NPC4) and oligodendrocyte formation (NPC5) were assessed after 120 h by immunostaining and high content imaging. (C) Schematic representation of UKN assays. Cell types used and exposure schemes are indicated. Viability and migration of the cells in all assays were determined simultaneously by automated high content imaging after staining of the cell cultures with calcein-AM and Hoechst H-33342. The UKN2 assay evaluated the migration of neural crest cells into an empty circular area. The UKN4/UKN5 assays evaluated neural outgrowth of central nervous system and peripheral nervous system immature neurons. Detailed descriptions of NPC and UKN assays are given in the ToxTemps.

2.7. Data accessibility

The full raw data set from the IVB-EU has been entered into the ToxCast data base and is available in a machine-readable format used by many computational toxicologists after the fall 2022 ToxCast release (US EPA ORD, 2022).

3. Results and discussion

3.1. The DNT in vitro battery (IVB)

A large panel of assays with direct or indirect relevance to DNT can be found in the literature. Criteria needed to be developed to select a prototype battery of assays that was large enough for the main objective of this study, i.e. providing a basis for preparation of a general technical guidance document on battery testing for regulatory applications. At the same time, reasons of feasibility and limited resources called for keeping the number of NAMs included in the test run low. Experts with a regulatory background (from the US and Europe) were involved in the selection. The overall plan was to start testing in some European laboratories on a core battery (IVB-EU) of fully ready NAMs, and then to combine data on the same set of compounds with tests established at the US EPA. The three main selection criteria for the DNT NAMs were: (i) complementarity, (ii) documentation, and (iii) the readiness level (Fig. 1A). The first point meant that the assays were selected in a way to fill gaps of knowledge and to cover many KNDPs. It was also considered here to use assays for overlapping biological functions to learn about their orthogonality for later designs of tiered testing and sub-batteries. The second point referred to the availability of method documentations useful at a regulatory level (i.e. defined by OECD guidance document GD211) for the use of NAMs. Linked to this was the third criterion which referred to the technical performance of the NAMs, and the level of confidence into their predictivity and relevance. These issues are in some legislations referred to as validation state (Leist et al., 2012; Hartung et al., 2013; Judson et al., 2013; Cote et al., 2016; Griesinger et al., 2016; Bal-Price et al., 2018; Andersen et al., 2019; Masjoshthmann et al., 2020). In the selection of assays for the IVB-EU, we used a more flexible definition, termed “readiness” (Krebs et al., 2020b; Patterson et al., 2021). The assays used here all had undergone such an evaluation (Bal-Price et al., 2018; Klose et al., 2021a; Koch et al., 2022).

An additional criterion important for development of additional assays, now recommended in the draft OECD DNT-IVB test guideline is use of a common pool of test compounds (Fig. 1A). Ten assays fulfilled all criteria, and they were considered to be suitable for forming the IVB-EU. In addition to the above points, all selected assays use human cells, cover four major KNDP, reflect seven different brain cell types and represent different neurodevelopmental stages (Fig. 1B and C; Fig. 2).

To obtain an overview of test battery relevance and predictivity, a gap analysis was performed. Comparison of the included tests with the known neurodevelopmental processes showed that some KNDP are currently not covered by the IVB-EU. These include very early developmental processes such as stem cell differentiation into neural progenitor cells and subsequent neural tube construction, as well as processes necessary for neuronal circuit building, like formation, maturation and function of neuronal networks. As such gaps may reduce the sensitivity of DNT predictions, we explored the availability of assays that fulfill the IVB-EU inclusion criteria and could become part of an expanded full battery (Fig. 2). Many assays for network formation have indeed already shown to be at high readiness, yet these are based on rat cortical cells (Carstens et al., 2022) calling for human cell-based neuronal network formation assays. The early embryonal stages of neural development may be covered by the UKN1 assay (Dresler et al., 2020; Meisig et al., 2020). Some functional endpoints related to non-neuronal cells are also desirable for the IVB, as these cells (astrocytes, microglia, myelinating oligodendrocytes, microvascular endothelial cells) do not only have support and immune function, but rather participate in multiple neurodevelopmental processes (Allen and Lyons, 2018). Several 3D systems have been described to include the necessary cell types (Brull et al., 2020; Chesnut et al., 2021; Nunes et al., 2022), but still need some development to meet basic inclusion criteria (set up of test methods, throughput, documentation) for the IVB. The same applies to dedicated assays to investigate neurotransmitter systems (e.g. glutamate and acetylcholine signaling) (Klima et al., 2021; Loser et al., 2021b). However, a large part of signaling systems is covered already by the recent development of neural network formation assays (Frank et al., 2017; Nimtz et al., 2020). An interesting endpoint to comprehensively capture neuronal differentiation is transcriptome profiling (Pallocca et al., 2016; Shinde et al., 2017; Simon et al., 2019; Dresler et al., 2020; Meisig et al., 2020; Hu et al., 2022). This was exemplified here by the UKN1 assay. Modern high throughput sequencing techniques (Simon

KNDPs	Precursor proliferation	Migration	Differentiation	Neurite outgrowth	Neural network formation and function	Cell activation & stimulation
	Death of specific cell populations					
In vitro methods	NPC1 NPCs	NPC2a radial glia	NPC3 neuron	NPC4 CNS neuron	MEA based assays *	transport activity
	hiPSC based NPCs *	NPC2b neuronal	NPC5 oligo-dendrocytes	UKN4 CNS neuron	maturation & synaptogenesis	neuro-transmitter signals *
	radial glia	NPC2c oligo-dendrocytes	(NPC6) oligo-dendrocyte maturation *	UKN5 PNS neuron	myelination	inflammation (glia activation)
		UKN2 neural crest	NEPs (UKN1 & RoFA) *			signal transduction gaps
			astrocytes & radial glia *			
	*established with good readiness level for toxicity testing, but not part of initial IVB-EU					
						Covered in IVB-EU
						Potential gap

Fig. 2. Key neurodevelopmental processes (KNDP) covered by IVB-EU. Categories of KNDPs, according to Bal-Price et al., 2018 are listed on top. Specific cell death in a neurodevelopmental sub-population may either be considered a KNDP or an adverse effect. As it is measured as endpoint in all assays of other KNDP, it was considered to be broadly covered by the IVB-EU without a dedicated own assay. The lower part of the figure indicates NAM (designated here: in vitro methods) that are related to the respective KNDP on top of each column. The coverage of KNDPs by assays that are part of the current IVB-EU is shown (bold). For some KNDPs, more than one test was available. The reason was that several distinct subpopulations e. g. migrate (radial glia, neurons, oligodendrocytes and neural crest cells) or grow neurites (different types of CNS and PNS neurons). Potential gaps of the current IVB-EU are shown as assays in the non-bold in vitro method boxes. Assays that have already been established in the co-authors' labs are indicated by asterisks. They may be included in an extended version of the IVB, once they fulfill all inclusion criteria (Fig. 1). CNS: central nervous system; hiPSC: human induced pluripotent stem cells; NEP: neuroepithelial precursor; NPC: neural progenitor cell; MEA: microelectrode array; PNS: peripheral nervous system; RoFA: rosette formation assay.

et al., 2019; Jaklin et al., 2022; Spreng et al., 2022) now allow sufficient throughput for screening applications and it is likely that such assays will add additional information to the IVB in the future.

3.2. Readiness overview

The readiness of the assays of the DNT IVB was assessed on two tiers: first, the readiness of individual assays, as assessed earlier in individual publications, was an inclusion criterion (Fig. 1) of the IVB-EU. Second, the readiness of the overall battery and the performance of the assays under screening conditions was evaluated.

Concerning the first point, the underlying considerations are briefly re-iterated here, as they impinge on the interpretation and on the overall confidence into data from the NAMs of the IVB-EU. As for all toxicological assays, relevance, predictivity and reliability/robustness were considered. A major focus was put on the latter point, as suggested earlier (Leist et al., 2014; Krebs et al., 2019; Pallocca et al., 2022b). Earlier publications (summarized in Masjosthusmann et al. (2020)), and the ToxTemp (suppl. file 1) give more background information. One aspect helping to keep typical sources of variability low is that the selected IVB-EU assays all used a fully automated data capturing and evaluation procedure. However, the ultimate proof of the pudding for robustness, a blinded inter-lab comparison study, still has to be done for the assays.

When simple methods for 1:1 replacement of acute toxicity endpoints were evaluated, relevance and predictivity have been defined as separate aspects of NAMs. However, this concept has been modified for complex endpoints and batteries. In such more complex cases, the predictivity of a single NAM (for a given regulatory endpoint derived from animal studies) cannot be calculated, and the aspects of predictivity and relevance are strongly intertwined (Escher et al., 2022). In such cases, a scientific validation process is suggested that builds on two pillars: (i) comparison of the biological basis of the test system to that of the modeled human biology, and (ii) comparison of pathway modulations that lead to endpoint changes in the NAM to pathway changes known to be relevant to the respective human pathophysiology (Hartung, 2007; Leist et al., 2012; Hartung et al., 2013; Bal-Price et al., 2018; Piersma et al., 2018; Patterson et al., 2021). For the NAMs included in the IVB-EU, the test systems have been extensively documented and compared to the respective human developing nervous system counterparts. This involved the levels of cell morphology, cell function, and cell markers (see ToxTemp; suppl. file 1). Moreover, the relevant systems were profiled for their respective transcriptomes (Krug et al., 2014; Hoelting et al., 2016; Pallocca et al., 2017; Gutbier et al., 2018; Masjosthusmann et al., 2018; Klose et al., 2021a, 2021b, 2022). Also, the responses of the NAMs to modulation of signaling pathways relevant for brain development have been investigated by the use of compounds known to specifically affect signaling pathways (for overview: Klose et al. (2021b); Koch et al. (2022); Krebs et al. (2020b); Masjosthusmann et al. (2020)). A high-level summary of the responses to such “mechanistic tool compounds” is summarized in Fig. S4. One example is the Notch pathway, which determines a crucial switch between neurogenesis and oligodendrogenesis in vivo. By using the Notch pathway inhibitor DAPT, we can mimic this differentiation switch also in vivo with the NPC3/5 tests (Koch et al., 2022). Another illustrative example is the Rho pathway, which is involved in neurite growth in vivo. Activation of the RhoA kinase by narciclasine decreases neurite outgrowth in the NPC4, UKN4 and UKN5 assays. This successful characterization of neurodevelopmentally-relevant signaling in the IVB-EU assays is considered as the physiological basis and qualitative evidence for relevance and predictivity.

While the above-mentioned steps were important for the selection of NAMs and for giving confidence into their individual function within the IVB-EU, we also engaged in an effort to obtain information on the validity of the entire IVB-EU, as a battery. We considered the key parameters robustness, predictivity and relevance (Hartung et al., 2004;

Pallocca and Leist, 2022). Concerning relevance, it was mainly considered how many cell types and how many signaling pathways important for brain development were covered. A gap analysis showed that there was a need for few additional cells (e.g. microglia) and for some additional functions (e.g. neuronal network formation, astrocyte function). Moreover, more coverage of signaling (e.g. BDNF pathway and nicotinic signaling pathway) would be desirable. However, most relevant cell types were already represented, and many pathways known to be affected by toxicants were shown to be identifiable by at least one assay

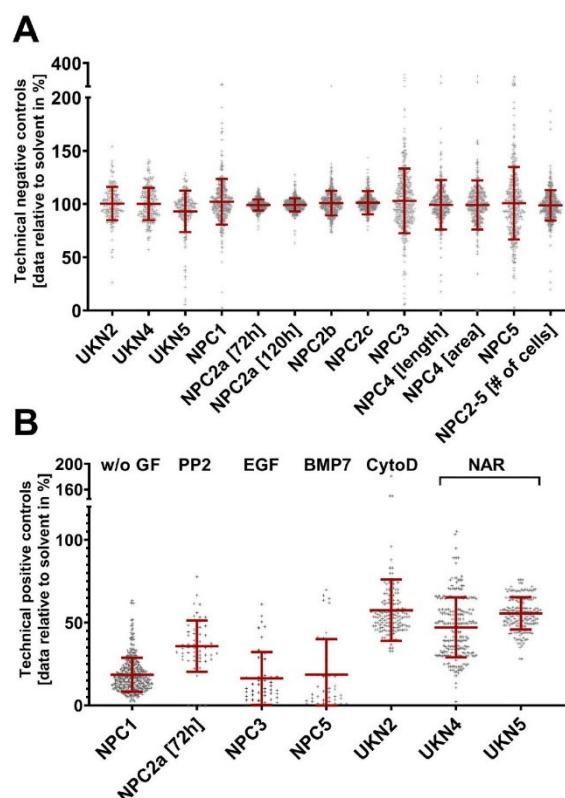


Fig. 3. Baseline noise and signal variation of acceptance controls in the IVB-EU assays. All tests were performed in a way so that each assay plate or experimental run contained wells with (i) negative controls, and at least one (ii) positive control. The reading of (ii) vs. (i) was used as acceptance criterion of the respective plate for UKN2, 4 and 5. If the positive control was not in a pre-specified range, the plate data were not included in screen results and measurements were repeated. Depending on the assay, plates contained different numbers of compounds. For some tests, the different concentrations of a given compound were on different plates. Thus, some plates contained the (iii) lowest concentration of a compound, and some did not. (A) To obtain a measure of inter-plate and intra-experimental variability of the baseline signal, the lowest concentration of each test compound (iii) was compared to the solvent control (i) on each plate. Altogether >200 data points were obtained for each IVB-EU endpoint from the testing campaign. For easier overview, the means \pm SD are indicated on top of the data points. (B) For each plate, the reading of the positive controls (ii) was compared to that of the negative controls (i) and normalized to negative control readings. The means \pm SD of data for positive controls are given for the IVB-EU endpoints. The compounds used to set acceptance criteria were as follows: w/o GF: without growth factor (omission of normally present growth factors in the positive control well); PP-2: SRC-kinase inhibitor; EGF: epidermal growth factor; BMP7: bone morphogenetic protein 7; CytoD: cytochalasin D; NAR: narciclasine. Details on concentrations are found in the ToxTemp (suppl. file 1).

(Fig. 2; Fig. S4).

One estimate for the robustness of screening results from the test battery is the baseline noise level of the NAM. As the results of all assays are normalized to solvent control data (which are set to 100%, and therefore do not vary by default), we used a surrogate baseline data set: from each concentration-response curve of the screen compounds, we selected the lowest concentration and assumed that this was in most cases a no-effect concentration. This assumption was consistent with the average of all these data points being about 100% for all assays. With this approach it was possible to visualize the baseline noise (as standard deviation around the average signal, Fig. 3A). From such data, we also calculated the assay-specific coefficients of variation (CoVs, see ToxTemp; suppl. file 1). As a second measure of robustness, we evaluated the responses of each test to the concurrent positive technical controls, which were run along on each plate/for every experiment during the screen (Fig. 3B). The positive controls were also used to determine acceptability of the respective plates/experiments for further evaluation. The plates/experiments, for which the acceptance criteria (see ToxTemp; suppl. file 1) were not met (<10% for all tests), were discarded.

3.3. Performance analysis

The predictivity of the IVB as a whole is a key feature of its regulatory applicability. This was examined as follows: First, all of the above discussed aspects of mechanistic validation were considered: the biology and pathophysiology covered by the entirety of assays of the IVB-EU suggested a high, but not perfect, biological applicability domain. This pointed at a sufficient predictivity for many purposes.

In a second step, we evaluated the capacity of the IVB-EU to correctly identify negative and positive controls. A list of 45 such calibration compounds was assembled from various literature references (Kadereit et al., 2012; Grandjean and Landrigan, 2014; Mundy et al., 2015; Aschner et al., 2017; Paparella et al., 2020; Crofton and Mundy, 2021). The challenges and shortcomings of this approach have been widely discussed (see above references), but our compound selection appeared to be a good compromise based on the present state of knowledge (Fig. 4A and B).

Prediction models for test batteries are an active field of research, and many possibilities exist (tiered approaches, Bayesian models, Boolean rules and decision trees). The difficulty to agree on the defined approaches for the small (3 NAM) battery used to predict dermal sensitization exemplifies these difficulties (Strickland et al., 2022). Here, we used a simple Boolean rule to define a battery hit as any compound that was a hit in one of the included DNT IVB-EU NAMs. A negative was defined as a compound not being a hit in any of the assays. This rule allows for a high transparency and simplicity. For statistical reasons, this battery prediction model may be associated with a high false discovery rate (testing for multiple endpoints considered to be independent). This was considered to be acceptable for screening and prioritization use. Moreover, the use of full concentration-response curves (instead of single data points) for definition of all positive hits reduced this problem. The false discovery rate was further reduced by our use of data from three independent experiments.

The 28 positive controls were used to obtain a preliminary measure of assay sensitivity (to be refined with time and the addition of more control compounds). We used different stringencies of hit definitions to obtain an estimate of the IVB-EU performance with respect to detection of DNT toxicants. When only the specific hits (compounds causing functional impairment at non-cytotoxic concentrations) were counted, the sensitivity of the IVB-EU was 68%. When borderline hits were included, this went up to 82%. When also cytotoxic compounds were included in the "hits", a further increase was observed. However, interpretation of cytotoxic compounds is presently not part of the IVB prediction model (Fig. 4A,C).

The 17 negative controls were used to obtain data on specificity.

When specific and borderline hits were counted, a value of 100% was obtained. Specificity dropped to 94%, when also cytotoxic effects were counted as "hit" (Fig. 4B,C).

Altogether, these preliminary performance estimates indicate that a balanced accuracy of about 80% or higher can be reached with the present IVB-EU. Based on the set of positive/negative control compounds, several additional performance measures were calculated (Fig. 4C) and it is particularly noteworthy that the IVB-EU had a high positive predictive value (PPV). This supports the conclusion that compounds identified as a hit should be prioritized for further evaluation of potential human hazard. Such data would also suggest that such chemicals better be excluded at early stages from further development (e.g. as a drug).

Nicotine serves as a good example for gaps in the IVB-EU, identified by the performance evaluation. It was identified as a false negative in the battery, and thus is indicative of a shortcoming with respect to sensitivity. The major action of nicotine is the stimulation of ionotropic acetylcholine receptors, and the IVB-EU does not (yet) include NAMs that would cover this biological function. This information is important when it comes to the interpretation of data from compounds that target nicotinic receptors, like neonicotinoid insecticides (Sheets et al., 2016; Loser et al., 2021a). Assays that fill these gaps are already under development (Fig. 2), and inclusion of assays based on zebra fish embryos and other model organisms (e.g. *C. elegans*) are considered an additional approach to close battery gaps (Atzei et al., 2021; Dasgupta et al., 2022).

Another limitation of the DNT IVB-EU is hard to overcome: the number of control compounds with clearly documented human effects is very limited, and also the compounds having been tested in DNT guideline studies in animals is small (Aschner et al., 2017). For this reason, performance metrics on the basis of currently-available control-compound predictivity will remain superficial. A way forward is to focus more on mechanistic validation approaches (Leist et al., 2012; Judson et al., 2013; Cote et al., 2016; Griesinger et al., 2016; Bal-Price et al., 2018; Andersen et al., 2019; Masjosthusmann et al., 2020) to gain further confidence into the predictivity of the battery for human adversities.

A final, but very important, consideration on predictivity is that this concept is highly context-dependent. In each sharply-defined use domain, it seems important to ask how far the battery is fit-for-purpose. Four issues need to be specified: (i) what regulatory problem is to be addressed (e.g. risk assessment of a new chemical, or prioritization of compounds for further testing); (ii) is there a focus on high positive predictivity or high negative predictivity; (iii) which type of chemicals is being examined (predictivity may be very high within certain compound groups, while it may be low for some compound classes); (iv) which types of biology (targets, pathways) play a role. It is likely that some adverse outcome pathways (AOP) are covered well, while others not at all. For example, acetylcholine esterase inhibitors may not be detected easily by the current IVB-EU, but this gap would be easily filled by an additional enzymatic assay (Li et al., 2017).

3.4. Compound testing and hit identification

In addition to the 45 compounds tested for the IVB-EU performance analyses, all 10 assays were challenged with additional 75 test compounds, so that the total screen comprised 120 chemicals (suppl. file 2). The result of the screen were benchmark concentrations (BMC) of effect (or no effect data within the used concentration range) for 120 compounds on ten functional and six viability endpoints, i.e. 1920 concentration response curves. A matrix including 405 BMCs for the IVB hits (with measures of uncertainty) was generated. To allow a better overview and focus, all compounds were compiled that affected at least one functional endpoint at a non-cytotoxic concentration ($n = 59$). To better visualize the activity profile of compounds, the endpoints for which toxicants had the highest potency (most sensitive endpoint(s)) were

A

Positive controls	specific + brdl. + cytotox	specific + brdl.	specific
Cadmium chloride	TP	TP	TP
Chlorpyrifos	TP	TP	FN
Dexamethasone	TP	TP	TP
Hexachlorophene	TP	TP	TP
Lead (II) acetate trihydrate	TP	TP	TP
Manganese (II) chloride	TP	TP	TP
Methylmercury chloride	TP	TP	TP
PBDE 47	TP	TP	TP
PBDE 99	TP	TP	FN
(±) Ketamine hydrochloride	FN	FN	FN
5,5-Diphenylhydantoin	FN	FN	FN
Acrylamide	TP	TP	TP
all-trans-Retinoic acid	TP	TP	TP
Chlorpromazine hydrochloride	TP	TP	TP
Deltamethrin	TP	TP	TP
Domoic acid	FN	FN	FN
Haloperidol	TP	TP	TP
Maneb	TP	TP	FN
Methylazoxymethanol acetate	TP	TP	TP
Nicotine	FN	FN	FN
Paraquat dichloride hydrate	TP	TP	TP
PFOA	TP	FN	FN
PFOSK	TP	TP	TP
Sodium valproate	TP	TP	TP
Tebuconazole	TP	TP	TP
Tributyltin chloride	TP	TP	TP
Trichlorfon	TP	TP	TP
Triethyl-tin bromide	TP	TP	FN

B

Negative controls	Acetaminophen	TN	TN	TN
	Amoxicillin	TN	TN	TN
	Aspirin	TN	TN	TN
	Buspirone	FP	TN	TN
	Chlorpheniramine maleate	TN	TN	TN
	D-Glucitol	TN	TN	TN
	Diethylene glycol	TN	TN	TN
	D-Mannitol	TN	TN	TN
	Doxylamine succinate	TN	TN	TN
	Famotidine	TN	TN	TN
	Ibuprofen	TN	TN	TN
	Metformin	TN	TN	TN
	Metoprolol	TN	TN	TN
	Penicillin	TN	TN	TN
	Saccharin	TN	TN	TN
	Sodium benzoate	TN	TN	TN
	Warfarin	TN	TN	TN

C

Performance [%]	Sensitivity	86	82	68
	Specificity	94	100	100
	Accuracy	89	89	80
	Balanced accuracy	90	91	84
	PPV	96	100	100
	F1 score	91	91	84
	MCC	78	80	67

(caption on next column)

Fig. 4. Performance overview of the test battery (IVB-EU). A set of predefined negative (n = 17) and positive (n = 28) control compounds was included in the set of screening compounds (n = 120). The rationale for their selection is given in Fig. S1 and S2. Note that the controls were randomly included in the overall screening workflow without being given any preferences or special treatment. This means that the standard prediction models of the assays were applied to them, so that they were classified as “no hit”, “cytotoxic”, “borderline (brdl)” or “specific hit” in individual NAM (see Fig. S3). A reference compound was considered to be a “positive” on the level of the overall IVB-EU, when it was an “alert” in at least one of the individual assays. The tabular display of the figure uses three definitions for an alert: anything that is not a “no hit” (first column), anything that was a specific hit or brdl (second column) or only specific hits (third column). (A) Alerts were considered true positives (TP), non-alerts were considered false negatives (FN). (B) Non-alerts were considered true negatives (TN), alerts were considered false positives (FP). (C) Performance parameters of the current DNT IVB-EU in percent. All parameters were calculated based on the TP, FN, TN, FP as indicated in (A) and (B). PPV: positive predictive value; MCC: Matthews correlation coefficient.

highlighted (Fig. 5). Compounds were considered to be about equally potent across test endpoints, when their activity did not differ by more than a factor of three. This is due to technical issues (the test concentrations were separated by a factor of three in the concentration-response curves), but also due to statistical considerations (the confidence intervals of BMCs separated by factor 3 overlapped in 85% of all cases).

Besides the 59 compounds that produced at least one specific hit (comprising 23 positive controls and 36 other compounds), there were also 61 compounds that had no specific hit in any of the 10 functional endpoints. Ten of these compounds were cytotoxic to one or more cell populations (Fig. S5A), while 51 compounds (including 16 negative controls) had no effect at all (Fig. S5B). This finding of 35 fully negatives (excluding the known negative controls) extends observations from the preliminary predictivity evaluation (using known negative control compounds) that showed that the IVB-EU, despite its large number of tests and endpoints, is not highly unspecific.

3.5. Hit patterns in the DNT IVB screen

Concerning the further analysis of battery hits, several strategies were followed. One approach was to select some individual hit compounds or groups of compounds for further toxicological evaluation. For instance, an expert group of EFSA and the OECD used IVB-EU data on deltamethrin and flufenacet for a case study within the OECD IATA program (EFSA PPR Panel, 2021). Another example is the group of flame retardants, for which the battery data were used to support a comprehensive hazard assessment (Klose et al., 2021a). Such specific toxicological follow-ups were beyond the scope of the present study. Instead, we analyzed general hit patterns of the screen to learn more about the relationship (complementarity/necessity) of the various assays and endpoints.

The first question was, how functional endpoints and specific hits related to the viability endpoints and cytotoxicity hits. To understand the overall data structure, we generated an overview, comparing for each specific hit compound the potency for the most sensitive functional endpoint in the battery (MSE) with the potencies for all cytotoxic effects across the battery test systems (cytotoxicity hits). There were 57 specific hits, plus two compounds (maneb and clorpyrifos), which were classified as borderline hits, and are being included here in the group of functional hits. Altogether 17 of the 59 compounds (29%) did not affect any of the battery's viability endpoints. For this subgroup, the functional endpoint provided a definite gain in sensitivity, compared to cytotoxicity assays. It is also very likely that the functional endpoint was directly affected by the test compounds, i.e. it was not an indirect effect of unspecific cytotoxicity.

As an alternative approach to understand the role of cytotoxicity, we

		M S E		within 3-fold MSE range	outside 3-fold range															
Compound / Test method ¹						UKN2	UKN4	UKN5	NPC1	NPC2a [72h]	NPC2a [120h]	NPC2b	NPC2c	NPC3	NPC4 [length]	NPC4 [area]	NPC5	NPC2-5 [# of cells]		
positive control compounds	Dexamethasone (DEX)								7.3					5.7						
	Tributyltin chloride (TTC)	6.9				7.2											6.8			
	Hexachlorophene (HCP)	6.4				4.9	6.9								7.1					
	Methylmercury(II) chloride (MMC)		7.0	6.4			6.2	6.3					6.3	6.7	6.3			6.1		
	Chlorpromazine hydrochloride (CPH)	5.2		5.5	6.0									5.4	5.3	6.4				
	Deltamethrin (DM)	4.8					6.0			5.4				4.9			6.4			
	Cadmium chloride (CCL)	6.3				5.3	5.8	5.6										5.5		
	Triethyltin bromide (TETB)				6.2															
	all-trans-Retinoic acid (RA)	5.3				5.7														
	Tebuconazole (TBN)																	5.6		
	Haloperidol (HPD)		4.9	4.9			4.9											5.6		
	PBDE-47 ²	4.3																5.4		
	Trichlorfon (TCF)														5.3			4.9		
	PFOSK ³	4.5												5.0						
	Maneb (MAB)	5.0																		
	PBDE 99 (BDE)		4.9																	
	Paraquat dichloride hydrate (PQ)	4.0	4.0	3.8			4.9													
	Manganese(II) chloride (MC)																	4.7		
	Lead(II) acetate trihydrate (LAT)	4.6																		
	Chlorpyrifos (CPF)	4.4	4.0																	
	Methylazoxymethanol acetate (MAM)					3.8														
	Sodium valproate (VPA)					3.3									3.3			3.3		
	Acrylamide (AAM)				2.9															
all other hits	Narciclasine (NAR)	7.3	7.9	7.7	7.8							7.2	8.3	7.9	8.0	7.8	7.6			
	Rotenone (RTN)	7.5	7.0	7.2			6.8	7.0					6.3	6.3	6.2			6.7		
	Glufosinate-ammonium (GFA)													7.1						
	Alpha-Endosulfan (AES)	5.1						4.8										6.4		
	Metaflumizone (MFM)																	6.3		
	Endosulfan sulfate (ESS)	5.4																6.3		
	Indoxacarb (IXC)	5.0			6.3													5.8		
	3,3',5,5'-Tetrabromobisphenol A (TBBPA)	5.9																6.3		
	gamma-Cyhalothrin (CHT)																	6.1		
	Emamectin benzoate (EMB)	6.0			6.0															
	2-Ethylhexyl diphenyl phosphate (EDP)	5.3			6.0													5.2		
	Beta-Cypermethrin (BCPM)																	6.0		
	Malathion (MLT)																	5.9		
	Endosulfan (EDS)							4.9										5.7		
	tau-Fluvalinate (TFV)																	5.7		
	tert-Butylphenyl diphenyl phosphate (BDP)	5.3						5.1										5.7		
	MPPs ⁴		5.7		5.6														5.6	
	Tris(1,3-dichloro-2-propyl) phosphate (TDCPP)							4.9											5.5	
	Flubendiamide (FBD)																		5.5	
	Fipronil (FPN)	4.8	5.0				4.9							4.9	4.9	4.9			4.8	
	Triphenyl phosphates isopropylated (TPI)	5.2																	5.4	
	Tri-o-tolyl phosphate (TOTP)	5.4						5.3											5.2	4.7
	Isodecyl diphenyl phosphate (IDDP)	5.0																	5.4	
	Penthiopyrad (PP)						4.8												5.3	
	Chlorpyrifos-oxon (CPFO)				5.2			5.0							4.9					
	Tris(2-butoxyethyl)phosphate (TBOEP)																		5.2	
	beta-Cyfluthrin (BCFT)																		5.2	
	Carbaryl (CBR)	5.0	4.8				5.1	5.2							5.1	5.0				
	Acibenzolar-S-methyl (ABM)				5.1															
	Tris(methylphenyl) phosphate (TTP)	5.1																		
	Clothianidin (CTN)														5.0					
	Triphenylphosphate (TPHP)	5.0																	4.9	
	Bisphenol A (BPA)	4.9																	4.9	
	Thiacloprid (TC)																		4.9	
	1-Naphthol (NT)	4.7						4.9												
	Azinphos-methyl (APM)							4.8	4.8											

¹all concentrations are given in -logM; ²2,2',4,4'-Tetrabromodiphenyl ether;

³Heptadecafluorooctanesulfonic acid potassium salt; ⁴1-Methyl-4-phenylpyridinium iodide,

Fig. 5. Hit summary of the IVB-EU screen. Overall, 120 compounds were screened in the current DNT IVB-EU. Screened substances were considered as “hits” when they were classified as a “specific hit” or a “borderline compound” in at least one assay of the battery (assays indicated on top of the columns). The upper section of the table shows all 23 hits amongst the 28 positive controls used in the screen (the remaining five positive controls were no hits). The lower section shows all additional 36 hits amongst the screened compounds. Within the groups, the compounds are ranked based on potency (indicated in units of -log [M]). The table includes all hits of the screen. For each compound, the most sensitive endpoint (MSE) is highlighted. In addition, hits of the respective chemical in other assays, which were of similar potency as in the MSE assay (within a 3-fold range), are also highlighted. The compounds that affected only viability endpoints in the IVB-EU are listed in Fig. S5A. The compounds that affected no endpoint at all are listed in Fig. S5B. Exact and complete screen data (including the uncertainties assessed as 95% confidence interval) are included in a suppl. file 2 – sheet 2 & 3.

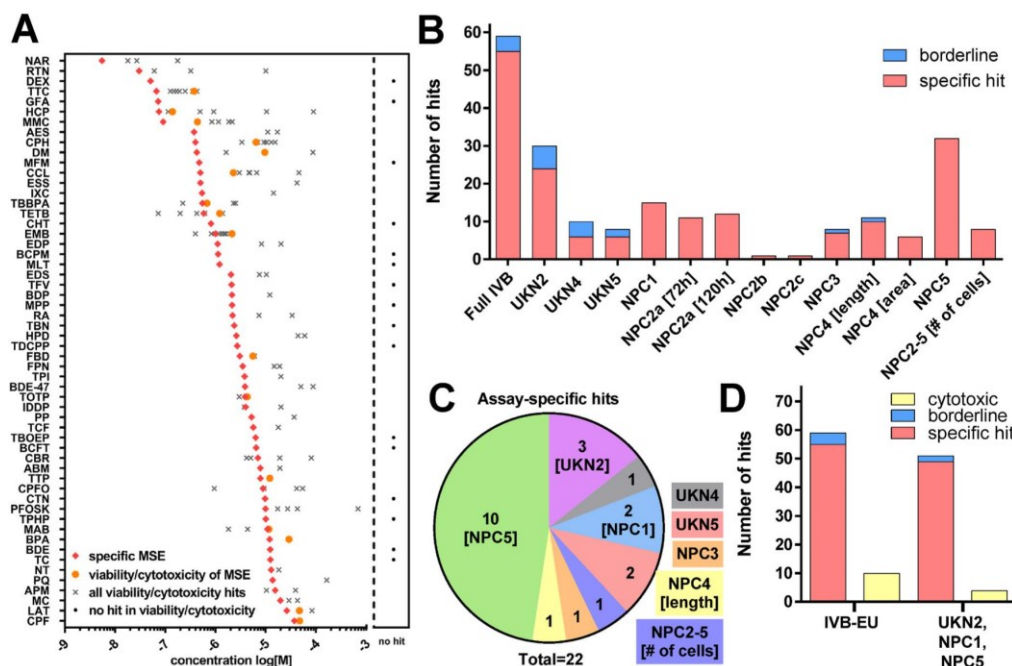


Fig. 6. Contribution of individual NAM to the overall IVB-EU. The screen was performed, hits were identified and the most sensitive endpoint (MSE) was defined for each compound as detailed in Fig. 5 (A). A potency overview of all hit compounds (see Fig. 5 for abbreviation) is displayed: The compounds are sorted according to the potency of their MSE. Note that all MSE data refer to a specific test endpoint (i.e. migration, differentiation, proliferation, neurite growth). In addition, the concentrations at which compounds were detected to be cytotoxic are indicated. Compounds that were not cytotoxic in any assay are indicated by a dot right of the dashed line. The cytotoxic concentration measured in the same assay as the MSE is given a separate symbol (filled circle) to allow an easy overview. Note that for many compounds, no cytotoxicity was measured in the assay that produced the MSE. For design reasons, three low potency compounds were not included in the figure: MAM (MSE = -3.8) orange point at x, 3 additional cytotoxic hits; VPA (MSE = -3.3) orange point at -2.7, four other cytotoxicity hits; AAM (MSE = -2.9) no other cytotoxic hit. All data are given in log(M). (B) The number of hits (out of 120 screen compounds) is indicated for each assay of the battery, and for the total IVB-EU (most leftward bar). The number of specific hits and of borderline hits can both be seen within one bar. The respective set of data for cytotoxic compounds in visualized in Fig. S7. (C) The number of compounds that were a hit in only one assay is displayed for all assays, e.g. 10 compounds were detected only in NPC5, but no other assay; one compound was detected only in UKN4 and no other assay. (D) The number of hits (separated in specific hits, borderline hits and cytotoxic-only compounds) was compared for the full IVB-EU and a hypothetical mini-battery consisting of 3 assays (UKN2, NPC1, NPC5). (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

asked, how the MSE concentration related to the cytotoxic potency in the same or in any other assay. There were only five compounds (8%) for which a cytotoxic endpoint was observed at higher (\geq factor 2) potency than the functional MSE (Fig. 6A). One example is carbaryl (CBR), which specifically inhibited neurite growth in the UKN4 assay (functional endpoint). It was particularly potent as cytotoxicant for peripheral neurons and mixed NPC cultures. This may indicate that CBR exerts a cell type-specific cytotoxicity for such neural cell populations. Such viability effects may be relevant for neurodevelopment, but further investigations would be required to allow clear conclusions.

We used a comparison to published data as one preliminary approach to test whether cytotoxicity hits of the IVB-EU are specific for neurodevelopmental cell types. We hypothesized that we may see a difference between cytotoxic potencies on conventional cell lines (HepG2, HEK293, etc.) and on the test systems used here, if a compound shows a developmental-stage specific cytotoxicity. Information on unspecific toxicity (called: cytotoxicity lower bound) was obtained from the ToxCast data base (Judson et al., 2016). For the 41 compounds, for which sufficient data was available, we found that cytotoxicity hit potency in the IVB-EU was at least 10-fold below the cytotoxicity lower bound for 7 compounds; 34 compounds showed no particular sensitivity in IVB-EU test systems compared to cell lines used for ToxCast screening (Fig. S6A). This may indicate that some, but not all cytotoxicity hits may be specific for neurodevelopmental cell types. To complete this

comparison, we also checked how the functional hits of the IVB-EU compared to the cytotoxicity lower bound. In general, the cytotoxicity threshold in ToxCast was often in the range of 5–20 μ M. Thus, the 17 IVB screen hits with MSEs <1 μ M (for which the cytotoxicity lower bound was available), seemed to separate clearly from general cytotoxicity except for TETB. The situation is complex for compounds with higher MSE potency in the IVB-EU. The data set is too small and compound behaviour is very heterogeneous. However, it is plausible, that specificity may be reduced (or lost) at higher screen concentrations (>20 μ M). It has been shown that unspecific baseline toxicity increases from this threshold on, due to membrane incorporation and alterations of protein conformations (Escher et al., 2019; Lee et al., 2021, 2022). Therefore, hits in a higher concentration range (e.g. MAM, VPA, AAM) need good justifications (e.g. clinically-observed plasma levels at hit concentration levels) and/or a detailed mechanistic follow-up providing a rationale for specific functional effects in the observed concentration range (Fig. S6B).

All these potency comparisons have an important caveat: the data we obtained are based on nominal concentrations, and these might differ from the free effective concentrations in the medium, and especially at the target sites (Kisitu et al., 2020). Especially, for comparisons to assays with tumor cell lines, it needs to be considered, that such systems usually use serum supplements containing protein and lipids, while most stem cell culture media used here had a low protein and lipid content. Under

the conditions used for the IVB-EU, the free concentrations are very close to the total concentrations in medium (Krebs et al., 2020b), while this is not necessarily the case for serum-containing media.

The second question we asked was, how the hits distributed over the different assays of the battery. Altogether 67 compounds affected at least one test endpoint: 57 specific, 2 borderline, 10 cytotoxic and 51 compounds affected no endpoint at concentrations up to 20 μ M (Fig. 6B, Fig. S5&S7). All cytotoxic compounds had potencies of ≥ 8 μ M (Fig. S5A). The number of hits obtained in each assay was also compiled. For instance, the NPC5 assay (examining the KNDP oligodendrocyte differentiation) identified the highest number ($n = 34$) of specific hits (Fig. 6B). Moreover, 10 compounds were hits only in this assay and would have been missed as potential toxicants without the NPC5 test as part of the IVB-EU (Fig. 6C). The second highest hit rate ($n = 30$) was found for the UKN2 assay (represents the KNDP of neural crest cell migration). Three compounds were unique hits in this test, i.e. not identified by another endpoint. Most other assays (UKN4, UKN5, NPC1, NPC2a, NPC3 and NPC4) identified 8–15 specific hits, and each of the assay identified at least one test compound that would have been missed by the other tests of the battery (Fig. 6C). This illustrates that the cell

types and endpoints assembled in the IVB-EU all differ in the pattern of toxicity pathways and targets they represent. This analysis also showed that the test methods are not redundant, even with this small number ($n = 120$) of screened chemicals. We anticipate that the broad coverage of cell types, developmental stages and endpoints of the IVB-EU will be even more required to ensure maximal sensitivity, when the chemical space is enlarged by broader test campaigns and a more-wide spread use of the battery.

A third question we asked dealt with resource optimization. Some assays, such as NPC2b/c (migration of neurons and oligodendrocytes) or UKN4 (neurite outgrowth) contributed relatively little to the overall hit rate, and one may consider them to be deleted from the battery or replaced. This would be a step towards a faster, more economical “mini-battery”, which would be expected to have a slightly reduced sensitivity, but not greatly reduced overall performance (accuracy; Matthews coefficient). However, in case of the neurosphere assay, individual read-outs are multiplexed, meaning that omission of one endpoint will not lead to saving resources, e.g. NPC2b/c are automatically assessed when NPC3/5 are evaluated. As NPC3 is multiplexed with NPC2 and 5, also this assay adds negligible extra time and costs to the overall assays

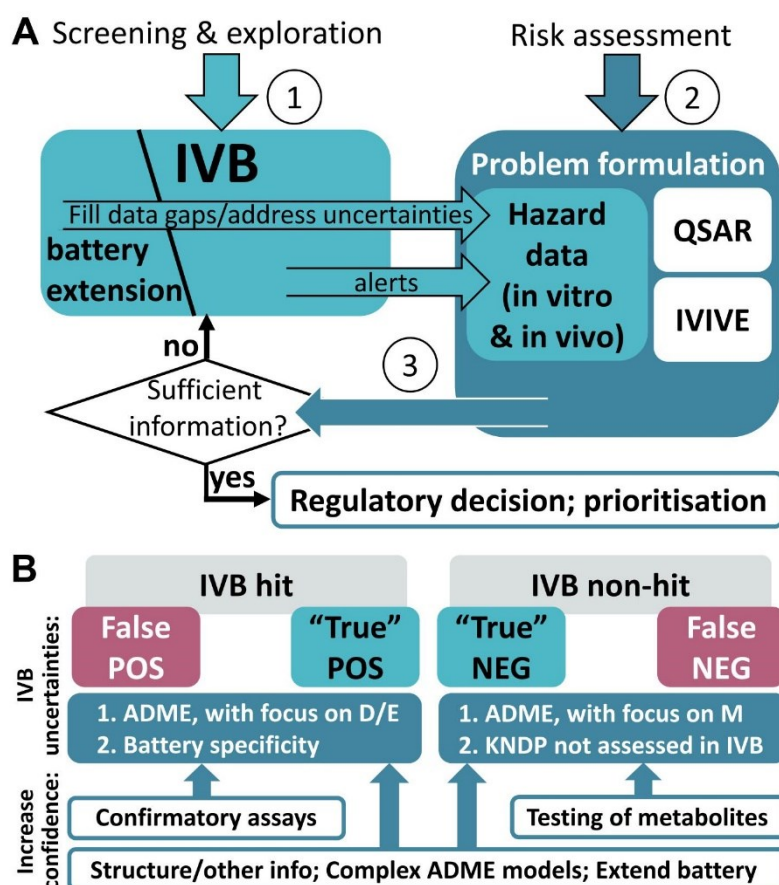


Fig. 7. Outlook on further uses and extensions of the IVB. (A) Incorporation of the IVB into an integrated approach to testing and assessment (IATA): Two different scenarios are depicted. In the first (1) the IVB will be used for screening of compound groups to generate hazard alerts (IVB hits). One way to follow up on these would be in the context of an IATA. In the second scenario (2), risk assessment of single chemicals would be performed in an IATA. This approach starts with a problem formulation (considering or not considering particular exposure situations). In this context all available data on hazard identification and characterization are collected. These may be extended via data of scenario (1). Quantitative structure activity relationships (QSAR) and in vitro-to-in vivo extrapolation (IVIVE) are shown as exemplary elements of the IATA framework. Further elements could include absorption, distribution, metabolism and excretion data (ADME) or an exposure assessment. If the hazard data of the assessed compound are considered not sufficient to derive a robust point of departure (PoD), further information could be obtained from the IVB. (3) In some cases, IVB extensions would be needed to fill data gaps and to reduce uncertainties, until sufficient information is available for regulatory action. (B) Each test method or battery has some uncertainties. The level of uncertainties that can be accepted depends on the problem formulation. For IVB hits and non-hits, one needs to consider that these may be either false positives/negatives, or compounds with a correctly identified hazard (“true” positives/negatives). One potential reason for misidentification is a lack of ADME features represented in the in vitro test systems. For example in vivo distribution and elimination (D/E) features may be misrepresented in the in vitro system. As a result, a compound never reaching the fetal brain because of the placental barrier may show effects on neurons in vitro. In contrast, some false negatives can be explained by a lack of metabolism (M) i.e. in vivo toxic metabolites which are not present in the IVB. Another reason is that a toxicant affects a key neurodevelopmental process (KNDP) that is not included in the IVB. In order to reduce the level of uncertainties and gain confidence into the results, further information can be added (low, white boxes). This includes information transfer

across tested compounds (grouping and readacross (RAx)), complex ADME models, confirmatory assays (battery extension), and direct testing of potential metabolites. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

NPC2-5. Hence, a mini-battery should only omit assays that practically save resources, i.e. individual assays. If one continues this line of thought, a minimal DNT IVB may consist of NPC1 (NPC proliferation), NPC2-5 and UKN2 (NCC migration) test methods (Fig. 6D). In our screen, this mini-battery would have identified 52 compounds (88% of all specific and borderline hits) of the 59 hits covered by the whole IVB-EU. Such a reduced approach may be used e.g. for quick/inexpensive pre-screens, e.g. in situations where sensitivity is of low importance, but compounds are to be ranked according to their priority for further testing. However, one may also consider adding an assay to a mini-battery that is not yet included in the IVB-EU. The gap analysis (Fig. 2) suggested that some biological domains are still poorly covered, and that an important gap would be filled by a neural network formation assay (Carstens et al., 2022). Thus, future batteries would need to consider the assays presented here, in addition to other established and emerging DNT NAM.

4. Conclusions and outlook

We have demonstrated here how NAMs with endpoints related to KNDP can be selected and assembled to an in vitro battery to screen for DNT hazard of chemicals. The technical feasibility and the implementation of solid reporting standards have been demonstrated by the use of 120 test compounds in a battery test-run that produced close to 2000 BMCs. These were used to provide battery performance estimates and to classify test compounds as specific hits, cytotoxicants or non-hits. The pattern of results was used to discuss the contribution of the assays and their endpoints to the overall IVB-EU and to define gaps still to be filled.

Pivotal questions for the future are (i) how battery hits would be further used and (ii) how the IVB-EU (or its future expanded version = IVB) could be implemented in a regulatory context (Fig. 7A and B). We anticipate that the first application of the IVB will be for screening of data-poor compounds to explore their DNT liabilities. As the overwhelming majority of chemicals lacks data on DNT hazard, compounds of particular concern (because of high exposure or structural alerts) may be screened first. The IVB would produce alerts for further testing. The underlying toxicological rationale is that disturbance of any KNDP covered by the IVB has the potential to lead to DNT. In a regulatory environment, the IVB data would provide a hazard characterization, and could be used as point-of-departure for further steps. In this context, physiology-based kinetic modelling (PBK) followed by in vitro-to-in vivo extrapolations (IVIVE) could be applied to convert the BMCs to estimated adverse doses (AEDs). These would be used to perform a risk assessment.

With growing experience and confidence into the IVB, its output could become a pivotal element of DNT risk assessment. Such a development is supported by the guidance document on the generation and use of the NAM-based DNT data (Crofton and Mundy, 2021). In a risk assessment situation with a defined problem formulation (e.g. for pesticide marketing re-approval in the EU, or during registration of a chemical in Japan) the compound to be evaluated would be run through the battery to provide hazard data. These might be clear and unambiguous. Or they may need to be complemented by additional rounds of testing in battery extensions. Together with the use of ADME data or other information (such as QSAR) and an IVIVE procedure, sufficient information for risk assessment would be generated (Fig. 7A).

One important aspect of using the battery data as hazard characterization is the interpretation and follow-up of hits. It is at present unclear, whether the number of positive battery endpoints correlates with the strength of DNT hazard. Hence, in the hazard characterization scenario one would be equally concerned if a compound produced one or several hits. However, the BMCs producing the hits have to be considered as multiple hits in the same order of magnitude suggest a higher concern than hits that only produce one low BMC. In the screening and prioritization scenario concern could be based on a

combination of BMC magnitude and number of hits similar to the approach practiced in Klose et al. (2021a) in the flame retardant case study. However, singleton-hit chemicals can be of high concern as exemplified by the illustrative example lead, which is one of the best-proven human DNT toxicants and only affected one functional endpoint of the IVB-EU.

For each battery hit, there is always the uncertainty, that it is either a true positive, i.e. that the battery results reflect real DNT hazard for humans, or that it is a false positive (FP). A reasons for the latter scenario may be toxicokinetic (ADME) properties. E.g. a compound may never reach the foetal or child brain because of barrier functions, but there is no such barrier in vitro. Some FP will also arise from test classification uncertainties (alpha error) and the IVB false discovery rate (FDR) due to the combination of a large number of assays. Fortunately, there are also ways to build confidence into the hit pattern and to reduce the uncertainty of a hit being a FP. The assays and their prediction models can be trimmed for high specificity (multiple test runs, full concentration-response curves, conservative thresholds for hit definition). Another powerful approach is to functionally group hit compounds and to use information on one compound to read across to others. This way, consistency and plausibility can be established and/or strengthened.

For some applications, also non-hits play an important role, e.g. for providing confidence to consumers on the safety of food constituents or contaminants. Non-hits may either be true (no hazard) or be false negatives (FN), i.e. have non-discovered toxic properties. The main sources of uncertainty on negatives are the gaps in the battery (KNDP or specific signaling pathway not covered) and toxicokinetic aspects. For instance, a tested parent compound may not be toxic, but a metabolite generated only in vivo may be a DNT toxicant. Fortunately, there are also strategies available to increase confidence in negative hits. If this is of particular importance, the sensitivity of assays can be increased by running a higher number of replicates. Also, a less conservative prediction model may be applied. This strategy is demonstrated here by the introduction of a borderline category, to capture toxic compounds that would otherwise have dropped out of the hit definition. Another major approach is the extension of the battery, e.g. by combination with the US EPA assays (Carstens et al., 2022). Last, but not least, grouping, and other information from data bases and the literature could be used for further evaluation of negative hits and decisions on potential extended testing (Fig. 7A).

Author contributions

All authors read, commented, and approved the manuscript. Jonathan Blum: study conception, investigation, data analysis, supervision, figure design, writing of article. Stefan Masjosthusmann: study conception, data analysis, supervision, figure design, writing of article. Kristina Bartmann: investigation. Farina Bendt: investigation. Xenia Dolde: investigation, data analysis. Anna-Katharina Holzer: investigation, data analysis. Ulrike Hübenthal: investigation. Sadiye Kilic: investigation. Jödis Klose: investigation. Melanie Pahl: investigation. Lynn-Christin Stürzl: investigation. Arif Dönmez: software development, data analysis. Nils Förster: software development, data analysis. Hagen Eike Keßel: software development, data analysis. Martin Scholz: software development, data analysis. Axel Mosig: software development, supervision. Iris Mangas: editing of article. Andrea Terron: editing of article. Kevin Crofton: editing of article. Marcel Leist: study conception, supervision, funding acquisition, project administration, figure design, writing of article. Ellen Fritsche: study conception, supervision, funding acquisition, project administration, figure design, writing of article.

Funding

This work was supported by the European Food Safety Authority (EFSA-Q-2018-00308), the Danish Environmental Protection Agency (EPA), Denmark, under the grant number MST-667-00205, the State

Ministry of Baden-Wuerttemberg, Germany, for Economic Affairs, Labour and Tourism (NAM-Accept), the project CERST (Center for Alternatives to Animal Testing) of the Ministry for culture and science of the State of North-Rhine Westphalia, Germany (file number 233-1.08.03.03- 121972/131-1.08.03.03-121972), the European Chemical Industry Council Long-Range Research Initiative (Cefic LRI) under the project name AIMT11 and the BMBF (NeuroTool). It has also received funding 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. 825759 (ENDpoiNTs).

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Ellen Fritsche, Kristina Bartmann, Arif Dönmez and Axel Mosig are shareholders of the company DNTOX that provides DNT-IVB assay services. The authors declare no potential conflicts of interest with respect to the research in this article. All other authors have no conflict of interest to declare.

Data availability

Data will be made available on request.

Acknowledgements

The authors are very grateful to Tanja Waldmann as well as Tim Shafer, Katie Paul Friedman and Kelly Carstens (all involved in compound selection, study design, data interpretation and transfer of data to the ToxCast data base). We would also like to thank R. Göttler for the experimental and J. Kapr for layout support. Furthermore, we thank T. Mayer, S. Müller, and the screening centre of the University of Konstanz for the experimental and technical support.

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.chemosphere.2022.137035>.

Abbreviations

AOP	– adverse outcome pathway
BMC	– benchmark concentration
BMCL	– lower limit of 95% confidence interval of BMC
BMCU	– upper limit of 95% confidence interval of BMC
DIV	– days <i>in vitro</i>
DNT	– developmental neurotoxicity
EFSA	– European Food Safety Authority
FDR	– false discovery rate
hNPC	– human neural progenitor cell
hiPSC	– human induced pluripotent stem cell
IVB	– <i>in vitro</i> battery
IVB-EU	– DNT IVB based on methods available in European laboratories
IVIVE	– <i>in vitro</i> to <i>in vivo</i> extrapolation
KNDP	– key neurodevelopmental process
MSE	– most sensitive endpoint
NAM	– new approach methods
PPV	– positive predictive value
IATA	– integrated approaches for testing and assessment
OECD	– Organisation for Economic Co-operation and Development
TN	– true negative
TP	– true positive
UKN	– University of Konstanz

US EPA – United States Environmental Protection Agency

References

- Allen, N.J., Lyons, D.A., 2018. Glia as architects of central nervous system formation and function. *Science* 362, 181–185.
- Andersen, M.E., McMullen, P.D., Phillips, M.B., Yoon, M., Pendse, S.N., Clewell, H.J., Hartman, J.K., Moreau, M., Becker, R.A., Clewell, R.A., 2019. Developing context appropriate toxicity testing approaches using new alternative methods (NAMs). *ALTEX* 36, 523–534.
- Aschner, M., Ceccatelli, S., Daneshian, M., Fritsche, E., Hasiwa, N., Hartung, T., Hogberg, H.T., Leist, M., Li, A., Mundi, W.R., Padilla, S., Piersma, A.H., Bal-Price, A., Seiler, A., Westerink, R.H., Zimmer, B., Lein, P.J., 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.
- Atzei, A., Jense, L., Zwart, E.P., Legradi, J., Venhuis, B.J., van der Ven, L.T.M., Heusinkveld, H.J., Hessel, E.V.S., 2021. Developmental neurotoxicity of environmentally relevant pharmaceuticals and mixtures thereof in a zebrafish embryo behavioural test. *Int. J. Environ. Res. Publ. Health* 18.
- Bal-Price, A., Crofton, K.M., Leist, M., Allen, S., Arand, M., Buetler, T., Delrue, N., FitzGerald, R.E., Hartung, T., Heinonen, T., Hogberg, H., Bennekou, S.H., Lichtensteiger, W., Oggier, D., Paparella, M., Axelstad, M., Piersma, A., Rached, E., Schilter, B., Schunck, G., Stoppini, L., Tongiorgi, E., Tiramini, M., Monnet-Tschudi, F., Wilks, M.F., Ylikomi, T., Fritsche, E., 2015. International Stakeholder Network (ISTNET): creating a developmental neurotoxicity (DNT) testing road map for regulatory purposes. *Arch. Toxicol.* 89, 269–287.
- Bal-Price, A., Hogberg, H.T., Crofton, K.M., Daneshian, M., FitzGerald, R.E., Fritsche, E., Heinonen, T., Høngaard Bennekou, S., Klima, S., Piersma, A.H., Sachana, M., Shafer, T.J., Terron, A., Monnet-Tschudi, F., Viviani, B., Waldmann, T., Westerink, R.H.S., Wilks, M.F., Witters, H., Zurich, M.G., Leist, M., 2018. Recommendation on test readiness criteria for new approach methods in toxicology: exemplified for developmental neurotoxicity. *ALTEX* 35, 306–352.
- Bal-Price, A.K., Coecke, S., Costa, L., Crofton, K.M., Fritsche, E., Goldberg, A., Grandjean, P., Lein, P.J., Li, A., Lucchini, R., Mundy, W.R., Padilla, S., Persico, A.M., Seiler, A.E., Kreysa, J., 2012. Advancing the science of developmental neurotoxicity (DNT): testing for better safety evaluation. *ALTEX* 29, 202–215.
- Baumann, J., Gassmann, K., Masjosthusmann, S., DeBoer, D., Bendt, F., Giersiefer, S., Fritsche, E., 2016. Comparative human and rat neurospheres reveal species differences in chemical effects on neurodevelopmental key events. *Arch. Toxicol.* 90, 1415–1427.
- Behl, M., Ryan, K., Hsieh, J.H., Parham, F., Shapiro, A.J., Collins, B.J., Sipes, N.S., Birnbaum, L.S., Bucher, J.R., Foster, P.M.D., Walker, N.J., Paulus, R.S., Tice, R.R., 2019. Screening for developmental neurotoxicity at the national toxicology program: the future is here. *Toxicol. Sci.* 167, 6–14.
- Bellinger, D.C., 2012. A strategy for comparing the contributions of environmental chemicals and other risk factors to neurodevelopment of children. *Environ. Health Perspect.* 120, 501–507.
- Bennett, D., Bellinger, D.C., Birnbaum, L.S., Bradman, A., Chen, A., Cory-Slechta, D.A., Engel, S.M., Fallin, M.D., Halladay, A., Hauser, R., Hertz-Picciotto, L., Kwiatkowski, C.F., Lanphear, B.P., Marquez, E., Marty, M., McPartland, J., Newschaffer, C.J., Payne-Sturges, D., Patisaul, H.B., Perera, F.P., Ritz, B., Sass, J., Schantz, S.L., Webster, T.F., Whyatt, R.M., Woodruff, T.J., Zoeller, R.T., Anderko, L., Campbell, C., Conry, J.A., DeNicola, N., Gould, R.M., Hirtz, D., Huffling, K., Landrigan, P.J., Lavin, A., Miller, M., Mitchell, M.A., Rubin, L., Schettler, T., Tran, H.L., Acosta, A., Brody, C., Miller, E., Miller, P., Swanson, M., Witherspoon, N.O., American College of O., Gynecologists, Child Neurology, S., Endocrine, S., International Neurotoxicology, A., International Society for Children's, H., the, E., International Society for Environmental, E., National, 2016. Council of asian pacific islander, P., national hispanic medical, A., national medical, A., In: Project TENDR: Targeting Environmental Neuro-Developmental Risks the TENDR Consensus Statement. *Environ Health Perspect.* vol. 124, pp. A118–A122.
- Brull, M., Spreng, A.S., Gutbier, S., Loser, D., Krebs, A., Reich, M., Kraushaar, U., Britschgi, M., Patsch, C., Leist, M., 2020. Incorporation of stem cell-derived astrocytes into neuronal organoids to allow neuro-glial interactions in toxicological studies. *ALTEX* 37, 409–428.
- Carstens, K.E., Carpenter, A.F., Martin, M.M., Harrill, J.A., Shafer, T.J., Paul Friedman, K., 2022. Integrating data from *in vitro* new approach methodologies for developmental neurotoxicity. *Toxicol. Sci.* 187, 62–79.
- Chesnut, M., Paschoud, H., Repond, C., Smirnova, L., Hartung, T., Zurich, M.G., Hogberg, H.T., Pamies, D., 2021. Human iPSC-derived model to study myelin disruption. *Int. J. Mol. Sci.* 22.
- Coecke, S., Goldberg, A.M., Allen, S., Buzanska, L., Calamandrei, G., Crofton, K., Hareng, L., Hartung, T., Knaut, H., Honegger, P., Jacobs, M., Lein, P., Li, A., Mundy, W., Owen, D., Schneider, S., Silbergeld, E., Reum, T., Tmoevec, T., Monnet-Tschudi, F., Bal-Price, A., 2007. Workgroup report: incorporating *in vitro* alternative methods for developmental neurotoxicity into international hazard and risk assessment strategies. *Environ. Health Perspect.* 115, 924–931.
- Cote, I., Andersen, M.E., Ankley, G.T., Barone, S., Birnbaum, L.S., Boekelheide, K., Bois, F.Y., Burgeon, L.D., Chiu, W.A., Crawford-Brown, D., Crofton, K.M., DeVito, M., Devlin, R.B., Edwards, S.W., Guyton, K.Z., Hattis, D., Judson, R.S., Knight, D., Krewski, D., Lambert, J., Maull, E.A., Mendrick, D., Paoli, G.M., Patel, C.J., Perkins, E.J., Poje, G., Portier, C.J., Rusyn, I., Schulte, P.A., Simeonov, A., Smith, M.T., Thayer, K.A., Thomas, R.S., Thomas, R., Tice, R.R., Vandenberg, J.J., Villeneuve, D.L., Wesselkamper, S., Whelan, M., Whittaker, C., White, R., Xia, M., Yauk, C., Zeise, L., Zhao, J., DeWoskin, R.S., 2016. The next generation of risk

- assessment multi-year study-highlights of findings, applications to risk assessment, and future directions. *Environ. Health Perspect.* 124, 1671–1682.
- Crofton, K.M., Mundy, W.R., 2021. External scientific report on the interpretation of data from the developmental neurotoxicity in vitro testing assays for use in integrated approaches for testing and assessment. EFSA Support. Pub. 18, 6924E.
- Crofton, K.M., Mundy, W.R., Lein, P.J., Bal-Price, A., Coecke, S., Seiler, A.E., Knauf, H., Buzanska, L., Goldberg, A., 2011. Developmental neurotoxicity testing: recommendations for developing alternative methods for the screening and prioritization of chemicals. *ALTEX* 28, 9–15.
- Crofton, K.M., Mundy, W.R., Shafer, T.J., 2012. Developmental neurotoxicity testing: a path forward. *Congenital. Anom.* 52, 140–146.
- Dasgupta, S., Simonich, M.T., Tanguay, R.L., 2022. Zebrafish behavioral assays in toxicology. *Methods Mol. Biol.* 2474, 109–122.
- Delp, J., Gutbier, S., Klima, S., Hoelting, L., Pinto-Gil, K., Hsieh, J.H., Aichem, M., Klein, K., Schreiber, F., Tice, R.R., Pastor, M., Behl, M., Leist, M., 2018. A high-throughput approach to identify specific neurotoxicants/developmental toxicants in human neuronal cell function assays. *ALTEX* 35, 235–253.
- Dresner, N., Madjar, K., Holzer, A.K., Kapitzka, M., Scholz, C., Kranaster, P., Gutbier, S., Klima, S., Kolb, D., Dietz, C., Trefzer, T., Meisig, J., van Thriel, C., Henry, M., Berthold, M.R., Bluthgen, N., Sachinidis, A., Rahnenfuhrer, J., Hengstler, J.G., Waldmann, T., Leist, M., 2020. Development of a neural rosette formation assay (RoFA) to identify neurodevelopmental toxicants and to characterize their transcriptome disturbances. *Arch. Toxicol.* 94, 151–171.
- Escher, B.L., Glauch, L., König, M., Mayer, P., Schlichting, R., 2019. Baseline toxicity and volatility cutoff in reporter gene assays used for high-throughput screening. *Chem. Res. Toxicol.* 32, 1646–1655.
- Escher, S.E., Partosch, F., Konzok, S., Jennings, P., Luijten, M., Kienhuis, A., de Leeuw, V., Reuss, R., Lindenmann, K.-M., Bennekou, S.H., 2022. Development of a roadmap for action on new approach methodologies in risk assessment. EFSA Support. Pub. 19, 7341E.
- Forster, N., Butke, J., Kessel, H.E., Bendt, F., Pahl, M., Li, L., Fan, X., Leung, P.C., Klose, J., Masjosthusmann, S., Fritsche, E., Mosig, A., 2022. Reliable identification and quantification of neural cells in microscopic images of neurospheres. *Cytometry* 101, 411–422.
- Frank, C.L., Brown, J.P., Wallace, K., Mundy, W.R., Shafer, T.J., 2017. From the cover: developmental neurotoxicants disrupt activity in cortical networks on microelectrode arrays: results of screening 86 compounds during neural network formation. *Toxicol. Sci.* 160, 121–135.
- Fritsche, E., Crofton, K.M., Hernandez, A.F., Hougaard Bennekou, S., Leist, M., Bal-Price, A., Reeves, E., Wilks, M.F., Terron, A., Solecki, R., Sachana, M., Goumlel, A., 2017. OECD/EFSA workshop on developmental neurotoxicity (DNT): the use of non-animal test methods for regulatory purposes. *ALTEX* 34, 311–315.
- Fritsche, E., Grandjean, P., Crofton, K.M., Aschner, M., Goldberg, A., Heinonen, T., Hessel, E.V.S., Hogberg, H.T., Bennekou, S.H., Lein, P.J., Leist, M., Mundy, W.R., Paparella, M., Piersma, A.H., Sachana, M., Schmuck, G., Solecki, R., Terron, A., Monnet-Tschudi, F., Wilks, M.F., Witters, H., Zurich, M.G., Bal-Price, A., 2018. Consensus statement on the need for innovation, transition and implementation of developmental neurotoxicity (DNT) testing for regulatory purposes. *Toxicol. Appl. Pharmacol.* 354, 3–6.
- Grandjean, P., Abdennebi-Najar, L., Barouki, R., Cranor, C.F., Etzel, R.A., Gee, D., Heindel, J.J., Hougaard, K.S., Hunt, P., Nawrot, T.S., Prins, G.S., Ritz, B., Soffritti, M., Sunyer, J., Weihe, P., 2019. Timescales of developmental toxicity impacting on research and needs for intervention. *Basic Clin. Pharmacol. Toxicol.* 125 (Suppl. 3), 70–80.
- Grandjean, P., Landrigan, P.J., 2006. Developmental neurotoxicity of industrial chemicals. *Lancet* 368, 2167–2178.
- Grandjean, P., Landrigan, P.J., 2014. Neurobehavioural effects of developmental toxicity. *Lancet Neurol.* 13, 330–338.
- Griesinger, C., Desprez, B., Coecke, S., Casey, W., Zang, V., 2016. Validation of alternative in vitro methods to animal testing: concepts, challenges, processes and tools. *Adv. Exp. Med. Biol.* 856, 65–132.
- Gutbier, S., May, P., Berthelot, S., Krishna, A., Trefzer, T., Behbehani, M., Efenova, L., Delp, J., Gstrauchthaler, G., Waldmann, T., Leist, M., 2018. Major changes of cell function and toxicant sensitivity in cultured cells undergoing mild, quasi-natural genetic drift. *Arch. Toxicol.* 92, 3487–3503.
- Harrill, J.A., Freudenrich, T., Wallace, K., Ball, K., Shafer, T.J., Mundy, W.R., 2018. Testing for developmental neurotoxicity using a battery of in vitro assays for key cellular events in neurodevelopment. *Toxicol. Appl. Pharmacol.* 354, 24–39.
- Hartung, T., 2007. Food for thought ... on validation. *ALTEX* 24, 67–80.
- Hartung, T., Bremer, S., Casati, S., Coecke, S., Corvi, R., Fortaner, S., Gribaldo, L., Halder, M., Hoffmann, S., Roi, A.J., Prieto, P., Sabbioni, E., Scott, L., Worth, A., Zang, V., 2004. A modular approach to the ECVAM principles on test validity. *Altern. Lab. Anim.* 32, 467–472.
- Hartung, T., Hoffmann, S., Stephens, M., 2013. Mechanistic validation. *ALTEX* 30, 119–130.
- Hoelting, L., Klima, S., Karremann, C., Grinberg, M., Meisig, J., Henry, M., Rotshteyn, T., Rahnenfuhrer, J., Bluthgen, N., Sachinidis, A., Waldmann, T., Leist, M., 2016. Stem cell-derived immature human dorsal root ganglia neurons to identify peripheral neurotoxicants. *Stem Cells Transl. Med.* 5, 476–487.
- Holzer, A.K., Suci, L., Karremann, C., Goj, T., Leist, M., 2022. Specific attenuation of purinergic signaling during bortezomib-induced peripheral neuropathy in vitro. *Int. J. Mol. Sci.* 23.
- Hu, W., Liu, C.W., Jimenez, J.A., McCoy, E.S., Hsiao, Y.C., Lin, W., Engel, S.M., Lu, K., Zylka, M.J., 2022. Detection of azoxystrobin fungicide and metabolite azoxystrobin acid in pregnant women and children, estimation of daily intake, and evaluation of placental and lactational transfer in mice. *Environ. Health Perspect.* 130, 27013.
- Jaklin, M., Zhang, J.D., Schafer, N., Clemann, N., Barrow, P., Kung, E., Sach-Peltason, L., McGinnis, C., Leist, M., Kustermann, S., 2022. Optimization of the TeraTox assay for preclinical teratogenicity assessment. *Toxicol. Sci.* 188, 17–33.
- Jensen, S.M., Kluxen, F.M., Streibig, J.C., Cedergreen, N., Ritz, C., 2020. bmd: an R package for benchmark dose estimation. *PeerJ* 8, e10557.
- Judson, R., Houck, K., Martin, M., Richard, A.M., Knudsen, T.B., Shah, L., Little, S., Wambaugh, J., Woodrow Setzer, R., Kothiy, P., Phuong, J., Filer, D., Smith, D., Reif, D., Rotroff, D., Kleinstreuer, N., Sipes, N., Xia, M., Huang, R., Crofton, K., Thomas, R.S., 2016. Editor's highlight: analysis of the effects of cell stress and cytotoxicity on in vitro assay activity across a diverse chemical and assay space. *Toxicol. Sci.* 152, 323–339.
- Judson, R., Kavlock, R., Martin, M., Reif, D., Houck, K., Knudsen, T., Richard, A., Tice, R., Whelan, M., Xia, M., Huang, R., Austin, C., Daston, G., Hartung, T., Fowle 3rd, J. R., Wooge, W., Tong, W., Dix, D., 2013. Perspectives on validation of high-throughput assays supporting 21st century toxicity testing. *ALTEX* 30, 51–56.
- Kaderit, S., Zimmer, B., van Thriel, C., Hengstler, J.G., Leist, M., 2012. Compound selection for in vitro modeling of developmental neurotoxicity. *Front. Biosci. (Landmark Ed)* 17, 2442–2460.
- Kappenberg, F., Brecklinghaus, T., Albrecht, W., Blum, J., van der Wurp, C., Leist, M., Hengstler, J.G., Rahnenfuhrer, J., 2020. Handling deviating control values in concentration-response curves. *Arch. Toxicol.* 94, 3787–3798.
- Keßel, 2022. Biostatistics and its impact on hazard characterization using in vitro developmental neurotoxicity assays. *ALTEX*. <https://doi.org/10.1101/2022.10.18.512648>. Submitted for publication.
- Kisitu, J., Hollert, H., Fisher, C., Leist, M., 2020. Chemical concentrations in cell culture compartments (C5) - free concentrations. *ALTEX* 37, 693–708.
- Klima, S., Brull, M., Spreng, A.S., Suci, L., Falt, T., Schwaborn, J.C., Waldmann, T., Karremann, C., Leist, M., 2021. A human stem cell-derived test system for agents modifying neuronal N-methyl-D-aspartate-type glutamate receptor Ca(2+)-signalling. *Arch. Toxicol.* 95, 1703–1722.
- Klose, J., Li, L., Pahl, M., Bendt, F., Hubenthal, U., Jungst, C., Petzsch, P., Schauss, A., Kohrer, K., Leung, P.C., Wang, C.C., Koch, K., Tigges, J., Fan, X., Fritsche, E., 2022. Application of the adverse outcome pathway concept for investigating developmental neurotoxicity potential of Chinese herbal medicines by using human neural progenitor cells in vitro. *Cell Biol. Toxicol.*
- Klose, J., Pahl, M., Bartmann, K., Bendt, F., Blum, J., Dolde, X., Forster, N., Holzer, A.K., Hubenthal, U., Kessel, H.E., Koch, K., Masjosthusmann, S., Schneider, S., Sturzl, L.C., Woeste, S., Rossi, A., Covaci, A., Behl, M., Leist, M., Tigges, J., Fritsche, E., 2021a. Neurodevelopmental toxicity assessment of flame retardants using a human DNT in vitro testing battery. *Cell Biol. Toxicol.*
- Klose, J., Tigges, J., Masjosthusmann, S., Schmuck, K., Bendt, F., Hubenthal, U., Petzsch, P., Kohrer, K., Koch, K., Fritsche, E., 2021b. TBPPA targets converging key events of human oligodendrocyte development resulting in two novel AOPs. *ALTEX* 38, 215–234.
- Koch, K., Bartmann, K., Hartmann, J., Kapr, J., Klose, J., Kuchovska, E., Pahl, M., Schlupmann, K., Zuhre, E., Fritsche, E., 2022. Scientific validation of human neurosphere assays for developmental neurotoxicity evaluation. *Front. Toxicol.* 4, 816370.
- Krebs, A., Nyffeler, J., Karremann, C., Schmidt, B.Z., Kappenberg, F., Mellert, J., Pallocca, G., Pastor, M., Rahnenfuhrer, J., Leist, M., 2020a. Determination of benchmark concentrations and their statistical uncertainty for cytotoxicity test data and functional in vitro assays. *ALTEX* 37, 155–163.
- Krebs, A., Nyffeler, J., Rahnenfuhrer, J., Leist, M., 2018. Normalization of data for viability and relative cell function curves. *ALTEX* 35, 268–271.
- Krebs, A., van Vugt-Lussenburg, B.M.A., Waldmann, T., Albrecht, W., Boei, J., Ter Braak, B., Brajnik, M., Braunbeck, T., Brecklinghaus, T., Busquet, F., Dinnyes, A., Dokler, J., Dolde, X., Exner, T.E., Fisher, C., Fluri, D., Forsby, A., Hengstler, J.G., Holzer, A.K., Janstova, Z., Jennings, P., Kisitu, J., Kobolaj, J., Kumar, M., Limonciel, A., Lundqvist, J., Mihalik, B., Moritz, W., Pallocca, G., Ulloa, A.P.C., Pastor, M., Rovida, C., Sarkans, U., Schimming, J.P., Schmidt, B.Z., Stober, R., Strassfeld, T., van de Water, B., Wilmes, A., van der Burg, B., Verfaillie, C.M., von Hellfeld, R., Vrieling, H., Vrijenhoek, N.G., Leist, M., 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.
- Krebs, A., Waldmann, T., Wilks, M.F., Van Vugt-Lussenburg, B.M.A., Van der Burg, B., Terron, A., Steger-Hartmann, T., Ruegg, J., Rovida, C., Pedersen, E., Pallocca, G., Luijten, M., Leite, S.B., Kustermann, S., Kamp, H., Hoeng, J., Hewitt, P., Herzler, M., Hengstler, J.G., Heinonen, T., Hartung, T., Hardy, B., Gantner, F., Fritsche, E., Fant, K., Ezendam, J., Exner, T., Dunkern, T., Dietrich, D.R., Coecke, S., Busquet, F., Braeuning, A., Bondarenko, O., Bennekou, S.H., Beilmann, M., Leist, M., 2019. Template for the description of cell-based toxicological test methods to allow evaluation and regulatory use of the data. *ALTEX* 36, 682–699.
- Krug, A.K., Balmer, N.V., Matt, F., Schonenberger, F., Merhof, D., Leist, M., 2013a. Evaluation of a human neurite growth assay as specific screen for developmental neurotoxicants. *Arch. Toxicol.* 87, 2215–2231.
- Krug, A.K., Gutbier, S., Zhao, L., Polt, D., Kullmann, C., Ivanova, V., Forster, S., Jagtap, S., Meiser, J., Leparc, G., Schildknecht, S., Adam, M., Hiller, K., Farhan, H., Brunner, T., Hartung, T., Sachinidis, A., Leist, M., 2014. Transcriptional and metabolic adaptation of human neurons to the mitochondrial toxicant MPP(+). *Cell Death Dis.* 5, e1222.
- Krug, A.K., Kolde, R., Gaspar, J.A., Rempel, E., Balmer, N.V., Meganathan, K., Vojnits, K., Baquie, M., Waldmann, T., Ensenat-Waser, R., Jagtap, S., Evans, R.M., Julien, S., Peterson, H., Zagoura, D., Kaderit, S., Gerhard, D., Sotiriadou, I., Heke, M., Natarajan, K., Henry, M., Winkler, J., Marchan, R., Stoppini, L., Bosgra, S.,

- Westerhout, J., Verwei, M., Vilo, J., Kortenkamp, A., Hescheler, J., Hothorn, L., Bremer, S., van Thriel, C., Krause, K.H., Hengstler, J.G., Rahnenfuhrer, J., Leist, M., Sachinidis, A., 2013b. Human embryonic stem cell-derived test systems for developmental neurotoxicity: a transcriptomics approach. *Arch. Toxicol.* 87, 123–143.
- Lee, J., Braun, G., Henneberger, L., König, M., Schlichting, R., Scholz, S., Escher, B.I., 2021. Critical membrane concentration and mass-balance model to identify baseline cytotoxicity of hydrophobic and ionizable organic chemicals in mammalian cell lines. *Chem. Res. Toxicol.* 34, 2100–2109.
- Lee, J., Escher, B.I., Scholz, S., Schlichting, R., 2022. Inhibition of neurite outgrowth and enhanced effects compared to baseline toxicity in SH-SY5Y cells. *Arch. Toxicol.* 96, 1039–1053.
- Lein, P., Locke, P., Goldberg, A., 2007. Meeting report: alternatives for developmental neurotoxicity testing. *Environ. Health Perspect.* 115, 764–768.
- Leist, M., Efremova, L., Karreman, C., 2010. Food for thought ... considerations and guidelines for basic test method descriptions in toxicology. *ALTEX* 27, 309–317.
- Leist, M., Hartung, T., Nicotera, P., 2008. The dawning of a new age of toxicology. *ALTEX* 25, 103–114.
- Leist, M., Hasiwa, N., Daneshian, M., Hartung, T., 2012. Validation and quality control of replacement alternatives – current status and future challenges. *Toxicology Research* 1, 8–22.
- Leist, M., Hasiwa, N., Rovida, C., Daneshian, M., Basketter, D., Kimber, I., Clewell, H., Gocht, T., Goldberg, A., Busquet, F., Rossi, A.M., Schwarz, M., Stephens, M., Taalman, R., Knudsen, T.B., McKim, J., Harris, G., Pamies, D., Hartung, T., 2014. Consensus report on the future of animal-free systemic toxicity testing. *ALTEX* 31, 341–356.
- Leontaridou, M., Urbisch, D., Kolbe, S.N., Ott, K., Mulliner, D.S., Gabbert, S., Landsiedel, R., 2017. The borderline range of toxicological methods: quantification and implications for evaluating precision. *ALTEX* 34, 525–538.
- Li, S., Huang, R., Solomon, S., Liu, Y., Zhao, B., Santillo, M.F., Xia, M., 2017. Identification of acetylcholinesterase inhibitors using homogenous cell-based assays in quantitative high-throughput screening platforms. *Biotechnol. J.* 12.
- Loser, D., Hinojosa, M.G., Blum, J., Schaefer, J., Brull, M., Johansson, Y., Suciu, I., Grillberger, K., Danker, T., Moller, C., Gardner, I., Ecker, G.F., Bennekou, S.H., Forsby, A., Kraushaar, U., Leist, M., 2021a. Functional alterations by a subgroup of neonicotinoid pesticides in human dopaminergic neurons. *Arch. Toxicol.* 95, 2081–2107.
- Loser, D., Schaefer, J., Danker, T., Moller, C., Brull, M., Suciu, I., Ucker, A.K., Klima, S., Leist, M., Kraushaar, U., 2021b. Human neuronal signaling and communication assays to assess functional neurotoxicity. *Arch. Toxicol.* 95, 229–252.
- Lotharius, J., Falsig, J., van Beek, J., Payne, S., Dringen, R., Brundin, P., Leist, M., 2005. Progressive degeneration of human mesencephalic neuron-derived cells triggered by dopamine-dependent oxidative stress is dependent on the mixed-lineage kinase pathway. *J. Neurosci.* 25, 6329–6342.
- Lupu, D., Andersson, P., Bornehag, C.G., Demenex, B., Fritsche, E., Gennings, C., Lichtensteiger, W., Leist, M., Leonards, P.E.G., Ponsonby, A.L., Scholze, M., Testa, G., Tresguerres, J.A.F., Westerink, R.H.S., Zalc, B., Ruegg, J., 2020. The ENDPOINTs project: novel testing strategies for endocrine disruptors linked to developmental neurotoxicity. *Int. J. Mol. Sci.* 21.
- Makris, S.L., Raffaele, K., Allen, S., Bowers, W.J., Hass, U., Alleve, E., Calamandrei, G., Sheets, L., Amicoff, P., Delrue, N., Crofton, K.M., 2009. A retrospective performance assessment of the developmental neurotoxicity study in support of OECD test guideline 426. *Environ. Health Perspect.* 117, 17–25.
- Masjosthusmann, S., Becker, D., Petzuch, B., Klose, J., Siebert, C., Deenen, R., Barenys, M., Baumann, J., Dach, K., Tigges, J., Hubenthal, U., Kohrer, K., Fritsche, E., 2018. A transcriptome comparison of time-matched developing human, mouse and rat neural progenitor cells reveals human uniqueness. *Toxicol. Appl. Pharmacol.* 354, 40–55.
- Masjosthusmann, S., Blum, J., Bartmann, K., Dolde, X., Holzer, A.-K., Stürzl, L.-C., Keßel, E.H., Förster, N., Dönmez, A., Klose, J., Pahl, M., Waldmann, T., Bendt, F., Kisitu, J., Suciu, I., Hubenthal, U., Mosig, A., Leist, M., Fritsche, E., 2020. Establishment of an a priori protocol for the implementation and interpretation of an in vitro testing battery for the assessment of developmental neurotoxicity. *EFSA Support. Pub.* 17, 1938E.
- Meisig, J., Dreser, N., Kapitza, M., Henry, M., Rotshteyn, T., Rahnenfuhrer, J., Hengstler, J.G., Sachinidis, A., Waldmann, T., Leist, M., Bluthgen, N., 2020. Kinetic modeling of stem cell transcriptome dynamics to identify regulatory modules of normal and disturbed neuroectodermal differentiation. *Nucleic Acids Res.* 48, 12577–12592.
- Modafferi, S., Zhong, X., Kleinsang, A., Murata, Y., Fagiani, F., Pamies, D., Hogberg, H. T., Calabrese, V., Lachman, H., Hartung, T., Smirnova, L., 2021. Gene-environment interactions in developmental neurotoxicity: a case study of synergy between chlorpyrifos and CHD8 knockout in human BrainSpheres. *Environ. Health Perspect.* 129, 77001.
- Mundy, W.R., Padilla, S., Breier, J.M., Crofton, K.M., Gilbert, M.E., Herr, D.W., Jensen, K. F., Radio, N.M., Raffaele, K.C., Schumacher, K., Shafer, T.J., Cowden, J., 2015. Expanding the test set: chemicals with potential to disrupt mammalian brain development. *Neurotoxicol. Teratol.* 52, 25–35.
- Nimtz, L., Hartmann, J., Tigges, J., Masjosthusmann, S., Schmuck, M., Kessel, E., Theiss, S., Kohrer, K., Petzsch, P., Adjaye, J., Wignmann, C., Wiczorek, D., Hildebrandt, B., Bendt, F., Hubenthal, U., Brockerhoff, G., Fritsche, E., 2020. Characterization and application of electrically active neuronal networks established from human induced pluripotent stem cell derived neural progenitor cells for neurotoxicity evaluation. *Stem Cell Res.* 45, 101761.
- Nimtz, L., Klose, J., Masjosthusmann, S., Barenys, M., Fritsche, E., 2019. The neurosphere assay as an in vitro method for developmental neurotoxicity (DNT) evaluation. In: Aschner, M., Costa, L. (Eds.), *Cell Culture Techniques*. Springer New York, New York, NY, pp. 141–168.
- Nunes, C., Singh, P., Mazidi, Z., Murphy, C., Bourguignon, A., Wellens, S., Chandrasekaran, V., Ghosh, S., Zana, M., Pamies, D., Thomas, A., Verfaillie, C., Culot, M., Dinnyes, A., Hardy, B., Wilmes, A., Jennings, P., Grillari, R., Grillari, J., Zurich, M.G., Exner, T., 2022. An in vitro strategy using multiple human induced pluripotent stem cell-derived models to assess the toxicity of chemicals: a case study on paraquat. *Toxicol. Vitro* 81, 105333.
- Nyffeler, J., Karreman, C., Leisner, H., Kim, Y.-J., Lee, G., Waldmann, T., Leist, M., 2017. Design of a high-throughput human neural crest cell migration assay to indicate potential developmental toxicants. *ALTEX* 34, 75–94.
- OECD, 2007. Test No. 426. Developmental Neurotoxicity Study.
- OECD, 2021. Guideline No. 497: Defined Approaches on Skin Sensitisation.
- Pallocca, G., Grinberg, M., Henry, M., Frickey, T., Hengstler, J.G., Waldmann, T., Sachinidis, A., Rahnenfuhrer, J., Leist, M., 2016. Identification of transcriptome signatures and biomarkers specific for potential developmental toxicants inhibiting human neural crest cell migration. *Arch. Toxicol.* 90, 159–180.
- Pallocca, G., Leist, M., 2022. On the usefulness of animals as a model system (part II): considering benefits within distinct use domains. *ALTEX* 39, 531–539.
- Pallocca, G., Mone, M.J., Kamp, H., Luijten, M., Van de Water, B., Leist, M., 2022a. Next-generation Risk Assessment of Chemicals - Rolling Out a Human-Centric Testing Strategy to Drive 3R Implementation: the RISK-Hunt3r Project Perspective. *ALTEX*.
- Pallocca, G., Nyffeler, J., Dolde, X., Grinberg, M., Gstraunthaler, G., Waldmann, T., Rahnenfuhrer, J., Sachinidis, A., Leist, M., 2017. Impairment of human neural crest cell migration by prolonged exposure to interferon-beta. *Arch. Toxicol.* 91, 3385–3402.
- Pallocca, G., Rovida, C., Leist, M., 2022b. On the usefulness of animals as a model system (part I): overview of criteria and focus on robustness. *ALTEX* 39, 347–353.
- Paparella, M., Bennekou, S.H., Bal-Price, A., 2020. An analysis of the limitations and uncertainties of in vivo developmental neurotoxicity testing and assessment to identify the potential for alternative approaches. *Reprod. Toxicol.* 96, 327–336.
- Patterson, E.A., Whelan, M.P., Worth, A.P., 2021. The role of validation in establishing the scientific credibility of predictive toxicology approaches intended for regulatory application. *Comput. Toxicol.* 17, 100144.
- Piersma, A.H., van Benthem, J., Ezendam, J., Kienhuis, A.S., 2018. Validation redefined. *Toxicol. Vitro* 46, 163–165.
- Pistolato, F., Carpi, D., Mendoza-de Gyves, E., Paini, A., Bopp, S.K., Worth, A., Bal-Price, A., 2021. Combining in vitro assays and mathematical modelling to study developmental neurotoxicity induced by chemical mixtures. *Reprod. Toxicol.* 105, 101–119.
- Products, E.P., Residues, t., Hernández-Jerez, A., Adriaanse, P., Aldrich, A., Berny, P., Coja, T., Duquesne, S., Focks, A., Marinovich, M., Millet, M., Pelkonen, O., Pieper, S., Tiktak, A., Topping, C., Widenfalk, A., Wilks, M., Wolterink, G., Crofton, K., Hougaard Bennekou, S., Paparella, M., Tzoulaki, I., 2021. Development of Integrated Approaches to Testing and Assessment (IATA) case studies on developmental neurotoxicity (DNT) risk assessment. *EFSA J.* 19, e06599.
- Ritz, C., Baty, F., Streibig, J.C., Gerhard, D., 2015. Dose-response analysis using R. *PLoS One* 10, e0146021.
- Ryan, K.R., Sirenko, O., Parham, F., Hsieh, J.H., Cronin, E.F., Tice, R.R., Behl, M., 2016. Neurite outgrowth in human induced pluripotent stem cell derived neurons as a high-throughput screen for developmental neurotoxicity or neurotoxicity. *Neurotoxicology* 53, 271–281.
- Sachana, M., Bal-Price, A., Crofton, K.M., Bennekou, S.H., Shafer, T.J., Behl, M., Terron, A., 2019. International regulatory and scientific effort for improved developmental neurotoxicity testing. *Toxicol. Sci.* 167, 45–57.
- Sachana, M., Willett, C., Pistolato, F., Bal-Price, A., 2021. The potential of mechanistic information organised within the AOP framework to increase regulatory uptake of the developmental neurotoxicity (DNT) in vitro battery of assays. *Reprod. Toxicol.* 103, 159–170.
- Schmidt, B.Z., Lehmann, M., Gutbier, S., Nembo, E., Noel, S., Smirnova, L., Forsby, A., Hescheler, J., Avci, H.X., Hartung, T., Leist, M., Kobalak, J., Dinnyes, A., 2017. In vitro acute and developmental neurotoxicity screening: an overview of cellular platforms and high-throughput technical possibilities. *Arch. Toxicol.* 91, 1–33.
- Schmuck, M.R., Temme, T., Dach, K., de Boer, D., Barenys, M., Bendt, F., Mosig, A., Fritsche, E., 2017. Omnisphero: a high-content image analysis (HCA) approach for phenotypic developmental neurotoxicity (DNT) screenings of organoid neurosphere cultures in vitro. *Arch. Toxicol.* 91, 2017–2028.
- Scholz, D., Polt, D., Genewsky, A., Weng, M., Waldmann, T., Schildknecht, S., Leist, M., 2011. Rapid, complete and large-scale generation of post-mitotic neurons from the human LUHMES cell line. *J. Neurochem.* 119, 957–971.
- Sheets, L.P., Li, A.A., Minnema, D.J., Collier, R.H., Creek, M.R., Peffer, R.C., 2016. A critical review of neonicotinoid insecticides for developmental neurotoxicity. *Crit. Rev. Toxicol.* 46, 153–190.
- Shinde, V., Hoelting, L., Srinivasan, S.P., Meisig, J., Meganathan, K., Jagtap, S., Grinberg, M., Liebing, J., Bluthgen, N., Rahnenfuhrer, J., Rempel, E., Stoerber, R., Schildknecht, S., Forster, S., Godoy, P., van Thriel, C., Gaspar, J.A., Hescheler, J., Waldmann, T., Hengstler, J.G., Leist, M., Sachinidis, A., 2017. Definition of transcriptome-based indices for quantitative characterization of chemically disturbed stem cell development: introduction of the STOP-Toxikon and STOP-Toxikon tests. *Arch. Toxicol.* 91, 839–864.
- Simon, J.M., Paranjape, S.R., Wolter, J.M., Salazar, G., Zylka, M.J., 2019. High-throughput screening and classification of chemicals and their effects on neuronal gene expression using RASL-seq. *Sci. Rep.* 9, 4529.
- Smirnova, L., Hogberg, H.T., Leist, M., Hartung, T., 2014. Developmental neurotoxicity - challenges in the 21st century and in vitro opportunities. *ALTEX* 31, 129–156.

- Spreng, A.S., Brull, M., Leisner, H., Suciu, I., Leist, M., 2022. Distinct and dynamic transcriptome adaptations of iPSC-generated astrocytes after cytokine stimulation. In: *Cells* 11.
- Stiegler, N.V., Krug, A.K., Matt, F., Leist, M., 2011. Assessment of chemical-induced impairment of human neurite outgrowth by multiparametric live cell imaging in high-density cultures. *Toxicol. Sci.* 121, 73–87.
- Strickland, J., Truax, J., Corvaro, M., Settivari, R., Henriquez, J., McFadden, J., Gullidge, T., Johnson, V., Gehen, S., Germolec, D., Allen, D.G., Kleinstreuer, N., 2022. Application of defined approaches for skin sensitization to agrochemical products. *Front Toxicol* 4, 852856.
- Thomas, R.S., Paules, R.S., Simeonov, A., Fitzpatrick, S.C., Crofton, K.M., Casey, W.M., Mendrick, D.L., 2018. The US Federal Tox21 Program: a strategic and operational plan for continued leadership. *ALTEX* 35, 163–168.
- Tohyama, C., 2016. Developmental neurotoxicity test guidelines: problems and perspectives. *J. Toxicol. Sci.* 41, SP69–SP79.
- Tsuji, R., Crofton, K.M., 2012. Developmental neurotoxicity guideline study: issues with methodology, evaluation and regulation. *Congenital. Anom.* 52, 122–128.
- Us Epa Ord, C.f.C.T., 2022. ToxCast Database: Invitrodb Version 3.5. The United States Environmental Protection Agency's Center for Computational Toxicology and Exposure.
- USEPA, 1998. Health Effects Test Guidelines OCSP 870.6300 Developmental Neurotoxicity Study. Washington, DC.
- van Thriel, C., Westerink, R.H., Beste, C., Bale, A.S., Lein, P.J., Leist, M., 2012. Translating neurobehavioural endpoints of developmental neurotoxicity tests into in vitro assays and readouts. *Neurotoxicology* 33, 911–924.
- Vinken, M., Benfenati, E., Busquet, F., Castell, J., Clevert, D.A., de Kok, T.M., Dirven, H., Fritsche, E., Geris, L., Gozalbes, R., Hartung, T., Jennen, D., Jover, R., Kandarova, H., Kramer, N., Krul, C., Luechtefeld, T., Masereeuw, R., Roggen, E., Schaller, S., Vanhaecke, T., Yang, C., Piersma, A.H., 2021. Safer chemicals using less animals: kick-off of the European ONTOX project. *Toxicology* 458, 152846.
- Worth, A.P., Balls, M., 2001. The importance of the prediction model in the validation of alternative tests. *Altern Lab Anim* 29, 135–144.

Blum & Masjosthusmann et al. (2022): In vitro battery for DNT testing

Supplementary information for

Establishment of a human cell-based in vitro battery to assess developmental neurotoxicity hazard of chemicals

Jonathan Blum^{1, #}, Stefan Masjosthusmann^{2, #}, Kristina Bartmann², Farina Bendt², Xenia Dolde¹, Arif Dönmez², Nils Förster⁴, Anna-Katharina Holzer¹, Ulrike Hübenthal², Hagen Eike Keßel², Sadiye Kilic¹, Jördis Klose², Melanie Pahl², Lynn-Christin Stürzl², Iris Mangas⁵, Andrea Terron⁵, Martin Scholze⁶, Axel Mosig⁴, Marcel Leist^{1, °}, Ellen Fritsche^{2, 3, °}

[#]these authors contributed equally; [°]these authors contributed equally

Table of Contents

Fig. S1	Commented list of positive controls used in the IVB-EU	page 2
Fig. S2	Commented list of negative controls used in the IVB-EU	page 3
Fig. S3	Classifications of test compounds as hits and alerts	page 4
Fig. S4	Overview of biological pathways known to contribute to the readouts of NAM used in the IVB-EU	page 5
Fig. S5	List of compounds that had only cytotoxic or no effects	page 6
Fig. S6	Screen hits of IVB-EU in comparison to ToxCast cytotoxicity assays	page 7
Fig. S7	Numbers of compounds detected by each assay of the IVB-EU as being cytotoxic	page 8
References		page 9
Annex I	ToxTemp NPC1	page 10
Annex II	ToxTemp NPC2-5	page 39
Annex III	ToxTemp UKN2	page 70
Annex IV	ToxTemp UKN4	page 91
Annex V	ToxTemp UKN5	page 119

Blum & Masjosthusmann et al. (2022): In vitro battery for DNT testing

	Positive controls (28)	Group	Reference	Comments
1	5,5-Diphenylhydantoin	pharmaceutical drug	[1]	Anti-seizure medication.
2	Acrylamide	industrial chemical	[2], [6]	Generated also in processed food; known neurotoxicant.
3	all-trans Retinoic acid	signalling molecule	[1]	Morphogen involved in brain development. Human evidence. Vitamin A metabolite.
4	Cadmium chloride	heavy metal	[1], [3]	DNT evidence based on animal and human data.
5	Chlorpromazine	pharmaceutical drug	[1]	First generation neuroleptic; multiple receptor inhibitions
6	Chlorpyrifos	pesticide	[1], [8]	Inhibitor of AChE. Evidence for DNT in humans.
7	Deltamethrin	pesticide	[4], [6]	DNT evidence (including epidemiological human studies) summarized by EFSA panel 2021.
8	Dexamethasone	pharmaceutical drug	[1]	Glucocorticoid
9	Domoic acid	environmental	[1], [7]	Causes shellfish poisoning. DNT evidence from animal studies.
10	Haloperidol	pharmaceutical drug	[1]	First generation neuroleptic; multiple receptor inhibitions
11	Hexachlorophene	industrial/pesticide	[1]	Human evidence. Disinfectant.
12	(±) Ketamine	pharmaceutical drug	[1]	NMDA receptor antagonist.
13	Lead (II) acetate	heavy metal	[1], [5]	Human evidence.
14	Maneb	pesticide	[1]	Thiourea fungicide containing manganese.
15	Manganese (II) chloride	metal	[1], [8]	Human evidence.
16	Methylazoxymethanol	environmental	[1]	Targets neuroblasts in CNS. Used in animal models to induce disease phenotypes.
17	Methylmercury chloride	metals	[1], [5]	Human evidence.
18	Nicotine	environmental	[1], [9]	Agonist of nAChRs. Human evidence.
19	Paraquat dichloride hydrate	pesticide	[1]	Herbicide. Linked to development of Parkinson's disease.
20	PBDE 47	industrial chemical	[1], [8]	Bromoaromatic flame retardant.
21	PBDE 99	industrial chemical	[1], [8]	Bromoaromatic flame retardant.
22	PFOA	industrial chemical	[1]	Used as industrial surfactant. Perfluorinated carboxylic acid.
23	PFOSK	industrial chemical	[1]	Used as industrial surfactant. Perfluorinated sulfonic acid.
24	Sodium valproate	pharmaceutical drug	[1]	Human evidence. Used to treat epilepsy and bipolar disorders.
25	Tebuconazole	pesticide	[6]	Triazole fungicide.
26	Tributyltin chloride	industrial chemical	[6]	Organotin compound. Inhibitor of mitochondrial ATP synthase.
27	Trichlorfon	pesticide	[6]	Inhibitor of AChE. Prodrug, which is activated non-enzymatically into dichlorvos (DDVP).
28	Triethyl-tin bromide	organotin	[1]	Neurotoxic organotin. Toxic to myelin.

Fig. S1: Commented list of positive controls used in the IVB-EU

A rough functional grouping of the 28 chemicals used as positive controls is provided. Details on the compounds (CAS-number, full name, abbreviation, etc.) are provided in the [suppl. Excel sheet](#). The reference column indicates the source of information used for classification of the compounds as positive controls. [1] Aschner et al. (2017); [2] Chain (2015); [3] Chandravanshi et al. (2021); [4] EFSA PPR Panel (2021); [5] Grandjean and Landrigan (2014); [6] Mundy et al. (2015); [7] Costa et al. (2010); [8] Grandjean and Landrigan (2006); [9] LeSage et al. (2006). Full citations are found in the [references chapter](#) of this suppl. document.

Blum & Masjosthusmann et al. (2022): In vitro battery for DNT testing

	Negative controls (17)	Group	Comments
1	Acetaminophen*	drug	Pain medication during pregnancy.
2	Amoxicillin*	antibiotic	Used to treat infections during pregnancy.
3	Aspirin	drug	Prescribed against pre-eclampsia.
4	Buspirone	drug	Anxiolytic
5	Chlorpheniramine	drug	Antihistamine
6	D-Glucitol*	sugar derivative	Converted in body to fructose.
7	Diethylene glycol*	solvent	Metabolite ethylen glycol is toxic at high conc.
8	D-Mannitol*	sugar derivative	Sweetener
9	Doxylamine	drug	Antihistamine
10	Famotidine	drug	Histamine H2-receptor antagonist (anti-ulcer).
11	Ibuprofen*	drug	Pain medication; COX-inhibitor.
12	Metformin	drug	Type 2 diabetes medication.
13	Metoprolol*	drug	β -receptor blocker
14	Penicillin	antibiotic	Used to treat bacterial infections.
15	Saccharin*	food additive	Sweetener
16	Sodium benzoate	food additive	Antioxidant
17	Warfarin*	drug	Reproductive toxicant, but not DNT.

Fig. S2: Commented list of negative controls used in the IVB-EU

A rough functional grouping of the 17 chemicals used as negative controls is provided. Details on the compounds (CAS-number, full name, abbreviation, etc.) are provided in the [suppl. Excel sheet](#). Note that the negative classification refers not only to the compounds as such, but to the compounds used in a concentration range of up to 20 μ M. In this range, literature data, and often clinical use suggest the absence of effects or of mechanisms relevant to DNT. *: suggested as negative control in [Aschner et al. \(2017\)](#).

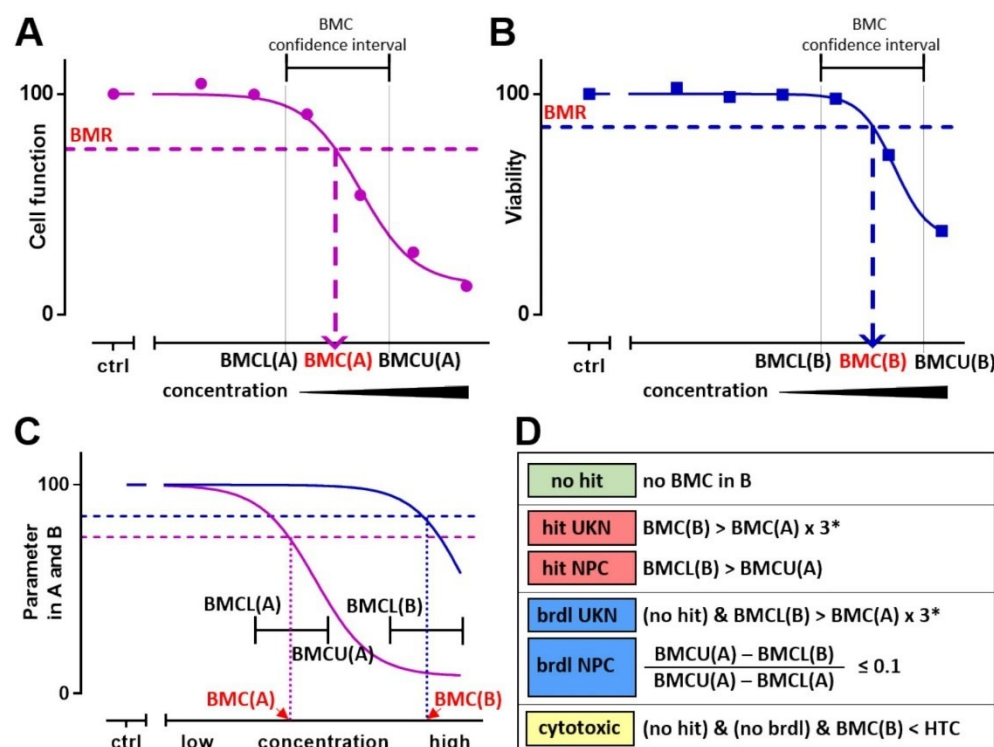


Fig. S3: Classifications of test compounds as hits and alerts

For each compound and each assay, two sets of concentration-response data were produced, one for the main functional endpoint of the NAM (e.g. migration, proliferation, neurite growth or differentiation) and one for the viability of the test system used. Summary data (e.g. the benchmark concentration (BMC) and its confidence interval) were produced from both data sets and used for classification of compounds. (A) Example of a data set on a functional endpoint. In each assay, a benchmark response (BMR) was defined (see [ToxTemps annexes](#)) as threshold between effect and no effect. The intersection of the BMR with the concentration-response curve defined the BMC. The uncertainty of the BMC was expressed by a confidence interval with the BMCL as lower limit and the BMCU as upper limit. BMC (A), BMCL(A) and BMCU(A) are the specific values of the example curve A. (B) Example of a data set on a viability endpoint. In each assay, a benchmark response (BMR) was defined as threshold between effect and no effect. Note that BMRs are assay-specific. BMC(B), BMCL(B) and BMCU(B) are the specific values of the example curve B. (C) An example is given for a data set for a compound that would be considered a screen hit: the BMC(A) and BMC(B) are separated by a large extent. For compounds with less separation, a borderline classification would result. Cytotoxic compounds would show no separation. Inactive compounds would have no responses. (D) Quantitative classification scheme according to the principle described qualitatively in (C): specific hits (hits), borderline hits (brdl) and cytotoxic hits (cytotoxic) would all be considered as “alerts”. They can be grouped in different ways for hit definitions and statistics (Fig. 4). The definitions are given for all assays (UKN = UKN2, UKN4, UKN5; NPC = NPC1-5) according to the respective assays’ ToxTemp description. *UKN assays are defined by ratios between summary data for functional endpoint and viability. A ratio of three is indicated here exemplarily and applies to the UKN5 test. Other ratios are part of the prediction models of UKN2 and UKN4.

Blum & Masjosthusmann et al. (2022): In vitro battery for DNT testing

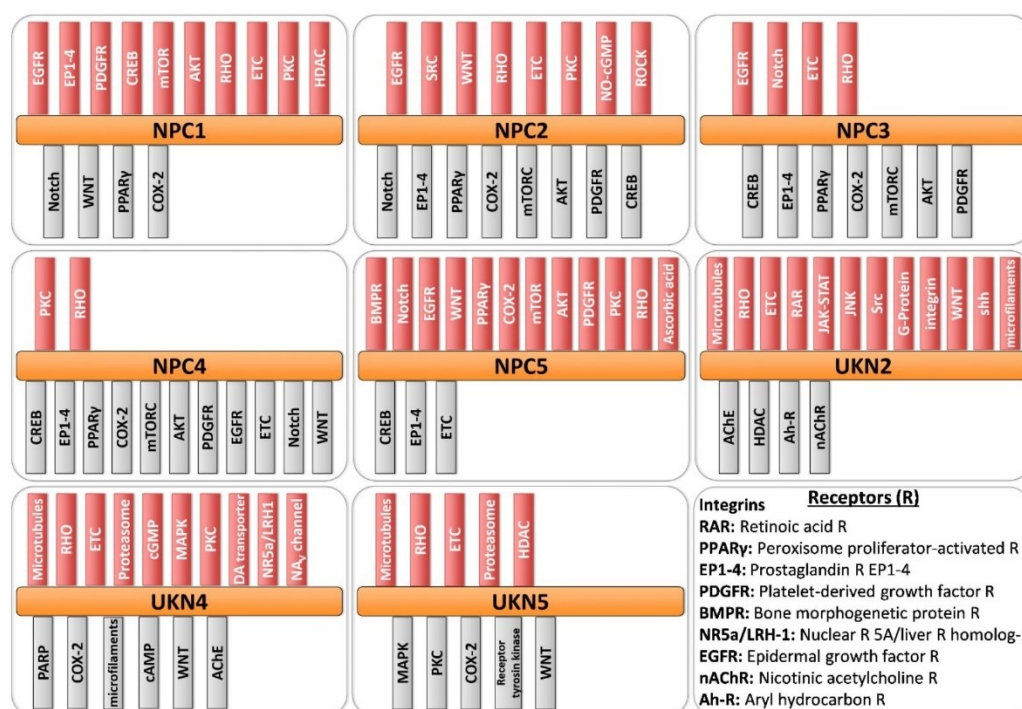


Fig. S4: Overview of biological pathways known to contribute to the readouts of NAM used in the IVB-EU

During the setup and readiness evaluation of the assays, pathway specific “tool compounds” were tested and evaluated for their effect on the test endpoint. These compounds are mostly pharmacological inhibitors or activators of enzymes/receptors/transporters with high specificity for their target. When they effected the test endpoint, it was assumed that the pathway or biochemical mechanism affected by these compounds played a role in the test system, so that it affected the overall readout. For instance, if modulators of the Rho/ROCK signaling cascade affected a test endpoint, it was concluded that toxicants that regulate this pathway would also be detected (displayed by bars on top of the assay name). If modulators of a pathway/target did not affect a test endpoint, it was concluded that a toxicant affecting the respective target or pathway would not be detected by the test (bars below the assay name). In many cases, pathway modulation was only tested in one direction (e.g. only inhibitors of the electron transport chain or only activators of the wnt pathway). This leaves open whether opposite modulator would also have an effect. For such details, original publications, (Koch et al., 2022) and (Masjosthusmann et al., 2020) give more details. Abbreviations of receptors are given in the figure. Notch: notch signalling pathway; COX-2: cyclooxygenase-2; CREB: cAMP response element-binding protein; mTOR: mammalian target of rapamycin; AKT: protein kinase B; ETC: electron transport chain; PKC: protein kinase C; HDAC: histone deacetylase; SRC: proto-oncogene tryosin-protein kinase Src; NO-cGMP: nitric oxide-cGMP sensitive kinase; ROCK: Rho-associated protein kinase; JNK: c-Jun N-terminal kinases; shh: sonic hedgehoc protein; AChE: acetylcholinesterase; PARP: Poly (ADP-ribose) polymerase; MAPK; mitogen-activated protein kinase; DA: dopamine; NA v : voltage gated sodium channel; WNT: wnt signaling; JAK-STAT: JAK-STAT signaling pathway; cGMP: cGMP-related signal transduction; cAMP: cAMP- related signal transduction

A

Compound	Viability assays					
	UKN2	UKN4	UKN5	NPC1	NPC2a [72 h]	NPC2-5 [120 h]
Buspirone		BMR nr*				
Ethylene Thiourea				5.1		
Flufenacet	4.1					
Glycerol		BMR nr*				
Malaoxon	4.1					
Mancozeb			BMR nr*			
Omethoate			4.1			
Parathion-methyl						5.0
Perfluorooctanoic acid		4.3	3.1			
Tri-allate		4.7				

B

compound	category	compound	category
1 (-)-Nicotine	positive control	27 Famotidine	negative control
2 (+)-Ketamine hydrochloride	positive control	28 Fenamidone	screening compound
3 5,5-Diphenylhydantoin	positive control	29 Ibuprofen	negative control
4 Acephate	screening compound	30 Imidacloprid	screening compound
5 Acetaminophen	negative control	31 Mepiquat chloride	screening compound
6 Acetamiprid	screening compound	32 Metformin	negative control
7 Aldicarb	screening compound	33 Methamidophos	screening compound
8 Amoxicillin	negative control	34 Methimazole	screening compound
9 Aspirin	negative control	35 Metoprolol	negative control
10 Bis-(2-butoxyethyl)phosphate	screening compound	36 Octamethylcyclotetrasiloxane	screening compound
11 Boscalid	screening compound	37 Penicillin VK	negative control
12 Captopril	screening compound	38 Pymetrozine	screening compound
13 Chlorpheniramine maleate	negative control	39 Saccharin	negative control
14 Chlorpyrifos-methyl	screening compound	40 Sodium benzoate	negative control
15 Cymoxanil	screening compound	41 Sodium chlorite	screening compound
16 Cypermethrin	screening compound	42 Sodium L-glutamate hydrate	screening compound
17 D-Glucitol	negative control	43 Sodium perchlorate	screening compound
18 Diazinon	screening compound	44 Spiroclufen	screening compound
19 Diethylene glycol	negative control	45 Tembotrione	screening compound
20 Dimethoate	screening compound	46 Terbutaline hemisulfate	screening compound
21 Dinotefuran	screening compound	47 Thiamethoxam	screening compound
22 Disulfoton	screening compound	48 Topramezone	screening compound
23 D-Mannitol	negative control	49 Tris(2-Chloroisopropyl)phosphate	screening compound
24 Domoic acid	positive control	50 Tris(chloroethyl)phosphate	screening compound
25 Doxylamine succinate	negative control	51 Warfarin	negative control
26 Etofenprox	screening compound		

Fig. S5: List of compounds that had only cytotoxic or no effects

Based on the screen results, all 61 compounds were selected that produced no specific hit on any of the assays (no functional endpoint affected at non-cytotoxic concentrations). (A) Ten compounds were found to be cytotoxic in at least one assay. The table lists the viability assessment belonging to the mentioned functional assays (e.g. UKN2 means the viability assay run within the functional testing of UKN2 and thus assessing effects on neural crest cells as test system). The cytotoxic potency is given in units of $-\log(M)$. BMR nr*: the concentration-response curve did not cross the BMR (defined in this assay at 75% for hit classification). But compounds reduced viability by more than 10%, which is defined in this assay as cytotoxicity alert. Therefore, compounds have no BMC value according to the classification scheme, but are still cytotoxic according to the assays own prediction model (at the highest screen concentration). (B) List of all 51 compounds that were neither cytotoxic nor produced any other alert across the IVB-EU.

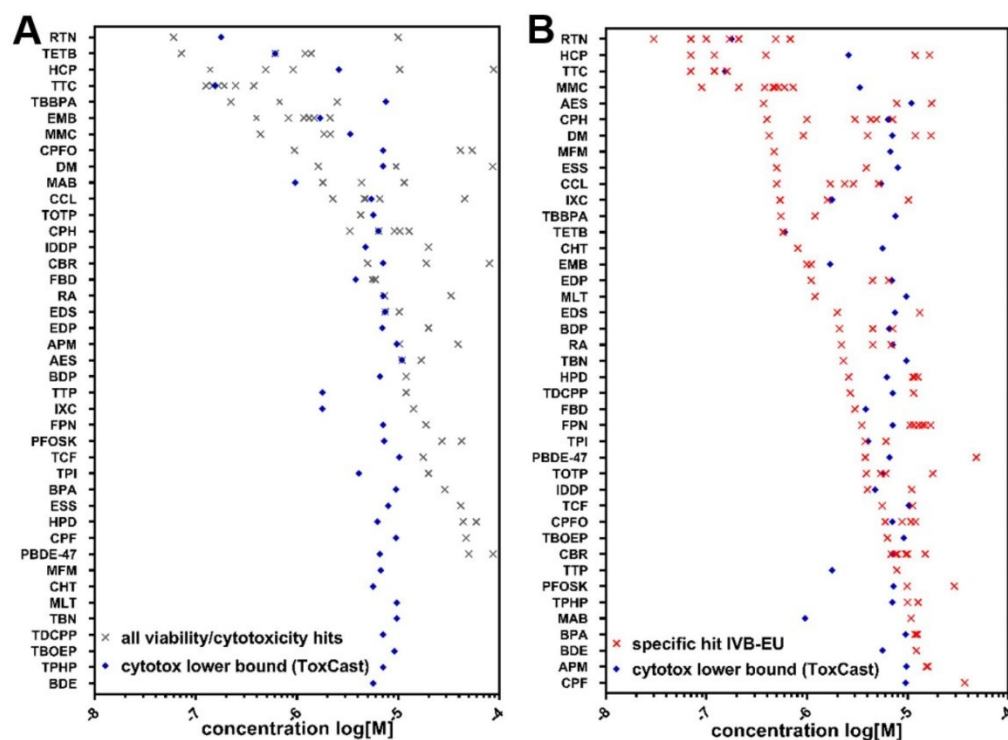


Fig. S6: Screen hits of IVB-EU in comparison to ToxCast cytotoxicity assays

Screen hits (see Fig. 5 for all screen hits and abbreviations) are compared to the cytotox lower bound (CBL) across all ToxCast cytotoxicity assays extracted from the EPA ToxCast Screening Library (<https://comptox.epa.gov/dashboard/chemical-lists/toxcast>). The CBL is calculated as 3-times the median absolute deviation below the median of all hits across the set of ToxCast cytotoxicity assays for each compounds with at least 2 cytotoxicity hits (Judson et al., 2016). Compound with less than 2 hits are left out. (A) Cytotoxic concentrations of all screen hits across the IVB-EU are compared to the CBL. Compounds are sorted by their cytotoxicity potency in the IVB-EU. The lower eight compounds have no cytotoxicity hit in the IVB-EU. (B) Screen hits of specific test endpoints (i.e. migration, differentiation, proliferation, neurite outgrowth) across the IVB-EU are compared to the CLB. Compounds are sorted according to the potency of their most sensitive endpoint. All data is given in log(M).

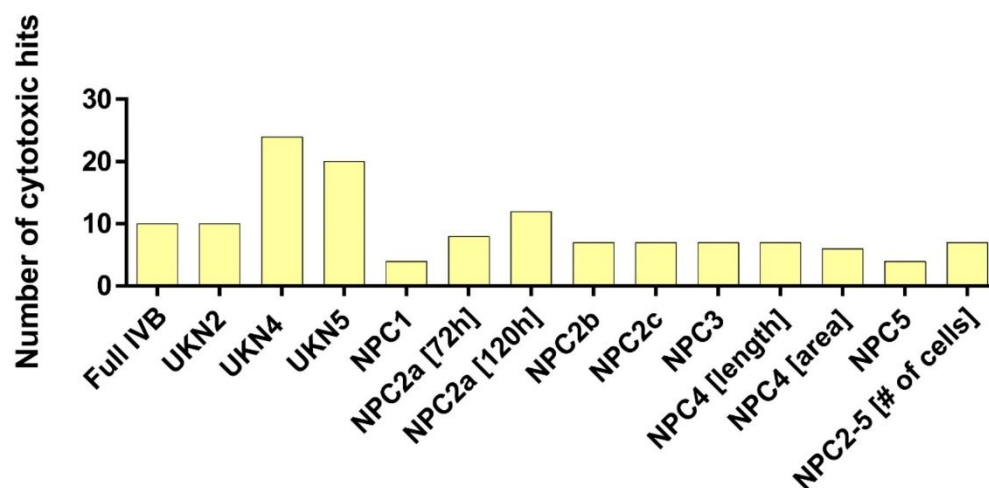


Fig. S7: Numbers of compounds detected by each assay of the IVB-EU as being cytotoxic

The screen was performed and hits were identified as detailed in [Fig. S3](#). The number of cytotoxic hits (out of 120 screen compounds) is indicated for each assay of the battery, and for the total IVB-EU (most leftward bar). The number of specific hits and of borderline hits can be seen in [Fig. 6](#).

Blum & Masjosthusmann et al. (2022): In vitro battery for DNT testing

References

- Aschner, M., Ceccatelli, S., Daneshian, M., Fritsche, E., Hasiwa, N., Hartung, T., Hogberg, H.T., Leist, M., Li, A., Mundi, W.R., Padilla, S., Piersma, A.H., Bal-Price, A., Seiler, A., Westerink, R.H., Zimmer, B., Lein, P.J., 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.
- Chain, E.P.o.C.i.t.F., 2015. Scientific Opinion on acrylamide in food. *EFSA Journal* 13, 4104.
- Chandravanshi, L., Shiv, K., Kumar, S., 2021. Developmental toxicity of cadmium in infants and children: a review. *Environ Anal Health Toxicol* 36, e2021003-2021000.
- Costa, L.G., Giordano, G., Faustman, E.M., 2010. Domoic acid as a developmental neurotoxin. *Neurotoxicology* 31, 409-423.
- Grandjean, P., Landrigan, P.J., 2006. Developmental neurotoxicity of industrial chemicals. *Lancet* 368, 2167-2178.
- Grandjean, P., Landrigan, P.J., 2014. Neurobehavioural effects of developmental toxicity. *Lancet Neurol* 13, 330-338.
- Judson, R., Houck, K., Martin, M., Richard, A.M., Knudsen, T.B., Shah, I., Little, S., Wambaugh, J., Woodrow Setzer, R., Kothiya, P., Phuong, J., Filer, D., Smith, D., Reif, D., Rotroff, D., Kleinstreuer, N., Sipes, N., Xia, M., Huang, R., Crofton, K., Thomas, R.S., 2016. Editor's Highlight: Analysis of the Effects of Cell Stress and Cytotoxicity on In Vitro Assay Activity Across a Diverse Chemical and Assay Space. *Toxicol Sci* 152, 323-339.
- Koch, K., Bartmann, K., Hartmann, J., Kapr, J., Klose, J., Kuchovska, E., Pahl, M., Schluppmann, K., Zuhr, E., Fritsche, E., 2022. Scientific Validation of Human Neurosphere Assays for Developmental Neurotoxicity Evaluation. *Front Toxicol* 4, 816370.
- LeSage, M.G., Gustaf, E., Dufek, M.B., Pentel, P.R., 2006. Effects of maternal intravenous nicotine administration on locomotor behavior in pre-weanling rats. *Pharmacol Biochem Behav* 85, 575-583.
- Masjosthusmann, S., Blum, J., Bartmann, K., Dolde, X., Holzer, A.-K., Stürzl, L.-C., Keßel, E.H., Förster, N., Dönmez, A., Klose, J., Pahl, M., Waldmann, T., Bendt, F., Kisitu, J., Suciu, I., Hübenthal, U., Mosig, A., Leist, M., Fritsche, E., 2020. Establishment of an a priori protocol for the implementation and interpretation of an in-vitro testing battery for the assessment of developmental neurotoxicity. *EFSA Supporting Publications* 17, 1938E.
- Mundy, W.R., Padilla, S., Breier, J.M., Crofton, K.M., Gilbert, M.E., Herr, D.W., Jensen, K.F., Radio, N.M., Raffaele, K.C., Schumacher, K., Shafer, T.J., Cowden, J., 2015. Expanding the test set: Chemicals with potential to disrupt mammalian brain development. *Neurotoxicol Teratol* 52, 25-35.
- Products, E.Panel o.P.P., Residues, t., Hernández-Jerez, A., Adriaanse, P., Aldrich, A., Berny, P., Coja, T., Duquesne, S., Focks, A., Marinovich, M., Millet, M., Pelkonen, O., Pieper, S., Tiktak, A., Topping, C., Widenfalk, A., Wilks, M., Wolterink, G., Crofton, K., Hougaard Bennekou, S., Paparella, M., Tzoulaki, I., 2021. Development of Integrated Approaches to Testing and Assessment (IATA) case studies on developmental neurotoxicity (DNT) risk assessment. *EFSA Journal* 19, e06599.

Establishment of a human cell-based in vitro battery to assess developmental neurotoxicity hazard of chemicals

Jonathan Blum, Stefan Masjosthusmann, Kristina Bartmann, Farina Bendt, Xenia Dolde, Arif Donmez, Nils Forster, Anna-Katharina Holzer, Ulrike Hübenthal, **Hagen Eike Keßel**, Sadiye Kilic, Jödis Klose, Melanie Pahl, Lynn-Christin Stürzl, Iris Mangas, Andrea Terron, Kevin M. Crofton, Martin Scholze, Axel Mosig, Marcel Leist, Ellen Fritsche

Journal:	Chemosphere
Impact factor:	7.086 (2021)
Contribution to the publication:	15%
	Biostatistics and image analysis software development, data analysis and evaluation, writing of manuscript
Type of authorship:	Co-authorship
Status of publication:	Published 31 st Oct 2022

2.4 Neurodevelopmental toxicity assessment of flame retardants using a human DNT in vitro testing battery

Jördis Klose, Melanie Pahl, Kristina Bartmann, Farina Bendt, Jonathan Blum, Xenia Dolde, Nils Förster, Anna-Katharina Holzer, Ulrike Hübenthal, **Hagen Eike Keßel**, Katharina Koch, Stefan Masjosthusmann, Sabine Schneider, Lynn-Christin Stürzl, Selina Woeste, Andrea Rossi, Adrian Covaci, Mamta Behl, Marcel Leist, Julia Tigges, Ellen Fritsche

Cell Biology and Toxicology

Aufgrund ihrer neurologischen Entwicklungstoxizität sind Flammschutzmittel (flame retardants; FRs) wie z.B. polybromierte Diphenylether verboten und durch alternative FRs wie z.B. Organophosphate ersetzt worden, deren toxikologisches Profil jedoch meistens unbekannt ist. Um ihre neurologische Entwicklungstoxizität einzuschätzen, haben wir diesbezüglich das Gefährdungspotential mehrerer FRs untersucht. Das verwendete Testset umfasste hierbei ausgemusterte polybromierte FRs und Organophosphate: 2,2',4,4'-Tetrabromdiphenylether (BDE47), 2,2',4,4',5-Pentabromdiphenylether (BDE-99), Tetrabromobisphenol A, Triphenylphosphat, Tris(2-butoxyethyl)phosphat und dessen Metabolit Bis-(2-butoxyethyl)phosphat, Isodecyl diphenyl phosphat, Isopropyliertes Triphenylphosphat, Trikresylphosphat, Tris(1,3-Dichlor-2-propyl)phosphat, Tert-Butylphenyl diphenyl phosphat, 2-Ethylhexyldiphenylphosphat, Tris(1-chlorisopropyl)phosphat und Tris(2-chlorethyl)phosphat. Hierfür verwendeten wir eine human basierte DNT in vitro Testbatterie, die eine Vielzahl von Endpunkten der neurologischen Entwicklung abdeckt. Die Potenz gemäß der jeweils empfindlichsten Benchmark-Konzentration (BMC) über die Batterie hinweg lag im Bereich von $< 1 \mu\text{M}$ (5 FRs), $1 < 10 \mu\text{M}$ (7 FRs) bis zum Bereich von $> 10 \mu\text{M}$ (3 FRs). Die Datenauswertung zur Priorisierung mit dem ToxPi-Tool ergab eine andere Rangfolge a) als mit den BMC Werten und b) im Vergleich zu den ToxCast-Daten, was darauf hindeutet, dass die DNT-Gefahr dieser FRs durch ToxCast-Assays nicht gut vorhergesagt wird. Die Extrapolation der BMC Werte ausgehend von der DNT in vitro Batterie auf die FRExposition des Menschen lediglich über die Muttermilch deutet auf ein eher geringes Risiko für einzelne Verbindungen hin. In Anbetracht der Tatsache, dass der Mensch jedoch Gemischen ausgesetzt ist, kann dies dennoch zu einem Risiko führen, insbesondere wenn verschiedene Chemikalien durch unterschiedliche Wirkmechanismen an gemeinsamen Endpunkten wie der Oligodendrozytendifferenzierung konvergieren. Diese FRs Fallstudie legt nahe, dass eine auf menschlichen Zellen basierende DNT in vitro Batterie ein vielversprechender Ansatz für die entwicklungsneurologische Gefahreinschätzung und die Priorisierung von Verbindungen bei der Risikobewertung darstellt.



Neurodevelopmental toxicity assessment of flame retardants using a human DNT in vitro testing battery

Jördis Klose · Melanie Pahl · Kristina Bartmann · Farina Bendt · Jonathan Blum · Xenia Dolde · Nils Förster · Anna-Katharina Holzer · Ulrike Hübenthal · Hagen Eike Keßel · Katharina Koch · Stefan Masjosthusmann · Sabine Schneider · Lynn-Christin Stürzl · Selina Woeste · Andrea Rossi · Adrian Covaci · Mamta Behl · Marcel Leist · Julia Tigges · Ellen Fritsche

Received: 16 November 2020 / Accepted: 11 March 2021
© The Author(s) 2021

Abstract Due to their neurodevelopmental toxicity, flame retardants (FRs) like polybrominated diphenyl ethers are banned from the market and replaced by alternative FRs, like organophosphorus FRs, that have mostly unknown toxicological profiles. To study their neurodevelopmental toxicity, we evaluated the hazard of several FRs including phased-out polybrominated FRs and organophosphorus FRs: 2,2',4,4'-tetrabromodiphenylether

(BDE-47), 2,2',4,4',5-pentabromodiphenylether (BDE-99), tetrabromobisphenol A, triphenyl phosphate, tris(2-butoxyethyl) phosphate and its metabolite bis(2-butoxyethyl) phosphate, isodecyl diphenyl phosphate, triphenyl isopropylated phosphate, tricresyl phosphate, tris(1,3-dichloro-2-propyl) phosphate, tert-butylphenyl diphenyl phosphate, 2-ethylhexyl diphenyl phosphate, tris(1-chloroisopropyl) phosphate, and tris(2-chloroethyl)

Highlights

- A human DNT in vitro testing battery was applied for assessing hazards of phased-out and alternative flame retardants (FR) for prioritization.
- Oligodendrocyte development was identified as a common key event for FR-induced DNT in vitro.
- Multiple modes-of-action seem to contribute to oligodendrocyte toxicity.
- Prioritization of FRs according to the DNT in vitro battery differs from FRs ranking using ToxCast assays.

J. Klose · M. Pahl · K. Bartmann · F. Bendt · U. Hübenthal · H. E. Keßel · K. Koch · S. Masjosthusmann · S. Schneider · L.-C. Stürzl · S. Woeste · A. Rossi · J. Tigges · E. Fritsche
IUF-Leibniz Research Institute for Environmental Medicine,
Auf'm Hennekamp 50, 40225 Duesseldorf, NRW, Germany

J. Blum · X. Dolde · A.-K. Holzer · M. Leist
Department of Biology, University of Konstanz, Universitätsstraße
10, 78464 Konstanz, BW, Germany

N. Förster
Faculty for Biology and Biotechnology, Bioinformatics Group,
RUB – Ruhr University Bochum, Bochum, Germany

A. Covaci
Toxicological Centre, Department of Pharmaceutical Sciences,
University of Antwerp, Universiteitsplein 1, 2610 Wilrijk, Belgium

M. Behl
Division of the National Toxicology Program, National Institute of
Environmental Health Sciences, Research Triangle Park, Durham,
North Carolina 27709, USA

E. Fritsche
Medical Faculty, Heinrich-Heine-University, Universitätsstraße 1,
40225 Duesseldorf, NRW, Germany
e-mail: ellen.fritsche@uni-duesseldorf.de

phosphate. Therefore, we used a human cell-based developmental neurotoxicity (DNT) in vitro battery covering a large variety of neurodevelopmental endpoints. Potency according to the respective most sensitive benchmark concentration (BMC) across the battery ranked from $<1 \mu\text{M}$ (5 FRs), $1<10 \mu\text{M}$ (7 FRs) to the $>10 \mu\text{M}$ range (3 FRs). Evaluation of the data with the ToxPi tool revealed a distinct ranking (a) than with the BMC and (b) compared to the ToxCast data, suggesting that DNT hazard of these FRs is not well predicted by ToxCast assays. Extrapolating the DNT in vitro battery BMCs to human FR exposure via breast milk suggests low risk for individual compounds. However, it raises a potential concern for real-life mixture exposure, especially when different compounds converge through diverse modes-of-action on common endpoints, like oligodendrocyte differentiation in this study. This case study using FRs suggests that human cell-based DNT in vitro battery is a promising approach for neurodevelopmental hazard assessment and compound prioritization in risk assessment.

Keywords Developmental neurotoxicity · Flame retardants · Human cell-based testing battery · 3D in vitro model · New approach methodologies · Hazard assessment

Introduction

Flame retardants (FRs) inhibit or delay the spread of fire by suppressing chemical reactions in the flame or by forming a protective layer on the material surface (Damerud et al. 2001). They are used in commercial products, such as electronics, furniture, and textiles. Since the 1970s, polybrominated diphenyl ether (PBDEs) had been in use as FRs. However, due to their accumulation in environmental samples, house dust, food, animal and human tissues (Damerud et al. 2001; De Wit 2002; Law et al. 2014) and their adversity for human health, particularly neurodevelopment (Chao et al. 2007; Roze et al. 2009; Shy et al. 2011; Eskenazi et al. 2013), the European Commission and the U.S. Environmental Protection Agency (US EPA) caused a phase out of PBDEs in 2004 (Blum et al. 2019). Despite their market ban, they are still present in the environment (Yogui and Sericano 2009; Ma et al. 2013; Law et al. 2014). With the phasing out, PBDEs were replaced by presumably safer and less persistent alternative FRs (aFRs), including organophosphorus FRs (OPFRs). Several aFRs were released onto the market,

although their kinetics and toxicities, specifically their neurodevelopmental hazards, have not been sufficiently investigated. Available data on the physico-chemical properties, environmental persistence, bioaccumulation, and toxicity of a subset of aFRs recently displayed large data gaps (van der Veen and de Boer 2012; Bergman et al. 2012; Waaijers et al. 2013). Similar to PBDEs, there has been growing evidence of widespread exposure to aFRs, as they were found in house dust, furniture foam, and baby articles (Stapleton et al. 2009; Sugeng et al. 2017), as well as in hand wipes and urine samples of children (Stapleton et al. 2014; Mizouchi et al. 2015; He et al. 2018a, b; Bastiaensen et al. 2019a). In general, children and especially toddlers are highly exposed towards FRs as they frequently spend their time close to the floor and exercise children-specific mouthing behavior (Fischer et al. 2006; Toms et al. 2009; Sugeng et al. 2017). Due to this high exposure and the fact that the developmental nervous system is a sensitive target organ for many FRs and organophosphorus pesticides (Muñoz-Quezada et al. 2013), which are structurally similar to OPFRs, it is essential to assess the developmental neurotoxicity (DNT) potential of aFRs (Hirsch et al. 2017).

Current DNT testing follows the in vivo guideline studies OECD 426 (OECD 2007) or EPA 870.6300 (EPA 1998) performed with rats. These studies are highly demanding with regard to time, money, and animals (Lein et al. 2005; Crofton et al. 2012) and are not suited for large scale DNT testing. Further limitations include their high variability and lack of reproducibility, as well as the uncertainty of extrapolation from animals to humans (Tsuji and Crofton 2012; Terron and Bennekou Hougaard 2018; Sachana et al. 2019). Therefore, regulators, academic, and industrial scientists recently agreed on a need for a new testing strategy to assess the DNT potential of chemicals (Crofton et al. 2014; Bal-Price et al. 2015; Fritsche et al. 2018b). A mechanistically informed, fit-for-purpose, human-relevant in vitro DNT test battery was suggested that covers different neurodevelopmental processes and stages (Andersen 2003; Bal-Price et al. 2018) and allows a faster and cheaper evaluation of substances for their DNT potential (EFSA 2013; Bal-Price et al. 2015, 2018; Fritsche et al. 2015, 2017, 2018a).

In this study, human-induced pluripotent stem cell (hiPSC)-derived neural crest cells (NCC), human mesencephalic cells (LUHMES), 3D human primary neural progenitor cell (NPC)-based neurospheres, as well as hiPSC-derived peripheral neurons were applied to study distinct neurodevelopmental key events (KEs)

in vitro. These KEs include NPC proliferation (NPC1), NCC (cMINC/UKN2), radial glia (NPC2a), neuronal (NPC2b) and oligodendrocyte (NPC2c) migration, differentiation into neurons (NPC3), neurite morphology (NPC4, NeuriTox/UKN4, PeriTox/UKN5), and oligodendrocyte differentiation (NPC5; Baumann et al. 2016; Barenys et al. 2017; Schmidt et al. 2017; Fritsche et al. 2018a; Masjosthusmann et al. 2018; Nimtz et al. 2019; Krebs et al. 2020b). These assays comprise a current DNT in vitro testing battery that was recently assembled to test 119 compounds (e.g., carbamates, metals, neonicotinoids, organochlorines/fluorines, and organophosphates pyrethroids) for regulatory purposes. Using selected known human DNT positive and negative compounds as benchmark, this battery performed with a sensitivity of 100% and a specificity of 88% (Masjosthusmann et al. 2020).

To study the neurodevelopmental hazard of FRs, we analyzed their adverse effects on the endpoints of this battery of human neurodevelopmental assays. FRs used include a set of phased-out and currently in use compounds. The phased-out FRs are PBDEs 2,2',4,4'-tetrabromodiphenylether (BDE-47) and 2,2',4,4',5-pentabromodiphenylether (BDE-99), while the current-use FRs include the organophosphorus FRs (OPFRs), such as triphenyl phosphate (TPHP), tris (2-butoxyethyl) phosphate (TBOEP) and its metabolite bis-(2-butoxyethyl) phosphate (BBOEP), isodecyl diphenyl phosphate (IDDPHP), triphenyl isopropylated phosphate (IPPHP), tricresyl phosphate (TCP), tris (1,3-dichloro-isopropyl) phosphate (TDCIPP), tert-butylphenyl diphenyl phosphate (t-BDPDHP), tri-O-cresyl phosphate (TOCP), 2-ethylhexyl diphenyl phosphate (EHDPHP), tris (1-chloro-isopropyl) phosphate (TCIPP), and tris (2-chloroethyl) phosphate (TCEP), as well as the brominated FR Tetrabromobisphenol A (TBBPA) (Table S1). The in vitro data were related to hazardous doses by toxicokinetic considerations. Moreover, such data were compared to potential exposure situations. Relating the phenomics of the in vitro methods to molecular signatures, we performed RNA sequencing analyses. This approach represents a case study for a new risk assessment paradigm for DNT by using phenotypic readouts of human cell-based assays that cover a variety of neurodevelopmental endpoints and studying their molecular signatures in response to different FRs.

Material and methods

Chemicals

TBBPA, BDE-99, TCEP, TPHP, TOCP, and TBOEP (for NPC assays) were purchased from Sigma-Aldrich and were dissolved as 50 mM and 20 mM stocks in dimethyl sulfoxide (DMSO; Carl Roth GmbH). The metabolite BBOEP (1500 ng/μL in Methanol) was custom synthesized by Dr. Vladimir Belov (Max Planck Institute, Göttingen, Germany) with a purity > 98% as measured by MS and NMR techniques. The FRs TCIPP, t-BDPDHP, and EHDPHP were obtained from ToxCast and are diluted in DMSO with stock concentration of 20 mM. All other flame retardants IDDPHP, IPPHP, TCP, TDCIPP, BDE-47 (for NPC assays) as well as TBBPA, BDE-47, BDE-99, TCEP, TPHP, IDDPHP, IPPHP, EHDPHP, t-BDPDHP, and TCP (for UKN assays) were provided by M. Behl from the National Toxicology Program, and stock solutions of 20 mM in DMSO were prepared. Solvent concentrations were 0.1% DMSO and 0.4% MeOH for BBOEP in dose-response experiments.

Cell culture

Human NPCs (hNPCs) from three different individuals (gestational week 16-19) were purchased from Lonza Verviers SPRL, Belgium. They were thawed and isolated as previously described (Baumann et al. 2016). hNPCs were cultured as free floating neurospheres in proliferation medium consisting of DMEM (Life Technologies) and Hams F12 (Life Technologies) (3:1) supplemented with 2% B27 (Life Technologies), 20 ng/mL EGF (Thermo Fisher), FGF (R&D Systems), and 1% penicillin and streptomycin (Pan-Biotech). Neurospheres were cultivated at 37 °C with 5% CO₂, passaged mechanically with a tissue chopper (McIlwain) once a week and thrice a week half of the medium was replaced.

For the cMINC assay (UKN2), NCCs are differentiated from the hiPSC line IMR90_clone #4 (WiCell, Wisconsin) by plating cells on Matrigel-coated 6-well plates (Falcon) at a density of 50000 cells/cm². One day prior differentiation, cells are cultivated in essential 8 (E8) medium (DMEM/F12 supplemented with 15 mM Hepes, 16 mg/mL L-ascorbic-acid, 0.7 mg/mL sodium selenite, 20 μg/mL insulin, 10 μg/mL holo-transferrin, 100 ng/mL bFGF, 1.74 ng/mL TGFβ) containing

10 μ M Rock inhibitor. Until 11 days in vitro (DIV), cells receive KSR medium (knock out DMEM, 15% knock out serum replacement, 1% GlutaMax, 1% MEM NEAA solution, 50 μ M 2-mercaptoethanol) which is gradually replaced by 25% increments of N2-S medium (DMEM/F12, 1.55 mg/mL glucose, 1% GlutaMax, 0.1 mg/mL apotransferrin, 25 μ g/mL insulin, 20 nM progesterone, 100 μ M putrescine, 30 nM selenium). From -1 DIV to 11 DIV, cells are cultured at 37 °C with 5% CO₂ and a daily medium change was performed. From 0 DIV to 2 DIV, medium is supplemented with 20 ng/mL Noggin. From 0 DIV to 3 DIV, it is supplemented with 10 μ M SB431542 and from 2 DIV to 11 DIV with 3 μ M CHIR 99021. After 11 DIV, cells are detached and resuspended in N2-S medium supplemented with 20 ng/mL EGF and 20 ng/mL FGF2 and seeded as droplets (10 μ L) on poly-L-ornithine (PLO)/laminin/fibronectin-coated 10-cm dishes. Until 39 DIV, cells are expanded by weekly splitting in N2-S medium supplied with EGF and FGF2 and a medium change is performed every other day. On 39 DIV, cells are detached, resuspended in freeze medium (FBS with 10% DMSO), and frozen at a concentration of 4×10^6 cells per mL at -80 °C overnight. After 24 h, cells are stored in liquid nitrogen until further use.

For the NeuroTox assay (UKN4), LUHMES cells are cultured and handled as described before (Lotharius et al. 2005; Scholz et al. 2011; Krug et al. 2013a). They are maintained in proliferation medium (PMed; AdvDMEM/F12 supplemented with 2 mM glutamine, 1 \times N2 supplement and 40 ng/mL FGF) at 37 °C with 5% CO₂. Cells are passaged every second or third day when reaching approximately 80% confluency. For pre-differentiation, 8×10^6 (45000 cells/cm²) cells are seeded one day before in PMed. Differentiation is started by switching to differentiation medium (DMed; AdvDMEM/F12 supplemented with 2 mM glutamine, 1 \times N2 supplement, 2.25 μ M tetracycline, 1 mM dibutyryl cAMP and 2 ng/mL GDNF).

For the PeriTox assay (UKN5), sensory neurons are differentiated from the hiPSC line SBAD2, which was derived and characterized at the University of Newcastle from Lonza fibroblasts CC-2511, Lot 293971 with the tissue acquisition number 24245 (Baud et al. 2017). Culturing, handling, and differentiation are performed according to standard protocols (Thomson et al. 1998; Chambers et al. 2013; Hoelting et al. 2016). Generation of sensory neurons is started on -2 DIV by resuspending hiPSCs in E8 medium containing 10 μ M

Rock inhibitor Y-27632. After replating cells at a density of 55000 cells/cm² on Matrigel coated 6-well plates (Falcon), a daily medium change is performed from -1 DIV until 10 DIV. E8 medium supplemented with rock inhibitor (10 μ M) is refreshed on -1 DIV. On 0 DIV, neural differentiation is initiated and until 10 DIV cells receive KSR medium which is, from 4 DIV onward, gradually replaced by 25% increments of N2-S medium. Until 4 DIV medium is supplied with 35 ng/mL Noggin, 600 nM dorsomorphin and 10 μ M SB431542 to initiate neutralization via dual-SMAD inhibition. From 2 DIV to 10 DIV, three further pathway inhibitors are added (1.5 μ M CHIR99021, 5 μ M SU5402, and 5 μ M DAPT). On 10 DIV, cells are detached, resuspended in freeze medium (FBS with 10% DMSO) and frozen at a concentration of 8×10^6 cells per mL at -80 °C overnight. After 24 h, cells are stored in liquid nitrogen until further use.

The “neurosphere assay”—NPC1-5

hNPCs were chopped to 0.2 mm 2–3 days before plating to reach a defined size of 0.3 mm. Each compound was tested in serial dilution (1:3) with 7 concentrations and a solvent control (SC) plated in five replicate wells per condition in 96-well plates (proliferation U-bottom, Falcon; differentiation flat bottom, Greiner). Each well contained one sphere in 100 μ L of the respective medium and FR/solvent(s) (proliferation medium (description in “Cell culture”); differentiation medium consisting of DMEM (Life Technologies), Hams F12 (Life Technologies) 3:1 supplemented with 1% of N2 (Life Technologies) and 1% penicillin and streptomycin (Pan-Biotech)). The 1:3 solution series and plate filling, LDH, CTB, and feeding step were performed automatically by STARlet 8 ML pipette robot system (MICROLAB STAR® M; Hamilton).

Proliferation

The proliferation by area (NPC1a) was assessed as slope of the increase in sphere size up to 3 DIV (0 h, 24 h, 48 h, and 72 h) measured by brightfield microscopy and using high content imaging (Cellomics Scan software, Version 6.6.0; Thermo Fisher Scientific). Proliferation by bromodeoxyuridine (BrdU; NPC1b) was analyzed after 3 DIV via a luminescence-based BrdU Assay (Roche) as previously published in Nimtz et al. (2019).

Immunocytochemical stainings

By plating neurospheres into 100 μ L differentiation medium on a poly-D-lysine (0.1 mg/mL, Sigma-Aldrich) and laminin (12.5 μ g/mL, Sigma-Aldrich)-coated 96-well plate (flat bottom, Greiner), spheres settle down and NPCs migrate radially out of the sphere core concurrently differentiating, into radial glia, neurons, and oligodendrocytes. After 5 days of migration and differentiation, human neurospheres were fixed with 4% paraformaldehyde (PFA, Merck) for 30 min at 37 °C and directly afterwards washed three times for 3 min with 250 μ L PBS (Biochrom) before stored at 4 °C until staining. Cells were always covered with 40 μ L PBS, and for staining, 10 μ L blocking solution (PBS, 50% Goat Serum (GS, Sigma-Aldrich) and 5% Bovines Serum Albumin (BSA, Serva Electrophoresis)) per well was added and incubated for 15 min at 37 °C. After removal of 10 μ L, cells were stained overnight at 4 °C with 10 μ L mouse IgM oligodendrocyte O4 antibody solution 1:400 (in PBS with 10% GS and 1% BSA; R&D System) followed by three 3-min washing steps by addition and removal of 250 μ L PBS. After the last washing step, 260 μ L was removed and 10 μ L secondary antibody solution in PBS (1:400 Alexa Fluor 488 anti-mouse IgM (Life Technologies), 10% GS, 5% BSA) was added for 30 min at 37 °C. After washing steps as previously described, cells were fixed a second time for 30 min at 37 °C in 4% PFA, followed by three 3-min washing steps and permeabilization in 0.5% PBS-T for 5 min at room temperature. Afterwards, cells were blocked for 15 min at 37 °C with 10 μ L PBS, 50% Rabbit Serum (RS, Sigma-Aldrich), and 5% BSA. For neuronal staining, neurospheres were incubated for 1 h at 37 °C with 10 μ L conjugated rabbit TUBB3 674 antibody (Abcam) 1:400 (in PBS with 10% RS, 1% BSA, and 5% Hoechst 33258 (Sigma-Aldrich)). After three additional 3-min washing steps, 250 μ L PBS was added to each well and the plates were stored in the dark at 4 °C. Images of immunochemical stainings of three channels (386 nm for Hoechst stained nuclei, 647 nm for β (III)tubulin stained neurons, 488 nm for O4 stained oligodendrocytes) were acquired with a 200-fold magnification and a resolution of 552 \times 552 pixel using the HCS Studio Cellomics software (version 6.6.0; Thermo Fisher Scientific).

Migration and differentiation

Radial glia migration distance (72 h, NPC2a) was analyzed by manual measurement of the radial migration

from the sphere core on brightfield images as number of pixels which is converted to micrometers. After 120 h, it is assessed by automatically identifying (Schmuck et al. 2016) the migration area of each sphere of Hoechst stained nuclei on fluorescence images. The migration distance of neurons (NPC2b) and oligodendrocytes (NPC2c) is defined as mean distance of all neurons/oligodendrocytes within the migrations area divided by radial glia migration distance after 120 h. The differentiation into neurons (NPC3) and oligodendrocytes (NPC5) is determined as number of all β (III)tubulin and O4-positive cells in percent of the total amount of Hoechst-positive nuclei in the migration area and is performed automatically using two convolutional neural networks (CNN) based on the Keras architecture implemented in Python 3, which were trained to identify both cell types. All neurons that were identified in NPC3 are analyzed for their morphology (NPC4) by characterizing the neurite length (in μ m) and area (amount of pixel). Detection of migration (120 h, NPC2) and morphological analysis (NPC4) is calculated automatically by high-content image analysis (HCA) tool Omnisphero (Schmuck et al. 2016). Migrating/differentiating neurospheres were exposed to FRs/solvent(s) for 5 days. On day 3, half of the exposure/solvent medium was exchanged and the supernatant was used to detect cytotoxicity by measuring lactate dehydrogenase (LDH) leakage.

“cMINC assay” UKN2

NCCs were thawed and seeded into 96-well plates in N2-S medium containing FGF2 and EGF according to the previously published protocol (Nyffeler et al. 2017). Cells were seeded around stoppers to create a circular cell-free area and after 24 h stoppers were removed to allow cell migration. One day later, cells were exposed to FRs/solvent(s) for 24 h. The number of migrated cells into the cell free zone was quantified 48 h after stopper removal and 24 h after treatment. Cells were stained with Calcein-AM and Hoechst-33342 (H-33342), and high content imaging was performed. Four images for migration were taken to cover the region of interest (ROI) using a high content imaging microscope (Cellomics ArrayScanVTI), and Calcein and H-33342 double-positive cell numbers were determined by an automated algorithm (RingAssay software; <http://invitro-tox.uni-konstanz.de>). For viability, four fields close to the well borders, i.e., outside the ROI,

were imaged. Viable cells were defined by double-positivity for H-33342 and calcein and determined by an automated algorithm as described before (Nyffeler et al. 2017). TBBPA, BDE-47, BDE-99, IDDPHP, TCP, t-BDPHP, and EHDPHP were tested in serial dilution (1:2) with 6 concentrations and SC, while TPHP and IPPHP were tested with 5 concentrations (Nyffeler et al. 2017). TCEP, TDCIPP, and TCIPP were negative within a 20- μ M pre-screening and therefore not tested further (data not shown). TBOEP, BBOEP, and TOCP were tested 1:3 with 6 concentrations and SC based on the method described in this study. Each compound concentration was plated in 4 replicate wells per condition.

“NeuriTox assay” UKN4

After 2 days of differentiation, 30000 LUHMES cells were reseeded into each well of a 96-well plate in DMED containing only tetracycline. After cells’ attachment for 1 h, they were exposed to FRs/solvent(s) for 24 h. One hour before read-out, cells were stained with Calcein-AM and H-33342 and imaged via a high-content imaging microscope (Cellomics ArrayScanVTI, Thermo Fisher Scientific) to assess neurite area. For neurite area determination, an automated algorithm was used, which calculates the area of the cell soma and subtracts this area from all calcein-positive pixels imaged (Stiegler et al. 2011; Krug et al. 2013a). To assess viability, all stained nuclei (H-33342 positive) are used to determine total cell number and H-33342 and calcein double-positive cells are defined as viable cells (Stiegler et al. 2011; Krug et al. 2013a). Each compound was tested in serial dilutions (1:3) with 10 concentrations starting at 20 μ M and SC plated in three replica wells per condition. Effects of TBBPA, BDE-47, BDE-99, IDDPHP, TCP, t-BDPHP, EHDPHP, TPHP, and IPPHP were assessed in a previous screening (Delp et al. 2018). TDCIPP, TOCP, and TCIPP were negative in a pre-screening at 20 μ M and therefore not tested any further (data not shown).

“PeriTox assay” UKN5

Differentiated sensory neurons were thawed and seeded in 25% KSR/75% N2-S medium supplemented with 1.5 μ M CHIR99021, 5 μ M SU5402, and 5 μ M DAPT into 96-well plates at a density of 100000 cells per cm^2 . After cells’ attachment for 1 h, they were exposed to

FRs/solvent(s) for 24 h. Assessments of neurite area and viability of the cells were performed as described above for the UKN4 assay. Each compound concentration was tested in three wells per plate (technical replicates) in a serial dilution (1:3) with 6 concentrations starting at 20 μ M and SC. Effects of TBBPA, BDE-47, BDE-99, IDDPHP, TCP, t-BDPHP, EHDPHP, TPHP, and IPPHP were assessed in a previous screening (Delp et al. 2018). TDCIPP, TOCP, and TCIPP were negative in a pre-screening at 20 μ M and therefore not tested any further (data not shown).

Viability and cytotoxicity

To distinguish compound effects from secondary effects due to loss of viability and cytotoxicity, respective assays were performed in parallel. Thereby, all viability and cytotoxicity assays are multiplexed within the respective assay. hNPC viability was assessed as mitochondrial activity by using an Alamar blue assay (CellTiter-Blue Assay (CTB); Promega) in the last 2 h of the respective compound treatment period (NPC1 at 3 DIV; NPC2-5 at 5 DIV). Cytotoxicity of treated hNPCs was detected by measuring LDH (CytoTox-ONE membrane integrity Assay; Promega) after 3 (NPC1; NPC2-5) and 5 (NPC2-5) DIV. It is of note that a reduced radial glia migration area causes a reduction in the CTB readout due to a diminished cell number without necessarily affecting cell viability (Fritsche et al. 2018a). Thus, when radial glia migration is inhibited by a compound, the LDH assay is solely the reference for DNT specificity of NPC2-5. Assessment of viability within the UKN assays was performed as described above.

RNA sequencing and RT-qPCR

For RNA sequencing (RNASeq) experiments, 1000 neurospheres per well with a defined size of 0.1 mm were plated onto PDL/laminin-coated 6-well plates and cultivated for 60 h in the presence and absence of selected FRs. The RNA isolation was performed using the RNeasy Mini Kit (Qiagen) according to the manufacturer’s protocol. Total RNA was analyzed for high quality using the Agilent High Sensitivity RNA ScreenTape System for Agilent 4150 TapeStation Bioanalyzer (Agilent Technologies) for human samples with an RNA integrity number (RIN) ≥ 8 . All samples in this study showed high-quality RINs ≥ 8.5 . For RNASeq,

1.0 µg total RNA was used for library preparation using the TruSeq RNA Sample Prep Kit v2 according to the manufacturer's protocol (Illumina). All steps of the protocol were performed as described in the Illumina kit. DNA library templates were quantified using the Qubit™ 4 Fluorometer and the Qubit 1× dsDNA HS Assay Kit (Thermo Fisher Scientific). Quality control and fragment size analysis were performed on Agilent 4150 TapeStation System and the Agilent D1000 Screen Tape System (Agilent Technologies). Sequencing was performed on a MiSeq instrument (Illumina) using v3 chemistry, resulting in an average of 50 million reads per library with 1×76 bp paired end setup.

Raw data were uploaded on BaseSpace Sequence Hub (Illumina) for FastQ generation. RNAseq analysis was performed using the Illumina pipeline (Illumina Annotation Engine 2.0.10.0). The resulting raw reads were assessed for quality, adapter content and duplication rates with the Illumina FASTQ file generation pipeline. Trimmed and filtered reads were aligned versus the *Homo sapiens* reference genome (UCSC hg19) using STAR Aligner (STAR_2.6.1a). Total number of reads was quantified using both TopHat2 and Salmon Quantification (0.11.2). Strelka Variant Caller (2.9.9) was used to detect somatic single nucleotide variants (SNVs).

Quantitative real-time polymerase chain reaction (RT-qPCR) was performed with the QuantiFast SYBR Green PCR Kit (Qiagen) within the Rotor Gene Q Cycler (Qiagen). Therefore, 250 ng RNA was transcribed into cDNA using the QuantiTect Reverse Transcription Kit (Qiagen) according to manufacturer's instructions. Analysis was performed using the software Rotor-Gene Q Series version 2.3.4 (Qiagen). Copy numbers (CN) of the genes of interest were calculated by using gene-specific copy number standards as described previously in detail (Walter et al. 2019) and normalized to the housekeeping gene *beta-actin*. Gene CN of solvent control and FR treated differentiated spheres were normalized to proliferative spheres, which are thought to express very low numbers of oligodendrocyte-specific mRNA. Here, the solvent control visualizes oligodendrocyte-related gene expression as a function of normal NPC development that can directly be compared to sphere development in presence of FRs.

Toxicological Priority Index

For relative toxicological ranking and hierarchical clustering, the BMC values of the tested FRs were integrated

and visualized by using the Toxicological Priority Index Graphical User Interface (ToxPi GUI) version 2.3 (Gangwal et al. 2012). In ToxPi, the BMC values across the data set of each endpoint were scaled with the formula $-\log_{10}(x)+6$ from 0 to 1, while 1 represents the lowest BMC and therefore the most potent compound. If BMC was not reached, a concentration of 10^6 was applied before, which became 0 upon scaling. Data are visualized in a pie chart, where every slice represents one DNT endpoint (Fig. 7). The farther the slice extends from its origin, the more potent the compound in this endpoint. In comparison, ToxCast data was used to give an initial idea on the general toxicity of these FRs across a variety of assays. Regarding ToxCast AC₅₀ (half-maximal activity concentration), values below a given cytotoxicity limit were used and scaled as described above. Each slide was assigned as one intended target family and contains several assays for respective endpoints.

Data analysis and statistics

All neurosphere experiments were performed with at least two different individuals. Experiments were defined as independent if they were generated with NPCs from different individuals or from a different passage of cells. For cMINC, NeuriTox, and PeriTox assays, biological replicates represent an independent experiment on another day with a different batch of NCCs, LUHMES cells, or 10 DIV sensory neurons thawed. If not otherwise indicated, results are presented as mean ± SEM. For dose-response curves, a sigmoidal (variable slope) or bell-shaped curve fit was applied using GraphPad Prism 8.2.1. Statistical significance was calculated using the same software and one-way ANOVA with Bonferroni's post hoc tests ($p \leq 0.05$ was termed significant).

BMC as well as upper and lower confidence intervals (CI) were calculated with GraphPad Prism 8.2.1. Based on overlap of confidence intervals of the BMCs calculated for the DNT-specific endpoints and the endpoints related to cytotoxicity/viability, NPC endpoints were classified as DNT-specific (no CI overlap), unspecific (CI overlap ≥ 10%), or borderline ($0 > \text{CI} < 10\%$; Masjosthusmann et al. 2020). The classification model applied for UKN assays is based on a ratio cutoff for the ratio between the BMC for cell viability and the specific endpoints (ratio BMC₁₀ viability/BMC₂₅ migration ≥ 1.3 in UKN2 assay; ratio BMC₂₅ viability/BMC₂₅

neurite area ≥ 4 in UKN4 assay or ≥ 3 in UKN5 assay). This is in line with the respective classification models suggested in previous publications (Krug et al. 2013b; Hoelting et al. 2016; Nyffeler et al. 2017).

Results

Experimental design of the human DNT testing battery

We assessed the neurodevelopmental hazard of 15 FRs (Table S1) and analyzed their adverse effects using a battery of human-based neurodevelopmental in vitro assays (Fig. 1). Within NPC assays, proliferation (NPC1), migration (NPC2), and differentiation into the main effector cells of the human brain, i.e., radial glia, neurons (NPC3), and oligodendrocytes (NPC5), were evaluated. NPC3 was multiplexed with NPC4, which quantifies neurite morphology by analyzing their length and area. The cMINC (UKN2) assay measures neural crest cell (NCC) migration and viability, while NeuroTox (UKN4) and PeriTox (UKN5) assays assess neurite morphology and viability of LUHMES cells and hiPSC-derived peripheral neurons, respectively. Finally, cytotoxicity was assessed after 3 (NPC1) and 5 (NPC2-5) DIV and cell viability was detected at the end of each assay. Additionally, RNA sequencing analyses provide further insight into the modes-of-action of FR toxicity.

Three out of the 15 analyzed FRs (BBOEP, TCIPP, and TCEP) did not produce significant effects in any of the tested endpoints up to a concentration of 20 μM . Therefore, the respective graphs are shown in supplementary Figs. S1–3.

hNPC proliferation is exclusively disturbed by alternative flame retardants

A fundamental neurodevelopmental KE is NPC proliferation. The analyzed PBDEs and aFRs did not affect sphere area increase over time (NPC1a; Fig. 2(a)). BrdU incorporation (NPC1b), however, as a direct measure of DNA synthesis has a higher sensitivity than NPC1a and EHDPHP and TCP reduced BrdU incorporation significantly (Fig. 2(b)) with EHDPHP being the more potent one with significant diminution of proliferation at 0.25 μM and 20 μM to $70.5 \pm 4.3\%$ and $37.4 \pm 2.7\%$ of the controls, respectively. TCP inhibited proliferation to $65.9 \pm 8.3\%$ and $58.5 \pm 6.8\%$ of controls at 6.6 μM and 20 μM ,

respectively. Neither viability nor cytotoxicity were altered by any of the analyzed FRs at the employed concentration levels, with the exception of IPPHP, which induced the mitochondrial activity at the highest concentration up to $121.1 \pm 4.9\%$ of control. The endpoint-specific control for NPC1 was hNPC cultivation in absence of growth factors causing significantly reduced proliferation (Suppl. Fig. 4(a, b)).

FRs affect migration in a cell type-specific manner

Next, we analyzed NCC (UKN2), radial glia (NPC2a), neuronal (NPC2b), and oligodendrocyte (NPC2c) migration in the presence and absence of FRs. NCC migration was affected by PBDEs, as well as organophosphorus aFRs and was significantly inhibited by 9 out of the 15 FRs tested (Fig. 3(a)). TBBPA reduced NCC migration to $52.6 \pm 9.2\%$ and $31.3 \pm 3.5\%$ of control at 2.5 μM and 5 μM , respectively (Fig. 3(a, c)). BDE-47, t-BPDHP, and TCP ($\geq 5 \mu\text{M}$) significantly reduced the number of migrating NCCs to $37.1 \pm 9.6\%$, $53.5 \pm 4.8\%$, and $56.6 \pm 4.4\%$ of controls, respectively. TOCP (6.67 μM) and BDE-99 (10 μM) significantly inhibited NCC migration to $43.2 \pm 7.6\%$ and $69.5 \pm 6.7\%$ of controls, respectively, while EHDPHP, IDDPHP, and TPHP disturbed NCC migration at the highest concentration to $31.8 \pm 23.1\%$, $52.7 \pm 10.6\%$, and $65.3 \pm 10.2\%$ of respective controls. NCC viability was significantly affected by 5 μM TBBPA ($81.1 \pm 1.7\%$); by $\geq 10 \mu\text{M}$ EHDPHP ($\leq 93.8 \pm 2.7\%$), TCP ($\leq 90.9 \pm 1.0\%$), and IPPHP ($\leq 93.1 \pm 1.2\%$); and by 20 μM BDE-47 ($86.6 \pm 5.5\%$) and TOCP ($63.3 \pm 10.2\%$; Fig. 3(b)). Cytochalasin D (200 nM) served as an endpoint specific control for UKN2 (data not shown). Similar to NCC migration, TBBPA is the most potent FR for hNPC migration inhibition, significantly disturbing radial glia (NPC2a), neuron (NPC2b), and oligodendrocyte (NPC2c) migration at concentrations $\geq 2.2 \mu\text{M}$ (Fig. 3(d, g)). Consequently, TBBPA decreased respective CTB values at concentrations $\geq 2.2 \mu\text{M}$ to $\leq 64.8 \pm 2.7\%$ of controls. However, also cytotoxicity was induced to $25.1 \pm 3.3\%$ (72 h) and $25.4 \pm 2.0\%$ (120 h) of the lysis control at concentrations $\geq 2.2 \mu\text{M}$ TBBPA (Fig. 3(e)).

The phased-out PBDEs did not affect migration behavior of differentiating hNPCs, while some OPFRs (TPHP, TDCIPP, IPPHP, and t-BPDHP) disturbed radial glia and oligodendrocyte migration selectively at the highest concentration of 20 μM . After 72 h, TPHP and TDCIPP inhibited radial glia migration to $86.3 \pm$

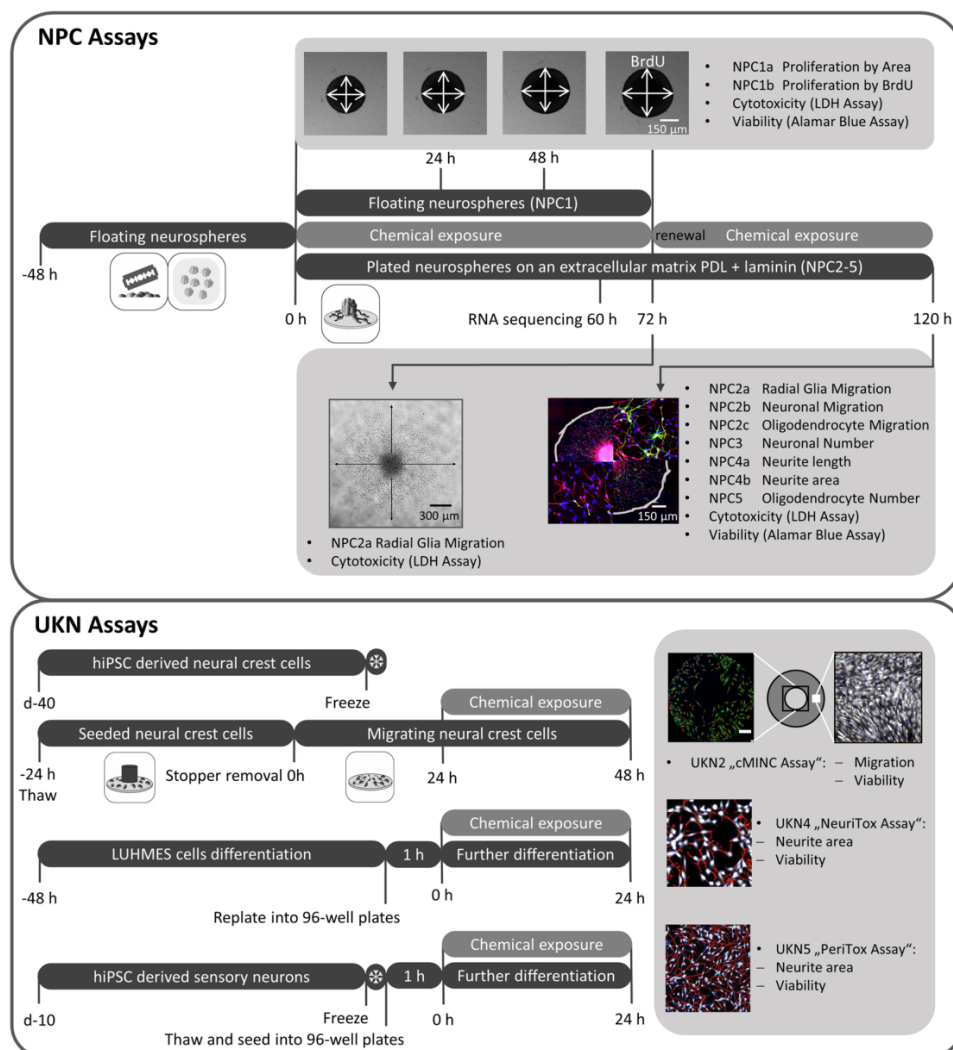


Fig. 1 Schematic overview of the battery of human-based neurodevelopmental in vitro assays. Experimental procedures for single assays are depicted schematically. Single endpoints

investigated by the battery assays are listed in gray boxes with their respective readout approach. PDL, poly-D-lysine; BrdU, bromodeoxyuridine; LDH, lactate dehydrogenase

2.9% and $90.5 \pm 2.5\%$ of controls, respectively (Fig. 3(f)). After 120 h, the influence of TPHP was reversed demonstrating the adaptive capabilities of the system. IPPHP, TDCIPP, and t-BDPHP inhibited radial glia migration (120 h) decreasing the distance to $85.6 \pm 8.1\%$, $82.2 \pm 3.8\%$, and $71.5 \pm 14.0\%$ of respective controls (Fig. 3(h)). None of the tested FRs altered neuronal migration distance (Fig. 3(i)), while oligodendrocyte migration was significantly shortened at 20 μ M of EHDPHP, IPPHP, and t-BDPHP to $83.6 \pm 3.5\%$,

$83.0 \pm 7.2\%$, and $73.1 \pm 8.3\%$ of respective controls (Fig. 3(j)). Both phased-out PBDEs and OPFRs did not impact cell viability/cytotoxicity at the conditions tested, except for TDCIPP (20 μ M) reducing mitochondrial activity (Fig. 3(k)). Strikingly, 6.6 μ M and 20 μ M IDDPHP increased cell viability to $133.2 \pm 4.9\%$ and $151.4 \pm 13.0\%$ of control, respectively, without affecting migration distance. The same effect was caused by 20 μ M EHDPHP (Fig. 3(h, k)). The endpoint-specific control for NPC2 was the src-kinase inhibitor PP2

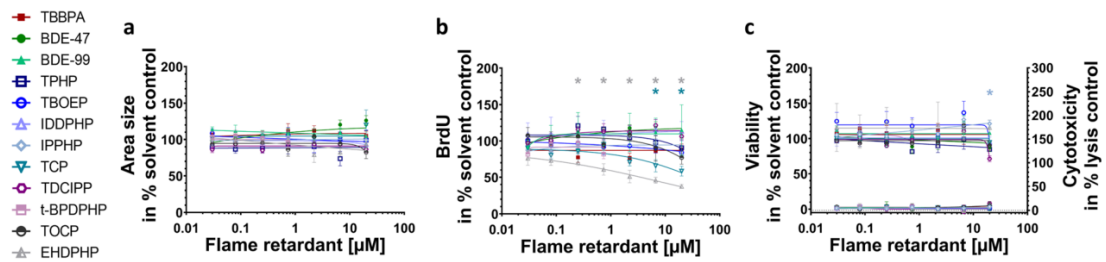


Fig. 2. Influence of FRs on proliferative hNPCs (NPC1). Spheres were plated in 96-well U-bottom plates and exposed to increasing FRs concentration over 72 h. Proliferation was studied by measuring the increase of sphere area (NPC1a) (a) and by quantifying BrdU incorporation (NPC1b) (b) into the DNA. In parallel, viability and cytotoxicity (c) were assessed by performing Alamar Blue Assay and LDH Assay. Data are represented as means \pm

SEM (except EHDHP in NPC1a and CTB $n=2$ mean \pm SD). Highest concentrations (≥ 2.2 μM) of t-BDPHP are not shown as spheres attached and differentiated. Statistical significance was calculated using one-way ANOVA followed by Bonferroni's post hoc tests ($p \leq 0.05$ was considered significant). BrdU, bromodeoxyuridine

significantly reducing migration to $36.9 \pm 29.9\%$ of control (Suppl. Fig. 4(c)).

Phased-out PBDEs and OPFRs do not interfere with neuronal differentiation and hardly affect neurite morphology

Within the migration area, hNPCs differentiate into different effector cells. In this study, 9.8% of the cells differentiated into neurons (Suppl. Fig. 4d). To analyze the influences of FRs on hNPC neuronal differentiation and neuronal morphology, NPC3 and NPC4 were performed. TBBPA (2.2 μM) reduced the total number of nuclei significantly to $60.8 \pm 7.0\%$ of control (Fig. 4(a, e)), which agrees with inhibition of radial glia migration (Fig. 3(d)). At higher TBBPA concentrations (6.6 μM and 20 μM), no nuclei and neurons were present (Fig. 4(a)) because migration was completely inhibited (Fig. 3(d)). The organophosphate-based IDDPHP (6.6 μM and 20 μM) increased the number of nuclei to $122.7 \pm 7.9\%$ and $133.4 \pm 6.2\%$ of controls, respectively (Fig. 4(c, e)) explaining the increased cell viability measures (Fig. 3(k)). All other FRs tested did not influence neuronal differentiation at concentrations up to 20 μM (Fig. 4(b, e)). For NPC3, the endpoint-specific control EGF significantly inhibited the total number of neurons to $1.0 \pm 0.2\%$ of total cell number (Suppl. Fig. 4(d)). The neurite length (NPC4) was significantly inhibited to $30.4 \pm 13.8\%$ of control by 20 μM TOCP only (Fig. 4(d)), while neurite area was not affected by any FR analyzed (Suppl. Fig. 3(f)). Additionally, LUHMES cells (UKN4) and hiPSC-derived peripheral neurons (UKN5) were used to analyze neurite morphology based

on two different cell types. Neurite outgrowth of both neuronal cell types (Fig. 4(f–h)) as well as their corresponding viability measures (Suppl. Fig. 3(i–j)) were not affected significantly by any of the FRs tested. As an endpoint-specific control for UKN4/5, cells were treated with 50 nM narciclasine which significantly reduced neurite outgrowth (data not shown).

Alteration of oligodendrocyte differentiation by all FR classes

Under differentiating conditions, 4.4% of the cells within the migration area differentiated into oligodendrocytes in this study (Suppl. Fig. 5c). Under the influence of TBBPA, differentiation into oligodendrocytes was specifically and significantly reduced starting from a concentration of 0.25 μM (to $66.2 \pm 8.9\%$ of control; Fig. 5(a, e)), as it was below the induction of cytotoxicity (Fig. 3(e)). BDE-47 significantly increased oligodendrocyte differentiation at low concentrations (0.03 μM to $147.4 \pm 4.1\%$; 0.08 μM to $172.5 \pm 6.4\%$ of control), whereas the highest concentration (20 μM) reduced their number to $10.9 \pm 5.9\%$ of control (Fig. 5(b, e)). Also, BDE-99 disturbed oligodendrocyte differentiation significantly at 2.5 μM to $35.2 \pm 11.7\%$, at 5 μM to $10.4 \pm 7.1\%$, and at 10 μM to $0.4 \pm 0.2\%$ (data taken from (Dach et al. 2017); Fig. 5(c, e)). The OPFR TDCIPP reduced the number of oligodendrocytes at 2.2 μM to $52.5 \pm 5.6\%$ of control (Fig. 5(d, e)). IDDPHP, TPHP, IPPHP, TOCP, and t-BDPHP produced similar results as they significantly affected oligodendrocyte differentiation at the two highest concentrations of 6.6 μM and 20 μM (Fig. 5(f, g, h, i, j, k, o)).

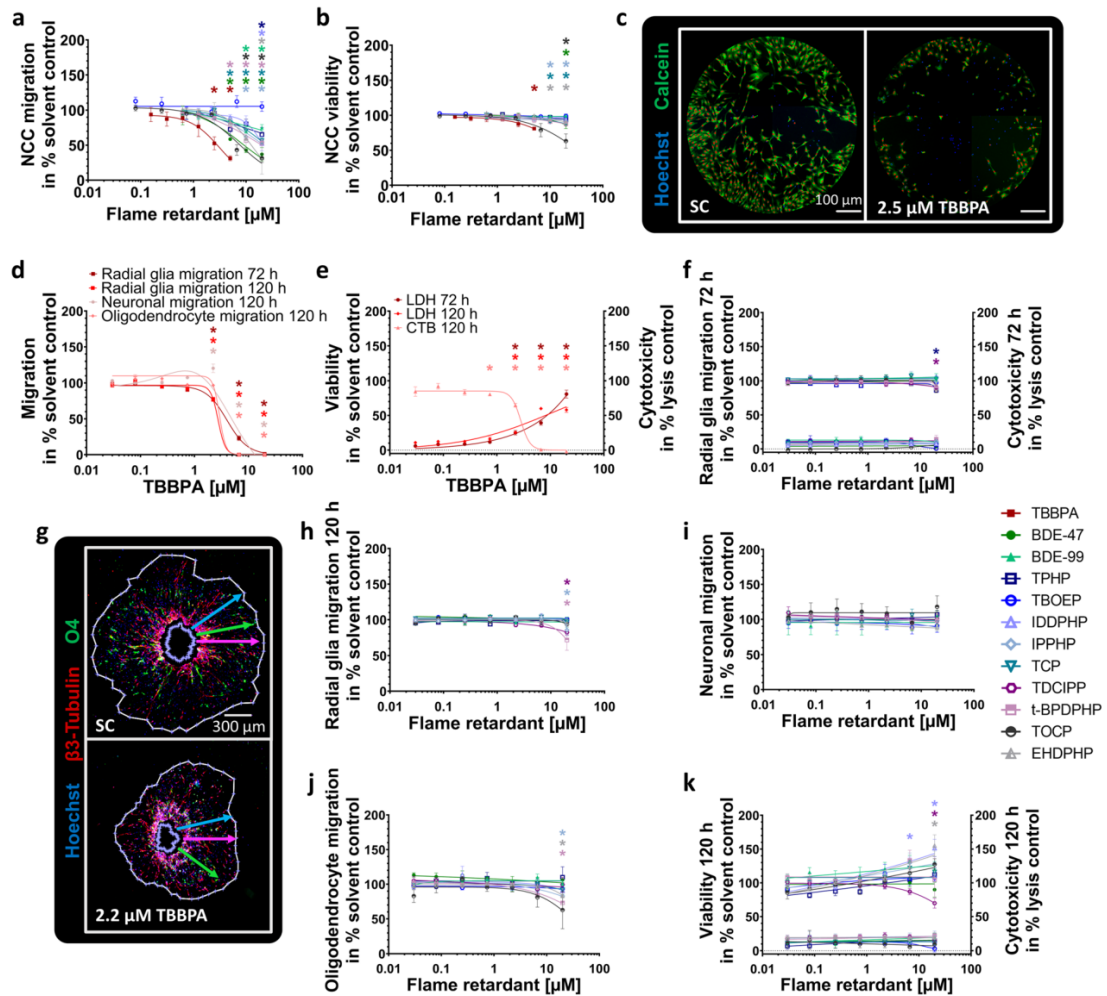


Fig. 3 Effects of FRs on different migration endpoints (NPC2, UKN2). NCCs were seeded around a stopper into 96-well plates. After stopper removal cells begin to migrate and were exposed to FRs/solvent(s) for 24 h. Cells were stained with Calcein-AM and H-33342, and the number of migrated cells (a) into the cell free zone was quantified using Cellomics ArrayScanVTI. Double-positive cell numbers were determined by an automated algorithm (marked with red dots, c). Viability was defined as the number of double-positive cells outside the ROI (b). Spheres were plated for hNPC migration analyses onto poly-D-lysine/laminin-coated 96-well plates in presence and absence of FRs for 120 h. Radial glia

migration (72 h) was determined by manually measuring the radial migration from the sphere core (d; f). After 120 h, the radial glia (d; h), neuronal (d; i), and oligodendrocyte migration (d; j) were assessed by automatically identifying (Omnisphero) the migration area of Hoechst stained nuclei, β (III)tubulin-stained neurons, and O4⁺ oligodendrocytes (g). In parallel, viability and cytotoxicity (e; f; k) were assessed by the Alamar Blue and the LDH Assay. Data are represented as means \pm SEM (except BDE-99 NPC2b; TOCP LDH 120 h, $n=2$, means \pm SD). Statistical significance was calculated using one-way ANOVA followed by Bonferroni's post hoc tests ($p \leq 0.05$ was considered significant). ROI, region of interest

Despite the fact that IDDPHP caused an increase in the number of nuclei (Fig. 4(c)), there were still less oligodendrocytes differentiated (Fig. 5(f, j)). EHDPPH, TCP, and TBOEP significantly reduced oligodendrocyte differentiation only at 20 μM to

$36.5 \pm 8.3\%$, $31.1 \pm 7.4\%$, and $24.8 \pm 9.0\%$ of controls, respectively (Fig. 5(l, m, n, o)). The endpoint-specific control BMP7 significantly reduced total number of oligodendrocytes to $0.4 \pm 0.1\%$ (Suppl. Fig. 4(e)).

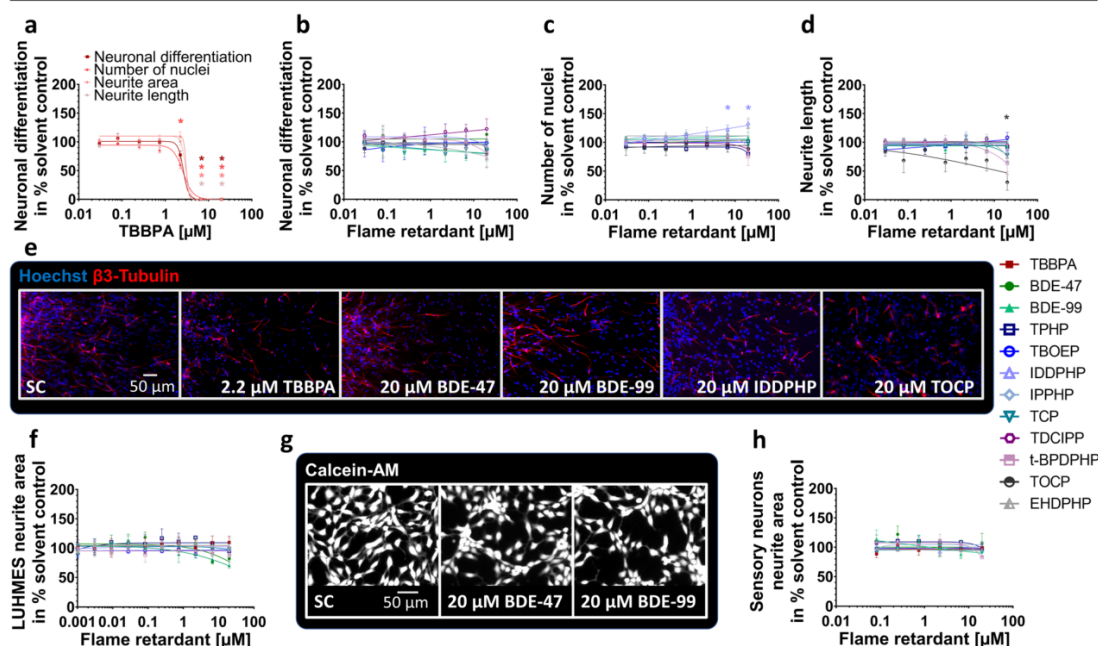


Fig. 4 Neuronal differentiation and morphology (NPC3, NPC4, UKN4, UKN5) in the presence and absence of FRs. Spheres were plated onto poly-D-lysine/laminin-coated 96-well plates in the presence and absence of FRs. Differentiation into neurons (**a**, **b**) was determined automatically by using a convolutional neural network (CNN) running on Keras implemented in Python 3. The number of all β (III)tubulin-positive cells (red) in percent of Hoechst positive nuclei (blue) in the migration area after 120 h of differentiation was calculated (**c**, **e**). Morphology (**d**) was determined automatically by using the software Omnisphero.

LUHMES cells and hiPSC derived sensory neurons were treated for 24 h in presence or absence of FRs and stained with Calcein-AM and H-33342 (**g**, LUHMES cells). An automated algorithm calculates the neurite area via subtraction of a calculated soma area from all calcein positive pixels (**f**, **h**). Data are represented as means \pm SEM (except BDE-99 NPC3, $n=2$, means \pm SD). Statistical significance was calculated using one-way ANOVA followed by Bonferroni's post hoc tests ($p < 0.05$ was considered significant)

Transcriptome changes in hNPCs

Since we identified 12 out of 15 FRs as disruptors of oligodendrocyte differentiation and for most of these compounds this endpoint was the only neurodevelopmental process disturbed in differentiating NPCs at these concentrations, we performed RNASeq analyses of neurospheres exposed to BMC₅₀ concentrations of selected FRs for 60 h. FR selection was based on DNTPi clustering choosing at least one FR from each DNTPi cluster (Fig. 7). For BDE-47, which produced a bell-shaped concentration-response curve, the highest significant concentration for the oligodendrocyte inducing effect was studied in addition. These experiments aimed at gaining understanding about similar or different modes-of-actions (MoA) underlying the observed endophenotype. The PCA analysis was based on 18941 genes and indicates the differences of individual FRs to the controls (Fig. 6(a)). The plot shows the highest

transcriptional variation in cells treated with EHDHP compared to the controls. Both phased-out PBDEs (higher concentration for BDE-47), TOCP and IDDPHP, and t-BDPHP, TDCIPP, and TBBPA clearly separated from the controls, while the low BDE-47 concentration did not lead to a separation from the controls. A hierarchical clustering of FRs based on their different gene expression levels was generated with Minkowski distance analyses (Fig. 6(b)). Similar to the PCA plot, EHDHP was the most distanced FR to control and IDDPHP, TOCP, as well as BDE-99 and the higher concentration of BDE-47 form two clusters in an independent manner to the control. BDE-47 (0.08 μM), TDCIPP, TBBPA, and t-BDPHP also form a cluster away from the controls, yet with less distance than the other compounds. This clustering is also reflected in the heatmap shown in Fig. 6(c). Here, the Z-score of up- and downregulated genes visually demonstrates that the pattern of BDE-47 (low), TDCIPP,

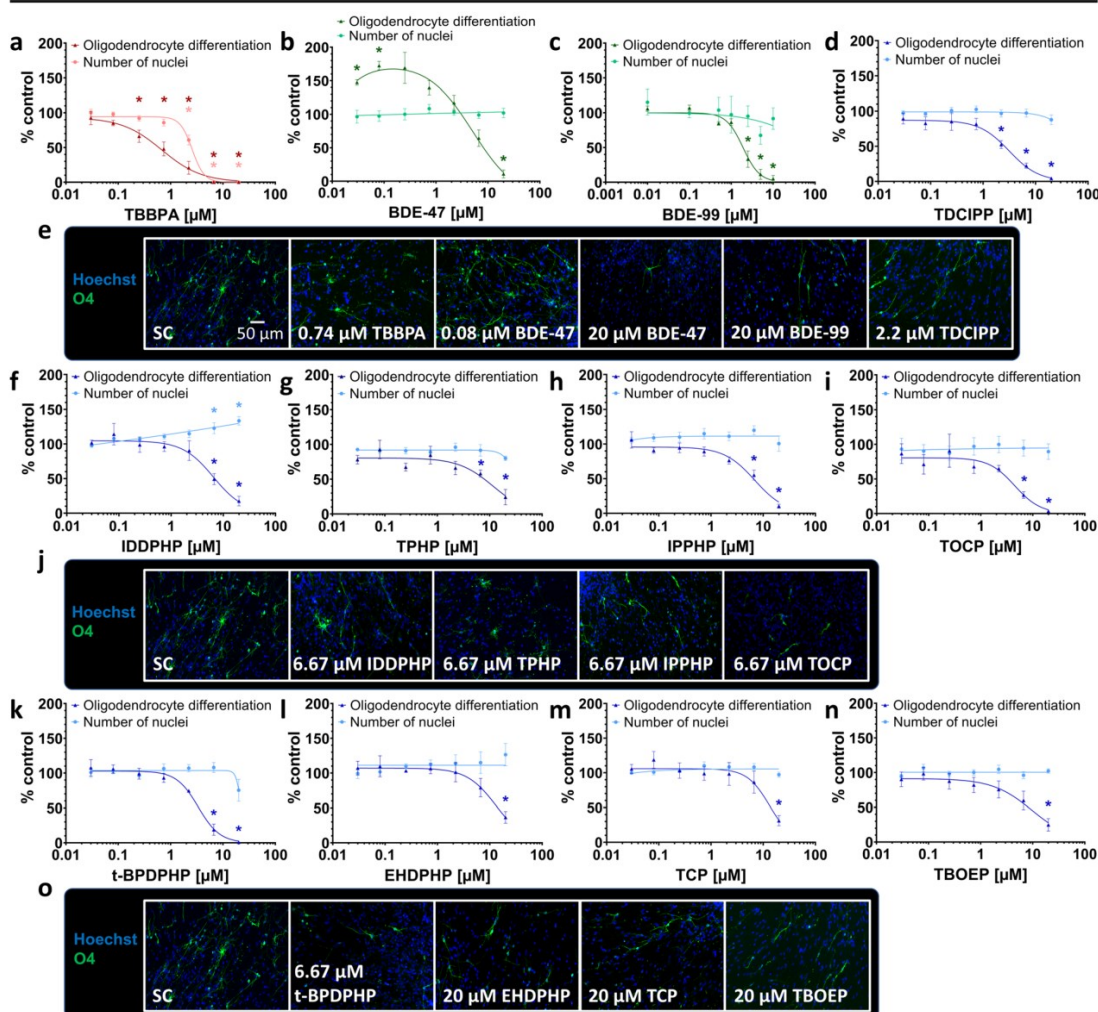


Fig. 5 Differentiation into oligodendrocytes (NPC5) in the presence and absence of FRs. Spheres were plated onto poly-D-lysine/laminin-coated 96-well plates in the presence and absence of FRs. Differentiation into oligodendrocytes was determined automatically based on immunocytochemical stainings (e, j, o) and by using a convolutional neural network (CNN) running on Keras

implemented in Python 3. The number of all O4-positive cells (green) in percent of Hoechst positive nuclei (blue) in the migration area after 120 h of differentiation was calculated (a, b, c, d, f, g, h, i, k, l, m, n). Data are represented as means \pm SEM. Statistical significance was calculated using one-way ANOVA followed by Bonferroni's post hoc tests ($p < 0.05$ was considered significant)

TBBPA, and t-BPDHP is similar to the pattern of controls. Equally to the PCA variance and Minkowski cluster, the patterns of IDDPHP and TOCP, as well as of both phased-out PBDEs, are visually similar to each other. Again, EHDPHP was clearly different from the controls and the other FRs.

To understand qualitative changes in gene expression related to FR effects on oligodendrocytes, we analyzed genes involved in selected pathways that relate to

toxicity of the oligodendrocyte lineage (Simons and Trajkovic 2006; Káradóttir et al. 2008; Volpe et al. 2011; Marinelli et al. 2016) listed in Fig. 6(d) and visualized those in respective heatmaps (Suppl. Fig. 6). Heatmap hierarchical clusters were used for classification into several levels. Level 1 (dark blue) describes the most distanced cluster from control, while the separation between samples and controls decreases in hierarchy up to > level 4 (white). In all pathways analyzed except for

NOTCH1 signaling (level 3), EHDPHP reached level 1 suggesting that EHDPHP interfered with a wide variety of oligodendrocyte-relevant cell signaling. Similarly, the phased-out PBDEs affected a broad variety of genes belonging to these pathway gene clusters. Here it is of interest that BDE-99 did not affect genes involved in cholesterol biosynthesis or mitochondrial calcium transport. TOCP and IDDPHP, which clustered in the previous analyses (Fig. 6(a, b)), also displayed a similar pattern in the pathway analyses. Both most strongly influenced NOTCH1 signaling and at a lower level affected almost all other pathways except for ROS detoxification. TDCIPP and t-BDPHP both exerted the least effects on the pathways as they disturb multiple pathways at level 4 without pathway overlap.

A special case in MoA seems to be TBBPA as it strongly and selectively affected cholesterol biosynthesis at level 2 and endoplasmic reticulum stress at level 4, while all other pathways are unaffected. These RNASeq data confirm previous Affymetrix microarray data identifying altered cholesterol metabolism as the predominant non-endocrine pathway affected by TBBPA in differentiating neurospheres (Klose et al. 2020). These data indicate that the studied FRs disturb a variety of pathways that influence amongst others oligodendrocyte differentiation. As this is a mixed culture, we cannot exclude that the signals produced by FRs are also derived from the other cell types in differentiated neurospheres, i.e., radial glia and neurons. It is to note that these RNASeq results are based on an $n=1$ each that give an orientation on similar or distinct MoA of the individual FR but need to be substantiated by more in-depth work in the future.

Due to the low percentage of oligodendrocytes (4.4%) within the migration area, the depth of RNASeq was not sufficient to detect transcription of oligodendrocyte-related genes in detail. Therefore, we performed RT-qPCR analyses of five oligodendrocyte-specific transcripts representing their different maturation stages (Fig. 6(e)). Gene expression data of the solvent control of differentiated spheres normalized to proliferating spheres reveal “normal” neurosphere development over a time course of 60 h (gray bars). These can be directly compared to the FR-treated samples (blue bars). Gene products chosen are representative for increasing oligodendrocyte maturation stages ($PDGFR\alpha < PLP < CNP < GALC < MBP$; Baumann and Pham-Dinh 2001; Kuhn et al. 2019), although these are an onsets of expression and the markers show considerable overlaps. All gene products were expressed at least twofold higher in differentiating versus

proliferating spheres supporting oligodendrocyte formation in the neurosphere system (Dach et al. 2017). FR exposure altered developmental gene expression changes from proliferating to 60 h differentiating neurospheres. Only t-BDPHP induced a twofold expression induction of $PDGFR\alpha$ mRNA, a gene expressed in oligodendrocyte progenitor cells (OPCs) and pre-oligodendrocytes (pre-OLs), but not in immature and mature oligodendrocytes (OLs), suggesting a delay in oligodendrocyte maturation. PLP is expressed in OPCs, pre-OLs, and OLs and was strongly reduced by TBBPA, BDE-99, TOCP, IDDPHP, BDE-47, and EHDPHP mirroring general reduction of OLs across maturation stages. In contrast, CNP and $GALC$ mRNA, which are expressed in all oligodendrocyte stages but the OPCs, were not affected by any of the compounds. MBP gene expression, one of the latest oligodendrocyte maturation markers, was reduced by BDE-47 (low concentration), TOCP, and EHDPHP (Fig. 6(e)). Interestingly, BDE-47 induced oligodendrocyte formation. These data demonstrate that despite the common phenotypical result of reduction in oligodendrocyte differentiation (besides BDE-47 low concentration), FRs’ molecular effects on oligodendrocyte marker expression patterns are compound-specific.

Compound classification based on BMC calculation

In order to provide a common metric of comparison across the different assays and substances, the benchmark dose (BMD) approach, which is recommended by the EFSA Scientific Committee (Hardy et al. 2017), was used. For in vitro toxicity testing, benchmark concentration (BMC) is comparable to the BMD (Krebs et al. 2020a) and derived from concentration-response information. The benchmark response (BMR) value was defined based on the variability of the respective endpoints (NPC1-5, Suppl. Fig. 4; UKN, Masjosthusmann et al. 2020). All BMCs calculated from all data points of the fitted concentration-response curves are listed in Table 1, with the respective upper and lower confidence intervals given in supplementary Table 2. From the FRs, which achieved BMCs, several questions can be drawn: (i) Are the observed effects DNT-specific or unspecific hits according to the classification models (Masjosthusmann et al. 2020)? (ii) What is the most sensitive endpoint (MSE) for each FR? And (iii) what is the potency ranking of the FRs? Most compound effects assessed by the battery are DNT-specific (Table 1), yet BBOEP, TCEP, and TCIIP did not reach DNT-specificity according to the

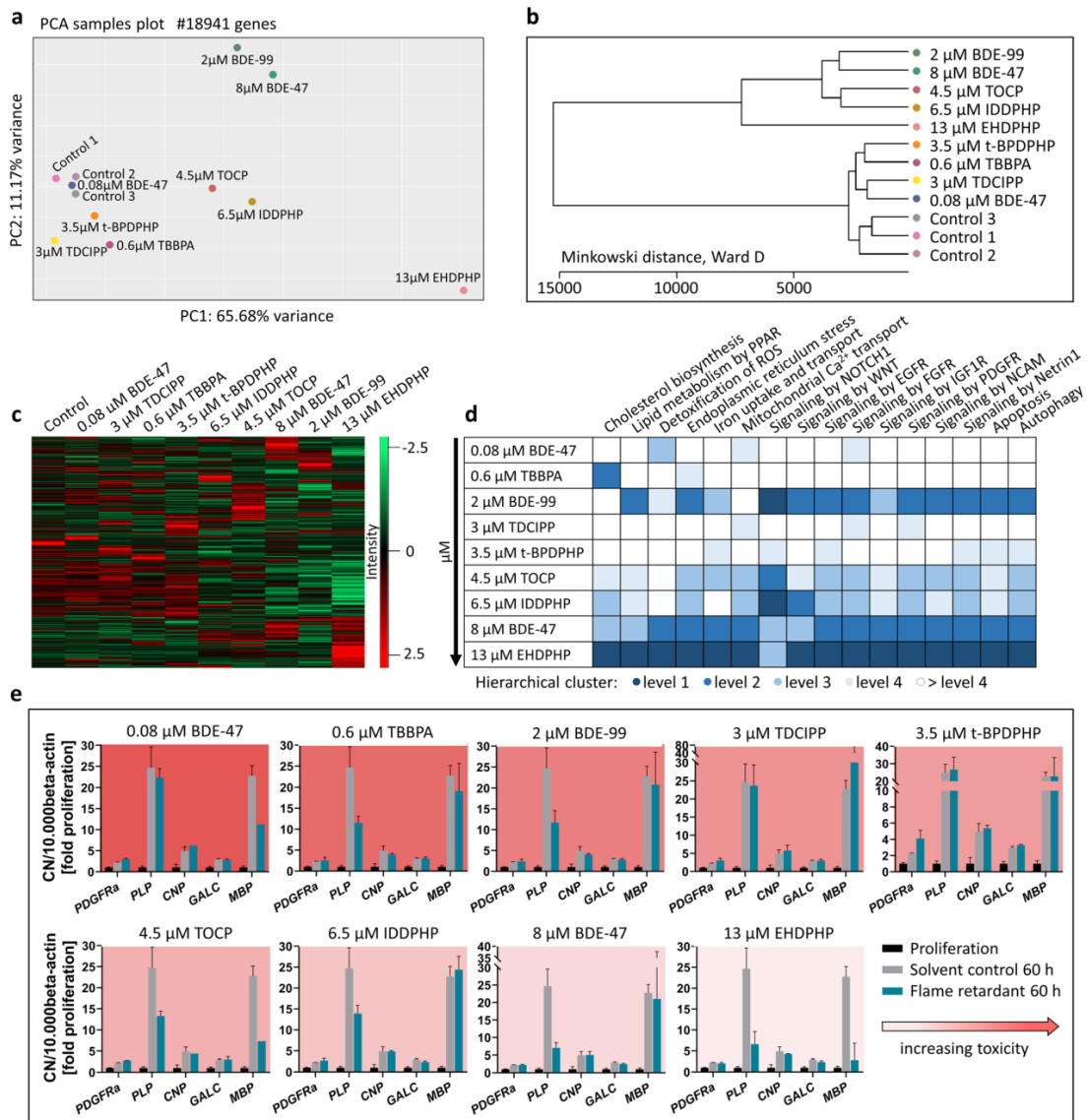


Fig. 6 RNA sequencing and RT-qPCR. Human NPCs differentiated for 60 h in the presence of 0.6 μ M TBBPA, 2 μ M BDE-99, 3 μ M TDCIPP, 3.5 μ M t-BPDHP, 4.5 μ M TOCP, 6.5 μ M IDDPHP, 8 μ M BDE-47, and 13 μ M EHDHP. These concentrations represent the BMC₅₀ values of oligodendrocyte differentiation inhibition. 0.08 μ M BDE-47 induced oligodendrocyte differentiation. Controls 1–3 represent spheres plated in solvent control 0.1% DMSO. PCA (**a**) and Minkowski distance plot (**b**) analyses were performed using the PCAGO online software (<https://pcago.bioinf.uni-jena.de/>) as previously described (Gerst and Hölzer 2019). Both plots were generated by normalizing the total number of reads of different samples to the Transcript per Kilobase Million (TPM) count. The heatmap (**c**) was generated using Perseus Version 1.6.2.2 (<https://www.maxquant.org/perseus/>).

Therefore, the Z-scores of TPM values were used with a cut-off of one valid value per condition. Classification of impact on oligodendrocyte differentiation-relevant pathways (**d**) was performed by expert judgment based on hierarchical clustering of pathway-related genes (Suppl. Fig. 6) and was categorized into four levels (level 1 as most and level 4 as least distanced to one merged control). Gene expression (**e**) of *platelet-derived growth factor receptor A* (PDGFR α), proteolipid protein (PLP), cyclic-nucleotide-phosphodiesterase (CNP), galactosylceramidase (GALC) and myelin basic protein (MBP) was assessed via RT-qPCR and normalized to the housekeeping gene beta actin (ACTB). In addition to solvent control (gray bars), proliferative spheres (black bars) were used as an internal control. Data are represented as mean \pm SD from 1 to 3 biological replicates

Table 1 Summary of BMCs across the DNT in vitro testing battery. Specific hits are highlighted bold and borderline hits are marked *cursive*. Red colored specifies most sensitive endpoints (MSEs). *Induced effects. Numbers are given in μM . No value assumes BMCs > 20 μM

		Brominated (BFRs)			Organophosphates (OPFRs)											
		TBBPA	BDE-47	BDE-99	TPHP	TBOEP	IDDPHP	IPPHP	TCP	TDCIPP	t-BPDHPHP	TOCP	EHDHPHP	BBOEP	TCEP	TCIPP
Proliferation by area	BMC ₂₀	-	-	-	-	-	-	-	0.86	-	-	17.2	0.02	-	18.9	-
Proliferation by BrdU		-	-	-	-	-	-	9.62 ⁺	-	19.2	-	-	-	-	-	-
Proliferation CTB		-	-	-	-	-	-	-	-	-	-	-	-	19.9	-	-
Proliferation LDH		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Radial glia migr. 72 h	BMC ₂₀	1.93	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Radial glia migr. 120 h		2.15	-	-	-	-	-	-	-	-	15.73	-	-	-	-	-
Neuronal migration		2.60	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Oligo. migration		2.23	-	-	-	-	-	-	-	-	12.54	8.12	-	-	-	-
LDH 72 h		1.75 ⁺	-	-	-	-	-	-	-	-	-	-	-	-	-	-
LDH 120 h		0.63 ⁺	-	-	-	-	-	-	-	-	-	-	-	-	-	-
CTB 120 h		1.38	-	3.56 ⁺	-	-	1.79 ⁺	5.50 ⁺	-	11.2	-	12.9 ⁺	5.88 ⁺	-	-	-
Neurite length		2.31	-	-	-	-	-	-	-	-	9.55	0.12	17.9	-	-	-
Neurite area		2.49	-	-	-	-	-	-	-	-	15.8	0.51	19.8	-	-	-
Number of nuclei		1.49	-	-	-	-	3.10 ⁺	-	-	-	19.8	-	8.72 ⁺	-	-	-
Number of neurons	2.18	-	-	-	-	-	-	-	12.8 ⁺	-	18.8	10.3	-	-	-	
Number of oligodendrocytes	BMC ₅₀	-	0.03 ⁺	-	-	-	-	-	-	-	-	-	-	-	-	
NCC migration	BMC ₂₅	0.55	8.00	1.91	6.39	7.62	6.45	6.88	13.2	3.13	3.37	4.49	13.1	-	-	
NCC viability	BMC ₁₀	1.56	2.71	15.8	10.0	-	14.1	6.66	7.99	-	4.05	3.32	6.46	-	-	
LUHMES neurite area	BMC ₂₅	2.78	14.2	-	-	-	-	-	16.9	-	14.0	3.44	11.4	-	-	
LUHMES viability		-	-	12.3	-	-	-	-	-	-	-	-	-	-	-	
Sensory N. neurite area		-	13.5	15.0	-	-	-	-	-	-	-	-	-	-	-	
Sensory N. viability		-	-	-	-	-	-	-	-	-	-	-	-	-	-	

classification models. For TBBPA, most endpoints were affected at concentrations also inducing cytotoxicity. Based on specific DNT hits, the MSE for each compound across the DNT battery was assessed. In most cases (7/12), it was oligodendrocyte differentiation (NPC5), followed by NCC migration (UKN2; 2/12), NPC proliferation (NPC1; 2/12), and neurite morphology (NPC4; 1/12). The other assays did not provide MSE. Potency ranking was as follows: EHDHP > BDE-47 > TOCP > TBBPA > TCP > BDE-99 > IDDPHP > TDCCP > t-BDPPHP > TPHP > IPPHP > TBOEP (Fig. 7).

Compound prioritization: ToxPi vs. DNTPi

Another currently propagated compound prioritization instrument is the Toxicological Prioritization Index (ToxPi) tool introduced by the US EPA (Reif et al. 2010; Marvel et al. 2018). Using this tool, FR testing results were visualized and prioritized according to their DNT profiles generated in this study by producing DNTpis (Fig. 7(b)), which are then compared to their toxicological profiles of the existing ToxCast data (ToxPis; Fig. 7(a); <https://www.epa.gov/chemical-research/toxicity-forecasting>). Here, the whole toxicological profiles are taken into account, i.e., also FR

effects on cell viability, and specific and non-specific hits are not distinguished. In general, the size of the Pi pieces does not reflect the actual BMC values but relates the BMCs for the studied compound to the BMCs of this endpoint across the highest and lowest values of the whole endpoint data set across all compounds irrespective of the values by distributing them between 0 and 1. Hence, it is a relative, not an absolute value. The ToxPi tool then hierarchically clusters the FRs within the ToxPis and the DNTpis according to their potency and assay hit patterns. Producing ToxPi information on compound clustering and ranking of a compound class for “general” (ToxPis) and “specific” toxicity, here DNT (DNTpis) gives information on the specificity of the compound effects.

Our ToxPi evaluation of the compound class of FRs clearly indicates that the Pi clustering is very different between the ToxPis and the DNTpis. For example, the two phased-out PBDEs are almost negative in the ToxCast assays and cluster collectively, while they evoke multiple responses in the DNT assays resulting in separate clusters. Similarly, e.g., TCIPP gives alerts in the ToxPi, while there is no effect in the DNTpis. Additionally, the program creates toxicity rankings and, in both rankings, TBBPA was classified as the most potent one. However, the overall ranking differs from each other, for example, t-

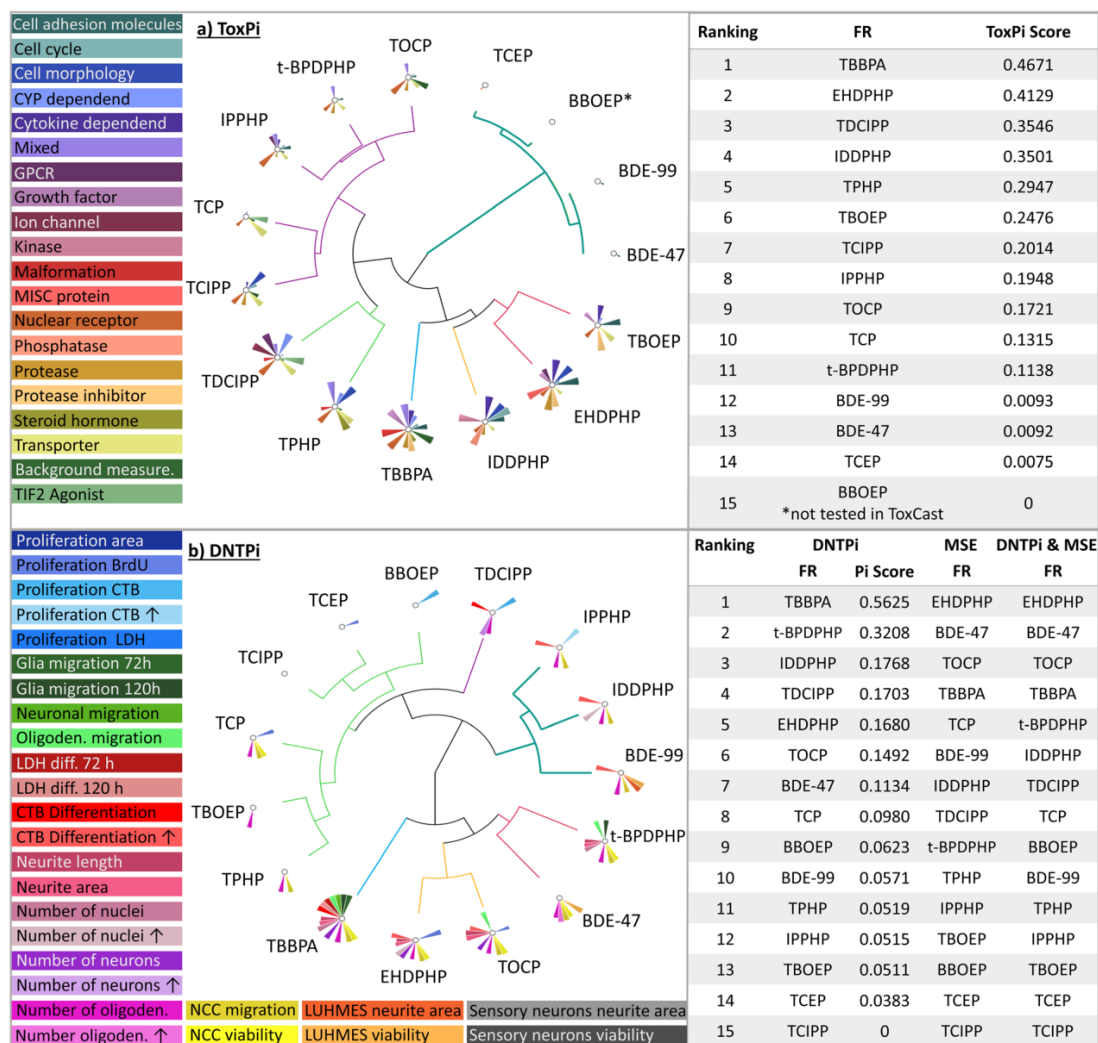


Fig. 7 Visualization and prioritization of FRs generated with ToxPi. ToxPis for general (a) and DNT-specific (b) toxicities using the ToxCast data and the results of the DNT in vitro battery, respectively. Graphs were produced with the Toxicological Prioritization Index (ToxPi) and Graphical User Interface (GUI) tool version 2.3. Size of pie slices represents relative strength of effect

on respective endpoint. For DNTPi and MSE ranking, first priority was given to MSE (Table 1); in the second line, ToxPi ranking was considered, e.g., for compounds with similar MSEs (starting from number 4 in the MSE analysis (Table 1; Suppl. Fig. 5a), due to overlapping 3-fold ranges for the MSE). *BBOEP was not tested in ToxCast

BPDHP ranks on number 2 in the DNTPi and on number 11 in the ToxPi. Similarly, TCIPP ranks on number 15 in the DNTPi and on number 7 in the ToxPi suggesting that general toxicity is not a good predictor for DNT. As the ToxPi tool does not distinguish between DNT-specific and non-DNT-specific effects and the ranking takes rather the number of modified endpoints than the effective concentrations, which relate to potency, into

account, we next combined the MSE-based (Table 1; Fig. 7; Suppl. Fig. 5(a)) with the ToxPi (Fig. 7) ranking. Therefore, the MSE with DNT-specific hits (Table 1; Suppl. Fig. 5(a)) was set to the first priority and, in the second line, DNTPi ranking was considered, e.g., for compounds with similar MSEs (starting from number 4 in the MSE analysis (Table 1; Suppl. Fig. 5(a)) due to overlapping 3-fold ranges for the MSE (Masjosthusmann

et al. 2020). Merging the two ranking strategies changes some of the FR ranking, yet not the four most potent compounds EHDPHP, BDE-47, TOCP, and TBBPA and results in the final ranking of FRs due to the data of this study (DNTPi and MSE; Fig. 7).

Discussion

In this study, we applied a human-based DNT in vitro battery of tests as a first case study for screening and prioritization of 15 data-poor compounds belonging to the class of FRs including phased out and alternative FRs. By using the BMC concept, specific DNT hits and most sensitive endpoints were identified across the endpoints of the battery. These scatter across the broad variety of neurodevelopmental processes investigated in this study.

TCP and EHDPHP

Two FRs, TCP, and EHDPHP inhibited NPC proliferation (NPC1) as the MSE at fairly low concentrations (BMC_{20} 0.86 and 0.02 μ M, respectively). Proliferation is a fundamental neurodevelopmental KE that, when altered, might cause microcephaly (Tang et al. 2016). This is the first time that the specific impact of TCP and EHDPHP on cell proliferation was shown in human cells. Previous work demonstrated neurodevelopmental behavioral adversities in a zebrafish model of these compounds at concentrations of 4 and 5 μ M lowest nominal effect levels, respectively (Alzualde et al. 2018). This model is well suited for informing on adverse outcomes but does not provide mechanistic information. A strong DNT potential for TCP was also identified in a recent study using a rat primary cell-based spheroid model. Concentrations as low as 0.1 μ M decreased the neurotransmitter content and affected genes related to neurotransmitter production after an exposure period of 14 days (Hogberg et al. 2020).

TOCP

TOCP was the only FR altering neurite length of young, primary fetal neurons as the MSE (BMC_{20} 0.12 μ M). Neurotoxicity of TOCP was previously observed in ferret (Stumpf et al. 1989) and in the hen sciatic nerve accompanied by a reduction in nerve calcium (Luttrell et al. 1993). Interference with neuronal calcium levels

could hint to a potential TOCP developmental MoA as calcium signaling is crucial for neurite outgrowth via regulating growth cone motility (Gasperini et al. 2017). TOCP was also identified as a neurotoxicant, as it disturbed the neural network activity in rat cortical neurons (Behl et al. 2015). Yet, these studies did not investigate neurodevelopment, but adult neurotoxicity.

IDDPHP

Interestingly, the OPFR IDDPHP induced the number of nuclei in the migration area as the MSE, probably due to excessive migration or proliferation of radial glia cells, the major and still proliferative cell type in the migration area. As IDDPHP did not alter radial glia migration distance, the action of IDDPHP on their proliferation seems to be the more probable explanation. However, this has to be substantiated by further experiments in the future. When it comes to radial glia, species specificities become crucial, as this cell type regulates evolutionary specificities of cortex formation (Zilles et al. 2013). Their proliferation and migration are tightly regulated processes orchestrating species-specific development of the cortex, with a special role in its folding in gyrencephalic species, like humans (Borrell and Götz 2014). Hence, interference with radial glia neural progenitors underlie a number of cortical malformations and cause mental retardation in genetic and infectious diseases (Guerrini and Dobyns 2014; Hu et al. 2014; Juric-Sekhar and Hevner 2019). In a recent study, IDDPHP triggered an increase of *nestin* expression, and this was interpreted as evidence of reactive astrogliosis (Hogberg et al. 2020). However, there may be alternative explanations, as changes in *nestin* may also point to effects on the radial glia and neural stem cell compartments. Zebrafish behavior was also affected by IDDPHP, yet at fairly high nominal lowest effect levels (40 μ M) with no knowledge on the underlying mechanisms (Alzualde et al. 2018).

IPPHP and t-BDPHP

NCC migration was the most sensitive endpoint (together with oligodendrocyte differentiation) upon IPPHP (BMC_{20} 6.66 μ M) and t-BDPHP (BMC_{20} 4.05) exposure. Disturbance of NCC migration causes, e.g., cleft palate or loss of functional hearing (Mayor and Theveneau 2013). Our data from human cells are in good agreement with model systems from other species:

micromolar concentrations of IPPHP and t-BPDPHP were also toxic for zebrafish (Behl et al. 2015; Alzualde et al. 2018), *Caenorhabditis elegans* (Behl et al. 2015; Boyd et al. 2016), rat cortical neurons (Behl et al. 2015), and 3D rat brain spheres (Hogberg et al. 2020). t-BPDPHP specifically inhibits neurite outgrowth of rat cortical neurons at 14.9 μM (Behl et al. 2015), an effect that we observe at similar concentrations in the NPC4, but not in the UKN4/5 assays. Similarly, IPPHP solely inhibits NCC, but not radial glia, neuronal or oligodendrocyte migration, while t-BPDPHP does alter other cell type migration at higher concentrations. Why different migration or neurite outgrowth assays yield different hits and are thus complementary to each other is probably due to different cell types, species, and neurodevelopmental timing represented in the assays. Hence, toxicity patterns across the battery reflect compounds different MoA by specifically altering certain targets.

Oligodendrocyte differentiation

Oligodendrocyte differentiation was the endpoint most frequently altered as the MSE upon cellular FR exposure with the following compound potency ranking: BDE-47 (low) > TBBPA > BDE-99 > TDCCP > t-BPDPHP > TPHP > IPPHP > TBOEP. Oligodendrocytes are necessary for proper brain functioning as they form and keep myelin sheaths around axons, thereby allowing rapid saltatory conduction of neuronal action potentials (Baumann and Pham-Dinh 2001; Kuhn et al. 2019). Hence, impaired oligodendrogenesis and resulting hypomyelination due to genetic disorders or severe brain injury contribute to functional adverse outcomes manifesting in neurological disorders such as the Alan-Hemdon-Dudley Syndrome (López-Espíndola et al. 2014; Tonduti et al. 2014) or periventricular leukomalacia (Back et al. 2001). Developing oligodendrocytes also exert a high susceptibility to stressors like reactive oxygen species and are sensitive to excitotoxicity and endoplasmic reticulum stress. They have a high energy and iron demand, are dependent on functional lipid metabolism, and their development and function are highly regulated by different hormones and growth factors (Bradl and Lassmann 2010; Volpe et al. 2011; Marinelli et al. 2016). Hence, developing oligodendrocytes can be concerned by a large variety of substances through a broad spectrum of MoA.

BDE-47 and oligodendrocyte differentiation

Since deviation from normal development into both directions, i.e., increase or decrease of a neurodevelopmental process, is considered adverse (Foti et al. 2013), the increase in oligodendrocyte differentiation by BDE-47 in the low nanomolar range needs attention. Consequences of increased oligodendrocyte differentiation are hypermyelination, an outcome observed for example in individuals with autism spectrum disorder (Ben Bashat et al. 2007; Wolff et al. 2013; Razeq et al. 2014). So far, BDE-47 was found to reduce mouse and human oligodendrocyte differentiation similar to the effects observed in this study at higher concentrations (Schreiber et al. 2010; Li et al. 2013). Li et al. (2013) did not test with BDE-47 concentrations that induced oligodendrocytes here (< 0.3 μM), whereas Schreiber et al. (2010) used concentrations as low as 0.1 μM . Here, inter-individual differences could explain the missing inducing oligodendrocyte effect as neurospheres used were generated from a different donor. Thus, it is increasing confidence that the data produced in this paper represents data from three different individuals. In addition, Schreiber et al. (2010) quantified oligodendrocytes by manual counting, while cells in this work here were identified using a convolutional neuronal network (CNN), which is more reliable, reproducible, and free of human counting bias. The induction mechanism of oligodendrocyte differentiation by BDE-47 is so far unknown. The performed RNASeq analyses did not reach a sufficient depth for such a cell type-specific molecular clarification. Interestingly, oligodendrocyte toxicity pathways are already triggered at 80 nM BDE-47 (Fig. 6(d)), probably resulting in loss of MBP-expressing more mature oligodendrocytes that is overridden by an unknown, oligodendrocyte-inducing trigger. In rat brain spheres, BDE-47 (0.1–5 μM) did not appear to affect *mbp* gene expression, but it caused a transient increase in myelin-associated glycoprotein (*mag*) transcript at 5 μM (Hogberg et al. 2020). Our previous species comparison of in vitro oligodendrogenesis found significant differences in timing, regulation of gene expression and response to toxicants between human and mouse oligodendrocytes (Dach et al. 2017; Klose et al. 2020). On the basis of these observations, it is likely that human neurospheres (as used here) will show differences to rat spheres. The difference in exposure schemes and readouts further complicates direct comparisons. A striking difference is for instance that none of the 15 FRs had any effect on human neuronal differentiation, while all 5 FRs tested in rat spheres reduced

neurofilament and other specifically neuronal markers (Hogberg et al. 2020).

TBBPA and oligodendrocyte differentiation

Similarly, TBBPA reduces oligodendrocyte differentiation. From the toxicity pathways analyzed by RNASeq, mainly genes relating to cholesterol biosynthesis were deregulated by TBBPA. This MoA was previously described as a putative adverse outcome pathway (Klose et al. 2020). TBBPA did not affect the number of corpus callosum CNP⁺ oligodendrocytes (Saegusa et al. 2009) or Ret⁺ oligodendrocytes (Fujimoto et al. 2013) in developmental rat studies. This might be due to the markers used in the in vivo study, as e.g., *CNP* expression did not, but only *PLP* expression changed upon TBBPA treatment in this study. Also, species (Dach et al. 2017) or brain regions with heterogeneous oligodendrocyte populations (Hayashi and Suzuki 2019) might have affected the results.

RNASeq analyses

In the Minkowski distance cluster and gene heatmap (Fig. 6(b, c)), the low concentration BDE-47, TBBPA, TDCIPP, and t-BPDPHP clustered together close to the controls. Different from TBBPA, the latter two change gene expression in variable oligodendrocyte toxicity pathways. These data suggest that either one specific pathway, like cholesterol metabolism for TBBPA, or multiple hits across distinct converging pathways like in the case of TDCIPP or t-BPDPHP, can summit in the same endophenotype. Minkowski cluster further demonstrates that TOCP, IDDPHP, PBDEs, and EHDPHP differ most from the controls and they strongly affect a large variety of oligodendrocyte toxicity pathways. Because oligodendrocytes provide just around 4% of the cells in the migration area, it is highly unlikely that these strong alterations in mRNA expression profiles can be attributed to oligodendrocytes only, but probably also derive from radial glia and/or neurons in the migration area. Because all other phenotypic endpoints of the neurosphere assay were not affected, these data clearly show the high susceptibility of oligodendrocytes towards alterations of these pathways and thus supports the notion of “just being an oligodendrocyte seems enough to put these cells at greater risk of damage” (Bradl and Lassmann 2010).

Compound prioritization

Such DNT in vitro battery data can be used for compound prioritization. Here, different methods are at hand. For one, BMC values with CI allow distinguishing between DNT-specific and DNT-unspecific hits (Masjosthusmann et al. 2020) giving objective potency ranking measures. However, this method takes only the MSE and not, e.g., the number of affected endpoints into consideration. To account for both, we merged the MSE method with the ToxPi approach by prioritizing for BMCs first and secondly adding the ToxPi ranking when BMCs of MSE of different compounds were located within their 3-fold ranges. In our opinion, prioritization for DNT only by ToxPi might include high uncertainty, because altering only one DNT endpoint can have detrimental effects on neurodevelopment, especially when it happens at low concentrations. Using this merged approach, our study revealed that BDE-47 and BDE-99, which are already banned due to their neurodevelopmental toxicity, rank as 2nd and 10th out of the 15 FRs investigated. Of the currently used aFRs, only TCIPP did not produce a hit in the battery according to the BMCs. However, also TCEP and BBOEP did not yield statistically significant hits, but just reached their BMC₂₀ values. Therefore, these three aFRs are rated as the least toxic with the DNT in vitro battery, while EHDPHP together with BDE-47 summit at the top as the most hazardous FR. These data indicate that the DNT in vitro battery is a useful tool for prioritizing compounds for their DNT hazard potential. It has to be noted that the battery applied here still has known gaps that need to be closed in the future. These include test methods for neuronal network formation (Frank et al. 2017; Shafer et al. 2019; Nimtz et al. 2020) including synaptogenesis (Pistollato et al. 2020), astrocyte, and microglia performance.

One question that arises is if such a DNT in vitro battery is at all necessary or if DNT might as well be predicted by the general ToxCast assays. To answer this question, FR DNT in vitro battery is compared to ToxCast data by ToxPi versus DNTPi assessment. The results demonstrate the uniqueness of the DNT in vitro battery for DNT hazard assessment. Such an approach has never been executed before and was shown here to be very helpful for assays' specificity analyses.

Moving from hazard to risk

When moving from hazard characterization to risk assessment, exposure data is crucial. Biomonitoring data for parent compounds currently available (Table 2; Cariou et al. 2008; Sundkvist et al. 2010; Kim et al. 2014; Tang and Zhai 2017; Beser et al. 2019; Ma et al. 2019; Chupeau et al. 2020) reveal a gap on human FR exposure data, especially for OPFRs. While phased-out PBDEs and TBPPA can be measured in human samples, most OPFRs metabolize fast and parent compounds cannot be detected, e.g., in cord blood or breast milk. Therefore, the occurrence of OPFR metabolites is measured in urinary samples of adults (Bastiaensen et al. 2019b; Gibson et al. 2019; Chupeau et al. 2020; Li et al. 2020) and children (He et al. 2018a, b; Bastiaensen et al. 2019a; Gibson et al. 2019; Chupeau et al. 2020) or in

hair (Kucharska et al. 2015; Chupeau et al. 2020). These studies clearly demonstrate the existence of OPFR metabolites in human samples, especially in children.

For relating such biomonitoring data to the studied in vitro hazards, we converted the internal FR concentrations from cord blood or breast milk given in nanograms per gram of fat to molarity by using a fat content of 5.8 g/L for serum (Akins et al. 1989; Phillips et al. 1989; Covaci et al. 2006; Rylander et al. 2006) and 33 g/L for breast milk (Kent et al. 2006; Prentice et al. 2016). Such in vitro–in vivo comparisons are very crude and do not account for in vitro kinetics or for actual fetal brain concentrations in vivo.

Hence, advanced kinetic modelling would be eventually needed to perform proper in vitro to in vivo extrapolation (IVIVE). Nevertheless, our crude evaluations revealed cord blood values for BDE-99, BDE-47, and TBBPA of

Table 2 Exposure data collected from published FR measurements in human breast milk and cord blood samples (Cariou et al. 2008; Sundkvist et al. 2010; Kim et al. 2014; Tang and Zhai 2017; Beser et al. 2019; Ma et al. 2019; Chupeau et al. 2020)

	Breast milk						Cord blood					
	BDE-99		BDE-47		TBBPA		BDE-99		BDE-47		TBBPA	
	ng/g lw		ng/g lw		ng/g lw		ng/g lw		ng/g lw		ng/g lw	
	μM		μM		μM		μM		μM		μM	
<i>Korea</i>	54.0	0.0316	31.0	0.0211	-	-	19.0	0.0020	36.0	0.0044	-	-
<i>China</i>	10.8	0.0063	27.5	0.0187	-	-	3.45	0.0004	8.49	0.0010	-	-
<i>Japan</i>	3.20	0.0019	4.90	0.0033	-	-	-	-	0.12	0.00001	-	-
<i>Philippines</i>	0.82	0.0005	3.60	0.0024	-	-	-	-	-	-	-	-
<i>Vietnam</i>	0.38	0.0002	0.40	0.0003	-	-	-	-	-	-	-	-
<i>USA</i>	6.40	0.0037	29.7	0.0202	-	-	23.3	0.0024	4.60	0.0006	-	-
<i>France</i>	0.53	0.0003	1.15	0.0008	4.1	0.0025	7.43	0.0008	-	-	103	0.0111
<i>Germany</i>	0.18	0.0001	0.45	0.0003	-	-	-	-	-	-	-	-
<i>UK</i>	0.80	0.0005	2.70	0.0018	-	-	-	-	-	-	-	-
<i>Sweden</i>	0.48	0.0003	2.28	0.0015	-	-	0.22	0.00002	3.4	0.0004	-	-
<i>Spain</i>	0.51	0.0003	0.54	0.0004	-	-	4.3	0.0004	3.3	0.0004	-	-
	Breast milk						Cord blood					
	TPHP		TBOEP		TCEP		TCIPP		EHDPHP		TCP	
	ng/g lw		ng/g lw		ng/g lw		ng/g lw		ng/g lw		ng/g lw	
	μM		μM		μM		μM		μM		μM	
<i>Japan</i>	1.40	0.0014	0.24	0.0002	0.14	0.0002	-	-	-	-	-	-
<i>Philippines</i>	19.0	0.0192	22.0	0.0182	42.0	0.0554	-	-	-	-	2.30	0.0021
<i>Vietnam</i>	4.90	0.0050	-	-	-	-	-	-	-	-	0.28	0.0003
<i>Sweden</i>	8.50	0.0086	4.70	0.0039	4.90	0.0065	45.0	0.0453	6.50	0.0059	0.80	0.0007
<i>Spain</i>	9.90	0.0100	14.8	0.0123	-	-	12.5	0.0126	-	-	19.0	0.0170
	Breast milk						Cord blood					
	TPHP		TBOEP		TCEP		TCIPP		EHDPHP		IDDPHP	
	ng/mL		ng/mL		ng/mL		ng/mL		ng/mL		ng/mL	
	μM		μM		μM		μM		μM		μM	
<i>USA</i>	0.15	0.0005	1.44	0.0036	0.04	0.0001	0.22	0.0005	0.02	0.00006	0.01	0.00003

0.002, 0.004, and 0.011 μM in a Korean (PBDEs) and French (TBBPA) cohort, respectively (Table 2). Breast milk concentrations calculated to 0.032 and 0.021 μM for BDE-99 and BDE-47 in Korea and 0.003 for TBBPA in France. OPFRs in breast milk occur with the highest measured values across all FRs with TCEP 0.055 μM , TPHP 0.019 μM , and TBOEP 0.018 μM (Philippines) and TCIPP 0.045 μM (Sweden). Assuming a breast milk intake of 1 L/day, exposure to these FRs approximates to 32 nmol/day BDE-99, 21 nmol/day BDE-47, 3 nmol/day TBBPA, 55 nmol/day TCEP, 19 nmol/day TPHP, 18 nmol/day TBOEP, and 45 nmol/day TCIPP. While the BMCs calculated for DNT in vitro hazard for BDE-99 and OPFRs are more than one order of magnitude lower than the estimated daily intake and cord blood concentrations, the BDE-47 BMC for the MSE is just one order of magnitude higher than the estimated exposure (suggesting a bioavailability of 100%, slow/no liver metabolism, perfect blood-brain-barrier (BBB) passage (1:1), and protein binding according to logP prediction model).

However, humans are generally exposed to compound mixtures including FRs, pesticides, pharmaceuticals, toxic metals, and other environmental contaminants. Therefore, individual compound exposure easily adds up to mixtures at relevant concentrations that might exert additive, synergistic, or antagonistic effects, especially when the same converging endpoint is affected. This is likely the case for oligodendrocytes because they seem to be the most susceptible cell type of the brain. Mixture experiments as well as sophisticated IVIVE are needed to substantiate these concerns.

Summary and conclusion

In summary, we tested 15 FRs including phased-out PBDEs, TBBPA and OPFRs for their neurodevelopmental toxicity in a human cell-based DNT in vitro battery. FR hazards across different neurodevelopmental endpoints were used for calculating BMC and CI leading to a potency ranking. Evaluation of the data with the ToxPi tool revealed a distinct ranking that we combined with the BMC ordering for final prioritization. In addition, comparison of DNT hazard ranking according to the ToxPi tool with the ToxCast data revealed DNT-specific hazard for this group of FRs that is not well predicted by ToxCast assays. Extrapolating DNT battery BMC to human FR exposure via breast milk suggests low risk for individual compounds but raises concern for mixture exposure, which

is the real-life situation. This is especially of apprehension when different compounds converge through diverse MoA on common endpoints like oligodendrocyte differentiation in this study.

This case study using FRs contextualized with the performance characteristics of the battery using diverse compound classes (Masjosthusmann et al. 2020) suggests that using a human cell-based DNT in vitro battery for hazard assessment for compound prioritization is a promising approach for future risk assessment procedures.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s10565-021-09603-2>.

Author contribution All authors read, commented, and approved the manuscript. Jödis Klose: study conception, investigation, data collection and analysis, figure design, writing. Melanie Pahl: investigation, data collection and analysis. Kristina Bartmann: investigation. Farina Bendt: investigation. Jonathan Blum: study conception, investigation, data collection and analysis. Xenia Dolde: study conception, investigation, data collection and analysis. Nils Förster: software. Anna-Katharina Holzer: study conception, investigation, data collection and analysis. Ulrike Hübenthal: investigation. Hagen Eike Keßel: software. Katharina Koch: study conception. Stefan Masjosthusmann: study conception and data analysis. Sabine Schneider: data analysis. Selina Woeste: investigation. Andrea Rossi: data analysis. Adrian Covaci: resources. Mamta Behl: resources. Marcel Leist: study conception, funding acquisition for all experiments performed with NCCs, LUHMES cells, and hiPSC-derived neurons; writing. Julia Tigges: study conception, supervision, project administration and writing. Ellen Fritsche: study conception, supervision, funding acquisition for all experiments performed with hNPCs, project administration, and writing.

Funding Open Access funding enabled and organized by Projekt DEAL. This work was supported by the FOKO (Forschungskommision of the medical faculty of the Heinrich-Heine-University) (2016-53), the EFSA (European Food Safety Authority) (OC/EFSA/PRAS/2017/01), CERST (Center for Alternatives to Animal Testing) of the Ministry for Culture and Science of the State of North-Rhine Westphalia, Germany) (file number 233-1.08.03.03-121972/131—1.08.03.03—121972), and the DFG Ursula M. Händel Tierschutzpreis to EF (DFG FR 1392/6-1). It has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No. 681002 (EU-ToxRisk) and No. 825759 (ENDpoiNTs).

Declarations

Ethics approval hNPCs were purchased from Lonza Verviers SPRL, Belgium, and work was approved by the ethics committee of the Heinrich-Heine University Duesseldorf.

Conflict of interest The authors declare that they have no conflicts of interest.

References

- Akins JR, Waldrep K, Bernert JT. The estimation of total serum lipids by a completely enzymatic 'summation' method. *Clin Chim Acta*. 1989;184:219–26.
- Alzualde A, Behl M, Sipes NS, Hsieh JH, Alday A, Tice RR, et al. Toxicity profiling of flame retardants in zebrafish embryos using a battery of assays for developmental toxicity, neurotoxicity, cardiotoxicity and hepatotoxicity toward human relevance. *Neurotoxicol Teratol*. 2018;70:40–50. <https://doi.org/10.1016/j.ntt.2018.10.002>.
- Andersen SL. Trajectories of brain development: point of vulnerability or window of opportunity? *Neurosci Biobehav Rev*. 2003;27:3–18. [https://doi.org/10.1016/S0149-7634\(03\)00005-8](https://doi.org/10.1016/S0149-7634(03)00005-8).
- Back SA, Luo NL, Borenstein NS, Levine JM, Volpe JJ, Kinney HC. Late oligodendrocyte progenitors coincide with the developmental window of vulnerability for human perinatal white matter injury. *J Neurosci*. 2001;21:1302–12. <https://doi.org/10.1523/jneurosci.21-04-01302.2001>.
- Bal-Price A, Crofton KM, Leist M, Allen S, Arand M, Buetler T, et al. International STakeholder NETwork (ISTNET): creating a developmental neurotoxicity (DNT) testing road map for regulatory purposes. *Arch Toxicol*. 2015;89:269–87. <https://doi.org/10.1007/s00204-015-1464-2>.
- Bal-Price A, Hogberg HT, Crofton KM, et al. Recommendation on test readiness criteria for new approach methods in toxicology: exemplified for developmental neurotoxicity. *ALTEX*. 2018;35:306–52. <https://doi.org/10.14573/altex.1712081>.
- Barenys M, Gassmann K, Baksmeier C, Heinz S, Reverte I, Schmuck M, et al. Epigallocatechin gallate (EGCG) inhibits adhesion and migration of neural progenitor cells in vitro. *Arch Toxicol*. 2017;91:827–37. <https://doi.org/10.1007/s00204-016-1709-8>.
- Bastiaansen M, Ait Bamai Y, Araki A, van den Eede N, Kawai T, Tsuboi T, et al. Biomonitoring of organophosphate flame retardants and plasticizers in children: associations with house dust and housing characteristics in Japan. *Environ Res*. 2019a;172:543–51. <https://doi.org/10.1016/j.envres.2019.02.045>.
- Bastiaansen M, Malarvannan G, Been F, Yin S, Yao Y, Huygh J, et al. Metabolites of phosphate flame retardants and alternative plasticizers in urine from intensive care patients. *Chemosphere*. 2019b;233:590–6. <https://doi.org/10.1016/j.chemosphere.2019.05.280>.
- Baud A, Wessely F, Mazzacova F, McCormick J, Camuzeaux S, Heywood WE, et al. A multiplex high-throughput targeted proteomic assay to identify induced pluripotent stem cells. *Anal Chem*. 2017;89:2440–8. <https://doi.org/10.1021/acs.analchem.6b04368>.
- Baumann N, Pham-Dinh D. Biology of oligodendrocyte and myelin in the mammalian central nervous system. *Physiol Rev*. 2001;81:871–927. <https://doi.org/10.1152/physrev.2001.81.2.871>.
- Baumann J, Dach K, Barenys M, et al. Application of the neurosphere assay for DNT hazard assessment: challenges and limitations. *Methods Pharmacol Toxicol*. 2016;49:1–29. https://doi.org/10.1007/978-94-007-653-4_9.
- Behl M, Hsieh JH, Shafer TJ, Mundy WR, Rice JR, Boyd WA, et al. Use of alternative assays to identify and prioritize organophosphorus flame retardants for potential developmental and neurotoxicity. *Neurotoxicol Teratol*. 2015;52:181–93. <https://doi.org/10.1016/j.ntt.2015.09.003>.
- Ben Bashat D, Kronfeld-Duenias V, Zachor DA, Ekstein PM, Hendler T, Tarrasch R, et al. Accelerated maturation of white matter in young children with autism: a high b value DWI study. *Neuroimage*. 2007;37:40–7. <https://doi.org/10.1016/j.neuroimage.2007.04.060>.
- Bergman Å, Rydén A, Law RJ, de Boer J, Covaci A, Alaee M, et al. A novel abbreviation standard for organobromine, organochlorine and organophosphorus flame retardants and some characteristics of the chemicals. *Environ Int*. 2012;49:57–82.
- Beser MI, Pardo O, Beltrán J, Yusá V. Determination of 21 perfluoroalkyl substances and organophosphorus compounds in breast milk by liquid chromatography coupled to orbitrap high-resolution mass spectrometry. *Anal Chim Acta*. 2019;1049:123–32. <https://doi.org/10.1016/j.aca.2018.10.033>.
- Blum A, Behl M, Bimbaum LS, Diamond ML, Phillips A, Singla V, et al. Organophosphate ester flame retardants: are they a regrettable substitution for polybrominated diphenyl ethers? *Environ Sci Technol Lett*. 2019;6:638–49. <https://doi.org/10.1021/acs.estlett.9b00582>.
- Borrell V, Götz M. Role of radial glial cells in cerebral cortex folding. *Curr Opin Neurobiol*. 2014;27:39–46. <https://doi.org/10.1016/j.conb.2014.02.007>.
- Boyd WA, Smith MV, Co CA, Pirone JR, Rice JR, Shockley KR, et al. Developmental effects of the ToxCast™ phase I and phase II chemicals in *Caenorhabditis elegans* and corresponding responses in zebrafish, rats, and rabbits. *Environ Health Perspect*. 2016;124:586–93. <https://doi.org/10.1289/ehp.1409645>.
- Bradl M, Lassmann H. Oligodendrocytes: biology and pathology. *Acta Neuropathol*. 2010;119:37–53. <https://doi.org/10.1007/s00401-009-0601-5>.
- Cariou R, Antignac J-P, Zalko D, Berrebi A, Cravedi JP, Maume D, et al. Exposure assessment of French women and their newborns to tetrabromobisphenol-A: occurrence measurements in maternal adipose tissue, serum, breast milk and cord serum. *Chemosphere*. 2008;73:1036–41. <https://doi.org/10.1016/j.chemosphere.2008.07.084>.
- Chambers SM, Qi Y, Mica Y, Lee G, Zhang XJ, Niu L, et al. Combined small molecule inhibition accelerates developmental timing and converts human pluripotent stem cells into nociceptors. *Nat Biotechnol*. 2013;30:715–20. <https://doi.org/10.1038/nbt.2249>.
- Chao HR, Wang SL, Lee WJ, Wang YF, Pöpke O. Levels of polybrominated diphenyl ethers (PBDEs) in breast milk from central Taiwan and their relation to infant birth outcome and maternal menstruation effects. *Environ Int*. 2007;33:239–45. <https://doi.org/10.1016/j.envint.2006.09.013>.

- Chupeau Z, Bonvallot N, Mercier F, le Bot B, Chevrier C, Glorennec P. Organophosphorus flame retardants: a global review of indoor contamination and human exposure in Europe and epidemiological evidence. *Int J Environ Res Public Health*. 2020;17:1–24. <https://doi.org/10.3390/ijerph17186713>.
- Covaci A, Voorspoels S, Thomsen C, van Bavel B, Neels H. Evaluation of total lipids using enzymatic methods for the normalization of persistent organic pollutant levels in serum. *Sci Total Environ*. 2006;366:361–6. <https://doi.org/10.1016/j.scitotenv.2006.03.006>.
- Crofton KM, Mundy WR, Shafer TJ. Developmental neurotoxicity testing: a path forward. *Congenit Anom (Kyoto)*. 2012;52:140–6. <https://doi.org/10.1111/j.1741-4520.2012.00377.x>.
- Crofton K, Fritsche E, Ylikomi T, Bal-Price A. International STakeholder NETwork (ISTNET) for creating a Developmental Neurotoxicity Testing (DNT) roadmap for regulatory purposes. *ALTEX*. 2014;31:223–4. <https://doi.org/10.14573/altex.1402121>.
- Dach K, Bendt F, Huebenthal U, Giersiefer S, Lein PJ, Heuer H, et al. BDE-99 impairs differentiation of human and mouse NPCs into the oligodendroglial lineage by species-specific modes of action. *Sci Rep*. 2017;7:1–11. <https://doi.org/10.1038/srep44861>.
- Damerud PO, Eriksen GS, Jóhannesson T, Larsen PB, Viluksela M. Polybrominated diphenyl ethers: occurrence, dietary exposure, and toxicology. *Environ Health Perspect*. 2001;109:49–68. <https://doi.org/10.1289/ehp.01109s149>.
- De Wit CA. An overview of brominated flame retardants in the environment. *Chemosphere*. 2002;46:583–624. [https://doi.org/10.1016/S0045-6535\(01\)00225-9](https://doi.org/10.1016/S0045-6535(01)00225-9).
- Delp J, Gutbier S, Klima S, et al. A high-throughput approach to identify specific neurotoxicants/developmental toxicants in human neuronal cell function assays. *ALTEX*. 2018;35:235–53. <https://doi.org/10.14573/altex.1712182>.
- EFSA. Scientific Opinion on the developmental neurotoxicity potential of acetamiprid and imidacloprid. *EFSA J*. 2013;11:1–47. <https://doi.org/10.2903/j.efsa.2013.3471>.
- EPA 1998. EPA test guidelines for pesticides and toxic substances. Health effects test guidelines: OPPTS 870.6300 developmental neurotoxicity study. <https://www.epa.gov>
- Eskenazi B, Chevrier J, Rauch SA, Kogut K, Harley KG, Johnson C, et al. In utero and childhood polybrominated diphenyl ether (PBDE) exposures and neurodevelopment in the CHAMACOS study. *Environ Health Perspect*. 2013;121:257–62.
- Fischer D, Hooper K, Athanasiadou M, Athanassiadis I, Bergman Å. Children show highest levels of polybrominated diphenyl ethers in a California family of four: a case study. *Environ Health Perspect*. 2006;114:1581–4. <https://doi.org/10.1289/ehp.8554>.
- Foti F, Menghini D, Mandolesi L, Federico F, Vicari S, Petrosini L. Learning by observation: insights from Williams syndrome. *PLoS One*. 2013;8:1–10. <https://doi.org/10.1371/journal.pone.0053782>.
- Frank CL, Brown JP, Wallace K, Mundy WR, Shafer TJ. Developmental neurotoxicants disrupt activity in cortical networks on microelectrode arrays: results of screening 86 compounds during neural network formation. *Toxicol Sci*. 2017;160:121–35. <https://doi.org/10.1093/toxsci/kfx169>.
- Fritsche E, Alm H, Baumann J, Geerts L, Håkansson H, Masjosthusmann S, et al. Literature review on in vitro and alternative developmental neurotoxicity (DNT) testing methods. *EFSA Support Publ*. 2015;12:1–186. <https://doi.org/10.2903/sp.efsa.2015.en-778>.
- Fritsche E, Crofton KM, Hernandez AF, et al. OECD/EFSA workshop on developmental neurotoxicity (DNT): the use of non-animal test methods for regulatory purposes. *ALTEX*. 2017;34:311–5. <https://doi.org/10.14573/altex.1701171s>.
- Fritsche E, Barenys M, Klose J, Masjosthusmann S, Nimtz L, Schmuck M, et al. Current availability of stem cell-based in vitro methods for developmental neurotoxicity (DNT) testing. *Toxicol Sci*. 2018a;165:21–30. <https://doi.org/10.1093/toxsci/kfy178>.
- Fritsche E, Grandjean P, Crofton KM, Aschner M, Goldberg A, Heinonen T, et al. Consensus statement on the need for innovation, transition and implementation of developmental neurotoxicity (DNT) testing for regulatory purposes. *Toxicol Appl Pharmacol*. 2018b;354:3–6. <https://doi.org/10.1016/j.taap.2018.02.004>.
- Fujimoto H, Woo G-H, Morita R, et al. Increased cellular distribution of vimentin and ret in the cingulum of rat offspring after developmental exposure to decabromodiphenyl ether or 1,2,5,6,9,10-hexabromocyclododecane. *J Toxicol Pathol*. 2013;26:119–29. <https://doi.org/10.1293/tox.26.119>.
- Gangwal S, Reif DM, Mosher S, Egeghy PP, Wambaugh JF, Judson RS, et al. Incorporating exposure information into the toxicological prioritization index decision support framework. *Sci Total Environ*. 2012;435–436:316–25. <https://doi.org/10.1016/j.scitotenv.2012.06.086>.
- Gasperini RJ, Pavez M, Thompson AC, Mitchell CB, Hardy H, Young KM, et al. How does calcium interact with the cytoskeleton to regulate growth cone motility during axon path-finding? *Mol Cell Neurosci*. 2017;84:29–35. <https://doi.org/10.1016/j.mcn.2017.07.006>.
- Gerst R, Hölzer M (2019) PCAGO: an interactive web service to analyze RNA-Seq data with principal component analysis. *bioRxiv*. <https://doi.org/10.1101/433078>
- Gibson EA, Stapleton HM, Calero L, Holmes D, Burke K, Martinez R, et al. Differential exposure to organophosphate flame retardants in mother-child pairs. *Chemosphere*. 2019;219:567–73. <https://doi.org/10.1016/j.chemosphere.2018.12.008>.
- Guerrini R, Dobyns WB. Malformations of cortical development: clinical features and genetic causes. *Lancet Neurol*. 2014;13:710–26. [https://doi.org/10.1016/S1474-4422\(14\)70040-7](https://doi.org/10.1016/S1474-4422(14)70040-7).
- Hardy A, Benford D, Halldorsson T, et al. Update: use of the benchmark dose approach in risk assessment. *EFSA J*. 2017;15:1–41. <https://doi.org/10.2903/j.efsa.2017.4658>.
- Hayashi C, Suzuki N. Heterogeneity of Oligodendrocytes and Their Precursor Cells. *Adv Exp Med Biol*. 2019;1190:53–62. https://doi.org/10.1007/978-981-32-9636-7_5
- He C, English K, Baduel C, Thai P, Jagals P, Ware RS, et al. Concentrations of organophosphate flame retardants and plasticizers in urine from young children in Queensland, Australia and associations with environmental and behavioural factors. *Environ Res*. 2018a;164:262–70. <https://doi.org/10.1016/j.envres.2018.02.040>.
- He C, Toms L-ML, Thai P, van den Eede N, Wang X, Li Y, et al. Urinary metabolites of organophosphate esters: concentrations and age trends in Australian children. *Environ Int*.

- 2018b;111:124–30. <https://doi.org/10.1016/j.envint.2017.11.019>.
- Hirsch C, Striegl B, Mathes S, Adlhart C, Edelmann M, Bono E, et al. Multiparameter toxicity assessment of novel DOPO-derived organophosphorus flame retardants. *Arch Toxicol*. 2017;91:407–25. <https://doi.org/10.1007/s00204-016-1680-4>.
- Hoelting L, Klima S, Karreman C, Grinberg M, Meisig J, Henry M, et al. Stem cell-derived immature human dorsal root ganglia neurons to identify peripheral neurotoxicants. *Stem Cells Transl Med*. 2016;5:476–87. <https://doi.org/10.5966/sctm.2015-0108>.
- Hogberg HT, de Cássia da Silveira E, Sá R, Kleensang A, et al. Organophosphorus flame retardants are developmental neurotoxicants in a rat primary brain sphere in vitro model. *Arch Toxicol*. 2020;95:207–28. <https://doi.org/10.1007/s00204-020-02903-2>.
- Hu WF, Chahrouh MH, Walsh CA. The diverse genetic landscape of neurodevelopmental disorders. *Annu Rev Genomics Hum Genet*. 2014;15:195–213. <https://doi.org/10.1146/annurev-genom-090413-025600>.
- Juric-Sekhar G, Hevner RF. Malformations of cerebral cortex development: molecules and mechanisms. *Annu Rev Pathol Mech Dis*. 2019;14:293–318. <https://doi.org/10.1146/annurev-pathmechdis-012418-012927>.
- Kárádóttir R, Hamilton NB, Bakiri Y, Attwell D. Spiking and nonspiking classes of oligodendrocyte precursor glia in CNS white matter. *Nat Neurosci*. 2008;11:450–6. <https://doi.org/10.1038/nn2060>.
- Kent JC, Mitoulas LR, Cregan MD, Ramsay DT, Doherty DA, Hartmann PE. Volume and frequency of breastfeedings and fat content of breast milk throughout the day. *Pediatrics*. 2006;117:387–95. <https://doi.org/10.1542/peds.2005-1417>.
- Kim JW, Isobe T, Muto M, Tue NM, Katsura K, Malarvannan G, et al. Organophosphorus flame retardants (PFRs) in human breast milk from several Asian countries. *Chemosphere*. 2014;116:91–7. <https://doi.org/10.1016/j.chemosphere.2014.02.033>.
- Klose J, Tigges J, Masjosthusmann S, et al. TBBPA targets converging key events of human oligodendrocyte development resulting in two novel AOPs. *ALTEX Prepr*. 2020;38:1–18. <https://doi.org/10.14573/altex.2007201>.
- Krebs A, Nyffeler J, Karreman C, et al. Determination of benchmark concentrations and their statistical uncertainty for cytotoxicity test data and functional in vitro assays. *ALTEX*. 2020a;37:155–63. <https://doi.org/10.14573/altex.1912021>.
- Krebs A, van Vugt-Lussenburg BMA, Waldmann T, Albrecht W, Boei J, ter Braak B, et al. The EU-ToxRisk method documentation, data processing and chemical testing pipeline for the regulatory use of new approach methods. *Arch Toxicol*. 2020b;94:2435–61. <https://doi.org/10.1007/s00204-020-02802-6>.
- Krug AK, Balmer NV, Matt F, Schönerberger F, Merhof D, Leist M. Evaluation of a human neurite growth assay as specific screen for developmental neurotoxicants. *Arch Toxicol*. 2013a;87:2215–31. <https://doi.org/10.1007/s00204-013-1072-y>.
- Krug AK, Kolde R, Gaspar JA, Rempel E, Balmer NV, Meganathan K, et al. Human embryonic stem cell-derived test systems for developmental neurotoxicity: a transcriptomics approach. *Arch Toxicol*. 2013b;87:123–43. <https://doi.org/10.1007/s00204-012-0967-3>.
- Kucharska A, Cequier E, Thomsen C, Becher G, Covaci A, Voorspoels S. Assessment of human hair as an indicator of exposure to organophosphate flame retardants. Case study on a Norwegian mother-child cohort. *Environ Int*. 2015;83:50–7. <https://doi.org/10.1016/j.envint.2015.05.015>.
- Kuhn S, Gritti L, Crooks D, Dombrowski Y. Oligodendrocytes in development, myelin generation and beyond. *cells* 2019;1424:1–23. <https://doi.org/10.3390/cells8111424>.
- Law RJ, Covaci A, Harrad S, Herzke D, Abdallah MAE, Fernie K, et al. Levels and trends of PBDEs and HBCDs in the global environment: status at the end of 2012. *Environ Int*. 2014;65:147–58. <https://doi.org/10.1016/j.envint.2014.01.006>.
- Lein P, Silbergeld E, Locke P, Goldberg AM. In vitro and other alternative approaches to developmental neurotoxicity testing (DNT). *Environ Toxicol Pharmacol*. 2005;19:735–44. <https://doi.org/10.1016/j.etap.2004.12.035>.
- Li T, Wang W, Pan YW, Xu L, Xia Z. A hydroxylated metabolite of flame-retardant PBDE-47 decreases the survival, proliferation, and neuronal differentiation of primary cultured adult neural stem cells and interferes with signaling of ERK5 map kinase and neurotrophin 3. *Toxicol Sci*. 2013;134:111–24. <https://doi.org/10.1093/toxsci/kft083>.
- Li M, Yao Y, Wang Y, Bastiaansen M, Covaci A, Sun H. Organophosphate ester flame retardants and plasticizers in a Chinese population: significance of hydroxylated metabolites and implication for human exposure. *Environ Pollut*. 2020;257:257. <https://doi.org/10.1016/j.envpol.2019.113633>.
- López-Espíndola D, Morales-Bastos C, Grijota-Martínez C, Liao XH, Lev D, Sugo E, et al. Mutations of the thyroid hormone transporter MCT8 cause prenatal brain damage and persistent hypomyelination. *J Clin Endocrinol Metab*. 2014;99:E2799–804. <https://doi.org/10.1210/jc.2014-2162>.
- Lotharius J, Falsig J, Van Beek J, et al. Progressive degeneration of human mesencephalic neuron-derived cells triggered by dopamine-dependent oxidative stress is dependent on the mixed-lineage kinase pathway. *J Neurosci*. 2005;25:6329–42. <https://doi.org/10.1523/JNEUROSCI.1746-05.2005>.
- Luttrell WE, Olajos EJ, Pleban PA. Change in hen sciatic nerve calcium after a single oral dose of tri-O-tolyl phosphate. *Environ Res*. 1993;60:290–4.
- Ma Y, Salamova A, Venier M, Hites RA. Has the phase-out of PBDEs affected their atmospheric levels? Trends of PBDEs and their replacements in the great lakes atmosphere. *Environ Sci Technol*. 2013;47:11457–64. [dx.doi.org/https://doi.org/10.1021/es403029m](https://doi.org/10.1021/es403029m).
- Ma J, Zhu H, Kannan K. Organophosphorus flame retardants and plasticizers in breast milk from the United States. *Environ Sci Technol Lett*. 2019;6:525–31. <https://doi.org/10.1021/acs.estlett.9b00394>.
- Marinelli C, Bertalot T, Zusso M, Skaper SD, Giusti P. Systematic review of pharmacological properties of the oligodendrocyte lineage. *Front Cell Neurosci*. 2016;10:1–27. <https://doi.org/10.3389/fncel.2016.00027>.
- Marvel SW, To K, Grimm FA, et al. ToxPi Graphical User Interface 2.0: dynamic exploration, visualization, and sharing of integrated data models. *BMC Bioinformatics*. 2018;19:1–7. <https://doi.org/10.1186/s12859-018-2089-2>.

- Masjosthusmann S, Becker D, Petzuch B, Klose J, Siebert C, Deenen R, et al. A transcriptome comparison of time-matched developing human, mouse and rat neural progenitor cells reveals human uniqueness. *Toxicol Appl Pharmacol*. 2018;354:40–55. <https://doi.org/10.1016/j.taap.2018.05.009>.
- Masjosthusmann S, Blum J, Bartmann K, et al. Establishment of an a priori protocol for the implementation and interpretation of an in-vitro testing battery for the assessment of developmental neurotoxicity. *EFSA J*. 2020;17:1938e. <https://doi.org/10.2903/sp.efsa.2020.EN-1938>.
- Mayor R, Theveneau E. The neural crest. *Dev Glance*. 2013;140:2247–51. <https://doi.org/10.1242/dev.091751>.
- Mizouchi S, Ichiba M, Takigami H, Kajiura N, Takamuku T, Miyajima T, et al. Exposure assessment of organophosphorus and organobromine flame retardants via indoor dust from elementary schools and domestic houses. *Chemosphere*. 2015;123:17–25. <https://doi.org/10.1016/j.chemosphere.2014.11.028>.
- Muñoz-Quezada MT, Lucero BA, Barr DB, Steenland K, Levy K, Ryan PB, et al. Neurodevelopmental effects in children associated with exposure to organophosphate pesticides: a systematic review. *Neurotoxicology*. 2013;39:158–68. <https://doi.org/10.1016/j.neuro.2013.09.003>.
- Nimt L, Klose J, Masjosthusmann S, et al. The neurosphere assay as an in vitro method for developmental neurotoxicity (DNT) evaluation. *Cell Cult Tech Neuromethods*. 2019;145:141–68. <https://doi.org/10.1007/978-1-4939-9228-7>.
- Nimt L, Hartmann J, Tigges J, Masjosthusmann S, Schmuck M, Keßel E, et al. Characterization and application of electrically active neuronal networks established from human induced pluripotent stem cell-derived neural progenitor cells for neurotoxicity evaluation. *Stem Cell Res*. 2020;45:101761. <https://doi.org/10.1016/j.scr.2020.101761>.
- Nyffeler J, Dolde X, Krebs A, Pinto-Gil K, Pastor M, Behl M, et al. Combination of multiple neural crest migration assays to identify environmental toxicants from a proof-of-concept chemical library. *Arch Toxicol*. 2017;91:3613–32. <https://doi.org/10.1007/s00204-017-1977-y>.
- OECD 2007. OECD guideline for the testing of chemicals: health effects. Test No. 426: developmental neurotoxicity study. <http://www.oecd.org/dataoecd/20/52/37622194.pdf>
- Phillips DL, Pirkle JL, Burse VW, Bernert JT Jr, Henderson LO, Needham LL. Chlorinated hydrocarbon levels in human serum: effects of fasting and feeding. *Arch Environ Contam Toxicol*. 1989;18:495–500. <https://doi.org/10.1007/BF01055015>.
- Pistolato F, De Gyves EM, Carpi D, et al. Assessment of developmental neurotoxicity induced by chemical mixtures using an adverse outcome pathway concept. *Environ Heal A Glob Access Sci Source*. 2020;19:1–26. <https://doi.org/10.1186/s12940-020-00578-x>.
- Prentice P, Ong KK, Schoemaker MH, Tol EAF, Vervoort J, Hughes IA, et al. Breast milk nutrient content and infancy growth. *Acta Paediatr Int J Paediatr*. 2016;105:641–7. <https://doi.org/10.1111/apa.13362>.
- Razek AA, Mazroa J, Baz H. Assessment of white matter integrity of autistic preschool children with diffusion weighted MR imaging. *Brain and Development*. 2014;36:28–34. <https://doi.org/10.1016/j.braindev.2013.01.003>.
- Reif DM, Martin MT, Tan SW, Houck KA, Judson RS, Richard AM, et al. Endocrine profiling and prioritization of environmental chemicals using toxcast data. *Environ Health Perspect*. 2010;118:1714–20. <https://doi.org/10.1289/ehp.1002180>.
- Roze E, Meijer L, Bakker A, van Braeckel KNJA, Sauer PJJ, Bos AF. Prenatal exposure to organohalogens, including brominated flame retardants, influences motor, cognitive, and behavioral performance at school age. *Environ Health Perspect*. 2009;117:1953–8. <https://doi.org/10.1289/ehp.0901015>.
- Rylander L, Nilsson-Ehle P, Hagmar L. A simplified precise method for adjusting serum levels of persistent organohalogen pollutants to total serum lipids. *Chemosphere*. 2006;62:333–6. <https://doi.org/10.1016/j.chemosphere.2005.04.107>.
- Sachana M, Bal-Price A, Crofton KM, Bennekou SH, Shafer TJ, Behl M, et al. International regulatory and scientific effort for improved developmental neurotoxicity testing. *Toxicol Sci*. 2019;167:45–57. <https://doi.org/10.1093/toxsci/kfy211>.
- Saegusa Y, Fujimoto H, Woo G-H, Inoue K, Takahashi M, Mitsumori K, et al. Developmental toxicity of brominated flame retardants, tetrabromobisphenol A and 1,2,5,6,9,10-hexabromocyclododecane, in rat offspring after maternal exposure from mid-gestation through lactation. *Reprod Toxicol*. 2009;28:456–67. <https://doi.org/10.1016/j.reprotox.2009.06.011>.
- Schmidt BZ, Lehmann M, Gutbier S, Nembo E, Noel S, Smirnova L, et al. In vitro acute and developmental neurotoxicity screening: an overview of cellular platforms and high-throughput technical possibilities. *Arch Toxicol*. 2017;91:1–33. <https://doi.org/10.1007/s00204-016-1805-9>.
- Schmuck MR, Temme T, Dach K, et al. Omnisphero: a high-content image analysis (HCA) approach for phenotypic developmental neurotoxicity (DNT) screenings of organoid neurosphere cultures in vitro. *Arch Toxicol*. 2016;1–12. <https://doi.org/10.1007/s00204-016-1852-2>.
- Scholz D, Pörtl D, Genewsky A, Weng M, Waldmann T, Schildknecht S, et al. Rapid, complete and large-scale generation of post-mitotic neurons from the human LUHMES cell line. *J Neurochem*. 2011;119:957–71. <https://doi.org/10.1111/j.1471-4159.2011.07255.x>.
- Schreiber T, Gassmann K, Götz C, Hübenthal U, Moors M, Krause G, et al. Polybrominated diphenyl ethers induce developmental neurotoxicity in a human in vitro model: evidence for endocrine disruption. *Environ Health Perspect*. 2010;118:572–8. <https://doi.org/10.1289/ehp.0901435>.
- Shafer TJ, Brown JP, Lynch B, Davila-Montero S, Wallace K, Friedman KP. Evaluation of chemical effects on network formation in cortical neurons grown on microelectrode arrays. *Toxicol Sci*. 2019;169:436–55. <https://doi.org/10.1093/toxsci/kfz052>.
- Shy C-G, Huang H-L, Chang-Chien G-P, Chao HR, Tsou TC. Neurodevelopment of infants with prenatal exposure to polybrominated diphenyl ethers. *Bull Environ Contam Toxicol*. 2011;87:643–8. <https://doi.org/10.1007/s00128-011-0422-9>.
- Simons M, Trajkovic K. Neuron-glia communication in the control of oligodendrocyte function and myelin biogenesis. *J Cell Sci*. 2006;119:4381–9. <https://doi.org/10.1242/jcs.03242>.
- Stapleton HM, Klosterhaus S, Eagle S, Fuh J, Meeker JD, Blum A, et al. Detection of organophosphate flame retardants in furniture foam and U.S. house dust. *Environ Sci Technol*. 2009;43:7490–5. <https://doi.org/10.1021/es9014019>.

- Stapleton HM, Misenheimer J, Hoffman K, Webster TF. Flame retardant associations between children's handwipes and house dust. *Chemosphere*. 2014;116:54–60. <https://doi.org/10.1016/j.chemosphere.2013.12.100>.
- Stiegler NV, Krug AK, Matt F, Leist M. Assessment of chemical-induced impairment of human neurite outgrowth by multiparametric live cell imaging in high-density cultures. *Toxicol Sci*. 2011;121:73–87. <https://doi.org/10.1093/toxsci/kfr034>.
- Stumpf AM, Tanaka D, Aulerich RJ, Bursian SJ. Delayed neurotoxic effects of tri-o-tolyl phosphate in the european ferret. *J Toxicol Environ Health*. 1989;26:61–73. <https://doi.org/10.1080/15287398909531233>.
- Sugeng EJ, Leonards PEG, van de Bor M. Brominated and organophosphorus flame retardants in body wipes and house dust, and an estimation of house dust hand-loadings in Dutch toddlers. *Environ Res*. 2017;158:789–97. <https://doi.org/10.1016/j.envres.2017.07.035>.
- Sundkvist AM, Olofsson U, Haglund P. Organophosphorus flame retardants and plasticizers in marine and fresh water biota and in human milk. *J Environ Monit*. 2010;12:943–51. <https://doi.org/10.1039/b921910b>.
- Tang J, Zhai JX. Distribution of polybrominated diphenyl ethers in breast milk, cordblood and placentas: a systematic review. *Environ Sci Pollut Res*. 2017;24:21548–73. <https://doi.org/10.1007/s11356-017-9821-8>.
- Tang H, Hammack C, Ogden SC, Wen Z, Qian X, Li Y, et al. Zika virus infects human cortical neural progenitors and attenuates their growth. *Cell Stem Cell*. 2016;18:587–90. <https://doi.org/10.1016/j.stem.2016.02.016>.
- Terron A, Bennekou Hougaard S. Towards a regulatory use of alternative developmental neurotoxicity testing (DNT). *Toxicol Appl Pharmacol*. 2018;354:19–23. <https://doi.org/10.1016/j.taap.2018.02.002>.
- Thomson JA, Itskovitz-Eldor J, Shapiro SS, et al. Embryonic stem cell lines derived from human blastocysts. *Science*. 1998;282(80):1145–7. <https://doi.org/10.1126/science.282.5391.1145>.
- Toms L-ML, Sjödin A, Harden F, Hobson P, Jones R, Edenfield E, et al. Serum polybrominated diphenyl ether (PBDE) levels are higher in children (2-5 years of age) than in infants and adults. *Environ Health Perspect*. 2009;117:1461–5. <https://doi.org/10.1289/ehp.0900596>.
- Tonduti D, Vanderver A, Berardinelli A, et al. MCT8 deficiency: extrapyramidal symptoms and delayed myelination as prominent features. *Child Neurol Psychiatry*. 2014;28:795–800. <https://doi.org/10.1177/0883073812450944.MCT8>.
- Tsuji R, Crofton KM. Developmental neurotoxicity guideline study: issues with methodology, evaluation and regulation. *Congenit Anom (Kyoto)*. 2012;52:122–8. <https://doi.org/10.1111/j.1741-4520.2012.00374.x>.
- van der Veen I, de Boer J. Phosphorus flame retardants: properties, production, environmental occurrence, toxicity and analysis. *Chemosphere*. 2012;88:1119–53. <https://doi.org/10.1016/j.chemosphere.2012.03.067>.
- Volpe JJ, Kinney HC, Jensen FE, Rosenberg PA. The developing oligodendrocyte: key cellular target in brain injury in the premature infant. *Int J Dev Neurosci*. 2011;29:423–40. <https://doi.org/10.1016/j.ijdevneu.2011.02.012>.
- Waaaijers SL, Kong D, Hendriks HS, et al. Persistence, Bioaccumulation, and Toxicity of Halogen-Free Flame Retardants. *Rev Environ Contam Toxicol*. 2013;222:1–71. <https://doi.org/10.1007/978-1-4614-6470-9>.
- Walter KM, Dach K, Hayakawa K, et al. Ontogenetic expression of thyroid hormone signaling genes: An in vitro and in vivo species comparison. *PLoS One*. 2019;14:1–26. <https://doi.org/10.1371/journal.pone.0221230>.
- Wolff JJ, Ph D, Gu H, et al. Differences in white matter fiber tract development present from 6 to 24 months in infants with autism. *Am J Psychiatry*. 2013;169:589–600. <https://doi.org/10.1176/appi.ajp.2011.11091447.Differences>.
- Yogui GT, Sericano JL. Polybrominated diphenyl ether flame retardants in the U.S. marine environment: a review. *Environ Int*. 2009;35:655–66. <https://doi.org/10.1016/j.envint.2008.11.001>.
- Zilles K, Palomero-Gallagher N, Amunts K. Development of cortical folding during evolution and ontogeny. *Trends Neurosci*. 2013;36:275–84. <https://doi.org/10.1016/j.tins.2013.01.006>.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Neurodevelopmental toxicity assessment of flame retardants using a human DNT in vitro testing battery

Jördis Klose, Melanie Pahl, Kristina Bartmann, Farina Bendt, Jonathan Blum, Xenia Dolde, Nils Förster, Anna-Katharina Holzer, Ulrike Hübenthal, **Hagen Eike Keßel**, Katharina Koch, Stefan Masjosthusmann, Sabine Schneider, Lynn-Christin Stürzl, Selina Woeste, Andrea Rossi, Adrian Covaci, Mamta Behl, Marcel Leist, Julia Tigges, Ellen Fritsche

Journal:	Cell Biology and Toxicology
Impact factor:	6.691 (2021-2022)
Contribution to the publication:	5%
	Biostatistics and image analysis software development, data analysis and evaluation
Type of authorship:	Co-authorship
Status of publication:	Published 10 th May 2021

3 Discussion

To improve human risk assessment and reduce animal testing, the US-National Research Council proposed a new strategy for toxicity testing in the 21st century, which is based on a shift from conventional *in vivo* toxicity testing to high throughput *in vitro* assays (NRC, 2007; Collins *et al.*, 2008). The assays and bioinformatic tools developed in the context of this shift should also help to close the knowledge gap for DNT hazard assessment of a large number of compounds (Crofton and Mundy 2021; Grandjean and Landrigan, 2014; Tsuji and Crofton, 2012). To close this gap and gain compound data fit for regulatory decision making, a biologically relevant DNT-IVB has been established. An integral part of this DNT-IVB is the human neurosphere assay that contributes unique endpoints to the battery (Crofton and Mundy 2021). Such neurospheres originate from the correct species for human risk assessment, can be grown in 3D (Alépée *et al.*, 2014), differentiate into multiple cell types (Moors *et al.* 2009) and thus allow evaluation of a large variety of endpoints in an organotypic manner (Fritsche *et al.*, 2015; Koch *et al.* 2022). As described in manuscript 2.3 (Blum *et al.*, 2022), the neurosphere model based on human neural progenitor cells (NPC) along with human induced pluripotent stem cell (hiPSC)-derived neural crest cells and sensory neurons (Nyffeler *et al.*, 2017; Hoelting *et al.*, 2016; Holzer *et al.*, 2022) as well as differentiated dopaminergic Luhmes cells (Delp *et al.*, 2018) were established as cell models for DNT hazard characterization. However, even the best cell model cannot be used for hazard assessment, if the data acquired from this model is not reliable. Bioinformatic tools not only need to bring along high-throughput capabilities to reduce time and resources compared to *in vivo* testing, but also to reliably process endpoint data and predict hazard on this basis. This is an especially challenging undertaking, since one of the key characteristics of computational high throughput workflows poses their ability to function without human supervision (or as minimal supervision as possible) for data generation and evaluation. For the workflow presented in this thesis, it thus was a demanding task to ensure that the algorithms employed for this purpose were (i) able to handle vast amounts of data from different assays, (ii) generate endpoint data while dealing with biological and technical aberrations, (iii) evaluate the data with as much certainty as possible while taking these aberrations into account and (iv) make precise hazard predictions on this basis, all fully automatized with as less human intervention as possible. The workflow we established incorporates all steps from image acquisition to hazard predictions (Fig. 6). This thesis aims to explore the different steps: It will be discussed, how the use of adequate algorithms and statistical methods can lead to robust and reliable data, by comparing the methods we implemented for this purpose with alternative, yet popular, approaches. By taking a closer look at employed algorithms and methods of this workflow, gained knowledge about their advantages and remaining concerns are discussed, including suggestions that we are able to make based on the gained knowledge. Finally, it will be discussed how different

parts (e.g. image acquisition as one and BMC estimation as another part) of the workflow are interconnected and thus affecting each other.

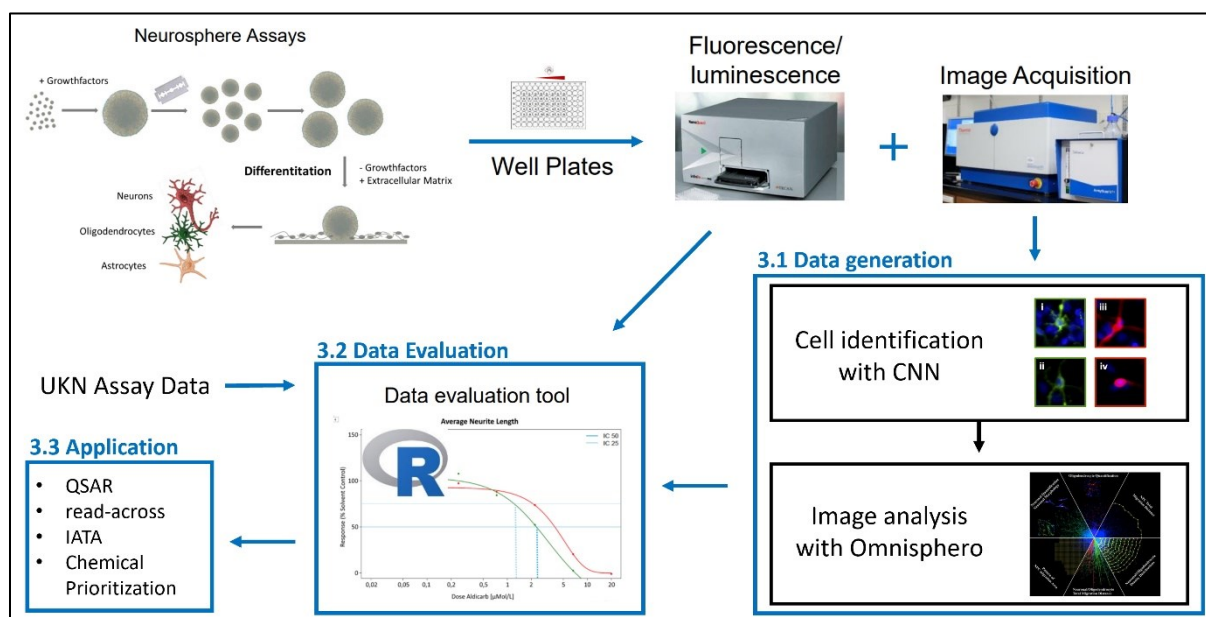


Figure 6: DNT-IVB workflow

Primary human neurospheres are exposed towards chemicals when plated into laminin-coated 96-well plates with one sphere (0,3 mm Ø) per well. Cytotoxicity and viability measurements are performed on the cell supernatants after 5 days of exposure with the LDH- and Alamar Blue Assays, respectively, using a fluorescence and luminescence multiplate reader. In parallel, remaining cells are fixed in paraformaldehyde and stained with the DNA intercalating dye Hoechst for nuclei and immunocytochemical reactions using antibodies against β III-tubulin (neurons) and O4 (oligodendrocytes). Fluorescence images are taken with the *ArrayScan V⁷ HCS Reader* (ThermoFisher Scientific). Neurons and Oligodendrocytes are subsequently identified by a CNN. Images and cell identities are imported into Omnisphero and analysed for endpoint data generation (discussed in section 3.1). Generated endpoint data is passed down to the data evaluation tool written in R for hazard characterization (3.2). Resulting evaluation can be used for QSAR, IATA, read-across approaches, chemical prioritization and other fields of application for hazard assessment (3.3).

3.1 Data generation

In order to extract endpoint data for a certain cell type, cells of this type must first be identified within the image. Omnisphero originally relied on overlap algorithms for cell identification and skeletonization for assessment of morphological endpoints such as neurite length or number of branching points. On the basis of these algorithms, it was already shown that Omnisphero vastly outperformed other cell identification tools, such as *Neuronal Profiling BioApplication version 4.1* (Schmuck *et al.*, 2017), which relies on overlap algorithms and superellipsoids for identification and morphology analyses.

Evaluation of additional cell donors, changes in image acquisition (new device, different camera) and staining protocols since the set-up of Omnisphero indicated that cell identification software tools are required, which go beyond the before developed algorithms. Therefore, the decision was made to replace Omnisphero's overlap algorithms with ML approaches for cell identification. ML approaches (often implemented as deep learning models) have already been used for a variety of applications in

life science. Some examples are the use of ML approaches for classification of disease, localization of organs and lesions, or segmentation of organic structures (Litjens *et al.*, 2017; Shen *et al.*, 2017), rendering ML technology as a promising approach for cell-type classification within the neurosphere model. Two supervised learning models were implemented as convolutional neuronal network (CNN) models to identify neurons and oligodendrocytes, respectively, within neurosphere fluorescence images. With these CNN models, a performance of a precision and recall with area-under-curve (AUC) values around 0.8 for both oligodendrocytes and neurons was achieved for the validation dataset and deemed as accurate enough to be applied for image analyses (manuscript 2.1 – Förster *et al.*, 2021). A direct comparison with the outdated overlap-algorithm (Figure 7) clearly reveals the superiority of the new approach, as it is able to handle confounders such as luminosity (7B), staining artifacts (7C) and overlapping cells (7D). This finding is in line with recent development in the field of life science, where ML based technology is often seen as a superior alternative to other algorithms in a variety of different *in vitro* studies (Shariff *et al.*, 2010; Ching *et al.*, 2018). So far, only neurons and oligodendrocytes are identified with our CNN models. However, as they show promising performance, more CNN models can be employed for other cell types in the future.

As the employment of ML approaches for *in vitro* studies is gaining increasing attention (Ching *et al.*, 2018; Villeneuve *et al.*, 2019), identifying guidelines and minimum standards for application of ML approaches is an important contribution for such ML methods to gain acceptance. With the knowledge gained by establishment of the novel algorithms for cell identification within the neurosphere model (manuscript 2.1 – Förster *et al.*, 2021), we are able to formulate guidelines for validation of employed algorithms. These do not only apply for the establishment of ML approaches for analysis within the neurosphere model, but are rather applicable for the development of ML approaches for *in vitro* studies in general.

We recommend external validation of CNN data, as over-fitting is an inherent danger of CNN models (Choi *et al.*, 2020) and can potentially lead to false positive or false negative concentration–response relationships in compound screening (manuscript 2.1 – Förster *et al.*, 2021). To counteract this, external validation can be done. A fully independent comparison (e.g. validation on data from different laboratories using different devices and a fully independent sample preparation) would be the ideal way to validate. If a CNN is only trained with data deriving from one laboratory, but subsequently is also able to perform correctly on the data of the same assay derived from other laboratories, it thus is validated as robust against differences in data origin. Welch *et al.* (2020) were already able to demonstrate this, as they trained a CNN model for classification of dental artifacts and successfully validated it by application on external datasets. We were able to show the benefit of such validation, by training our CNN models with cells from different individuals, which in a validation step led the models to be more robust against inter-individual differences (manuscript 2.1 – Förster *et al.*, 2021).

These findings indicate that a crucial part of CNNs in screening applications is the identification of sources of bias and confounding, in order to systematically validate against these confounders. Furthermore, to ensure reliability of CNN model predictions used for studying the effects of substances on neurospheres, it is important to validate the resulting concentration-response relationship of these substances. To do this, it is necessary to compare the concentration-response pattern gained by the CNN to the expected pattern of a substances with known toxicity behavior, which is very similar to the concept of assay performance discussed in manuscript 2.2 (Keßel *et al.*, 2022, preprint).

Data annotation is often seen as the bottleneck of ML based approaches (Shariff *et al.*, 2010) and often a major issue for employment of ML approaches in life science, as it is a time- and sometimes even resource-consuming procedure (Zheng *et al.*, 2017). It is because of this, that ML approaches are often considered with skepticism by regulatory decision makers, as many ML models lack sufficient training data (Ching *et al.*, 2018). We therefore trained the models with a set of 10,945 cells (containing 1,114 annotated neurons and 718 annotated oligodendrocytes) of different chemical treatment and added augmented data (Dhungel *et al.*, 2015), ensuring well trained models fit for regulatory purposes (manuscript 2.1 – Förster *et al.*, 2021).

As ML approaches consistently improved over the last years (Ching *et al.*, 2018), the question arises, how well they perform compared to true human evaluation. Our experience (gained by visual assessment of random samples) has shown, that the employed CNN models matched human

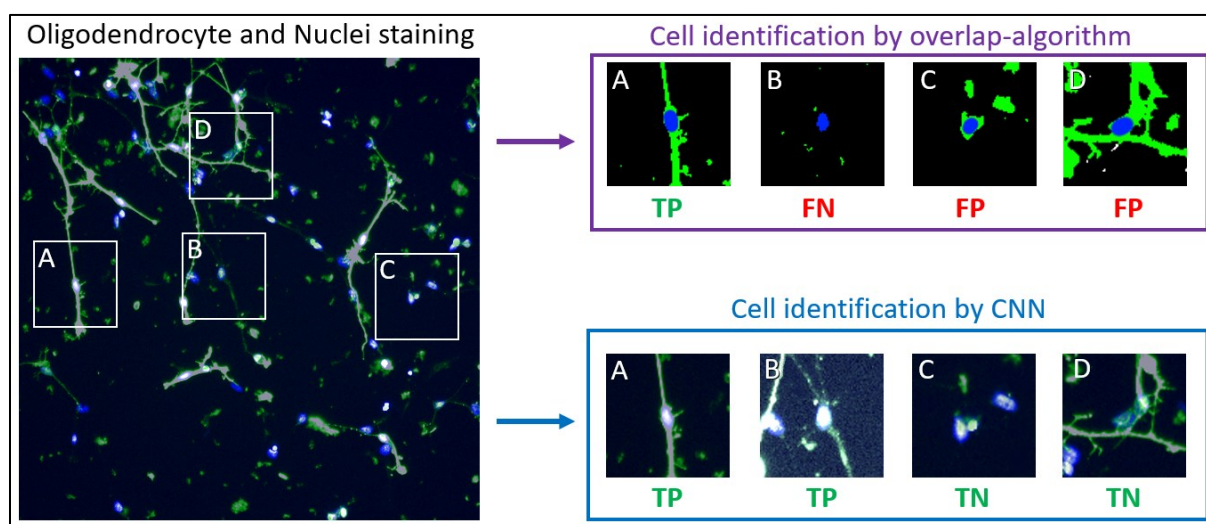


Figure 7: Oligodendrocyte identification with different bioinformatical approaches

Staining of nuclei (Hoechst, blue) and oligodendrocytes (O4, green) are depicted as displayed in Omnisphero. Two algorithms are compared for their performance of identifying oligodendrocytes. Four cells are picked as examples (A-D). The overlap algorithm binarizes the staining images, where every pixel with brightness values above a set threshold is translated into a binarized signal. This is done for both staining channels, resulting in a binarized image with both channels overlapping. Several parameters define how much overlap the oligodendrocyte pixels (green) must have with the according nuclei pixels (blue) so that the according nuclei is identified as an oligodendrocyte. For the CNN, both channels are merged into one image and contrast-normalized image tiles (one tile is corresponding to one nucleus) are evaluated for oligodendrocytes.

evaluation and sometimes even outperformed it. This is in line with a study from Kooi *et al.*, (2017), where they found no significant differences between the performance of a network model and certified screening radiologists detecting mammographic lesions.

3.2 Data evaluation

To translate the generated endpoint data deriving from image analyses tools into evaluations which enable statements about the potential toxicity, i.e. hazard, of tested compounds, biostatistical processing and evaluation of the endpoint data is required. In the course of a compound screening project performed on behalf of an EFSA procurement during the years 2017-2020 (OC/EFSA/PRAS/2017/01), a bioinformatics workflow was developed which enables processing and subsequent DNT evaluation of NPC and UKN assay data (manuscript 2.2 – Keßel *et al.*, 2022, preprint), which is the basis of an OECD guidance document on use and interpretation of the DNT-IVB (Crofton and Mundy, 2021). As part of the workflow, a biostatistics pipeline was employed and also used for a study analyzing the impact of common biostatistical concentration-response methods on the overall DNT-IVB performance. As *in vitro* methods have been gaining complexity over the last decade, i.e. from reporter gene assays towards organotypic cultures, the hypothesis if the selection of a biostatistical method can affect the performance of the DNT-IVB was tested. Therefore, a comparative assessment of different biostatistical methods on the BMC estimation, DNT hit classification and DNT-IVB performance was performed (manuscript 2.2 - Keßel *et al.*, 2022, preprint). A standard data evaluation protocol for DNT-IVB data was defined and by changing statistical methods as part of the protocol, the impact on BMC estimation, the uncertainty of a BMC (expressed as the width of the central 95% confidence interval of a BMC estimation), the endpoint-specific hazard classification of the compound and the final assay performance were quantified and compared across the various specific assay endpoints. Five key aspects of HTS data evaluation were identified and evaluated for their impact on hazard identification: i) The impact of different methods for experimental data averaging. Only minor differences in BMC estimations (Fig. 8A and B) and hazard classification outcomes (Fig. 8I-K) were observed by comparing the two approaches, with relatively few data sets, where a strong difference on the estimated BMC (Fig. 8C) was observed. ii) The impact of different data normalization approaches. Very different BMC estimations (Fig. 8A) were often observed for the both methods and furthermore, where the BMC is not supported by the data in extreme cases (Fig. 8D). Although the majority of data sets did not necessarily require a control-renormalization, a change to the standard control normalization still changed the hit category for approx. 5% of all endpoint-specific DNT hazard classifications and reduced the performance of the DNT-IVB's predictivity (Fig. 8I-K). iii) The impact of different approaches for concentration-response regression modelling. The best-fit model approach responded more flexible to data sets and therefore resulted often to BMC estimations that differed

significantly from those derived by the sole application of one predefined three-parameter log-logistic model (Figure 8E). Furthermore, the sole application of the Hill model occasionally prevented the estimation of a BMC and its uncertainty, leading to less data sets for which a hazard identification could be performed. Comparison between inverse regression and model averaging for BMC estimation showed no big differences between both methods (Fig. 8A). iv) The impact of different approaches for estimation of BMCs and their uncertainty. There are three general types of BMCs and uncertainty estimation methods: inverse regression, asymptotic approaches and bootstrapping approaches. Inverse regression estimates the uncertainty directly from the regression fit around the BMC (Buckley, Piegorsch & West, 2009; Fang, Piegorsch & Barnes, 2015) and was found to be the most reliable method. The delta method is an asymptotic approach which combines information of the estimated model parameters to derive a Wald-type interval (Jensen *et al.*, 2020). This approach often led to an unreliable CI spanning the entire range of test concentrations or failed entirely (Fig. 8F). Based on the study outcome, this method is deemed as unfit for an automatic HTS data evaluation. Both bootstrapping and model averaging are based on computer-intensive statistical resampling techniques that resample the original dataset to create a huge number of simulated samples (Jensen *et al.*, 2019). These methods put strong emphasis on the given data for the resampling and are thus vulnerable to biased interval estimations if the data shows high variability between tested concentrations, i.e. mode of the resampled BMC distribution differs from the original BMC estimation. Furthermore, due to the small number of biological replicates, given assay designs are not optimal for regression resampling. Thus it is not surprising that bootstrapping often resulted in very different estimations compared to inverse regression. The CI was often vastly wider and sometimes even failed to produce an estimation (Fig. 8G). v) The impact of measured BMRs. The use of BMR50 (BMR set at 50% response) has been a common practice for years and is still used in recent publications, despite not having any biological reasoning. As an alternative, an endpoint-specific BMR that is adjusted to the baseline noise of the according endpoint can be used. A larger BMR leads to a higher BMC and the consequence for all data sets with a much lower data variability is that their substance responses observed at concentration ranges below the BMC are ignored. In a hazard identification context, this is problematic, since it contradicts the intended regulatory meaning of a benchmark concentration. Furthermore, it would also rule out those data sets for a BMC estimation where the observed maximal responses are below the BMR and thus a BMC cannot be established. As a consequence, hazard classifications change, with a change mainly from specific DNT hit classifications to no hit classifications, which lastly affects the assay performance as well (Fig. 8I-K). Thus, the use of the most common descriptor for concentration response data in pharmacology and *in vitro* toxicology, an IC₅₀ or EC₅₀, cannot be recommend as surrogate for a BMC for endpoints of the DNT-IVB.

A broad variety of free software packages for the statistical analysis of dose-response data and dose-response modelling are available, with PROAST (RIVM National Institute for Public Health and the Environment), BMDS (US EPA), ToxCast pipeline (tcpl, Filer *et al.* 2017) or BMCEasy (Krebs *et al.*, 2019) posing some of the many options. Similar to the R packages we use (drc and bmd, Ritz *et al.*, 2015 and Jensen *et al.*, 2020), most of these software packages provide a variety of options in order to respond as flexible as possible to the various data scenarios a user can possibly face, yet always require a certain degree of statistical (and sometimes also coding) knowledge from the user. Similar to the tcpl pipeline we became interested in an automated data evaluation platform with no required user intervention and addressing the specific features of DNT data or other data from organotypic cultures. The standout feature of our data evaluation platform is the integration of a sophisticated endpoint-specific hazard classification model, including flagging systems for uncertain cases, which none of the software packages mentioned above offer. Rather than just relying on one benchmark value (or the lower limit of such; Filer *et al.* 2017, Jensen *et al.*, 2020), our classification model involves confidence intervals for different hazard classifications. We consider it crucial for the hazard assessment to differentiate between general cell toxicity and specific DNT hits. None of the aforementioned software programs do inherit any classification model and require the data to be exported into another software to gain classifications. This poses another barrier and potential pitfall to overcome to gain reliable classifications, as the experimenter needs to operate a separate software which is not guaranteed to handle the data derived from external software appropriately (again, statistical and coding knowledge is required). With the classification model integrated into our pipeline, all classifications are fully automated, appropriate for the data format and require no prior statistical or coding knowledge, thus reducing the human handling error and making it a reliable and accessible option for experimenters. These findings point out the relevance of careful employment of statistical approaches in DNT data evaluation. Each method and software comes with its own advantages and disadvantages, where finding the method and software that is best suited for the given data and purpose is key.

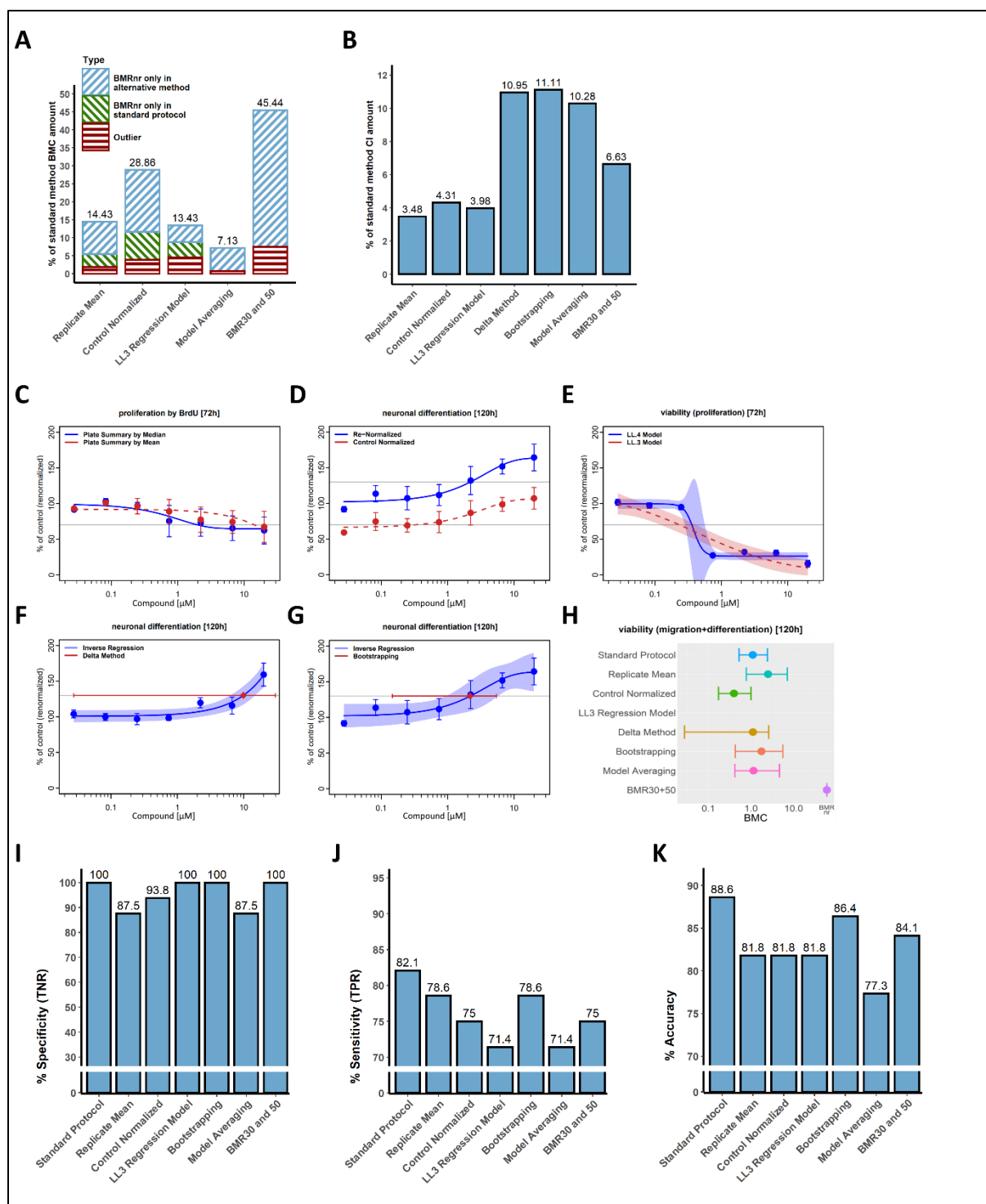


Figure 8: Methodological changes in the data evaluation on BMC estimation and assay performance

A)-B) Distribution of (A) BMC fold-changes and (B) CI width fold-changes in response to statistical method changes from the standard protocol. Box whisker plot show the median (horizontal line), interquartile range (box), 5% and 95% percentile values (whisker), and extreme values (black dots). C)-H) For several different steps of the data analysis and evaluation, the data resulting from the standard protocol (blue) is compared to the data deriving from the alternative protocol (red). Error bars show the SEM between summarized experiment data. Horizontal grey lines indicate the BMR. C) Experiment summarization by median and by mean. D) Re-normalized data and control-normalized data. E) Best fit approach and use of only a LL3 regression model. CI is displayed as confidence band around the fit model. Both models are applied to the data shown in blue. F) Inverse regression and delta method. CI of the alternative method is shown as red bar and BMC as red square. G) Inverse regression and bootstrapping. H) All method changes and their resulting BMC (displayed as dots) and CI values (displayed as bars) are shown for one exemplary dataset. I)-K) Bar graphs show the results of the predictive capability of the DNT-IVB for 28 substances of known DNT and 17 negative control substances in terms of specificity, sensitivity and accuracy.

3.3 Data application

While data support for alternative testing approaches are given, regulatory jurisdictions are lacking behind. Thus, creating a framework that is fit-for-purpose to incorporate the DNT-IVB into regulatory processes attending regulatory questions is recommended. Yet, it was already shown to be applicable for several hazard and risk decisions (Crofton and Mundy, 2021). One example poses the case study of manuscript 2.4 (Klose *et al.*, 2021), in which the DNT-IVB was used as a first case study for screening and prioritization of 15 data-poor compounds belonging to the class of flame retardants including phased out and alternative flame retardants, further closing the gap of data knowledge. By estimating BMCs and CIs, as well as subsequent classification, specific DNT hits were identified across the endpoints of the battery, giving information about potential hazard deriving from these flame retardants and enabling compound prioritization. For instance, triphenyl isopropylated phosphate and tert-butylphenyl diphenyl phosphate both were identified as toxicants affecting the migration of neural crest cells. This finding is in line with observations that were made with other models such as zebrafish or rat cortical neurons (Behl *et al.*, 2015), revealing the capability of the DNT-IVB to replicate known DNT effects. However, results from this study also revealed new phenomena: For the first time, specific toxic effects on proliferation were shown for tricresyl phosphate and 2-ethylhexyl diphenyl phosphate in human cells, hinting at the possibility of the DNT-IVB to uncover new knowledge about DNT attributes of compounds. These results are of great relevance for human risk assessment, considering that this finding was made with complex 3D human cell models rather reflecting the human system than 2D animal cell models. It furthermore is to note that the DNT-IVB was not able to replicate all observations made for other models. For example, none of the 15 FRs tested in the study showed any effect on human neuronal differentiation, while all 5 FRs tested in rat neurospheres affected neuronal differentiation (Hogberg *et al.*, 2020).

With the BMC and uncertainty values at hand, compounds can be prioritized, for example by ranking the magnitude of compound effect (BMCs, “ToxPi Scores”, manuscript 2.4 – Klose *et al.*, 2021). Another extrapolation that can be done with BMC data is the transition from *in vitro* systems to *in vivo* risk estimations. This is done by converting a given compound concentration from an *in vitro* system into an estimated internal concentration. Due to their low demand on time and resources, the DNT-IVB can also be applied for screening of data-poor compounds for DNT. It is recommended to prioritize compounds with high human exposure or have structural similarities to known DNT compounds. If any KNDP is affected by a compound, this could be taken as a point-of-departure for further steps such as kinetic modelling, QSAR, IVIVE and estimation of adverse doses. Another example for the use of NAMs for regulatory purposes is the use of data from multiple DNT-IVB assays for weight-of-evidence estimation for organophosphates (USEPA 2020b). Furthermore, *in vitro* data was used by EFSA to

develop IATA case studies for deltamethrin and flufenacet (EFSA *et al.* 2021), which resulted in an AOP-informed DNT risk assessment, comprehending available information from a broad variety of DNT assays (e.g. *in vitro*, *in vivo*, toxicokinetics and epidemiology).

While the DNT-IVB shows promising results for application in regulatory decision making (manuscript 2.3 – Blum *et al.*, 2022), there are remaining concerns about its predictivity for human hazard. The assay performance analysis has shown 82.1% sensitivity for the standard protocol (Fig. 8J). It thus is clear that not every compound inducing human DNT hazard is also evaluated as a DNT hazard by the DNT-IVB resulting in a false negative classification. This can be due to the lack of important KNDP like neuronal network formation in the evaluated data set, and the discussed differences between *in vitro* systems and the far more complex biology of the human body. However, it can also go in the other direction: a specific DNT hit in the DNT-IVB is not necessary reflecting a real DNT hazard for humans resulting in a false positive classification. Both of these circumstances further point to the relevance of incorporating data from multiple test systems covering a large variety of KNDP to gain as much information about potential hazards as possible for safe decision making. A similar concern exists for no hits, since they may either be true (no hazard) or false negatives (compound has toxic properties but they were not detected by the DNT-IVB). The main sources of uncertainty on negatives are the gaps in the battery, i.e. KNDP or specific signalling pathways not covered, and toxicokinetic aspects.

Estimations of assay performance with control compounds is an instrument for DNT-IVB validation, where there is a trade-off between sensitivity and specificity. If the DNT-IVB is very specific, but not very sensitive, the positive hit calls have high certainty. This is because the high specificity showed that compounds not inducing any DNT effects are reliably detected as negatives. It thus is less likely, that unknown DNT negatives are considered as (false) positive by the DNT-IVB. The opposite case would be a DNT-IVB with a high sensitivity, but lower specificity. Here, it would be more likely for the DNT-IVB to cover more DNT hazards, on the cost of producing more false positives. Putting this trade-off in the context of hazard assessment, both options come with their advantages in disadvantages: A higher specificity would mean more confident hit calls, resulting in more relevance for hazard assessment, i.e. the hit calls have a higher precision and are more reliable to predict potential human hazard. Yet, more additional information from other assays would be needed to be incorporated to cover the DNT positives not detected by the very specific DNT-IVB. A higher sensitivity would mean that more potential DNTs are detected, i.e. the DNT-IVB expresses more sources of potential human hazard. For hazard assessment in the regulatory context, this is the more desirable option, since it is more favourable for human health to identify harmless substances as harmful rather than identifying harmful substances as harmless. Yet, it comes with less certainty, reducing the relevance of hit calls

and also increasing the number of follow-up testing required to separate false positives from true positives.

3.4 Connecting the dots: How data generation affects the evaluation and what it means for the application in hazard assessment

As depicted in Figure 6, the process from cell model exposure to hazard assessment is a chain of subsequent steps, where each step depends on the outcome of the last. It thus is obvious that changes in one of the steps affects all the subsequent ones. Consequently, for reliable information on potential DNT hazard, each step must be carefully established and validated. This is even the more crucial for an automatized workflow, since the experimenter is supposed to give the biological material (cell cultures) as input and receives the evaluated hazard data as output – rendering the entire process in between as a “black box”. This black box must be trusted to fulfil each function with precision and robustness against data variability, which neurosphere assay readouts are susceptible to. It therefore is a very insightful case study to examine how changes of not only in one, but several of the different workflow steps (in our case employment of a new algorithm for cell identification and application of different biostatistical methods for hazard characterization) impact all subsequent steps. In section 3.2 it was already discussed, how changes in biostatistics impact the outcome of hazard characterization. This section takes the same principle of action and consequence, but with the entire DNT-IVB workflow as scope. More precisely, it will be explored how the changes of image analysis algorithms for cell identification impacts the data that is evaluated and thus also the hazard assessment. For illustrative purposes, the assessment of oligodendrocyte differentiation for a positive control compound is taken as an example to follow along (Fig. 9). Three scenarios are displayed for demonstration. In terms of data generation, Omnisphero originally relied on overlap algorithms to identify neurons. This principle transferred to oligodendrocyte identification is depicted in scenario A. As discussed in 3.1.1, the overlap algorithm is prone to misclassification: It misidentifies both nuclei with O4 staining artefacts and nuclei overlapping with branches from an oligodendrocyte as positives, while misidentifying cells as negatives where the marker expression value was too low for the set binarization-threshold. If the acquired neurosphere images have many artefacts and/or varying staining brightness, counted oligodendrocytes (i.e. the measured response for oligodendrocyte differentiation) per well have high abbreviations and are not in line of what a human experimenter would observe. In this example, the algorithm did clearly fail at reaching the human ground truth of manual annotation and the response readouts are fairly scattered over the test concentration range. While evaluating the data, the employed regression model does not support any effects, a BMR is not reached. This leads to a “no hit” classification and, if no other endpoint is classified as “borderline” or “specific hit”, to a false

negative call for the compound. This demonstrates how employment of suboptimal image analysis algorithms impairs the sensitivity, since toxic effects are not assessed correctly by the image analysis resulting in false negative classifications. With a weak sensitivity, the DNT-IVB is not able to reliably detect DNT effects. There are ways to retroactively counteract the high data variability caused by flawed image analysis. Different BMR levels could be chosen, for example a BMR50. This would ensure that the BMR is not measuring the fluctuations given by the suboptimal image analysis. However, a larger BMR leads to a higher BMC to be estimated (manuscript 2.2 – Keßel *et al.*, 2022, preprint). Another retrospective counteract would be the employment of replicate outlier criteria e.g. with truncated outlier filtering (Costa, 2014). This, however, is problematic for the small sample size given in the DNT-IVB and would also raise the follow-up-question on how to deal appropriately with outlier values (e.g., removing, winsorization, trimming). As a last retrospective counteract, a less conservative classification model could be used. However, all of these countermeasures would rather cure the symptom and not the cause.

As a flawed image analysis is identified as the cause of misclassification in scenario A, in scenario B a well-trained CNN model is employed for identification of oligodendrocytes. It mimics human evaluation well and the generated endpoint data shows clear toxic effects on oligodendrocyte differentiation, as the applied regression model reaches the BMR. In scenario B, bootstrapping is chosen for BMC uncertainty estimation. Compared to inverse regression, bootstrapping often results in wider CIs for the given DNT-IVB experiment design (manuscript 2.2 – Keßel *et al.*, 2022, preprint). The classification model relates to BMC uncertainty and the high uncertainty led to a “unspecific hit” classification. Consequently, the compound is possibly identified as a false negative again. A similar result would be expected, if a poor regression model is chosen and over-parametrization leads to a wider CI. In scenario C, the BMC uncertainty of the CNN-generated data is estimated by inverse regression. Inverse regression by trend results in narrower CI widths for the DNT-IVB experiment design (manuscript 2.2 – Keßel *et al.*, 2022, preprint). With less uncertainty in the data, the classification model identifies the endpoint as “specific hit” and the compound consequently as a true positive. The different outcome between scenario B and C illustrates how statistical methods and the logic of a classification model need to be established with regard to the experimental design: Most of the DNT-IVB assays from our lab are prone to higher data variability than conventional cell systems, which renders bootstrapping as a less optimal choice to get reliable hazard hit calls, as it produces very large CIs, leading up to false negatives and thus a poor sensitivity. The illustrated example only shows how differences in bioinformatical and -statistical methodology impacts sensitivity. However, the same principle can be applied to specificity as well: If cells are not identified correctly, a control compound with no toxicity can misclassified as a toxicity hit compound (e.g. if the algorithm failed to detect

neurons in higher test-concentrations and thus a reduction of neuronal differentiation is falsely identified). And if a poor statistical method is chosen (e.g. leading to a reached BMC for data that does not support hints for toxic effects in the concentration-response pattern), the classification model might identify the compound as toxic as well, thus also leading to a poor specificity. In the examples above, only the BMC uncertainty estimation step is discussed as a critical point where hit calls can depend on. But as discovered in manuscript 2.2 (Keßel *et al.*, 2022, preprint), all choices in biostatistical methodology can impact the subsequent hit calls, hazard characterization and assay performance. Because of this, the choice of data evaluation software and choice of software parameters are another concern to be taken into account for hazard identification (Jensen *et al.*, 2020; manuscript 2.2 – Keßel *et al.*, 2022, preprint). In the common practice of *in vitro* testing, these software packages are often used by non-statisticians and inexperienced experimenters often rely on the default settings of given software to perform data analysis and evaluation. It therefore should be given that the default settings are the ones that can be applied to most data scenarios for sound evaluation. I.e., the default settings should be a good compromise between robustness against different data scenarios from the abundance of different assays, while still maintaining precision in their estimations. One example for challenges deriving from software choice poses our experience with the ToxCast pipeline (tcpl, Filer *et al.*, 2017), a software package able to evaluate data from a broad variety of different assays due to flexible algorithms and options. While being a capable software tool for data evaluation that is a good fit for many *in vitro* assays, our experience has shown that the data derived from it did not match the requirements for regulatory acceptance of our assays. This is because the neurosphere assay data comes with individual requirements such as assay-specific pre-processing (Schmuck *et al.*, 2017; Manuscript 2.2 – Keßel *et al.*, 2022, preprint) or characteristics such as high fluctuation in response data (Manuscript 2.2 – Keßel *et al.*, 2022, preprint), which all need to be considered carefully during statistical data evaluation and require more options as provided by tcpl. Furthermore, tcpl does not provide a classification model, which is mandatory for hazard assessment. Thus, a sufficient data analysis with the tcpl software would require several additional algorithms, both up- and downstream, overcomplicating the entire process and making one comprehensive software tool (as presented in this work) a far more desirable solution. In the end, there is always the danger of an automatized data analysis and evaluation workflow not being prepared to deal with an unusual data set. These are scenarios that most likely can only be avoided by either analyzing each data set individually by an expert (which is very counterintuitive for high content analysis and evaluation, especially in an automated format), or implementation of flagging systems that are able to detect “problematic” data samples. This approach was taken for the

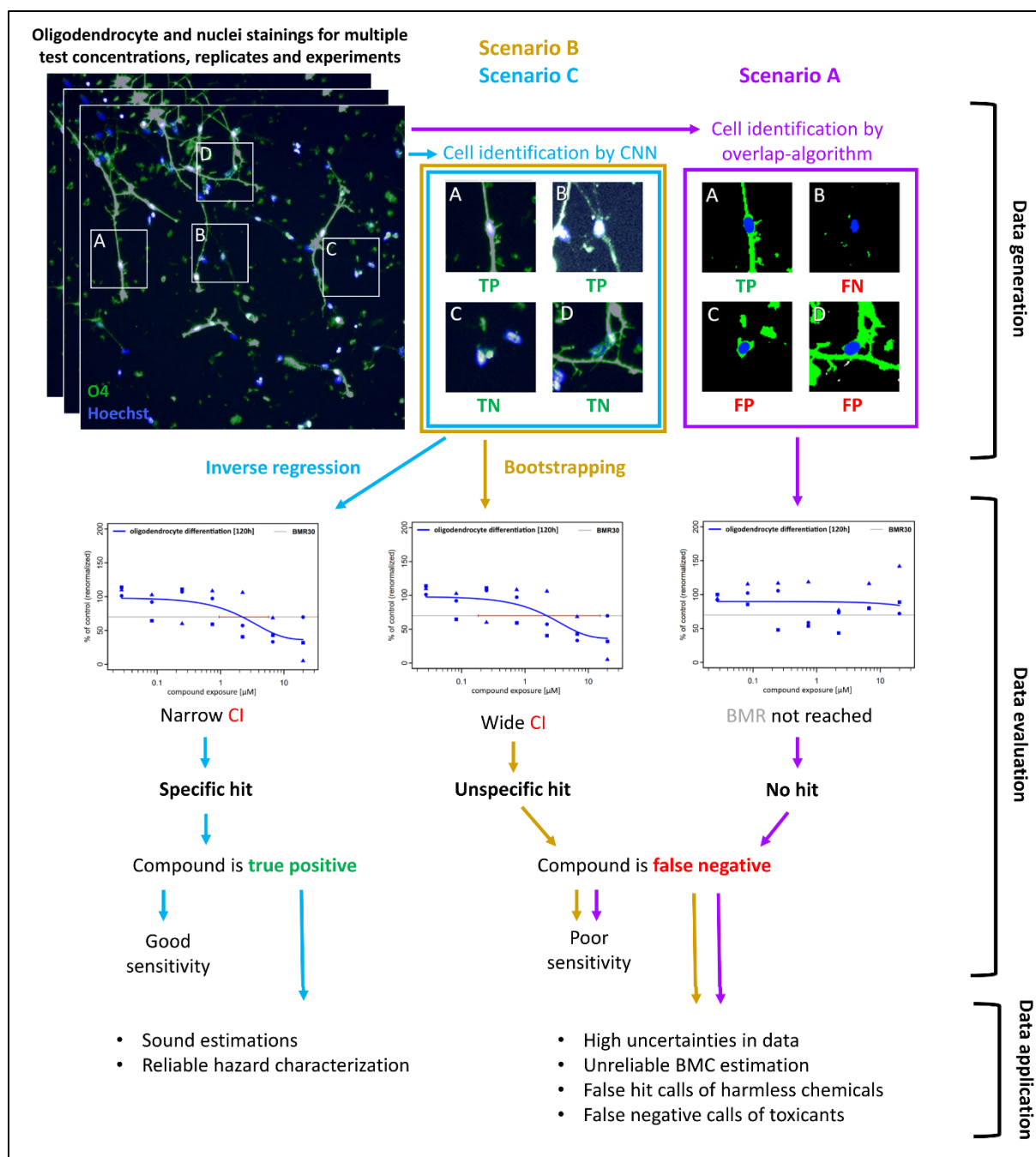


Figure 9: Consequences of change in methodology within the DNT-IVB bioinformatics workflow

Schematic overview showing the aftermath of methodology change in either the image analysis and biostatistics for hazard characterization and assessment. In an example, oligodendrocyte and nuclei stainings were exposed with different concentrations of a known DNT toxicant. The stainings are analyzed in Omnisphero for oligodendrocyte identification. In scenario A, the overlap algorithm is used for oligodendrocyte identification. Due to the proneness of the overlap algorithm to misclassifications of cells, the data has very high uncertainty. The employed regression model did not reach a BMR, thus no effect on oligodendrocyte differentiation is observed. Consequently, the endpoint was classified as a “no hit”, which in this example is a false negative for the DNT toxicant. This leads to a poor sensitivity. In this scenario, the data shows high uncertainties and can be seen as unreliable. This would raise the question if the DNT-IVB is adequate for the use for regulatory purposes. In a different scenario, oligodendrocytes are reliably identified by a well-trained CNN model. The resulting data shows clear effects, a BMC was estimated. In scenario B, bootstrapping was employed for uncertainty estimation, resulting in a wide CI. This high uncertainty of BMC leads to a “unspecific hit” classification and thus is also a false negative. Despite having sound image analysis and data support for a DNT toxic effect, employed biostatistics were chosen poorly and result in misleading statements about potential hazard. In scenario C, inverse regression is employed for BMC uncertainty estimations and leads to a narrower CI. With higher certainty in the data, the endpoint is evaluated as a “specific hit” and thus validates the chosen methodology. With enough validation, this methodology can then be employed for more compound testing and hazard characterization. Data shown in this figure were artificially created for demonstration purposes.

classification model in which an elaborate flagging system was implemented, as described in manuscript 2.2 (Keßel *et al.*, 2022, preprint). With this flagging system, problematic cases were identified and classification of these was done based on expert judgement for reliable hazard identification of such cases.

3.5 Conclusion

In this thesis, the progression of DNT data from data generation by image acquisition to application of evaluated data for regulatory purposes was observed. This progression can be divided into several individual, yet interconnected steps (generation, evaluation, application). Each step was described and thoroughly discussed, where it also was pointed out, how the steps are interconnected with each other. It has become apparent, that there is a strong dependency between them. Methodological changes in one step always affect subsequent ones – sometimes to a drastic degree. By taking the insights gained from manuscript 2.2 and 2.4 into consideration, it was clearly demonstrated, how the method of BMC estimation can decide the fate of compound hit calls and what consequences that might have for regulatory acceptance of the DNT-IVB. For example, a compound's toxicity can either be not detected or falsely identified, if the methods are chosen poorly.

With constant improvement of existing and development of novel technologies, conventional approaches can become redundant and sometimes even outright negligent to be kept in use (Ching *et al.*, 2017; Judson *et al.* 2017; Jensen *et al.*, 2019; Villeneuve *et al.*, 2019). The change of technology for cell type identification showed how much improvement can be achieved by application of a novel technology, where cells were identified with significantly higher precision. Reliable generation of data is a mandatory basis, since the data evaluation that follows the generation happens under the assumption that given data reliably reflects the biological phenomena. As for the evaluation, there is no simple way to tell which methods are the best for concentration-response data evaluation in general. Rather, there is a strong need to choose the methods which fit the given data structure/experimental design best. Otherwise, misinterpretation of data is an inherent danger. Comparison of performance, measured by control compounds, can be a feasible way to quantify the adequacy of different approaches for their purpose. Furthermore, it has become clear that a hazard classification method with focus on endpoint relationships is essential for a reliable identification of hazard alerts. DNT-specific endpoints should always take general cell health into account to enable distinction between general and DNT-specific effects. With precisely generated data reflecting biological phenomena well and biostatistical approaches being fit-for-purpose, reliable assumptions and predictions can be made on the basis of that data.

Discussion

Although this study was conducted with data from the DNT-IVB, we assume many of the conclusions can be generalized to data from other assays and even fields in life science. It demonstrates how novel technology can better reflect biological phenomena, that statistical decisions which seem to be of minor importance can become impactful, how the precision of different approaches can be quantified and how the data can finally be used for regulatory decision making. At this point, the DNT-IVB testing cannot replace the use of the OECD TG426 for hazard-based decisions. Yet, it has the potential to be a powerful tool for regulatory needs using an IATA framework.

4 Summary

Neurodevelopmental toxicants can affect early brain development and therefore present a long-underestimated health risk to our society. Conventional *in vivo* developmental neurotoxicity (DNT) testing methods are very resource- and time-intensive and were only performed for a limited amount of chemicals. This leaves a data gap concerning the DNT potential of most chemicals. In general, there is consensus that more chemicals need to be tested for their potential to induce DNT. A promising approach is the use of new approach methods (NAMs), set up in a DNT *in vitro* battery (IVB) that can evaluate chemical effects on major neurodevelopmental key events and overcome several limitations of *in vivo* testing. Neural progenitor cells (NPCs) cultivated as 3D neurospheres are one promising NAM used in the current DNT-IVB, since they mimic key processes of brain development such as cell proliferation, migration and differentiation in a 3D context.

To extract relevant and reliable information on the DNT of many chemicals from 3D neurospheres, an automatized workflow containing bioinformatic and -statistical medium-throughput pipelines was developed. This allows image analysis for cell biological endpoints and facilitates a biostatistical data analysis for DNT hazard classification of chemicals in a regulatory context. In this thesis, the process from generation to evaluation and finally application of *in vitro* DNT testing data is explored. It is furthermore demonstrated how the application of different data analysis methods affects the final DNT hazard classification of a chemical.

To generate endpoint data, the high content image analysis software 'Omnisphero' was developed previously. Omnisphero uses fluorescence-based images acquired with a high content imaging device for quantification of cell type-specific endpoints such as migration or neuronal and oligodendrocyte differentiation. Originally, Omnisphero relied on overlap-algorithms for cell-type identification. However, these algorithms did not meet the cell type identification accuracy required for regulatory application. As part of this thesis machine learning (ML) approaches were developed, which strongly outperform the overlap algorithm in terms of accuracy and flexibility. The endpoint data deriving from image analysis need to be further analyzed and evaluated, to enable DNT classification of chemicals. For this purpose, in this thesis a biostatistical software tool was developed in R, which transforms data from different assays into a uniform format and applies several statistical methods relevant for final data interpretation. For this evaluation, a variety of biostatistical approaches are employed, which are all interconnected. The choice of which methods to employ has been shown to be impactful for the final hazard classifications. It thus became a necessity to carefully evaluate a multitude of different biostatistical approaches with regard to their application in DNT hazard identification. Depending on which approach is employed, the data evaluation accuracy, measured by expected behavior of control chemicals, varied between 77.3% and 88.6%. Statements on DNT deriving from the data evaluation methods developed in this thesis can subsequently be used in combination with other data. Examples are the discovery of a potential DNT hazard, prioritization of compound testing or integration into the Adverse Outcome Pathway concept.

In summary, significant progress was made in both development and application of DNT NAM approaches. Attention was raised on how important the choice of bioinformatic and -statistical methodology can be for DNT classification of chemicals, as well as how mandatory careful selection and validation of these methods is to gain reliable information.

5 Zusammenfassung

Entwicklungsneurotoxische Chemikalien können die frühe Gehirnentwicklung in utero beeinträchtigen und stellen daher ein Gesundheitsrisiko für unsere Gesellschaft dar. Herkömmliche *in vivo* Methoden zur Testung von Entwicklungsneurotoxizität (developmental neurotoxicity, DNT) sind sehr ressourcen- und zeitintensiv und wurden nur für eine begrenzte Anzahl von Chemikalien durchgeführt. Es besteht internationaler Konsens darüber, dass mehr Chemikalien auf ihr DNT-Potenzial getestet werden müssen. Ein vielversprechender Ansatz ist die Verwendung neuartiger Test-Methoden (NAMs) im Rahmen einer DNT *in vitro* Batterie (IVB), mit welcher die Auswirkungen von Chemikalien auf wichtige Schlüsselereignisse der Gehirnentwicklung bewertet und mehrere Einschränkungen von *in vivo* Tests überwunden werden können. Neuronale Vorläuferzellen (NPCs), welche als 3D-Neurosphären kultiviert werden, sind vielversprechende NAMs, da sie Schlüsselprozesse der Gehirnentwicklung wie Zelldifferenzierung und Migration in einem 3D-Kontext nachahmen.

Um relevante und zuverlässige Informationen über das DNT Potential vieler Chemikalien aus 3D-Neurosphären zu extrahieren, wurde ein automatisierter bioinformatischer und biostatistischer Workflow entwickelt. Dieser Workflow ermöglicht eine Bildanalyse für zellbiologische Endpunkte sowie biostatistische Datenanalyse für die DNT-Gefährdungsklassifizierung von Chemikalien in einem regulatorischen Kontext. In dieser Arbeit wird der Prozess von der Generierung über die Auswertung bis hin zur Anwendung von *in vitro* DNT-Testdaten untersucht. Darüber hinaus wird gezeigt, wie sich die Anwendung verschiedener Datenanalysemethoden auf die endgültige DNT-Gefahrenklassifizierung von Chemikalien auswirkt.

Zur Generierung von Endpunktdaten wurde vormals die Bildanalysesoftware „Omnisphero“ entwickelt. Omnisphero verwendet fluoreszenzbasierte Bilder, um zelltypspezifische Endpunkte wie Migration oder Differenzierung von Neuronen und Oligodendrozyten zu quantifizieren. Ursprünglich stützte sich Omnisphero zur Identifizierung von Zelltypen auf Überlappungsalgorithmen. Diese Algorithmen erreichten jedoch nicht die für regulatorische Anwendungen erforderliche Genauigkeit. Im Rahmen dieser Arbeit wurden mithilfe des maschinellen Lernens (ML) neue Ansätze entwickelt, welche die Überlappungsalgorithmen in Bezug auf Genauigkeit und Flexibilität deutlich übertreffen. Die so gewonnenen Daten zur Zellidentifizierung müssen weiter prozessiert werden, um eine DNT-Klassifizierung von Chemikalien zu ermöglichen. Zu diesem Zweck wurde eine biostatistische Software in R entwickelt, welche Daten aus verschiedenen Assays in ein einheitliches Format umwandelt und mehrere biostatistische Methoden anwendet, die für die endgültige Datenauswertung relevant sind. Die Methoden sind dabei alle miteinander verknüpft. Es hat sich gezeigt, dass die Wahl der anzuwendenden Methoden einen Einfluss auf die endgültigen Gefahrenklassifizierungen hat. Daher wurde es notwendig, eine Vielzahl verschiedener biostatistischer Ansätze im Hinblick auf ihre Anwendung bei der Identifizierung von DNT-Gefahren sorgfältig zu bewerten. Die Genauigkeit der Datenauswertung, gemessen am erwarteten Verhalten von Kontrollchemikalien, lag je nach Ansatz zwischen 77,3 % und 88,6 %. Aussagen zu ENT, welche sich aus den in dieser Arbeit entwickelten Datenauswertungsmethoden ergeben, können anschließend in Kombination mit anderen Daten verwendet werden. Beispiele hierfür sind die Entdeckung eines DNT-Gefährdungspotentials, die Priorisierung von Substanztests oder die Integration in das *Adverse Outcome Pathway* Konzept.

Zusammenfassend kann gesagt werden, dass sowohl bei der Entwicklung als auch bei der Anwendung von DNT NAM Ansätzen erhebliche Fortschritte erzielt wurden. Es konnte gezeigt werden, wie wichtig die Wahl der bioinformatischen und -statistischen Methoden für die DNT-Klassifizierung von Chemikalien sein kann und wie wichtig eine sorgfältige Auswahl und Validierung dieser Methoden ist, um zuverlässige Informationen zu erhalten.

List of abbreviations

2D	Two-dimensional
3D	Three-dimensional
AOP	Adverse outcome pathway
AUC	Area-under-curve
BMC	Benchmark concentration
BMCL	Lower benchmark response confidence limit
BMCU	Upper benchmark response confidence limit
BMR	Benchmark response
CI	Confidence interval
CNN	Convolutional Neural Network
DNT	Developmental neurotoxicity
EC	Effective Concentration
EFSA	European Food Safety Authority
ENT	Entwicklungsneurotoxizität
EPA	Environmental Protection Agency
FN	False negative
FP	False positive
FR	Flame retardants
HCI	High content imaging
HCIA	High content image analyses
HTS	High throughput screening
IATA	Integrated Approaches to Testing and Assessment
IVB	IVB in vitro battery
KNDP	Key neurodevelopmental process
LC	Lethal Concentration
MIE	Molecular initiating event
ML	Machine learning
MOA	Mode of action
NAM	New approach method
NPC	Neural progenitor cell
NPC	Neural progenitor cells
NRC	National research council
OECD	Organisation for Economic Co-Operation and Development
QSAR	Quantitative structure–activity relationship
TN	True negative
TP	True positive
UKN	University of Konstanz

References

Alépée N, Bahinski A, Daneshian M, De Wever B, Fritsche E, Goldberg A, Hansmann J, Hartung T, Haycock J, Hogberg H, Hoelting L, Kelm JM, Kadereit S, McVey E, Landsiedel R, Leist M, Lübberstedt M, Noor F, Pellevoisin C, Petersohn D, Pfannenbecker U, Reisinger K, Ramirez T, Rothen-Rutishauser B, Schäfer-Korting M, Zeilinger K, Zurich MG. State-of-the-art of 3D cultures (organs-on-a-chip) in safety testing and pathophysiology. *ALTEX*. 2014;31(4):441-77. doi: 10.14573/altex.1406111. Epub 2014 Jul 14. PMID: 25027500; PMCID: PMC4783151.

Al-Kofahi K.A. *et al.*, "Rapid automated three-dimensional tracing of neurons from confocal image stacks," in *IEEE Transactions on Information Technology in Biomedicine*, vol. 6, no. 2, pp. 171-187, June 2002, doi: 10.1109/TITB.2002.1006304.

Bai X., L. J. Latecki and W. -y. Liu, "Skeleton Pruning by Contour Partitioning with Discrete Curve Evolution," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 3, pp. 449-462, March 2007, doi: 10.1109/TPAMI.2007.59.

Bal-Price A, Crofton KM, Sachana M, Shafer TJ, Behl M, Forsby A, Hargreaves A, Landesmann B, Lein PJ, Louisse J, Monnet-Tschudi F, Paini A, Rolaki A, Schratzenholz A, Suñol C, van Thriel C, Whelan M & Fritsche E (2015) Putative adverse outcome pathways relevant to neurotoxicity, *Critical Reviews in Toxicology*, 45:1, 83-91, DOI: 10.3109/10408444.2014.981331

Bal-Price A, Hogberg HT, Crofton KM, Daneshian M, FitzGerald RE, Fritsche E, Heinonen T, Hougaard Bennekou S, Klima S, Piersma AH, Sachana M, Shafer TJ, Terron A, Monnet-Tschudi F, Viviani B, Waldmann T, Westerink RHS, Wilks MF, Witters H, Zurich MG, Leist M. Recommendation on test readiness criteria for new approach methods in toxicology: Exemplified for developmental neurotoxicity. *ALTEX*. 2018;35(3):306-352. doi: 10.14573/altex.1712081. Epub 2018 Feb 23. Erratum in: *ALTEX*. 2019;36(3):506. PMID: 29485663; PMCID: PMC6545888.

Bas G. H. Bokkers, Wout Slob, A Comparison of Ratio Distributions Based on the NOAEL and the Benchmark Approach for Subchronic-to-Chronic Extrapolation, *Toxicological Sciences*, Volume 85, Issue 2, June 2005, Pages 1033–1040, <https://doi.org/10.1093/toxsci/kfi144>

Behl M, Hsieh JH, Shafer TJ, Mundy WR, Rice JR, Boyd WA, Freedman JH, Hunter ES, Jarema KA, Padilla S, Tice RR, Use of alternative assays to identify and prioritize organophosphorus flame retardants for potential developmental and neurotoxicity, *Neurotoxicology and Teratology*, Volume 52, Part B, 2015, Pages 181-193, ISSN 0892-0362, <https://doi.org/10.1016/j.ntt.2015.09.003>.

Bennett M, Gilroy DW. Lipid Mediators in Inflammation. *Microbiol Spectr*. 2016; 4(6):10.1128/microbiolspec.MCHD-0035-2016. doi:10.1128/microbiolspec.MCHD-0035-2016

Blum J, Masjosthusmann S, Bartmann K, Bendt F, Dolde X, Dönmez A, Förster N, Holzer AK, Hübenthal U, Keßel HE, Kilic S, Klose J, Pahl M, Stürzl LC, Mangas I, Terron A, Crofton KM, Scholze M, Mosig A, Leist M, Fritsche E, Establishment of a human cell-based in vitro battery to assess developmental neurotoxicity hazard of chemicals, *Chemosphere*, Volume 311, Part 2, 2023, 137035, ISSN 0045-6535, <https://doi.org/10.1016/j.chemosphere.2022.137035>.

References

Buckley, B.E., Piegorsch, W.W., West, R.W. Confidence limits on one-stage model parameters in benchmark risk assessment. *Environ Ecol Stat* 16, 53–62 (2009). <https://doi.org/10.1007/s10651-007-0076-2>

Carusi A, Davies MR, Grandis GD, Escher BI, Hodges G, Leung KMY, Whelan M, Willett C, Ankley GT, Harvesting the promise of AOPs: An assessment and recommendations, *Science of The Total Environment*, Volumes 628–629, 2018, Pages 1542-1556, ISSN 0048-9697, <https://doi.org/10.1016/j.scitotenv.2018.02.015>.

Ching T, Himmelstein DS, Beaulieu-Jones BK, et al. Opportunities and obstacles for deep learning in biology and medicine. *J R Soc Interface*. 2018;15(141):20170387. doi:10.1098/rsif.2017.0387

Choi RY, Coyner AS, Kalpathy-Cramer J, Chiang MF, Campbell JP. Introduction to machine learning, neural networks, and deep learning. *Trans Vis Sci Tech*. 2020;9(2):14, <https://doi.org/10.1167/tvst.9.2.14>

Collins FS, Gray GM, Bucher JR. 2008. Toxicology. Transforming environmental health protection. *Science* 319:906-907; doi:10.1126/science.1154619.

Costa PJ, 2014. Truncated Outlier Filtering, *Journal of Biopharmaceutical Statistics*, 24:5, 1115-1129, DOI: 10.1080/10543406.2014.926366

Crofton KM, Mundy WR, Shafer TJ. 2012. Developmental neurotoxicity testing: A path forward. *Congenit. Anom. (Kyoto)*. 52:140-146; doi:10.1111/j.1741-4520.2012.00377.x.

Crofton, K.M., Mundy, W., R, (2021). External Scientific Report on the Interpretation of Data from the Developmental Neurotoxicity In Vitro Testing Assays for Use in Integrated Approaches for Testing and Assessment. *EFSA supporting publication*; 18(10):EN-6924. 42 pp. doi:10.2903/sp.efsa.2021.EN-6924

Davis A. J., Jeffrey S. Gift, Q. Jay Zhao, Introduction to benchmark dose methods and U.S. EPA's benchmark dose software (BMDS) version 2.1.1, *Toxicology and Applied Pharmacology*, Volume 254, Issue 2, 2011, Pages 181-191, ISSN 0041-008X, <https://doi.org/10.1016/j.taap.2010.10.016>.

Delp J, Gutbier S, Klima S, Hoelting L, Pinto-Gil K, Hsieh JH, Aiche M, Klein K, Schreiber F, Leist M, 2018. *A high-throughput approach to identify specific neurotoxicants/ developmental toxicants in human neuronal cell function assays*. In: *Alternatives to Animal Experimentation : ALTEX*. **35**(2), pp. 235-253. ISSN 1868-596X. eISSN 1868-8551. Available under: doi: 10.14573/altex.1712182

Dhungel N, Carneiro G, Bradley AP. 2015 Deep learning and structured prediction for the segmentation of mass in mammograms. In 18th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI), Munich, Germany, October. *Lecture Notes in Computer Science*, vol. 9349. Cham, Switzerland: Springer.

EFSA PPR Panel (EFSA Panel on Plant Protection Products and their Residues), Hernández-Jerez, A, Adriaanse P, Aldrich A, Berny P, Coja T, Duquesne S, Focks A, Marinovich M, Millet M, Pelkonen O, Pieper S, Tiktak A, Topping C, Widenfalk A, Wilks M, Wolterink G, Crofton K, Hougaard Bennekou S, Paparella M and Tzoulaki I, 2021. Scientific Opinion on the development of Integrated Approaches to Testing and Assessment (IATA) case studies on developmental neurotoxicity (DNT) risk assessment. *EFSA Journal* 2021;19(6):6599, 63pp. <https://doi.org/10.2903/j.efsa.2021.6599>

EFSA Scientific Committee, Hardy, A., Benford, D., Halldorsson, T., Jeger, M.J., Knutsen, K.H., More, S., Mortensen, A., Naegeli, H., Noteborn, H., Ockleford, C., Ricci, A., Rychen, G., Silano, V., Solecki, R.,

References

- Turck, D., Aerts, M., Bodin, L., Davis, A., Edler, L., Gundert-Remy, U., Sand, S., Slob, W., Bottex, B., Abrahantes, J.C., Marques, D.C., Kass, G. and Schlatter, J.R., 2017. Update: Guidance on the use of the benchmark dose approach in risk assessment. *EFSA Journal* 2017;15(1):4658, 41 pp. <https://doi.org/10.2903/j.efsa.2017.4658>.
- Escher, SE, Partosch, F, Konzok, S, Jennings, P, Luijten, M, Kienhuis, A, de Leeuw, V, Reuss, R, Lindemann, K-M, Hougaard Bennekou, S, 2022. Development of a Roadmap for Action on New Approach Methodologies in Risk Assessment. 19(6): 153 pp. doi:10.2903/sp.efsa.2022.EN-7341
- Fang, Q., Piegorsch, W. W., and Barnes, K. Y. (2015) Bayesian benchmark dose analysis. *Environmetrics*, 26: 373– 382. doi: 10.1002/env.2339.
- Fischer I, Nickel AC, Qin N, *et al.* Different Calculation Strategies Are Congruent in Determining Chemotherapy Resistance of Brain Tumors In Vitro. *Cells*. 2020;9(12):2689. Published 2020 Dec 15. doi:10.3390/cells9122689
- Förster, N., Butke, J., Keßel, H.E., Bendt, F., Pahl, M., Li, L., *et al.* Reliable identification and quantification of neural cells in microscopic images of neurospheres. *Cytometry*. 2022; 101: 411– 422 <https://doi.org/10.1002/cyto.a.24514>
- Fourches D, Sassano MF, Roth BL, Tropsha A. HTS navigator: freely accessible cheminformatics software for analyzing high-throughput screening data. *Bioinformatics*. 2014;30(4):588-589. doi:10.1093/bioinformatics/btt718
- Frommolt P, Thomas RK. Standardized high-throughput evaluation of cell-based compound screens. *BMC Bioinformatics*. 2008;9:475. Published 2008 Nov 12. doi:10.1186/1471-2105-9-475
- Grandjean P, Landrigan PJ. 2006. Developmental neurotoxicity of industrial chemicals. *Lancet* 368:2167-2178; doi:10.1016/S0140-6736(06)69665-7.
- Grandjean P, Landrigan PJ. 2014. Neurobehavioural effects of developmental toxicity. *Lancet Neurol*. 13:330-338; doi:10.1016/S1474-4422(13)70278-3.
- Hoelting L, Klima S, Karreman C, Grinberg M, Meisig J, Henry M, Rotshteyn T, Rahnenführer J, Blüthgen N, Sachinidis A, Waldmann T, Leist M, Stem Cell-Derived Immature Human Dorsal Root Ganglia Neurons to Identify Peripheral Neurotoxicants, *Stem Cells Translational Medicine*, Volume 5, Issue 4, April 2016, Pages 476–487, <https://doi.org/10.5966/sctm.2015-0108>
- Holzer A-K, Suci I, Karreman C, Goj T, Leist M. Specific Attenuation of Purinergic Signaling during Bortezomib-Induced Peripheral Neuropathy In Vitro. *International Journal of Molecular Sciences*. 2022; 23(7):3734. <https://doi.org/10.3390/ijms23073734>
- Jensen, SM., Kluxen, FM., Ritz, C. A Review of Recent Advances in Benchmark Dose Methodology. *Risk Anal*. 2019 Oct;39(10):2295-2315. doi: <https://doi.org/10.1111/risa.13324> Epub 2019 May 2. PMID: 31046141.
- Jensen, SM., Kluxen, FM., Streibig, JC., Cedergreen, N., Ritz, C. (2020). *bmd*: an R package for benchmark dose estimation. *PeerJ* 8:e10557 <https://doi.org/10.7717/peerj.10557>.

References

- Judson R.S., Houck K.A., Watt E.D., Thomas R.S. (2017). On selecting a minimal set of in vitro assays to reliably determine estrogen agonist activity, *Regulatory Toxicology and Pharmacology*, Volume 91, 2017, Pages 39-49, ISSN 0273-2300, <https://doi.org/10.1016/j.yrtph.2017.09.022>.
- Kooi T, Litjens G, van Ginneken B, Gubern-Me'rida A, Sa'nchez CI, Mann R, den Heeten A, Karssemeijer N. 2017 Large scale deep learning for computer aided detection of mammographic lesions. *Med. Image Anal.* 35, 303– 312. (doi:10.1016/j.media.2016.07.007)
- Kopp-Schneider A, Prieto P, Kinsner-Ovaskainen A, Stanzel S, Design of a testing strategy using non-animal based test methods: Lessons learnt from the ACuteTox project, *Toxicology in Vitro*, Volume 27, Issue 4, 2013, Pages 1395-1401, ISSN 0887-2333, <https://doi.org/10.1016/j.tiv.2012.08.016>.
- Krewski, D., Andersen, M.E., Tyshenko, M.G. *et al.* Toxicity testing in the 21st century: progress in the past decade and future perspectives. *Arch Toxicol* 94, 1–58 (2020). <https://doi.org/10.1007/s00204-019-02613-4>
- Leist, M., Hasiwa, N., Rovida, C., Daneshian, M., Basketter, D., Kimber, I., Clewell, H., Gocht, T., Goldberg, A., Busquet, F., Rossi, A.-M., Schwarz, M., Stephens, M., Taalman, R., Knudsen, T. B., McKim, J., Harris, G., Pamies, D. and Hartung, T. (2014) "Consensus report on the future of animal-free systemic toxicity testing", *ALTEX - Alternatives to animal experimentation*, 31(3), pp. 341–356. doi: <https://doi.org/10.14573/altex.1406091>.
- Leontaridou, M., Urbisch, D., Kolle, S. N., Ott, K., Mulliner, D. S., Gabbert, S. and Landsiedel, R. (2017) "The borderline range of toxicological methods: Quantification and implications for evaluating precision", *ALTEX - Alternatives to animal experimentation*, 34(4), pp. 525–538. doi: <https://doi.org/10.14573/altex.1606271>.
- Litjens G et al. 2017 A survey on deep learning in medical image analysis. *Med. Image Anal.* 42, 60 – 88. (doi:10.1016/j.media.2017.07.005)
- Masjosthusmann, S., Blum, J., Bartmann, K., Dolde, X., Holzer, A.-K., Stürzl, L.-C., Hagen, Keßel H. E., Förster, N., Dönmez, A., Klose, J., Pahl, M., Waldmann, T., Bendt, F., Kisitu, J., Suciu, I., Hübenthal, U., Mosig, A., Leist, M., Fritsche, E., 2020. Establishment of an a priori protocol for the implementation and interpretation of an in-vitro testing battery for the assessment of developmental neurotoxicity. *EFSA supporting publication* 2020: 17(10): EN-1938. 152 pp. doi: 10.2903/sp.efsa.2020.EN-1938
- Moors M, Rockel TD, Abel J, Cline JE, Gassmann K, Schreiber T, *et al.* 2009. Human neurospheres as three-dimensional cellular systems for developmental neurotoxicity testing. *Environ. Health Perspect.* 117:1131-8; doi:10.1289/ehp.0800207.
- NRC. 2007. Toxicity Testing in the 21st Century: A Vision and a Strategy I The National Academies Press.; doi:10.17226/11970.
- Nyffeler J, KarremanC, Leisner H, Kim YJ, Lee G, Waldmann T, Leist M, 2017. *Design of a high-throughput human neural crest cell migration assay to indicate potential developmental toxicants*. In: *Alternatives to Animal Experimentation: ALTEX*. 34(1), pp. 75-94. ISSN 0946-7785. eISSN 1868-8551. Available under: doi: 10.14573/altex.1605031

References

OECD (2006), *Current Approaches in the Statistical Analysis of Ecotoxicity Data: A guidance to application (annexes to this publication exist as a separate document)*, OECD Series on Testing and Assessment, No. 54, OECD Publishing, Paris, <https://doi.org/10.1787/9789264085275-en>.

Paparella M, Bennekou SH, Bal-Price A, An analysis of the limitations and uncertainties of in vivo developmental neurotoxicity testing and assessment to identify the potential for alternative approaches, *Reproductive Toxicology*, Volume 96, 2020, Pages 327-336, ISSN 0890-6238, <https://doi.org/10.1016/j.reprotox.2020.08.002>.

Prieto P, Kinsner-Ovaskainen A, Stanzel S, Albella B, Artursson P, Campillo N, Cecchelli R, Cerrato L, Díaz L, Di Consiglio E, Guerra A, Gombau L, Herrera G, Honegger P, Landry C, O'Connor JE, Páez JA, Quintas G, Svensson R, Turco L, Zurich MG, Zurbano MJ, Kopp-Schneider A, The value of selected in vitro and in silico methods to predict acute oral toxicity in a regulatory context: Results from the European Project ACuteTox, *Toxicology in Vitro*, Volume 27, Issue 4, 2013, Pages 1357-1376, ISSN 0887-2333, <https://doi.org/10.1016/j.tiv.2012.07.013>.

Russell WMS, Burch RL. 1959. *The Principles of Humane Experimental Technique*.

Sachana M, Bal-Price A, Crofton KM, Bennekou SH, Shafer TJ, Behl M, Terron A, International Regulatory and Scientific Effort for Improved Developmental Neurotoxicity Testing, *Toxicological Sciences*, Volume 167, Issue 1, January 2019, Pages 45–57, <https://doi.org/10.1093/toxsci/kfy211>

Schmuck, M.R., Temme, T., Dach, K., *et al.* Omnisphero: a high-content image analysis (HCIA) approach for phenotypic developmental neurotoxicity (DNT) screenings of organoid neurosphere cultures in vitro. *Arch Toxicol.* 2017;91(4):2017-2028. doi: <https://doi.org/10.1007/s00204-016-1852-2>

Shariff A, Kangas J, Coelho LP, Quinn S, Murphy RF. Automated Image Analysis for High-Content Screening and Analysis. *Journal of Biomolecular Screening*. 2010;15(7):726-734. doi:10.1177/1087057110370894

Shen D, Wu G, Suk H. 2017 Deep learning in medical image analysis. *Annu. Rev. Biomed. Eng.* 19, 221 – 248. (doi:10.1146/annurev-bioeng071516-044442)

Tsuji, R., & Crofton, K. M. (2012). Developmental neurotoxicity guideline study: issues with methodology, evaluation and regulation. *Congenital Anomalies*, 52(3), 122-128.

Tyrrell, J. A., di Tomaso, E., Fuja, D., Tong, R., Kozak, K., Jain, R. K., & Roysam, B. (2007). Robust 3-D modeling of vasculature imagery using superellipsoids. *IEEE transactions on medical imaging*, 26(2), 223-237.

USEPA Guidelines for ecological risk assessment. EPA/630/R-95/002F. Risk assessment forum U.S. Environmental Protection Agency, Washington, DC, USA (1998)

USEPA. 2000. Supplementary Guidance for Conducting Health Risk Assessment of Chemical Mixtures.

USEPA 2021. New Approach Methods Work Plan (v2). U.S. Environmental Protection Agency, Washington, DC. EPA/600/X-21/209.

Villeneuve, D. L., Crump, D., Garcia-Reyero, N., Hecker, M., Hutchinson, T. H., LaLone, C. A., ... & Whelan, M. (2014). Adverse outcome pathway (AOP) development I: strategies and principles. *Toxicological Sciences*, 142(2), 312-320.

References

Villeneuve, D.L., Coady, K., Escher, B.I., Mihaich, E., Murphy, C.A., Schlekat, T., Garcia-Revero, N. High-throughput screening and environmental risk assessment: State of the science and emerging applications. *Environ Toxicol Chem.* 2019 Jan;38(1):12-26. doi: doi.org/10.1002/etc.4315. Epub 2018 Dec 20. PMID: 30570782; PMCID: PMC6698360.

Welch ML, McIntosh C, Traverso A, et al. External validation and transfer learning of convolutional neural networks for computed tomography dental artifact classification. *Phys Med Biol.* 2020;65(3):035017. Published 2020 Feb 5. doi:10.1088/1361-6560/ab63ba

Wheeler, M.W., Park, R. M., Bailer A.J., Whittaker, C., (2015) Historical Context and Recent Advances in Exposure-Response Estimation for Deriving Occupational Exposure Limits, *Journal of Occupational and Environmental Hygiene*, 12:sup1, S7-S17, DOI: 10.1080/15459624.2015.1076934

Zheng T, Xie W, Xu L, He X, Zhang Y, You M, Yang G, Chen Y. 2017 A machine learning-based framework to identify type 2 diabetes through electronic health records. *Int. J. Med. Inform.* 97, 120–127. (doi:10.1016/j.ijmedinf.2016.09.014)

Acknowledgements

At this point, I would like to acknowledge all people, who enabled this dissertation.

I would like to thank my supervisor Prof. Dr. Ellen Fritsche for the invaluable opportunity to perform my thesis in her working group and for the productive support during the entire dissertation.

I would like to thank Dr. Stefan Masjosthusmann for his great supervision throughout the entire dissertation. I always valued his dedication to support and motivate me. It truly was a pleasure, to working with him.

I would like to thank Axel Mosig for his help with the bioinformatics and also providing me an office in the Ruhr-University Bochum throughout my entire dissertation. I would like to thank Martin Schmuck and Thomas Temme for their help with Omnisphero. I would like to thank Signe Marie Jensen for her quick help with the programming. I would like to thank Martin Scholze for giving me great support with the statistics and writing process.

I would like to very kindly thank all of my colleagues in both working groups for their support and for enabling a working atmosphere, I was feeling very comfortable in. It was always a stimulating experience to exchange during fruitful discussions. I will always fondly remember hour-long chats, kicker-sessions, after-work free time activities and the new friendships that formed over the years. Special thanks go to Nils Förster, who started the dissertation along with me. He was always a reliable and kind colleague, never hesitating to help me out. Together, we went through thick and thin.

I would like to thank my entire family for supporting me throughout the entire dissertation. Special thanks go to my mother, who provided me a home in Düsseldorf and gave me emotional support during the more stressful phases.

I would also like to thank all of my other friends for all the support, advice and stimulating discussions.

From the bottom of my heart,

thank you!

Eidesstattliche Erklärung

Hiermit versichere ich an Eides statt, dass ich die vorliegende Arbeit „*In silico* generation, evaluation and application of developmental neurotoxicity data derived from high throughput screening assays" selbständig verfasst und ausschließlich die von mir angegebenen Hilfsmittel verwendet habe. Die Dissertation wurde in der vorgelegten oder einer ähnlichen Form noch bei keiner anderen Institution eingereicht. Ich habe bisher keine erfolglosen Promotionsversuche unternommen.

I declare that I have developed and written the enclosed Thesis „*In silico* generation, evaluation and application of developmental neurotoxicity data derived from high throughput screening assays" completely by myself, and have not used sources or means without declaration in the text. Any thoughts from others or literal quotations are clearly marked. The Thesis was not used in the same or in a similar version to achieve an academic grading elsewhere.

Hagen Eike Keßel

Düsseldorf, April 2023

