Food for Thought ...

Developing Context Appropriate Toxicity Testing Approaches Using New Alternative Methods (NAMs)

Melvin E. Andersen[1], Patrick D. McMullen[1], Martin B. Phillips[1], Miyoung Yoon[2,3], Salil N. Pendse[1], Harvey J. Clewell[2,4], Jessica K. Hartman[1], Marjory Moreau[1], Richard A. Becker[5] and Rebecca A. Clewell[2,6]

[1]ScitoVation LLC, Durham, NC, USA; [2]ScitoVation LLC, Research Triangle Park, NC, USA; [3]ToxStrategies, Cary, NC, USA; [4]Ramboll, Research Triangle Park, NC, USA; [5]American Chemistry Council, Washington, DC, USA; [6]21st Century Tox Consulting, Chapel Hill, NC, USA

Abstract

In the past 10 years, the public, private, and non-profit sectors have found agreement that hazard identification and risk assessment should capitalize on the explosion of knowledge in the biological sciences, moving away from in life animal testing toward more human-relevant *in vitro* and *in silico* methods, collectively referred to as new approach methodologies (NAMs). The goals for implementation of NAMs are to efficiently identify possible chemical hazards and to gather dose-response data to inform more human-relevant safety assessment. While work proceeds to develop NAMs, there has been less emphasis on creating decision criteria or showing how risk context should guide selection and use of NAMs. Here, we outline application scenarios for NAMs in different risk contexts and place different NAMs and conventional testing approaches into four broad levels. Level 1 relies solely on computational screening; Level 2 consists of high throughput *in vitro* screening with human cells intended to provide broad coverage of possible responses; Level 3 focuses on fit-for-purpose assays selected based on presumptive modes of action (MOA) and designed to provide more quantitative estimates of relevant dose responses; Level 4 has a variety of more complex multi-dimensional or multi-cellular assays and might include targeted *in vivo* studies to further define MOA. Each level also includes decision-appropriate exposure assessment tools. Our aims here are to (1) foster discussion about context-dependent applications of NAMs in relation to risk assessment needs and (2) describe a functional roadmap to identify where NAMs are expected to be adequate for chemical safety decision-making.

1 Introduction

The National Academy of Sciences (NAS) report in 2007, "Toxicity Testing in the 21st Century: A Vision and A Strategy," proposed fundamental changes in chemical risk assessment, including moving to human cells, tissues, or cell lines, developing high-throughput methods for evaluating large numbers of chemicals more efficiently, and using various computational chemistry and bioinformatic tools for data analysis and prediction of risk (NRC, 2007). The National Center for Computational Toxicology (NCCT) at the US EPA had previously developed a plan to incorporate many of these approaches in toxicity testing, as described in "A Framework for a Computational Toxicology Research Strategy" (Kavlock et al., 2003) and after publication of the NAS report, moved forward to outline a broadly collaborative program (Collins et al., 2008) with other United States federal agencies to implement recommendations from the NAS report. A multi-stakeholder program, SEURAT (Safety Evaluation Ultimately Replacing Animal Testing)[1] that focused on implementation of non-animal methods also began in Europe. Following SEURAT, EU-ToxRisk, a large-scale project funded by the European Commission's Horizon 2020 program[2], is now driving the European research efforts on alternative testing methods.

These *in vitro* and computational technologies, together with application of existing tools to new data streams (e.g., read-across), are collectively referred to as new approach methodologies – NAMs (US EPA, 2018b). The US EPA under the new Toxic Substances Control Act (TSCA) in section 4(h) is required

[1] https://www.ifado.de/toxicology/2015/12/04/seurat-1-painting-the-future-animal-free-safety-assessment-of-chemical-substances/

[2] https://www.eu-toxrisk.eu/page/en/about-eu-toxrisk.php

Received June 26, 2019; Accepted October 2, 2019;
Epub October 12, 2019; © The Authors, 2019.

ALTEX 36(4), 523-534. doi:10.14573/altex.1906261

Correspondence: Melvin E. Andersen, PhD
ScitoVation LLC, 100 Capitola Drive Suite 106
Durham, NC 27713, USA
(mandersen@scitovation.com)

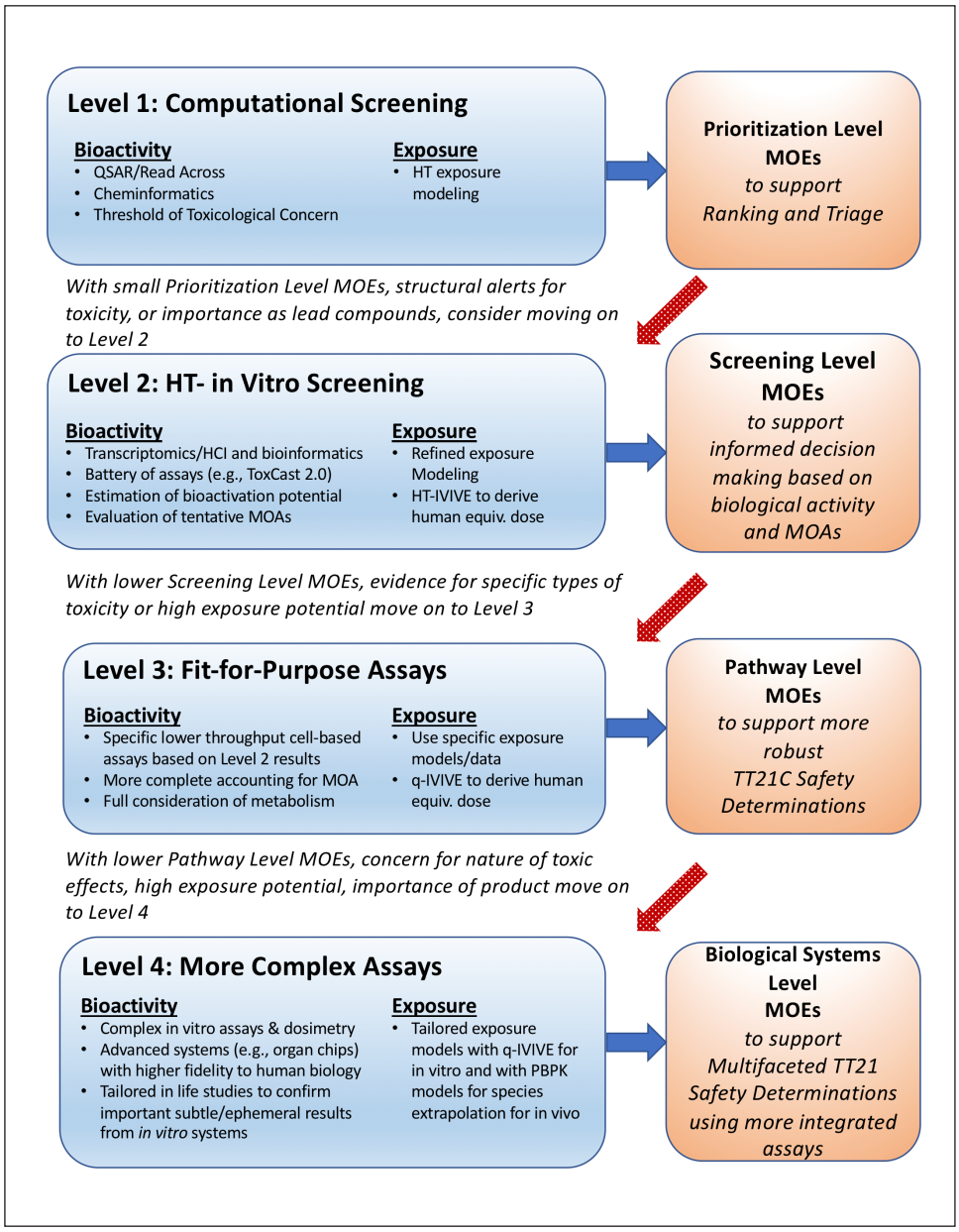

Fig. 1: A multi-level strategy for using new alternative methods and higher-throughput exposure tools for context dependent safety assessments
This tiered testing strategy is problem formulation-driven. The level of information available about chemicals will guide the particular testing required for any use conditions. The progression through different levels (orange arrows) is governed by decision context. Depending on the margin-of-exposure (MOE) estimated at each level, a decision-maker might still regard the information at any specific level to be insufficient, leading to consideration of higher-level testing to refine the analysis and have a greater degree of confidence in any decision. More detail on the NAMs at each level is in the text.

to maintain a list of acceptable NAMs. More recently, a US EPA memorandum published in September 2019 officially announced a commitment to reduce its requests for and funding of live mammal studies by 30% by 2025 and to eliminate all live mammal study requests by 2035 (Grimm, 2019).

The primary testing initiatives following the release of the 2007 NAS report focused on screening large numbers of compounds with existing high-throughput assays (e.g., ToxCast, Tox21), many of which were repurposed from pharmaceutical applications (Judson et al., 2010; Reif et al., 2010). These efforts developed the infrastructure necessary for collection and analysis of large-scale data and determining the utility of existing methods for supporting chemical safety decisions. Now, with lessons learned about the practicalities of high-throughput screening, together with significant advancements in data science and *in vitro* technologies, the toxicology and risk assessment communities are better equipped to approach the realities of NAM-based risk assessments. New assays examining a broader palette of possible responses (e.g., "omic" technologies and high-content imaging) are now being discussed, and tiered approaches are being developed for the use of these test platforms to streamline toxicity testing (Thomas et al., 2013, 2019).

One aspect emphasized in the 2007 NAS Toxicity Testing report was that collecting data on hazard should be tied to risk decision contexts. The original figure (S-1) from the report (NRC, 2007) describing components of the vision had three parts – Chemical Characterization, Toxicity Testing (including both toxicity pathway evaluations and targeted testing) and Dose-Re-

sponse and Extrapolation Modeling. These three were surrounded by a circle identified as: "Risk Contexts" and "Population and Exposure Data." In 2009, the NAS produced another report on opportunities in exposure science (NRC, 2009). The linkage between assessing toxicity and biological activity with NAMs and the use of higher-throughput methods for exposure assessment provide the basis for developing testing approaches more suited to answering diverse risk assessment questions.

2 NAMs and risk-based decisions

Converting computational approaches and *in vitro* test results to expected potency of test compounds for specific in life responses is more complex than with traditional animal tests. The more straight-forward uses of these tests are to provide indications of expected in life responses based on chemical properties or *in vitro* "hits" or to predict expected exposures. For these reasons, most recommendations for early implementation of NAMs focused on prioritization: identifying chemicals with higher potential for toxicity for more in-depth evaluation or removing chemicals entirely from further consideration. Higher priority compounds – those with some perceived hazard for specific types of adverse responses or with higher exposure potential – might be escalated to additional testing or, in some cases, to traditional animal-based methods depending on the decision context. Conversely, materials with higher perceived risks might simply be dropped from further consideration for development or removed from commercial use. It is important in developing more explicit prioritization schemes to include criteria that allow decision-making that avoids simply having lower priority compounds set aside awaiting extensive testing once the higher priority compounds move through more comprehensive testing of toxicity pathways or on to in life testing. For instance, after identifying and testing the high priority compounds, what is the strategy for moving on to those defined as lower priority? Our goal in developing these different test levels has been to minimize the need for higher tier testing through adherence to a risk context-based implementation of NAMs.

Figure 1 depicts four levels of testing that focus on use of NAMs (including both assays of biological activity and higher-throughput exposure methods) to develop information sufficient for decision-making. Each level has NAMs for assessing both biological activity and expected human exposure. Relative safety of product usage is estimated by calculating margins-of-exposure (MOEs), a ratio of a measure of expected potency divided by a measure of expected exposure in a population. This paper considers the question of when information available from any one of these levels would be considered sufficient for risk-based decisions and the kinds of decisions possible at the various levels. Light blue boxes to the left describe level-appropriate approaches for assessing bioactivity and for estimating expected exposure. The information provided by NAMs at each level allow calculation of an MOE (orange boxes to the right). Inferences about likely risks or safe usage of compounds at each of these levels underpin decisions regarding safety for specific use conditions or the need for further testing. A variety of considerations, such as the magnitude of the MOE and the accuracy, regulatory acceptance and biological coverage of the assays populating the level, would have to be considered in deciding if higher-level testing would be necessary.

3 Looking at decisions at each level

The levels are characterized, in part, by the investment in information generated by each level and by confidence in the results. Gaining confidence in their use will be essential to make NAMs acceptable by regulatory agencies. The US EPA strategic plan based on the TSCA has three important components that are (1) identifying, developing and integrating NAMs for TSCA decisions; (2) building confidence that the NAMs are scientifically reliable and relevant for TSCA decisions; and (3) implementing the reliable and relevant NAMs for TSCA decisions[3]. Approaches to establish confidence (validation) will need to be developed. Issues of confidence affect NAMs at all four of our testing levels.

Output of testing and exposure assessment at each of these levels are potentially suitable for informing different decisions. In "Level 1 – Computational screening," high-throughput predictions of exposure, putative toxicity and expected metabolites, etc., are obtained using computational methods and would support chemical categorization and decision-making in limited contexts. For instance, when choosing among several lead chemicals for any particular application, compounds predicted to have higher exposures or carrying possible toxicity liabilities as determined by computational methods could be dropped from further consideration. NAMs populating this level are purely computational and support prioritization for further study or decisions that the chemical is unsuitable for intended applications (Thomas et al., 2013).

With indications of higher expected exposures or indications of possible toxicity at Level 1, testing would be completed and compounds triaged or deprioritized. For some compounds already in commerce or those moving forward in development due to favorable use characteristics, further work would be required to refine exposure potential and determine activity in "Level 2 – High throughput (HT) *in vitro* screening." This level would be populated by rapid, high-throughput dose-response screening of compound bioactivity and high-throughput *in vitro-in vivo* extrapolation (HT-IVIVE) (Yoon et al., 2015). HT-IVIVE converts active concentrations from an *in vitro* assay (e.g., Tox21, ToxCast) to a human equivalent dose, i.e., to a human dose or human exposure that would be expected to produce concentrations in an exposed person equal to the active concentration from the *in vitro* study (Rotroff et al., 2010; Sipes et al., 2017; Casey et al., 2018; Wambaugh et al., 2018). The NAMs populating Level 2 would optimally provide quantitative measures of response,

[3] https://www.epa.gov/sites/production/files/2018-06/documents/epa_alt_strat_plan_6-20-18_clean_final.pdf

such as AC50 (a concentration causing 50% of maximal change in the assay results) or LEC (the lowest effective concentration) and, based on active concentrations in these assays, permit estimation of human equivalent doses (HEDs) (Wetmore, 2015). The MOE would be the ratio of a measure of the HED divided by expected exposure levels. A decision-maker would have more confidence in the MOEs arising from these studies than the comparisons of estimated exposure and putative risks predicted from Level 1. Nonetheless, it would be difficult to be entirely comfortable making risk assessment decisions at Level 2 for compounds for which the estimated MOE was either not sufficiently large; where the presumed MOAs inferred from these assays – e.g., MOAs - reproductive, developmental, carcinogenic potential, etc. – increased the level of concern; or where high exposures were expected in a potential target population.

The testing at Level 2 with read-outs from multiple HT *in vitro* assays or from limited broad-coverage assays should be designed to provide information on MOAs. Based on presumptions of MOAs from these assays, "Level 3 – Fit-for-purpose assays and safety assessment" would apply human-relevant fit-for-purpose (FFP) assays to provide more in-depth examinations of MOA-related cellular perturbations in cell systems (Clewell et al., 2016). FFP assays would be designed with read-outs that represent key signaling processes for cellular pathways associated with a chemical's MOA and ideally include markers that correlate with or directly measure adversity. Optimally, dose-response data from the FFP assays, together with computational pathway models (Bhattacharya et al., 2011), could provide a mechanistic understanding of the shape of the dose-response curve and support more informed extrapolation to relevant human exposure. Quantitative IVIVE (QIVIVE), accounting for human relevant metabolism coupled with dose-response relationships from the FFP assays, would provide more confidence in estimated MOEs and determination of regions of safety, i.e., exposure concentrations at which no increased risk is expected in a human. Depending on the MOE obtained with these more comprehensive FFP assays and better knowledge of use-specific exposures, a decision-maker might still deem this information to be insufficient, leading to consideration of more complex assays and more detailed compound-specific exposure information at "Level 4 – More intact systems."

The opportunity to accumulate relevant and context-dependent information at each level should substantially reduce the number of chemicals tested in these more complex NAMs. And, when necessary, alternative multi-dimensional and multi-cellular assays would assess human tissue-based dose responses rather than moving to studies in animal models. In addition, MOA information provided from HT testing and FFP assays could support more limited testing in animal models targeted to the specific MOA. It bears emphasis that, depending on the decision context, e.g. lead candidate selection, prioritization for remediation, ranking liabilities of possible substitutes, estimating risks with compounds lacking signals for endpoints of high regulatory concern and formal regulatory decision-making, etc., compounds would not have to be tested sequentially at all four levels in a tiered approach. While the emphasis here is on NAMs for human health risk-based decision-making, the principles and approaches discussed here also apply to ecological risk evaluations, recognizing that the assays, IVIVE tools and exposure evaluations would need to be tailored to support ecological risk assessment decision contexts.

Our context-dependent testing approach is applicable to a wide chemical space. However, the structure and physicochemical properties of compounds may be challenging in short-term NAMs and will determine which assays are possible and how testing should proceed. For example, modeling or prediction of compounds that are highly lipophilic and hence slowly eliminated from the body is extremely difficult. These compounds are also difficult to keep in solution for testing, as they adsorb on surfaces or form micelles. Most of the training sets for *in silico* tools were developed based on pharmaceuticals with a narrow range of physicochemical/metabolic properties, not chemicals with much broader physicochemical properties like the ones in ToxCast (Moreau et al., in preparation). Measuring metabolic rates of slowly cleared compounds is also challenging, and better approaches need to be developed with an eye toward defining a chemical space for which existing computational tools can predict metabolism with reasonable accuracy and an understanding about when these tools are inadequate for low-tier decision-making (Moreau et al., in preparation). Nonetheless, the testing and exposure assessment across these four levels should work with most compounds.

3.1 Level 1: Computational screening

We can next ask what NAMs might be involved at each of these levels. The optimal suite of computational tools in Level 1 should estimate physical properties, infer possible toxicity (e.g., QSAR platforms, threshold of toxicological concern, etc.), predict or take into account likely metabolites, and assess exposures that would be expected to arise during anticipated use conditions. The goal in Level 1 is to have computational tools that are developed with as large a range of compounds as possible in order to have confidence when calculating similar properties for a new compound or new class of compounds. The need for breadth of coverage is a challenge for model developers when data are not available to create models for specific endpoints or when data covering the domain of structural applicability is sparse. Even when data are available, they are often in need of curation, which is costly and time consuming.

Currently a variety of tools are available for estimating physicochemical properties and environmental fate endpoints: e.g., EPA's EPI Suite and ECOSAR (USEPA, 2018a), for predicting toxicological endpoints: TIMES (Mekenyan et al., 2004) and Leadscope (Roberts et al., 2000), for metabolites likely formed *in vivo*: (Leonard et al., 2018), Meteor Nexus (Marchant et al., 2008), BioTransformer (Djoumbou-Feunang et al., 2019), and ADMET Predictor®, and for both thresholds of toxicological concern (TTC) and possible exposure levels: (Patlewicz et al., 2018). With the collaborative estrogen receptor activity prediction project (CERAPP), large-scale modeling using 32,464 structures showed the possibility of screening large libraries of chemicals using a consensus of different *in silico* approaches (Mansouri et al., 2016). This approach has also been used to identify androgen active chemicals (Manganelli et al., 2019).

The computational models that apply algorithms to estimate TTCs and exposures permit estimation of approximate margins of safety – MOS (Wambaugh et al., 2013; Nicolas et al., in preparation). These TTCs are derived from *in vivo* toxicity datasets and include a 100-fold safety factor. Due to the use of the 100-fold safety factor, MOS values, i.e., the TTC divided by expected exposure, are more conservative than MOEs derived using a HED. Similar approaches have been applied to the large CERAPP dataset to calculate both TTCs and exposures, thereby providing approximate estimates of MOEs. Other tools at Level 1 include *in silico* metabolite identification (met-ID) using tools such as the OECD QSAR Toolbox metabolism profiler[4] or ACD/Labs MetaSense biotransformation map software[5]. The results of computational scrutiny of a compound or group of compounds could provide compelling results, i.e., very low predicted exposures or lack of signals for expected toxic liabilities, leading either to a much-reduced level of concern or exemption of the chemical(s) from further study.

Some combination of higher expected exposures, indications of specific types of toxicity from QSAR methods, or physicochemical properties that indicate long half-lives in a target species or the environment, would raise flags, indicating a higher priority for considering further testing or, in the case of new compound development, possibly a decision to discontinue further development. In cases where Level 1 analyses fail to provide a sufficiently large MOS, the next step would be to use Level 2 NAMs to test for biological responses in high-throughput cellular or subcellular assay platforms, first considering NAMs that target possible toxicity identified in Level 1.

3.2 Level 2: High throughput *in vitro* screening

Level 2 comprises assays that can be easily run on a large number of compounds in high-throughput mode. Examples here are the Tox21 and ToxCast assays from NIH and EPA, respectively, which could be run on essentially any compound that is soluble in water or DMSO and is not highly volatile. Ideally, Level 2 assays should be tailored toward endpoints of regulatory concern, with a good understanding of how the assay fits within known MOAs or adverse outcome pathways (AOPs). The original intentions of Tox21 and ToxCast were to generate directly comparable data for a large number of chemicals to facilitate grouping of chemicals by MOA, ranking of chemicals within a particular MOA by potency, prioritization of these chemicals for risk assessment by regulatory bodies, and ultimately providing a platform where unknowns could be subjected to the same battery of assays and MOAs assigned based on the pattern of activity seen in the results. Even though many of these goals were later shown to be out of reach, at least for the assays chosen to be part of the Phase I and Phase II efforts, these assays still show promise for grouping chemicals as part of a process known as "biological read across." This approach is similar to hazard identification in the traditional risk assessment process, where activity in Level 2 assays and/or similarity to chemicals with known hazard profiles can be used to justify a finding that further study is or is not warranted. Furthermore, in our conception of Level 2, ADME (absorption, distribution, metabolism, elimination) data would also be integrated with bioactivity assay data to convert AC50 or LEC values to HEDs using *in vitro* to *in vivo* extrapolation methods and compared with exposure estimates to generate an MOE. An important challenge in interpreting responses in Level 2 assays relates to distinguishing biologically relevant pathway responses from the "burst effect" that can arise from substances that lack specific affinity for cellular pathways and that, at relatively high concentrations, elicit broad low-affinity non-covalent interactions, trigger cell stress pathways, or cause physical disruption of proteins or membranes (Judson et al., 2016; Shah et al., 2016).

There are now wider discussions about using cell-based assays designed to broadly examine gene expression using high-throughput transcriptomic analysis (Grimm et al., 2016; McMullen et al., 2019) and to assess cellular morphology using high-content imaging (HCI) (Vantangoli et al., 2016). Gene expression platforms such as BioSpyder[6] have reduced the cost for whole genome differential gene expression analysis (DGEA) (Yeakley et al., 2017). Benchmark dose analysis (Thomas et al., 2007) and pathway visualization methods for MOA analysis, such as mode of action visualization software MoAviz (Andersen et al., 2018; McMullen et al., in press), assess both potency and biological functions/pathways affected by treatment. HCI platforms can be automated (Feng et al., 2009; Bray et al., 2016, 2017) to query a wide range of cellular phenotypes, and linking the two assay platforms DGEA and HCI could provide the necessary link between transcriptomic signatures, cellular phenotype and MOAs. The DGEA and HCI can be regarded as pathway-agnostic methods; analysis of results of the assays gives an indication of MOA rather than using inferences from Level 1 to design more MOA-targeted assays.

While there is significant enthusiasm for broad coverage assays that are not directly based on knowledge of MOAs, there is as yet no consensus on cell types or duration of exposure for these transcriptomic studies. To the extent that these second-generation assays (Thomas et al., 2019) are successful and their use to assess affected biological pathways becomes more widespread, the developed information can be merged with available databases, including those from ToxCast, and be used to develop DGE-signatures for the well-studied compounds in ToxCast Phase I and Phase II. The results of broad coverage, pathway-agnostic assays in Level 2 should ideally allow for (a) identification of AOPs/MOAs activated by a test substance and (b) IVIVE approaches to convert the active concentration to a HED. Through this process, confidence in subsequent MOE calculations will increase and some Level 2 results may be accepted for decision-making, partially due to narrowing uncertainty regarding the MOE and knowledge of likely MOAs.

[4] https://qsartoolbox.org/

[5] https://www.acdlabs.com/solutions/metasense/

[6] https://www.biospyder.com

3.3 Level 3: Fit-for-purpose assays and safety assessment

Level 3 assays would be designed so that the output from the NAMs in Level 2 and from more refined exposure assessments increase confidence in the estimated MOE and support formal risk assessment without moving on to more complex assays. These applications require FFP assays (Clewell et al., 2016). These FFP platforms are targeted cellular assays that are developed based on an understanding of human biology. The landscape of FFP assay platforms in Level 3 is diverse with many more under active development. Complex CNS tissues – so-called mini-brains – are just one example (Pamies et al., 2017; Boutin et al., 2018). Three-dimensional cultures of cells derived from various tissues are also being used to develop more relevant platforms and are increasingly integrated with HCI technologies in order to simultaneously assess multiple phenotypic endpoints – the combination of these platforms provides a form of cellular pathology (Kabadi et al., 2015).

One of the most publicized successes of the ToxCast program to date, the prediction of *in vivo* rodent uterotrophic results using *in vitro* assay data, used a computational model (Judson et al., 2015). This early success has energized efforts in the field of endocrine disruption to try to replicate this success for other, related MOAs, such as using androgen receptor data to predict *in vivo* Hershberger assay results. For estrogenic mode of action, over one hundred of the 1812 evaluated chemicals were predicted to be endocrine-active based on this computational model (Browne et al., 2015; Judson et al., 2015). These results appear robust as model results demonstrated that the method worked well for a set of reference chemicals by correctly identifying agonist, antagonist and inactive compounds with high sensitivity and specificity. Using HT-IVIVE and the results of the assays allowed estimation of pathway-altering doses. While this estrogenicity model utilized a variety of molecular and cellular endpoints, all were based on signaling through two estrogen nuclear receptors (ESR1, also referred to as ER66/ERα, and ESR2/ERβ). The type of analysis is akin to Level 2 in our scheme.

The analysis of estrogenicity using information on only the two receptors ERα and ERβ ignores other pertinent information on estrogen signaling. FFP assays for estrogenicity should be designed to account for human biology, focus on specific cellular outcomes, and assure that the cell system has the molecular pathway components necessary to recapitulate the cellular read-outs relevant to an adverse response *in vivo*. The coordination of estrogen responses in any tissue integrates the action of at least five estrogen receptors, including both classical nuclear and membrane-bound receptors: ER66, ER46, ER36, ERβ, and GPR30 (Miller et al., 2017). The goal in FFP assays for estrogenicity is to generate appropriate cellular response assays, quantitative IVIVE approaches, refined exposure information, and inferences about possible bioactive metabolites to create a package of information sufficient for MOA-based risk assessment without resorting to *in vivo* testing.

The ToxCast estrogenicity assays were conducted with cell-free and pathway-overexpression systems and using a phenotypic assay measuring proliferation in a breast cancer cell line (T47D). None of the ToxCast *in vitro* assays evaluate uterine response, even though the uterus is a critical target tissue for estrogenic compounds and there are differences in breast and uterine responses to various estrogenic compounds (Barakat, 1995). Our approach for creating a FFP assay for estrogenicity (Miller et al., 2016, 2017; Beames et al., in press) has relied on using a human adenocarcinoma cell line, i.e. Ishikawa cells, and confirming that the cells retained all components of the estrogen signaling network involved in the control of cell proliferation. To test the utility of the *in vitro* model to predict quantitative dose-response relationships in the species of interest, we tested the assay output for endogenous estrogen and known human uterotrophic drugs against clinical and epidemiological data. The FFP *in vitro* uterine assay consistently predicted chemical concentrations associated with human estrogenicity. Further, the assay predicted activity at lower concentrations than any of the ToxCast HT assays (Miller et al., 2016).

Based on these studies, we had confidence that the assay was sufficiently sensitive to predict safe levels of human exposure for uterotrophic compounds. To test the application of the assay for the broader universe of environmental compounds, we ran dose-response curves for 116 chemicals (Beames et al., in press), including chemicals that were determined to be estrogenic (n = 106) or non-estrogenic (n = 10) in the EPA estrogen model (Browne et al., 2015; Judson et al., 2015), and possible metabolites of 5 parent compounds from the ToxCast library. The Ishikawa assay was compared to ToxCast assay results, as well as *in vivo* rodent uterotrophic and two-generation reproductive study data. Active concentrations in the uterine proliferation assay were consistently among the lowest of the test models, whether comparing *in vivo* or *in vitro* results, indicating that observed activity in the *in vitro* model would provide a sufficiently protective point of departure for risk assessment. However, when compared to animal studies, approximately 41% of the compounds that caused uterotrophic response in guideline-like rodent studies did not show proliferative activity in the human cell-based assay.

This disconnect between the human *in vitro* assay and the rodent *in vivo* assay highlights an important issue in validation of NAMs, i.e. selection of the benchmark or comparator. When comparing *in vivo* studies in rats with assays in human-relevant cells, challenges arise related to decisions about defining compounds as positive or negative. Should animal studies be considered the "gold standard" when we know that many substances cause an effect in rats but not in humans, or *vice versa*? Or should we not even attempt to "predict" rodent responses and benchmark NAM predictivity against human data only? This is an open question regarding validation, and one that will (at least for the time being) necessarily be addressed on a case-by-case basis. Here, the FFP estrogen model faithfully reproduced human response for the admittedly few compounds for which clinical data was available (n = 4), but only showed activity for just over half of the rodent uterotrophic compounds. While we neither have nor ever expect to have human data with most chemicals, it nonetheless bears emphasis that the responses in test animals *in vivo* frequently differ from those in human populations, and animal tox-

icity results should not necessarily be considered the gold standard for comparison (Blaauboer and Andersen, 2007).

Other examples of FFP assays have been pursued with p53 mediated DNA damage (Clewell et al., 2014, 2016; Adeleye et al., 2015; Clewell and Andersen, 2016), PPARα signaling (McMullen et al., 2014, 2019) and adipocyte differentiation (Foley et al., 2017; Hartman et al., 2018). Our work with these FFP assays has helped establish criteria for the cellular read-outs to ensure applicability of the results for assessing adversity. These assays can be particularly informative of human relevance. For example, with the PPARα assay, criteria have been established for comparing results from human cells to *in vivo* outcomes in rodents that appear to be qualitatively different from responses expected in humans (McMullen et al., in press). Another opportunity arising from development of FFP assays is the possibility of examining the signaling networks controlling various cellular responses to develop computational systems biology modeling tools to assess the biological basis of cellular dose-response behaviors, including a better understanding of threshold behaviors at the cellular and organism level (Zhang et al., 2014, 2015; Tyson and Novák, 2015; Clewell and Andersen, 2016).

3.4 Level 4: More intact systems

Ultimately, the goal of defining multiple levels for context-appropriate testing is that over time the tools used for targeted *in vivo* studies in test animals will be regarded as studies of last resort. Problems arising from high dose animal studies in relation to human relevance and kinetic non-linearities at high doses are well-documented and these high dose rodent studies frequently raise more issues than they resolve. Instead of simply moving to *in vivo* studies, over time Level 4 should become populated with more complex assays, including multi-cellular and multi-dimensional assays, human-on-a-chip (Zhang and Radisic, 2017; Zhang et al., 2018), linked tissue surrogates with provisions for liver metabolism and inter-tissue circulation of metabolites, and inclusion of metabolite generating cells or subcellular fractions within the assay platforms (Zhang and Radisic, 2017; Zhang et al., 2018), providing a variety of biologically inspired test systems for conducting more integrated toxicity testing (Marx et al., 2016).

4 Dosimetry, extrapolation and MOE considerations

While the preceding description of the risk context-related levels focused more on assessing the biological targets, MOAs and dose-response, each level also requires consideration of dosimetry, IVIVE and exposure assessment in order to estimate MOSs or MOEs to place results from NAMs in an appropriate risk/safety context (NRC, 2007).

At Level 1, computational methods permit HT predictions of exposure and metabolism including estimation of intrinsic clearance (CL_{int}) and unbound fraction (F_u) based on chemical structure. These metabolism prediction tools are currently best suited to predict parent chemical clearance, and the domain of applicability is centered in the pharmaceutical compound space. However, recent efforts are testing their suitability for use as first tier metabolism predictions of environmental compounds (Casey et al., 2018; Moreau et al., in preparation). A proof of concept study was completed that integrated TTC values with HT exposure modeling to provide prioritization level MOEs for close to 7000 substances (Patlewicz et al., 2018). More recently, TTC values derived for approximately 40,000 substances (Nicolas et al., in preparation) have been disseminated publicly as a searchable table on the internet[7]. Advances in HT exposure modeling have now yielded median human intake rates and credible upper bound intervals for more than 450,000 chemicals in various U.S. population demographic age groups (Ring et al., 2019). As one component of the priority setting in the US EPA's projected approach for conducting risk-based prioritization of existing chemicals under TSCA, the Agency intends to use HT exposure modeling and TTC values to calculate TTC-to-exposure ratios (US EPA, 2018c).

MOE calculations for Level 2 screening assay results with consideration of exposure began with simple kinetic models, assuming steady-state oral exposure and determination of HEDs (Wetmore et al., 2013). IVIVE methods have also been used with AC50 values from ToxCast estrogenic assays to generate HEDs and MOEs by comparing these HEDs to predicted human exposures. In this way, these ToxCast estrogen assay-derived MOEs could be used as stand-alone risk-based screening values or compared to the MOE of the ubiquitous dietary phytoestrogen to provide additional context (Becker et al., 2015).

We recently proposed an alternative dosimetry measure for fruit and vegetable mixtures (Wetmore et al., 2019). The dose measure was related to daily intake of the juices in relation to their bioactivity in the BioMap® assay platform. This measure of activity was then compared with the equivalent adjusted daily intake of agrichemical residues found in these produce materials in relation to their potency. While this measure of dose does not account for pharmacokinetics of the juices, which are complex mixtures, the adjusted daily intake allows comparison of the degree of assay activity expected from the produce and the agrichemicals. The contribution from most of the produce juices was more than 1000-fold greater than the contribution of bioactivity associated with agrichemicals used in growing this produce. This examination of fruit and vegetable juices falls into Level 2 testing with mixtures using a total intake dosimeter. More extensive examination of mixture kinetics could follow with identification of major components or fractionation into different chemical subclasses that could be studied individually. Depending on the test materials, especially for mixtures and chemical substances of unknown or variable composition (e.g., biological products, herbal medicines and dietary supplements, foods), dose measures other than HEDs will need to be considered and evaluated.

Whether dealing with mixtures with known constituents or single chemicals, available computational tools can predict likely metabolites (met-ID) and infer possible toxicity of test com-

[7] https://scitovation.shinyapps.io/TTCApplet/

pounds (QSAR). These predicted values could, in theory, be used as an early estimate of bioactivity and MOE, though metabolite prediction software is presently more qualitative than quantitative. Level 2 would include measurements of CL_{int} and F_u in HT assays to estimate steady-state concentrations expected from continuous daily exposures. The ratio of the HED and actual human exposure provides the MOE at Level 2. Currently, the HTS assay systems focus on clearance of the parent chemical, assuming that metabolism is an inactivating step for the chemical. This assumption provides a first-order estimate of risk based on parent chemical but leaves bioactivation via metabolism unaddressed. Efforts are currently underway to address this gap by incorporation of metabolism into HTS screens, through the addition of hepatocytes, cellular fractions (S9) or recombinant enzymes (DeGroot et al., 2018). Unfortunately, the broader testing community rarely accounts for activation *via* metabolism in HTS evaluation.

The role of metabolism in toxicity will more likely be addressed in Level 3, where metabolic competence can be incorporated into the FFP assay designs. FFP assays conducted in the absence of components that ensure production of metabolites allow for the assessment of the bioactivity of the test compound itself, although its activity in an organism would depend on metabolism and bioavailability. With FFP assays at Level 3 it becomes particularly important to account for metabolism, either by incorporating metabolically competent preparations into the FFP-assay or by procuring potential metabolites and testing them (Beames et al., 2019).

For Level 3 studies to be regarded as sufficient for risk assessment, it may be necessary to estimate HEDs for more diverse exposure conditions and for multiple routes of exposure. By combining computational approaches for metabolism and pharmacokinetics (IVIVE, PBPK) with *in vitro* readouts for the suite of metabolites expected in the blood for a given exposure, it should be possible with more advanced kinetic models to develop a combined estimate of potency that is predictive of *in vivo* experience for oral, dermal and inhalation exposures and for multiple compounds. An example of assessment of parent compounds and active metabolites was completed in a case study looking at combined exposure to the multiple blood metabolites expected from exposures to both diethylhexyl phthalate and dibutyl phthalate. Here, *in vitro* assays evaluated potency of both parent phthalates and active metabolites, and PBPK modeling was used to predict serum metabolites at expected human exposures (Clewell et al., submitted).

Broad screening of possible MOAs along with Level 1 chemical characterization may indicate that responses are due to direct chemical reactivity or broad low-affinity non-covalent interactions (Judson et al., 2016) rather than interaction with more specific biological targets. In these cases, no observed transcriptional effect levels (NOTELs) coupled with HT-IVIVE can support decisions about MOEs. Many of the same considerations for Level 3 assays also apply for the more complex assays in Level 4. Of course, decisions based on *in vivo* studies would use pharmacokinetic (especially PBPK) modeling for assessing internal doses and for selecting extrapolation methods (e.g., threshold models augmented with use of uncertainty factors for non-cancer risk evaluations).

An integrated approach with bioactivity testing and exposure assessment for assays at Level 4 (Webster et al., 2019) employed an MOE approach referred to as a bioactivity exposure ratio. Results from HTS assays (ToxCast), *in vivo* screening level assays, and *in vivo* apical tests of adverse effects were used to inform the need for conducting additional testing. Importantly, this case example involved several data-rich substances and showed that *in vitro* MOE values were actually lower than the *in vivo* MOE values, an observation "that this health protective approach could facilitate a substance's prioritization or deprioritization for further action, including the need for comprehensive *in vivo* testing."

5 Domain of applicability

The goal of organizing NAMs within these four levels was to consider when data from any of the levels would return adequate information to determine product safety for intended uses. The organization then provides a focus on the risk context, not simply the types of assays and computational tools available. Its applicability to particular chemistries or industrial sectors depends on the end-uses of products and whether the value of the product is associated with some biological activity. With environmental compounds, where the functionality is not related to specific biological activity, these tools offer significant promise for safety assessments based on measures of MOEs (TTC or AC50 divided by exposure). This approach is more safety assessment-based and differs from risk assessment procedures over the last 40 years where there was an attempt to estimate a human dose (exposure) that would be expected to produce some low incidence of response in a human population. This difference, i.e. a safety assessment versus risk assessment emphasis, was highlighted as a key point in applying TT21C information rather than *in vivo* animal studies for decision-making (Andersen and Krewski, 2010).

The transformation from traditional risk assessment approaches to this problem-oriented, safety assessment approach based on the use of NAMs across the different levels should be appropriate for environmental compounds, GRAS substances, cosmetics and food additives (Rovida et al., 2015; Hartung, 2018). The use with functional food additives or cosmetics with targeted biological activity poses challenges depending on the nature of the biological activity, the level of exposure from the intended uses of products, and on the possibility of inappropriate use conditions leading to excessive exposures. These two classes, i.e. functional foods and bioactive cosmetics, are intermediate between environmental compounds and those marketed because of end-use bioactivity.

Pharmaceuticals and pesticides pose challenges in that intrinsic biological activity is essential to efficacy for their intended uses. These classes of compounds can have both excessive on-target and unanticipated off-target biological activity. The discussed scheme for using NAMs for safety assessment would likely need to be customized for pharmaceuticals and pesticides.

The challenges of developing NAM-based approaches with bioactive compounds was highlighted recently in a multi-stakeholder meeting aiming to establish readiness criteria for assessing developmental neurotoxicity (Bal-Price et al., 2018). The approaches with these bioactive compounds need to be fashioned to capture multiple possible MOAs and encourage use of integrated assessment approaches (IATAs) (Tollefsen et al., 2014) that have undergone some level of mechanistic validation (Hartung et al., 2013). Nevertheless, the challenges in pursuing NAM-based safety assessment with pesticides and pharmaceuticals do not diminish the promise of their more rapid application with these classes of products.

6 Summary

With the explosion of available NAMs in the past decade and changes in the regulatory environment afforded by various initiatives such as the Frank R. Lautenberg Chemical Safety for the 21st Century legislation[8], it is an opportune moment to assess how information developed using NAMs will shape approaches for various risk assessment decisions. In looking over the possibilities for their use, there is no one-size-fits-all solution; rather, the context of the decision needs to drive the selection of NAMs used in any risk assessment. This contribution organizes NAMs into different levels, emphasizing the types of decisions that can follow from completion of studies at each of the levels. Importantly, most risk-based decisions do not require bringing compounds or classes of compounds through a tiered strategy (i.e., going lockstep from Level 1 through Level 4). Moving through just one or two of these levels should allow decisions about relative risks of products, including absence or low degree of potential anticipated toxicity and low expected exposure (i.e., very high MOSs or MOEs). Level 2 and 3 assays should provide the necessary information for assessing MOAs, AC50s or LECs and, when combined with improved human exposure assessment methodologies, should become preferred approaches for most safety assessments. The context-dependent applications of NAMs and the functional roadmap we describe may be useful in motivating additional case examples documenting the utility of, and confidence in, using a defined set of NAMs for specific decisions. In addition, the framework and roadmap can also help to identify where additional scientific research is needed to build greater confidence in various NAMs so that they can be used in the future with the necessary degree of confidence.

References

Adeleye, Y., Andersen, M., Clewell, R. et al. (2015). Implementing toxicity testing in the 21st century (TT21C): Making safety decisions using toxicity pathways, and progress in a prototype risk assessment. *Toxicology 332*, 102-111. doi:10.1016/j.tox.2014.02.007

Andersen, M. E. and Krewski, D. (2010). The vision of toxicity testing in the 21st century: Moving from discussion to action. *Toxicol Sci 117*, 17-24. doi:10.1093/toxsci/kfq188

Andersen, M. E., Pendse, S. N., Black, M. B. and McMullen, P. D. (2018). Application of transcriptomic data, visualization tools and bioinformatics resources for informing mode of action. *Curr Opin Toxicol 9*, 21-27. doi:10.1016/j.cotox.2018.05.003

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

Barakat, R. R. (1995). The effect of tamoxifen on the endometrium. *Oncology 9*, 129-134; discussion 139-140, 142.

Beames, T., Roberts, A., Moreau, M. et al. (in press). The role of fit-for-purpose assays within tiered testing approaches: An example evaluating estrogen-active compounds in a human adenocarcinoma cell line. *Toxicol Appl Pharmacol.*

Becker, R. A., Friedman, K. P., Simon, T. W. et al. (2015). An exposure:activity profiling method for interpreting high-throughput screening data for estrogenic activity – Proof of concept. *Regul Toxicol Pharmacol 71*, 398-408. doi:10.1016/j.yrtph.2015.01.008

Bhattacharya, S., Zhang, Q., Carmichael, P. L. et al. (2011). Toxicity testing in the 21 century: Defining new risk assessment approaches based on perturbation of intracellular toxicity pathways. *PLoS One 6*, e20887. doi:10.1371/journal.pone.0020887

Blaauboer, B. J. and Andersen, M. E. (2007). The need for a new toxicity testing and risk analysis paradigm to implement REACH or any other large scale testing initiative. *Arch Toxicol 81*, 385-387. doi:10.1007/s00204-006-0175-0

Boutin, M. E., Kramer, L. L., Livi, L. L. et al. (2018). A three-dimensional neural spheroid model for capillary-like network formation. *J Neurosci Methods 299*, 55-63. doi:10.1016/j.jneumeth.2017.01.014

Bray, M. A., Singh, S., Han, H. et al. (2016). Cell Painting, a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes. *Nat Protoc 11*, 1757-1774. doi:10.1038/nprot.2016.105

Bray, M. A., Gustafsdottir, S. M., Rohban, M. H. et al. (2017). A dataset of images and morphological profiles of 30 000 small-molecule treatments using the Cell Painting assay. *Gigascience 6*, 1-5. doi:10.1093/gigascience/giw014

Browne, P., Judson, R. S., Casey, W. M. et al. (2015). Screening chemicals for estrogen receptor bioactivity using a computational model. *Environ Sci Technol 49*, 8804-8814. doi:10.1021/acs.est.5b02641

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

[8] https://www.epa.gov/laws-regulations/summary-toxic-substances-control-act

Clewell, R. A., Sun, B., Adeleye, Y. et al. (2014). Profiling dose-dependent activation of p53-mediated signaling pathways by chemicals with distinct mechanisms of DNA damage. *Toxicol Sci 142*, 56-73. doi:10.1093/toxsci/kfu153

Clewell, R. A. and Andersen, M. E. (2016). Approaches for characterizing threshold dose-response relationships for DNA-damage pathways involved in carcinogenicity in vivo and micronuclei formation in vitro. *Mutagenesis 31*, 333-340. doi:10.1093/mutage/gev078

Clewell, R. A., McMullen, P. D., Adeleye, Y. et al. (2016). Pathway based toxicology and fit-for-purpose assays. *Adv Exp Med Biol 856*, 205-230. doi:10.1007/978-3-319-33826-2_8

Clewell, R. A., Leonard, J., Nicolas, C. et al. (submitted). Application of a combined aggregate exposure pathway and adverse outcome pathway (AEP-AOP) approach to inform a cumulative risk assessment: A case study with phthalates. *Toxicol In Vitro*.

Collins, F. S., Gray, G. M., Bucher, J. R. (2008). Toxicology. Transforming environmental health protection. *Science 319*, 906-907. doi:10.1126/science.1154619

DeGroot, D. E., Swank, A., Thomas, R. S. et al. (2018). mRNA transfection retrofits cell-based assays with xenobiotic metabolism. *J Pharmacol Toxicol Methods 92*, 77-94. doi:10.1016/j.vascn.2018.03.002

Djoumbou-Feunang, Y., Fiamoncini, J., Gil-de-la-Fuente, A. et al. (2019). BioTransformer: A comprehensive computational tool for small molecule metabolism prediction and metabolite identification. *J Cheminform 11*, 2. doi:10.1186/s13321-018-0324-5

Feng, Y., Mitchison, T. J., Bender, A. et al. (2009). Multi-parameter phenotypic profiling: Using cellular effects to characterize small-molecule compounds. *Nat Rev Drug Discov 8*, 567-578. doi:10.1038/nrd2876

Foley, B., Doheny, D. L., Black, M. B. et al. (2017). Editor's highlight: Screening ToxCast prioritized chemicals for PPARG function in a human adipose-derived stem cell model of adipogenesis. *Toxicol Sci 155*, 85-100. doi:10.1093/toxsci/kfw186

Grimm, D. (2019). EPA plan to end animal testing splits scientists. *Science 365*, 1231. doi:10.1126/science.365.6459.1231

Grimm, F. A., Iwata, Y., Sirenko, O. et al. (2016). A chemical-biological similarity-based grouping of complex substances as a prototype approach for evaluating chemical alternatives. *Green Chem 18*, 4407-4419. doi:10.1039/c6gc01147k

Hartman, J. K., Beames, T., Parks, B. et al. (2018). An in vitro approach for prioritization and evaluation of chemical effects on glucocorticoid receptor mediated adipogenesis. *Toxicol Appl Pharmacol 355*, 112-126. doi:10.1016/j.taap.2018.05.016

Hartung, T., Hoffmann, S. and Stephens, M. (2013). Mechanistic validation. *ALTEX 30*, 119-130. doi:10.14573/altex.2013.2.119

Hartung, T. (2018). Rebooting the generally recognized as safe (GRAS) approach for food additive safety in the US. *ALTEX 35*, 3-25. doi:10.14573/altex.1712181

Judson, R. S., Houck, K. A., Kavlock, R. J. et al. (2010). In vitro screening of environmental chemicals for targeted testing prioritization: The ToxCast project. *Environ Health Perspect 118*, 485-492. doi:10.1289/ehp.0901392

Judson, R. S., Magpantay, F. M., Chickarmane, V. et al. (2015). Integrated model of chemical perturbations of a biological pathway using 18 in vitro high-throughput screening assays for the estrogen receptor. *Toxicol Sci 148*, 137-154. doi:10.1093/toxsci/kfv168

Judson, R., Houck, K., Martin, M. et al. (2016). Analysis of the effects of cell stress and cytotoxicity on in vitro assay activity across a diverse chemical and assay space. *Toxicol Sci 153*, 409. doi:10.1093/toxsci/kfw148

Kabadi, P. K., Vantangoli, M. M., Rodd, A. L. et al. (2015). Into the depths: Techniques for in vitro three-dimensional microtissue visualization. *Biotechniques 59*, 279-286. doi:10.2144/000114353

Kavlock, R. J., Ankley, G. T., Blancato, J. N. et al. (2003). A framework for a computational toxicology research program in ORD. Washington, DC: U.S. Environmental Protection Agency EPA600/R-03/65.

Leonard, J. A., Stevens, C., Mansouri, K. et al. (2018). A workflow for identifying metabolically active chemicals to complement in vitro toxicity screening. *Comput Toxicol 6*, 71-83. doi:10.1016/j.comtox.2017.10.003

Manganelli, S., Roncaglioni, A., Mansouri, K. et al. (2019). Development, validation and integration of in silico models to identify androgen active chemicals. *Chemosphere 220*, 204-215. doi:10.1016/j.chemosphere.2018.12.131

Mansouri, K., Abdelaziz, A., Rybacka, A. et al. (2016). CERAPP: Collaborative estrogen receptor activity prediction project. *Environ Health Perspect 124*, 1023-1033. doi:10.1289/ehp.1510267

Marchant, C. A., Briggs, K. A., Long, A. (2008). In silico tools for sharing data and knowledge on toxicity and metabolism: Derek for Windows, Meteor, and Vitic. *Toxicol Mech Methods 18*, 177-187. doi:10.1080/15376510701857320

Marx, U., Andersson, T. B., Bahinski, A. et al. (2016). Biology-inspired microphysiological system approaches to solve the prediction dilemma of substance testing. *ALTEX 33*, 272-321. doi:10.14573/altex.1603161

McMullen, P. D., Bhattacharya, S., Woods, C. G. et al. (2014). A map of the PPARα transcription regulatory network for primary human hepatocytes. *Chem Biol Interact 209*, 14-24. doi:10.1016/j.cbi.2013.11.006

McMullen, P. D., Pendse, S., Black, M. et al. (2019). Addressing systematic inconsistencies between vitro and in vivo mode of action signatures. *Toxicol In Vitro 58*, 1-12. doi:10.1016/j.tiv.2019.02.014

McMullen, P. D., Bhattacharya, S., Woods, C. et al. (in press). Identifying qualitative differences in PPARα signaling networks in human and rat hepatocytes and their significance for next generation chemical risk assessment methods. *Toxicol In Vitro*. doi:10.1016/j.tiv.2019.02.017

Mekenyan, O. G., Dimitrov, S. D., Pavlov, T. S. and Veith, G. D. (2004). A systematic approach to simulating me-

tabolism in computational toxicology. I. The TIMES heuristic modelling framework. *Curr Pharm Des 10*, 21. doi:10.2174/1381612043452596

Miller, M. M., Alyea, R. A., LeSommer, C. et al. (2016). Editor's highlight: Development of an in vitro assay measuring uterine-specific estrogenic responses for use in chemical safety assessment. *Toxicol Sci 154*, 162-173. doi:10.1093/toxsci/kfw152

Miller, M. M., McMullen, P. D., Andersen, M. E. and Clewell, R. A. (2017). Multiple receptors shape the estrogen response pathway and are critical considerations for the future of in vitro-based risk assessment efforts. *Crit Rev Toxicol 47*, 564-580. doi:10.1080/10408444.2017.1289150

Moreau, M., Mallick, P., Smeltz, M. et al. (in preparation). Incorporation of metabolite exposure in high-throighput in vitro to in vivo extrapolation (HT-IVIVE).

NRC – National Research Council (2007). *Toxicity Testing in the 21st Century: A Vision and a Strategy*. Washington, DC: The National Academies Press.

NRC (2009). *Exposure Science in the 21st Centruy: A Vision and a Strategy*. Washington, DC: The National Academies Press.

Nicolas, C. L., Minto, M. S., Mansouri, K. et al. (in preparation). Estimating provisional margins of exposure for data-poor chemicals using high-throughput computational methods.

Pamies, D., Barreras, P., Block, K. et al. (2017). A human brain microphysiological system derived from induced pluripotent stem cells to study neurological diseases and toxicity. *ALTEX 34*, 362-376. doi:10.14573/altex.1609122

Patlewicz, G., Wambaugh, J. F., Felter, S. P. et al. (2018). Utilizing threshold of toxicological concern (TTC) with high throughput exposure predictions (HTE) as a risk-based prioritization approach for thousands of chemicals. *Comput Toxicol 7*, 58-67. doi:10.1016/j.comtox.2018.07.002

Reif, D. M., Martin, M. T., Tan, S. W. et al. (2010). Endocrine profiling and prioritization of environmental chemicals using ToxCast data. *Environ Health Perspect 118*, 1714-1720. doi:10.1289/ehp.1002180

Ring, C. L., Arnot, J. A., Bennett, D. H. et al. (2019). Consensus modeling of median chemical intake for the U.S. population based on predictions of exposure pathways. *Environ Sci Technol 53*, 719-732. doi:10.1021/acs.est.8b04056

Roberts, G., Myatt, G. J., Johnson, W. P. et al. (2000). LeadScope: Software for exploring large sets of screening data. *J Chem Inf Comput Sci 40*, 1302-1314. doi:10.1021/ci0000631

Rotroff, D. M., Wetmore, B. A., Dix, D. J. et al. (2010). Incorporating human dosimetry and exposure into high-throughput in vitro toxicity screening. *Toxicol Sci 117*, 348-358. doi:10.1093/toxsci/kfq220

Rovida, C., Asakura, S., Daneshian, M. et al. (2015). Toxicity testing in the 21st century beyond environmental chemicals. *ALTEX 32*, 171-181. doi:10.14573/altex.1506201

Shah, I., Setzer, R. W., Jack, J. et al. (2016). Using ToxCast data to reconstruct dynamic cell state trajectories and estimate toxicological points of departure. *Environ Health Perspect 124*, 910-919. doi:10.1289/ehp.1409029

Sipes, N. S., Wambaugh, J. F., Pearce, R. et al. (2017). An intuitive approach for predicting potential human health risk with the Tox21 10k library. *Environ Sci Technol 51*, 10786-10796. doi:10.1021/acs.est.7b00650

Thomas, R. S., Allen, B. C., Nong, A. et al. (2007). A method to integrate benchmark dose estimates with genomic data to assess the functional effects of chemical exposure. *Toxicol Sci 98*, 240-248. doi:10.1093/toxsci/kfm092

Thomas, R. S., Philbert, M. A., Auerbach, S. S. et al. (2013). Incorporating new technologies into toxicity testing and risk assessment: Moving from 21st century vision to a data-driven framework. *Toxicol Sci 136*, 4-18. doi:10.1093/toxsci/kft178

Thomas, R. S., Bahadori, T., Buckley, T. J. et al. (2019). The next generation blueprint of computational toxicology at the U.S. Environmental Protection Agency. *Toxicol Sci 169*, 317-332.

Tollefsen, K. E., Scholz, S., Cronin, M. T. et al. (2014). Applying adverse outcome pathways (AOPs) to support integrated approaches to testing and assessment (IATA). *Regul Toxicol Pharmacol 70*, 629-640. doi:10.1016/j.yrtph.2014.09.009

Tyson, J. J. and Novák, B. (2015). Models in biology: Lessons from modeling regulation of the eukaryotic cell cycle. *BMC Biol 13*, 46. doi:10.1186/s12915-015-0158-9

US EPA – United States Environmental Protection Agency (2018a). Estimation Programs Interface Suite™ for Microsoft® Windows, v 4.11.

US EPA (2018b). Strategic Plan to Promote the Development and Implementation of Alternative Test Methods Within the TSCA Program. http://tiny.cc/tufdez

US EPA (2018c). A Working Approach for Identifying Potential Candidate Chemicals for Prioritization. Washington, DC. http://tiny.cc/bsfdez

Vantangoli, M. M., Madnick, S. J., Wilson, S. and Boekelheide, K. (2016). Estradiol exposure differentially alters monolayer versus microtissue MCF-7 human breast carcinoma cultures. *PLoS One 11*, e0157997. doi:10.1371/journal.pone.0157997

Wambaugh, J. F., Setzer, R. W., Reif, D. M. et al. (2013). High-throughput models for exposure-based chemical prioritization in the ExpoCast project. *Environ Sci Technol 47*, 8479-8488. doi:10.1021/es400482g

Wambaugh, J. F., Hughes, M. F., Ring, C. L. et al. (2018). Evaluating in vitro-in vivo extrapolation of toxicokinetics. *Toxicol Sci 163*, 152-169. doi:10.1093/toxsci/kfy020

Webster, F., Gagne, M., Patlewicz, G. et al. (2019). Predicting estrogen receptor activation by a group of substituted phenols: An integrated approach to testing and assessment case study. *Regul Toxicol Pharmacol 106*, 278-291. doi:10.1016/j.yrtph.2019.05.017

Wetmore, B. A., Wambaugh, J. F., Ferguson, S. S. et al. (2013). Relative impact of incorporating pharmacokinetics on predicting in vivo hazard and mode of action from high-throughput in vitro toxicity assays. *Toxicol Sci 132*, 327-346. doi:10.1093/toxsci/kft012

Wetmore, B. A. (2015). Quantitative in vitro-to-in vivo extrapolation in a high-throughput environment. *Toxicology 332*, 94-101. doi:10.1016/j.tox.2014.05.012

Wetmore, B. A., Clewell, R. A., Cholewa, B. et al. (2019). Assessing bioactivity-exposure profiles of fruit and vegetable extracts in the BioMAP profiling system. *Toxicol In Vitro 54*, 41-57. doi:10.1016/j.tiv.2018.09.006

Yeakley, J. M., Shepard, P. J., Goyena, D. E. et al. (2017). A trichostatin A expression signature identified by TempO-Seq targeted whole transcriptome profiling. *PLoS One 12*, e0178302. doi:10.1371/journal.pone.0178302

Yoon, M., Blaauboer, B. J., Clewell, H. J. (2015). Quantitative in vitro to in vivo extrapolation (QIVIVE): An essential element for in vitro-based risk assessment. *Toxicology 332*, 1-3. doi:10.1016/j.tox.2015.02.002

Zhang, B. and Radisic, M. (2017). Organ-on-a-chip devices advance to market. *Lab Chip 17*, 2395-2420. doi:10.1039/C6LC01554A

Zhang, B., Korolj, A., Lun-Lai, B. and Radisic, M. (2018). Advances in organ-on-a-chip engineering. *Nature Reviews Materials 3*, 257-278. doi:10.1038/s41578-018-0034-7

Zhang, Q., Bhattacharya, S., Conolly, R. B. et al. (2014). Molecular signaling network motifs provide a mechanistic basis for cellular threshold responses. *Environ Health Perspect 122*, 1261-1270. doi:10.1289/ehp.1408244

Zhang, Q., Bhattacharya, S., Pi, J. et al. (2015). Adaptive post-translational control in cellular stress response pathways and its relationship to toxicity testing and safety assessment. *Toxicol Sci 147*, 302-316. doi:10.1093/toxsci/kfv130

Conflict of interest

The authors declare that they have no conflicts of interest.

Acknowledgements

Many of our efforts to develop coherent approaches for decision-context appropriate testing with non-animal methods, including computational methods, IVIVE techniques and bespoke assay development, have received multi-year support at the Hamner Institutes and now ScitoVation from the Long Range Research Initiative of the American Chemistry Council (ACC-LRI). In addition, another program – TT21C: Toxicity Pathways and Network Biology – pursued three case studies to develop an understanding of the use of *in vitro* fit-for-purpose assays in human cells as the basis for risk assessment. This broad TT21C program received support from Dow Chemical, ExxonMobil Foundation, Dow Corning, Unilever, Agilent, Crop Life America and 3M Company. We are grateful for the support from all these organizations over the life of this research effort and the opportunity to integrate these efforts to pursue context dependent strategies for toxicity evaluation and risk assessments.