Integration of New Approach Methods for the Assessment of Data Poor Chemicals

Katie Paul Friedman[1*], Russell S. Thomas[1], John F. Wambaugh[1], Joshua A. Harrill[1], Richard S. Judson[1], Timothy J. Shafer[1], Antony J. Williams[1], Jia-Ying Joey Lee[2], Lit-Hsin Loo[2], Matthew Gagné[3], Alexandra S. Long[3], Tara S. Barton-Maclaren[3], Maurice Whelan[4], Mounir Bouhifd[5], Mike Rasenberg[5], Ulla Simanainen[5], Tomasz Sobanski[5]

[1]Center for Computational Toxicology and Exposure, Office of Research and Development, U.S. Environmental Protection Agency, Research Triangle Park, North Carolina, 27711, USA

[2]Innovations in Food and Chemical Safety Programme and Bioinformatics Institute, Agency for Science, Technology and Research, Singapore

[3]Healthy Environments and Consumer Safety Branch, Health Canada, Ottawa, Canada

[4]European Commission, Joint Research Centre (JRC), Ispra, Italy

[5]Computational Assessment Unit, European Chemicals Agency, Helsinki, Finland

*Corresponding author: Katie Paul Friedman, paul-friedman.katie@epa.gov

***Disclaimer**: The United States Environmental Protection Agency (U.S. EPA) through its Office of Research and Development has subjected this article to Agency administrative review and approved it for publication. Mention of trade names or commercial products does not constitute endorsement for use. The views expressed in this article are those of the authors and do not necessarily represent the views or policies of A*STAR, US EPA, ECHA, EFSA, Health Canada, or the JRC.*

Acronyms

ACC = active concentration at the cutoff defined for a positive response from the ToxCast pipeline (see https://CRAN.R-project.org/package=tcpl);

APCRA = Accelerating the Pace of Chemical Risk Assessment;

AQC = analytical quality control;

AUC = area under the curve score, as in the ToxCast estrogen and androgen receptor pathway models;

BER = bioactivity:exposure ratio;

BPAC = minimum biological pathway altering concentration;

Css = steady-state plasma concentration;

HTTK = high-throughput toxicokinetics (capitalized for data and models; lower case and italicized in reference to the R package);

HTPP = high-throughput phenotypic profiling;

HTTr = high-throughput transcriptomics;

IVIVE = *in vitro* to *in vivo* extrapolation;

Log10 POD ratio = ratio of log10 POD_{trad} to log10 POD_{NAM}, i.e., log10(POD_{trad}) – log10(POD_{NAM});

MEA = microelectrode array;

MBC = minimum bioactive concentration *in vitro*;

NAM = new approach method;

NBA = NAM-based assessment;

NOEL = no observable effect level;

NOAEL = no observable adverse effect level;

POD = point-of-departure;

POD_{NAM} = *in vitro* NAM-based POD;

POD_{SUB} = *in vivo* POD value subset to include only subchronic studies;

POD_{TTC} = threshold POD assigned using the TTC;

POD_{trad} = traditional *in vivo* POD value used to compare to the POD_{NAM};

TTC = thresholds of toxicological concern

Abstract

Use of new approach methods (NAMs), including high-throughput, *in vitro* bioactivity data, in setting a point-of-departure (POD) will accelerate the pace of human health hazard assessments. Combining hazard and exposure predictions into a bioactivity:exposure ratio (BER) for use in risk-based prioritization and utilizing NAM-based bioactivity flags to indicate potential hazards of interest for further prediction or mechanism-based screening together comprise a prospective approach for management of substances with limited traditional toxicity testing data. In this work we demonstrate a NAM-based assessment case study conducted via the Accelerating the Pace of Chemical Risk Assessment (APCRA) initiative, a consortium of international research and regulatory scientists. The primary objective was to develop a reusable and adaptable approach for addressing chemicals with limited traditional toxicity data using a NAM-based POD, BER, and bioactivity-based flags for indication of putative endocrine, developmental, neurological, and immunosuppressive effects via data generation and interpretation for 200 substances. Multiple data streams, including *in silico* and *in vitro* NAMs, were used. High-throughput transcriptomics and phenotypic profiling data, as well as targeted biochemical and cell-based assays, were combined with generic high-throughput toxicokinetic models parameterized with chemical-specific data to estimate dose for comparison to exposure predictions. This case study further enables regulatory scientists from different international purviews to utilize efficient approaches for prospective chemical management, addressing hazard and risk-based data needs, while reducing the need for animal studies. This work demonstrates the feasibility of using a battery of toxicodynamic and toxicokinetic NAMs to provide a NAM-based POD for screening-level assessment.

Keywords

New approach methods; high-throughput toxicokinetics; NAM-based assessment; data-poor chemicals

Highlights

- *In silico* tools should be used to determine chemical amenability to *in vitro* screening and support putative targets for toxicity.
- Prospective application of a battery of broad profiling and targeted *in vitro* NAMs provides sufficient data to develop a point-of-departure (POD_{NAM}) that is generally protective.
- POD_{NAM} are typically within ± 2 log10-mg/kg/day of existing repeat dose POD data from animal models.
- Combining POD_{NAM} with exposure estimates to derive a bioactivity:exposure ratio can help identify data-poor chemicals for additional consideration.
- In the future, *in silico* models could combine information from POD_{NAM} and existing animal-based POD values for consensus POD prediction.

Introduction

Internationally, chemicals regulation by statutory authorities proceeds with different requirements that vary by country or jurisdiction, but there are several unifying elements to chemical assessment needs whether it be for new chemical submissions or to address those that are already in commerce. There is the need to address 100s to 1000s of chemicals or chemical submissions effectively; the need to provide an approach that appropriately addresses the hazard and risk of chemicals that may be characterized as "data-poor;" the need to provide data-driven, health-protective decisions on these chemical submissions and assessments; and, the need to deliver these decisions on a relatively short time scale. Chemical submissions under the Registration, Evaluation, Authorisation, and Restriction of Chemicals (REACH) in the European Union (EU) (EuropeanCommission, 2007) are associated with a dossier of required studies, including repeated dose as well as reproductive and prenatal developmental toxicity studies (which may be combined as Organisation for Economic Co-operation and Development [OECD] test guideline 422). Other non-experimental information is also used, including (quantitative) structure-activity relationships ((Q)SARs), as well as grouping and read-across, with read-across as the preferred option for filling gaps in repeated dose toxicity information (ECHA, 2023). In the United States, the Toxic Substances Control Act (TSCA), implemented by the Environmental Protection Agency (EPA), does not establish a minimum data set or require that certain tests be conducted prior to the submission of a new chemical notice. If a determination is made that available information is insufficient to evaluate health effects, then EPA may require development of test data (Lautenberg, 2016). Indeed, new chemical reviews under TSCA have utilized (Q)SARs and other predictive models and tools coupled with category-based approaches (USEPA, 2022b) to inform decisions about new chemical substances. Recently, the TSCA New Chemicals Collaborative Research Program was designed with the aim of augmenting the currently available (Q)SAR, read-across, and predictive approaches with newly developed chemical groupings, systematized read-across approaches, newly validated (Q)SARs (OECD,

2014; OECD, 2023), and information from *in vitro* NAMs (USEPA, 2022a) to inform rapid new chemical assessments. New chemical submissions in Canada follow the New Substance Notification Regulations (NSNR) (Canada, 2018) under the Canadian Environmental Protection Act 1999 (CEPA) and for certain volume triggers, require the submission of a repeated dose toxicity study among other requirements. The use of NAMs can be accommodated to meet technical information requirements prescribed by the NSNR when determined to provide a scientifically valid measure of the endpoint under investigation. For the Existing Substances program under CEPA, data generated from NAM-based approaches have been increasingly used to support various decision-making contexts including for grouping and read-across, prioritization and to address data needs for risk assessment under Canada's Chemicals Management Plan (CMP). A goal in Canada is to continue to use *in silico* and *in vitro* NAMs to address challenges associated with putative hazard identification and assessments of chemicals that lack traditional toxicity data (Barton-Maclaren *et al.*, 2022; Beal *et al.*, 2023; Beal *et al.*, 2022; Bhuller *et al.*, 2021; HC/ECCC, 2023; Johnson *et al.*, 2022; Kulkarni *et al.*, 2016; Zwickl *et al.*, 2022). Thus, there is already precedent for the use of NAMs in chemical assessment, but the approaches and particularly use of *in vitro* bioactivity NAM data, as well as the specific applications (e.g., prioritization, replacement of animal studies, informing a fully NAM-based assessment) vary between regulatory contexts and statutory authorities.

Despite the use of largely *in silico* predictions and read-across to manage or prioritize many data-poor chemicals, repeated dose animal studies continue to be a requirement and provide an anchor for current human health risk assessment practices in many regulatory contexts. As such, we perceived the need to demonstrate how a battery of *in silico* and *in vitro* NAMs could provide a replacement for, or in some cases a bridge to, currently used repeated dose animal studies. Repeated dose studies in animals are considered a definitive source for determining a point-of-departure (POD) for risk assessment and for hazard identification in the current international regulatory environment in the event existing data

gap filling techniques like read-across are not suitable. Repeated dose studies such as the subchronic study (90-day exposure) may also demonstrate effects to support a hazard indication or concern, which under REACH chemical management typically triggers further testing to assure availability of appropriate information for chemical safety assessment. Such triggered testing may include studies intended to assess potential chemical effects such as carcinogenicity, reproductive toxicity, developmental toxicity, neurotoxicity, and/or immunotoxicity. In other words, repeated dose toxicity testing is recognized as covering a potentially wide range of toxicological effects from repeated exposure.

However, in some regulatory paradigms, data-poor chemicals may not have associated repeated dose study data, and if these data are available, they may be considered insufficient for characterizing specific hazards such as endocrine, developmental, or reproductive toxicity, especially for humans. Given that animal studies may provide higher negative predictive value than positive predictive value for human clinical effects (i.e., that a negative result in animal studies is more predictive of a negative human clinical effect than a positive result animal result is predictive of a positive human clinical effect) (Monticello *et al.*, 2017), and that concordance of animal and human effects may vary based on specific combinations of organ system and species used (Clark and Steger-Hartmann, 2018), it follows that use of animals for the derivation of a dose level that is protective in terms of POD and specific hazards may be of greatest value for safety assessment (Browne *et al.*, 2024). The use of other human-based mechanistic models, i.e., *in vitro* NAMs using human cells or tissues, may provide insight into specific potential human health effects in toxicology applications (Pognan *et al.*, 2023). With tens of thousands of chemicals in commerce, it is unlikely that the use of a repeated dose animal study, e.g., the subchronic study, will be the predominant tool for evaluating the hazard potentially presented by the vast landscape of chemical exposures (Isaacs *et al.*, 2023). Given the economic value of obtaining both protective and reliable information on POD and hazard for risk assessment on a shorter timescale (Hagiwara *et al.*, 2023), the task before international regulatory authorities is: which NAMs should

comprise a flexible battery to evaluate data-poor chemicals, and how can specific frameworks that utilize NAMs of increasing complexity inform decisions with confidence and timeliness? And, further, what are plausible alternatives for repeated dose animal studies, including the subchronic study, which may not always be available for chemical evaluations?

Building alternatives to repeated dose toxicity studies requires a multi-faceted approach. Previously, short-term animal studies coupled with transcriptomic assessment have demonstrated success in providing predictive and/or protective values with respect to animal studies of longer exposure duration with apical toxicity measures, within a quantitative factor of 10 (Gwinn *et al.*, 2020; Pham *et al.*, 2020; Thomas *et al.*, 2013b; USEPA, 2023b), which is also within the quantitative variability of the PODs from these animal studies (Paul Friedman *et al.*, 2023; Pham*, et al.*, 2020). An integrated NAMs battery could provide a data-driven selection of chemicals for short-term transcriptomic enhanced animal studies, other repeated dose studies, or perhaps more biologically complex NAMs of lower throughput that recapitulate organ or systems functions (Thomas *et al.*, 2019). Additionally, NAMs could provide direct mechanistic, human-relevant flags of potential hazard based on (Q)SARs or bioactivity that short-term and subchronic animal studies might not be able to provide or might provide with a degree of uncertainty in terms of human relevance. Combining multiple NAMs into a framework to inform chemical assessment, and any additional data gathering, is the essence of NAM-based assessment (NBA) and the primary aim of the case study described herein.

In this work, the state of the science in applying NAMs to inform NBA and/or selection of chemicals for *in vivo* study is demonstrated in a transparent approach that can be adapted in the future for new types of data or for application to specific regulatory contexts. These latter activities are beyond the scope of this case study, due to the maturity of NBA for diverse chemistries and the need to define contexts of use within each regulatory framework. Rather, the objective of this research was to develop an understanding of the expectations, i.e. the protective and/or predictive nature, of using NAMs in an

NBA to produce a quantitative estimate of POD and qualitative, putative indicators of hazard. Existing NBA workflows incorporate several key pieces of information: estimates of exposure; structure alerts or (Q)SAR results; information from broad profiling of biological targets; information from targeted NAMs; toxicokinetic NAMs; and, potentially, NAMs of greater biological complexity to evaluate specific hazard types of interest. The proliferation of exposure-led NBA workflows (typically referenced as next generation risk assessment) has emanated in part from evaluation of cosmetic ingredients in the EU (Baltazar *et al.*, 2020; Basketter *et al.*, 2012; Dent *et al.*, 2021; Gilmour *et al.*, 2023; Hisaki *et al.*, 2015; Ouedraogo *et al.*, 2022; Reynolds *et al.*, 2021), wherein new animal testing is not permitted. However, NBA has other potential applications within the regulatory toxicology framework internationally, including new chemical assessment and to effectively address the many data-poor chemicals that are already on the market and in products available to consumers. A primary goal of the Accelerating the Pace of Chemical Risk Assessment (APCRA) initiative is for scientists engaged in solving regulatory toxicology problems across different regulatory systems to demonstrate more rapid and human-relevant approaches to health protective chemical risk assessment through case studies. Previously, our team engaged in a retrospective examination of how health-protective PODs based on NAMs (POD_{NAM}) were compared to PODs from traditional animal studies (POD_{trad}) and the utility of a bioactivity:exposure ratio (BER) to prioritize chemicals for further study, as part of an early and extensible NBA workflow (Paul Friedman *et al.*, 2020). The main objectives of this previous work were to develop a simple framework for using NAMs to demonstrate derivation of PODs; inform selection of chemicals for further study; and, to prioritize chemicals on the basis of a BER. This previously published retrospective case study furthered discourse in the toxicology community regarding how to construct POD_{NAM} and expectations on their protectiveness with respect to animal-based POD_{trad} values. It also encountered legitimate criticisms, including that the NBA approach was demonstrated largely with highly studied chemicals, especially enriched with pesticide active ingredients; the approach used all of the ToxCast database to

inform a POD_{NAM}, which was unlikely to be obtained for new chemicals; the approach using a highly protective POD_{NAM} may have been too conservative; and, that the approach was POD-focused within a risk context and agnostic to possible hazard indications. In the case study herein, we demonstrate an initial and straightforward approach for developing an integrated NAM dataset to prioritize chemicals for further evaluation as a key component of the bridge to future use of an NBA workflow that addresses some of the limitations of our previous work. As such, this work expands upon Paul Friedman *et al.* (2020), and colleagues throughout the field, by pursuing the combined use of broad, profiling NAMs and targeted NAMs; the use of a refined targeted NAM battery that is more tractable for prospective use; evaluation of multiple *in vitro* to *in vivo* extrapolation (IVIVE) decisions; demonstration of how different sets of *in vitro* NAMs could inform different POD_{NAM} estimates; evaluation of the performance of these POD_{NAM} as both protective and predictive of POD_{trad} to inform expectations for POD_{NAM}; development of putative flags for specific hazards of interest using a combination of *in vitro* and *in silico* NAMs; and, expansion of an NBA approach to chemicals that are data-poor. The primary goal of our case study was to demonstrate a NBA workflow that included chemical selection, bioactivity, exposure, and combined outputs, including a POD_{NAM}, using a refined battery of assays and methods, a BER, and hazard flags to give some indication of potential target toxicity that could be useful in selecting additional information to pursue. In developing the POD_{NAM} for this case study, we illustrate the impact of selecting different assays on the POD_{NAM}. In doing so, we provide information relevant to expectations of POD_{NAM} with respect to their ability to predict or be protective of a traditional, animal-based POD_{trad} estimate.

Methods

Cheminformatics

Chemical Selection

In contrast to the previously published "retrospective" case study[1] (Paul Friedman*, et al.*, 2020), chemicals were selected for this case study (referred to as the "prospective case study") to include more industrial and "data-poor" chemicals where additional data would be of interest to one or more of the case study partners (approximately 100 chemicals); chemicals that overlapped with the previous retrospective case study and could inform potential improvements in POD prediction (96 "data-rich" chemicals); and, finally, all chemicals selected needed to be available within the existing EPA ToxCast chemical library to conserve resources in conducting this case study. The 96 data-rich chemicals from the previous retrospective case study were selected to include approximately equal numbers of chemicals that demonstrated NAM-based PODs that were under-protective, approximately equal to, and over-protective from the previous retrospective case study POD_{trad}. In this case study, "data poor" was defined as chemicals lacking traditional repeated dose toxicity studies (narrowly defined here as subchronic or chronic studies). This narrow definition underscores that "data-poorness" is a context-specific determination. Some of the chemicals selected in this case study were already associated with toxicokinetic assay data that could inform *in vitro* to *in vivo* extrapolation (IVIVE) using a high-throughput toxicokinetic (HTTK) approach as well as analytical quality control (AQC) information (Richard *et al.*, In review.) that was collected as part of the Tox21 project. These data were only recently (after our chemical selection and screening had been executed) interpreted by analytical chemists, including evaluation across multiple analytical methods applied, to help inform amenability considerations for *in vitro* screening. Samples within the Tox21 project were solvated in dimethyl sulfoxide (DMSO) and aliquoted to dosing plates and stored at room temperature in ambient conditions, from which one or more analyses (liquid or gas chromatography coupled to mass spectrometry or

[1] https://comptox.epa.gov/dashboard/chemical-lists/APCRARETRO

nuclear magnetic resonance) were performed at 0 and 4 months (Richard*, et al.*, In review.). For more detail on the AQC flags available for DMSO-solvated samples, see Supplemental File 1, Table S1.

Assigned and/or predicted exposure pathways used in total population exposure predictions (Ring *et al.*, 2019) and physicochemical property predictions generated by OPERA version 2.6 (Mansouri *et al.*, 2018) for the CompTox Chemicals Dashboard (Williams *et al.*, 2017) were investigated as measures of chemical diversity to demonstrate the breadth of chemistries included in this case study and in the previously published "retrospective" case study[2] (Paul Friedman*, et al.*, 2020). Manual, expert review of the chemical list was performed at the inception of the case study, but some chemicals were included that are unlikely to be amenable to *in vitro* screening, as discussed further in the Results (see Figure 1). However, some chemicals that did not fully "pass" AQC were included in the chemical set screened, as these AQC information were not fully available for the ToxCast chemical library at the beginning of this work. We defined an AQC "pass" specifically for this work, summarized as follows. For AQC grades at time 0 (T0) of A, B, or C (molecular weight confirmed and purity greater than 90, 75-90, or 50-75%, respectively) regardless of chemical level stability call over time; grades at T0 of A, B, or C and with a chemical level stability call of "stable" or some "physical loss"; or, no data available [2 of the 201 chemicals], AQC was considered "passing" (samples for 178 of 201 chemicals). For all other grades, the chemical was considered "not passing" for this case study (samples for 23 of 201 chemicals). For a summary of available grades and calls, see Supplemental File 1, Table S2. Stability of chemical samples from T0 to 4 months in DMSO at room temperature (T4) was not required to pass, as this represents a fairly extreme handling of samples as chemical plates are typically stored in freezers between experiments. However, the summarization of the AQC for this case study is permissive, and as such, some chemicals labeled "passing" may have some stability issues over time in the DMSO solvated

[2] https://comptox.epa.gov/dashboard/chemical-lists/APCRARETRO

sample, but without any empirical data to fully characterize the degradants that might be present and drive bioactivity. In an effort to understand if we could identify chemicals that would not pass our permissive AQC filter using physicochemical properties, a Uniform Manifold Approximation and Projection (UMAP) (McInnes *et al.*, 2018) to reduce the feature dimensionality of molecular weight and predicted logP, vapor pressure, and water solubility was performed.

QSARs

In silico NAMs were applied in two different ways for this case study: (1) for POD estimation (threshold of toxicological concern, TTC); and, (2) for qualitative prediction of hazard. Quantitative TTC values, i.e., daily intake amounts below which there is a low probability of risk to human health, have previously been proposed as a rapid screening and prioritization tool (EFSA, 2012; HealthCanada, 2016; Nicolas *et al.*, 2022; Patlewicz *et al.*, 2018; Paul Friedman*, et al.*, 2020). In this work, each chemical was assigned a TTC value using the software ToxTree [v2.6.6] which implements the TTC decision-tree based on chemical structures, as described in Kroes *et al.,* 2004 (HealthCanada, 2016; Patlewicz*, et al.*, 2018). TTC value assignment was based on structural classes (i.e. Cramer classification or organophosphates/carbamates). These TTC values relate most closely with the comparisons made in this case study as they were derived from known distributions of *in vivo* POD values. TTC values developed for potential genotoxicants were not used. Kroes *et al.* (2004) specified exclusionary structural classes where TTC values are known not to be applicable (e.g. steroids), and a TTC value was not assigned for these types of substances. Comparison of POD_{trad} to TTC was intended to provide insight on how protective the POD ratio is for the POD_{trad} to POD_{NAM} comparison, given that TTC values, based on their derivation, are expected to be highly protective of POD_{trad} (Paul Friedman*, et al.*, 2020). We expected that the median POD_{trad}:TTC ratio would be much greater than the median POD_{trad}:POD_{NAM} ratio. Qualitative hazard predictions included endocrine activity prediction (the Collaborative Estrogen Receptor Activity (CERAPP) and Collaborative Modeling Project for Androgen Receptor Activity

(COMPARA) consensus QSARs (Mansouri *et al.*, 2016; Mansouri *et al.*, 2020)) and developmental toxicity prediction (DEV TEST) (USEPA, 2020) as part of the cheminformatic portion of the NBA-workflow (Figure 1). CERAPP, COMPARA, and DEV TEST were utilized to describe endocrine and developmental toxicity, which were then combined into hazard flags for developmental and reproductive toxicity (DART), as described later in Methods and Results.

Bioactivity NAMs

Previously, a tiered scheme for *in vitro* screening has been suggested wherein Tier 1 NAMs broadly profile chemical-induced effects on transcriptomic signatures and cell morphometry, with the ability to inform both minimum *in vitro* bioactive concentration (MBC) and putative hazard (Thomas*, et al.*, 2019). Herein, a battery of bioactivity NAMs was selected as a demonstration of a putative minimal assay set for a combined prospective Tier 1 and 2 screening that could inform estimates of the MBC. This putative minimal assay set includes broad profiling methods (Tier 1) and targeted screening for specific bioactivities of interest (Tier 2) for regulatory toxicology, including endocrine, developmental, immunosuppressive, and neurological bioactivity as well as target cell type bioactivity for kidney, liver, and lung cells, as described in Figure 1 and detailed further in Table 1. Broad profiling assays included: high-throughput phenotypic profiling (HTPP) in U-2 OS cells and in the HIPPTox platforms at ASTAR (described in more detail below under *Hazard Flags*) and high-throughput transcriptomic (HTTr) assessment using the Templated Oligo with Sequencing Readout (TempO-Seq) whole transcriptome assay in U-2 OS, HepaRG, and MCF7 cells (Harrill *et al.*, 2021). The HTPP in U-2 OS, and the HTTr data from all 3 cell lines, were used to inform quantitative estimates of the MBC; the specific biological pathways or putative molecular targets that may be suggested by these broad profiling NAMs were not evaluated in this case study. The minimum phenotype-altering concentration (PAC) from a global and category level analysis of all 1300 features measured was used to summarize the MBC for HTPP in U-2OS (Nyffeler *et al.*, 2021). The minimum biological pathway altering concentration (BPAC) associated with a

super target signature for each cell line were used as quantitative MBC estimates for the HTTr assays by cell line (Harrill*, et al.*, 2021) (all signature concentration-response data for these three cell lines are available for public download at the CompTox Chemicals Dashboard).

A refined targeted NAMs dataset was constructed with the intention to cover key molecular initiating events (MIEs) or processes in order to demonstrate broad biological coverage, rather than using any assay available in the ToxCast database. The targeted NAM data used in this case study are from invitrodb version 3.5 (USEPA, 2022c). ToxCast data are made publicly available via releases of the ToxCast database (https://www.epa.gov/comptox-tools/exploring-toxcast-data) and in the CompTox Chemicals Dashboard (https://comptox.epa.gov/dashboard/). The key MIEs or processes covered included: nuclear receptor and oxidative stress pathways in Attagene (ATG) (Houck *et al.*, 2021; Martin *et al.*, 2010; Medvedev *et al.*, 2018); models of human pathophysiology to include immunosuppression in BioMAP (formerly, BioSeek or BSK) (Houck *et al.*, 2023); *in vitro* pharmacology profiling including cell-free receptor, enzyme, transporter, and ion channel targets in NovaScreen (NVS) (Knudsen *et al.*, 2011; Sipes *et al.*, 2013); acute neuronal microelectrode array (MEA) assays developed within CCTE (Kosnik *et al.*, 2020; Strickland *et al.*, 2018; Valdivia *et al.*, 2014); and, an assay model of developmental toxicity from Stemina (STM) (Zurlinden *et al.*, 2020) (Figure 1). The MBC (in µM units) per targeted assay was defined as the minimum active concentration at the cutoff defined for a positive response (ACC) from the ToxCast Pipeline (Filer *et al.*, 2017) (version 2.1.0). For derivation of the MBC, BPAC and ACC were considered comparable, as they both represent the concentration at the threshold for bioactivity. Some data cleaning steps were taken when leveraging the *in vitro* bioactivity data from invitrodb v3.5, as detailed in the available code (see *Software and Supplemental File Descriptions,* below). Similar to previous work (Paul Friedman*, et al.*, 2020), ToxCast data were filtered to exclude curve-fits with both borderline efficacy and potency values lower than the concentration range screened (as denoted by fit categories 36 and 45 from within level 5 of invitrodb v3.5) as well as curve-fits with greater than equal to

3 caution flags (from level 6 within invitrodb v3.5), as this simple filtering was found to exclude curve-fits with less quantitative reproducibility.

Table 1. Bioactivity NAM Application.

The application of bioactivity NAMs in this case study is described, including what kind of biology is informed, if the NAM is part of a hazard flag, or if the potency is used in calculation of the minimum bioactive concentration (MBC).

Data	Biology informed?	Hazard flag?	Potency used in MBC?	Potency type used to inform MBC
CERAPP, COMPARA, ToxCast ER/AR models	Informing an ER/AR hazard flag	Yes	No	NA
TEST DEV model	Informing a developmental toxicity flag	Yes	No	NA
HIPPTox: HepG2, BEAS-2B, HK-2	Informing target cell type predictions (liver, lung, kidney)	Yes	Yes	EC10
HTPP: U-2 OS	Broad profiling	No	Yes	Minimum PAC
HTTr: U-2 OS, HepaRG, MCF7	Broad profiling	No	Yes	Minimum BPAC for super target signatures per cell line
Attagene (ATG)	Multiplexed pathway profiling platform (nuclear receptors and stress response)	No	Yes	Minimum ACC value by assay from ToxCast database (invitrodb v3.5)
BioMAP	Complex primary cell and co-culture models of inflammation, fibrosis, tissue remodeling, and immune function	Yes	Yes	
NovaScreen (NVS)	Suite of *in vitro* pharmacology, including cell-free binding and biochemical assays, denoted as enzyme activity (ENZ), nuclear receptor (NR) ligand binding, and absorption, distribution, metabolism and excretion (ADME) as indicated by CYP inhibition	No	Yes	
Microelectrode array (MEA)	Indication of acute effects on neuronal cells and their electrical function	Yes	No	
Stemina (STM)	Stem-cell based screening with metabolomic indicator of developmental toxicity	Yes	Yes	

Hazard flags

Endocrine activity

Previously published QSAR and *in vitro* bioactivity models for estrogen and androgen receptor interactions were combined to create flags for qualitative predictions of ER and AR activity. When available, the ToxCast Estrogen Receptor (ER) and Androgen Receptor (AR) models, systems biology models based on *in vitro* bioactivity data from ToxCast (Judson *et al.*, 2015; Kleinstreuer *et al.*, 2017), were used to indicate potential ER and/or AR interactions. When unavailable, consensus QSAR models were used to indicate potential ER and/or AR interactions. The flags developed reflect qualitative values. For each receptor, if the agonist or antagonist modes of the ToxCast ER or AR model were available and positive, a score of 1 was assigned; if consensus QSAR models (Mansouri*, et al.*, 2016; Mansouri*, et al.*, 2020) were positive in any mode (agonist, antagonist, binding), a score of 0.5 was assigned; and if all models were negative, a score of zero was assigned. For the purpose of this qualitative approach, equivocal results for the ToxCast ER model (area under curve [AUC] score < 0.1) and the ToxCast AR model (AUC score < 0.1 or antagonist AUC score > 0.1 but confidence score was ≤ 2) were grouped with negative results for these models.

Developmental toxicity

A flag for developmental toxicity utilized an existing, publicly available QSAR for developmental toxicity from the Toxicity Estimation Software Tool (TEST) (USEPA, 2020) and *in vitro* assay data from the Stemina (STM) devTOX quickPredict (devTOX^{qp}) platform (Palmer *et al.*, 2013) to provide *in silico* and/or *in vitro* indicators of potential developmental toxicity. The QSAR component of the developmental toxicity flag was the TEST developmental toxicity model; this open-source model was developed using a training set of 285 unique chemicals with defined structures associated with human and/or animal developmental toxicity data obtained from the Teratogen Information System and Food and Drug Administration study data (Cassano *et al.*, 2010). In the current TEST developmental toxicity model (USEPA, 2020), a consensus model that averages results from hierarchical clustering, multilinear

regression, and nearest neighbor approaches achieved a 77% balanced accuracy for an external test set. The applicability domain for each model is described in the TEST User's Guide (https://www.epa.gov/sites/default/files/2016-05/documents/600r16058.pdf). The TEST developmental toxicity model is positive if the predicted value is ≥ 0.5. It should be noted that this model has not undergone formal validation, and the training set for this model is small (227 chemicals) and biased toward positive predictions (69% of the training set compounds are positive for developmental toxicity). The TEST DEV model is available in the CompTox Chemicals Dashboard via Batch Search and in the Predict module (https://comptox.epa.gov/dashboard/predictions). The *in vitro* assay component of the developmental toxicity flag was from the STM devTox assay (Palmer*, et al.*, 2013; Zurlinden*, et al.*, 2020) that utilized undifferentiated H9 human embryonic stem cells (hESCs) and indicates developmental toxicity based on decreased ornithine/cystine ratio as a biomarker of teratogenicity potential, likely related to glutathione synthesis and redox balance pathways. The data from STM were processed via the ToxCast Data Pipeline (tcpl) into the ToxCast database (invitrodb, version 3.5) (Filer*, et al.*, 2017; USEPA, 2022c), as described previously by Zurlinden and colleagues (Zurlinden*, et al.*, 2020). Briefly, log2-transformed data for the ornithine/cystine ratio and cell viability were normalized to the vehicle control wells, and responses greater than 3 times the baseline median absolute deviation were considered active; for multi-concentration series, these active data were fit using tcpl curve-fitting (v2.1.0) to derive estimates of potency, including the ACC, which was used as the *in vitro* MBC for STM. The flag for developmental toxicity is qualitative and multi-component and indicates whether the STM assay biomarker was positive (at any potency); selectively positive (i.e., positive at concentrations 0.3 log10-µM lower than concentrations that affected cell viability); and/or the TEST QSAR for developmental toxicity was positive.

Neuroactivity

Acute chemical exposures of primary rat neurocortical cells cultured on microelectrode arrays (MEAs) were used to inform a semi-quantitative flag for neuroactivity that reflects chemical effects on neuronal network activity. Previously, this assay has been shown to be responsive to various classes of neuroactive chemicals, illustrated by changes in the electrical spikes and groups of spikes (bursts) that result from extracellularly-measured neuronal action potentials (Kosnik*, et al.*, 2020; Martin *et al.*, 2024; Valdivia*, et al.*, 2014). In its current configuration, as made available in invitrodb version 3.5, the MEA acute assay contains 3 assay endpoints related to neuronal firing, 5 assay components related to neuronal network bursting activity, 7 assay components related to neuronal network connectivity, where each of these components have 2 endpoints to reflect the ability to increase ("up") or decrease ("dn") activity (see Supplemental File 1). Additionally, 2 assay components related to assessing cell viability are available. Due to the sensitivity of this multi-component acute MEA assay; lack of full coverage for all chemicals in the case study; and the lack of a blood-brain barrier in the bioactivity model or the IVIVE model, a neuroactivity flag was assigned when the estimated minimum neuroactive potency in the MEA (5^{th} percentile of ACC values for all MEA assay endpoints) was the minimum potency observed for a given chemical in the case study. This flag is only assigned based on potency in the MEA (when available) relative to other assay results. Further, the MEA flag was only applied if > 3 assay endpoints in the MEA assay endpoint list were positive in the same direction.

Target cell types from HIPPTox Profiling

For this case study, *in vitro* bioactivity data from the high-throughput imaging-based phenotypic profiling toxicity (HIPPTox) platform were generated based on three human cell lines, namely a bronchial epithelial cell line, BEAS-2B (Lee *et al.*, 2018), a hepatocarcinoma cell line, HepG2, and proximal tubule cell line, HK-2 (Su *et al.*, 2016). A total of 156 phenotypic readouts were quantified from the images of each cell line using the cellXpress software v2.2.2 (Laksameethanasan *et al.*, 2013). For

each cell line, a series of support vector machine models were used to distinguish between chemical- and DMSO-treated cells based on the phenotypic readouts for each of the seven chemical concentrations (0.87-500 µM). These classifiers automatically identified the most discriminative features for each cell line and provided a series of classification accuracy values for all the tested concentrations (Loo *et al.*, 2007). The values were then fit using a standard log-logistic model and a flat constant model, where the best fitted curve was determined using the Akaike Information Criterion as previously described (Miller and Loo, 2020). *In vitro* HIPPTox points-of-departure were defined as the 10% effect concentration (EC_{10}) levels based on the best fitted curves as used as the minimum *in vitro* bioactive concentration for each HIPPTox cell type. HIPPTox values less than 400 µM can be interpreted as positive hits for the respective cell types. In addition to inclusion in derivation of the POD_{NAM}, the presence of *in vitro* positives for lung, liver, or kidney cell toxicity in the HIPPTox models were interpreted semi-quantitatively as flags for putative target tissues, using the potency of the positive hit as the flag value.

Immunosuppression

Assay platform description. The BioMAP panel, now comprised of 12 different assay systems, has been used previously, largely in preliminary toxicity profiling of pharmaceutical and consumer chemicals (Betts *et al.*, 2018; Hammitzsch *et al.*, 2015; O'Mahony *et al.*, 2018; Shah *et al.*, 2017; Simms *et al.*, 2021; Singer *et al.*, 2019). These 12 assay systems include models of autoimmune disease, chronic (vascular) inflammation, allergy, monocyte activation, lung inflammation and fibrosis, cardiovascular inflammation, dermatitis, and wound healing (Kleinstreuer et al. 2014; Houck et al. 2022). Log10-fold change data from the 12-assay BioMAP panel were received via contract from DiscoveRx and processed using the ToxCast Data Pipeline (tcpl) for public release, as previously described (Houck et al. 2022). Briefly, tcpl was used to determine the lowest effective concentrations in the BioMAP panel. Due to the low number of concentrations and replicates used, lowest effective concentrations for these data were defined as the concentration where activity was greater than the threshold cutoff for a positive, and

these values were used in place of a calculated ACC as the minimum *in vitro* bioactive concentration for the BioMAP panel. This threshold cutoff was defined as the maximum of either: three times the median absolute deviation of wells that represented baseline or a 1.2 $\log_{10}$-fold change.

Within the BioMAP panel, specific readouts within 3 different model assay systems were identified as immunosuppression relevant for a semi-quantitative flag of potential immunosuppression. Details of these models were published previously (Kleinstreuer et al. 2014; Houck et al. 2022), including an analysis of the results produced by four immunosuppressive drugs (Houck *et al.*, 2022). The three model systems included in the immunosuppression flag were:

(1) Sag system (T cell activation "super-antigen" model; intended to model autoimmune or chronic inflammation states relevant to T-cell dependent conditions; uses co-cultured primary human peripheral blood mononuclear cells [PBMC] and human umbilical vein endothelial cells [HUVEC] stimulated with superantigens, i.e., T cell receptor [TCR] antigens);

(2) BT system (T cell dependent B cell activation; intended to model autoimmune, allergy, or asthma, or oncology disease states where B-cell activation and antibody production are relevant; uses co-cultured PBMC and CD19+-B cells stimulated with TCR antigens and anti-IgM); and,

(3) Mphg System (macrophage activation response; intended to model chronic inflammation and macrophage activation relevant to conditions involving cardiovascular inflammation, restenosis, and arthritis; uses co-cultured HUVEC cells and macrophages stimulated using toll-like receptor 2 [TLR2] ligands derived from yeast).

The Sag and BT systems provide information on the innate and adaptive immune responses, whereas the Mphg system provides information on macrophage activation; as such, these three systems

are relevant off-the-shelf assays for evaluating immunosuppression-relevant activity of chemicals. Effects on PBMC viability in the Sag and BT systems; decreased B cell proliferation in BT system; decreased T cell proliferation in the Sag system; decreased soluble IgG production in the BT system; and decreased IL-10 production in the Mphg systems were the measured endpoints from these systems considered immunosuppression relevant (assay endpoint identifiers 313, 315, 2810, 2812, 2814, and 2928 in invitrodb v3.5). These immunosuppression-relevant endpoints were selected for their biological relevance and response to four pharmacological immunosuppressive drugs. Azathioprine, methotrexate, and cyclosporin A decreased B cell proliferation and IgG in the BT system (among other cytokines); cyclosporin A decreased cytokine production and T cell proliferation in the Sag system; and dexamethasone decreased soluble interleukin 10 (IL-10) production in the Mphg system (Houck *et al*., 2022). The details and limitations of the BioMAP system for indicating putative immunosuppression-relevant bioactivity were described in Houck *et al.* (2022). Other potential specific pathways or pleiotropic modes of action resulting in immunosuppression may not be captured by the BioMAP panel and cannot be ruled out as contributing to potential immunosuppression *in vivo*. As such, the semi-quantitative flag for potential immunosuppression-relevant activity provides additional information but may not indicate all aspects of immune biology.

Immunosuppression flag. For the analysis herein, a simplified, semi-quantitative flag was developed to reflect selective immunosuppressive activity *in vitro* at the endpoints specified as "immunosuppression-relevant," where selectivity is defined as immunosuppression-relevant bioactivity occurring at concentrations lower than those that elicit overt cytotoxicity in the confluent, adherent BioMAP systems (11 of the 12 systems, excluding the BT system), as evaluated using sulforhodamine B (SRB) staining as a marker of total protein levels. The lowest effective concentration among the immunosuppression-relevant endpoints was subtracted from the minimum lowest effect concentration among the SRB cell viability endpoints. If there were no chemical effects on the SRB cell viability endpoints,

the immunosuppression-relevant lowest effective concentration was subtracted from 3 (equivalent to 1000 µM on a log10-µM scale).

Toxicokinetic NAMs

Data collection

R library *httk* (Pearce *et al.*, 2017) (version 2.3.0) was used for *in vitro* to *in vivo* extrapolation (IVIVE) of human administered equivalent doses (AEDs) in mg/kg/day units from *in vitro* bioactive concentrations in micromolar units utilizing a HTTK approach (Breen *et al.*, 2021). A generalized toxicokinetic model coupled to Monte Carlo simulation of human physiology provided consideration of population variability using data from the US population (Ring *et al.*, 2017). Predictions were made chemical-specific through consideration of structure-based physicochemical parameter predictions (such as hydrophobicity) and chemical-specific *in vitro* toxicokinetic measurements (fraction unbound in plasma and intrinsic hepatic clearance). At the inception of this case study *in vitro* HTTK data already existed within the *httk* R library for many (but not all) of the 201 case study chemicals.

For approximately 51 of the 201 chemicals that lacked empirical HTTK data, two HTTK assays were performed: plasma protein binding and hepatic metabolic clearance assays. These assays were performed via contract to GVK Biosciences (Hyderabad, India) (Paini *et al.*, 2020). The plasma protein binding assay used human plasma (freshly frozen, pooled, mixed gender, with heparin) treated with test chemicals (5 µM) in a 96-well rapid equilibration dialysis (RED) assay platform incubated at 37°C for 4 hr, with three technical and three biological replicates, as described previously (Wetmore *et al.*, 2015; Wetmore *et al.*, 2012). Samples were preserved for analytical detection via high-performance liquid-chromatography with mass spectrometric detection (LC-MS). Percent (%) plasma protein bound was calculated based on the amount of chemical that was dialyzed into the sample side of the RED assay (Waters *et al.*, 2008; Wetmore*, et al.*, 2012). The rate of hepatic metabolism of parent chemical was determined via a time course of incubation of chemical (at 1 and 10 µM) with suspended human

cryopreserved, pooled hepatocytes (0, 15, 30, 60, 90, 120 min), with 3 technical replicates and 3 biological replicates for each chemical concentration. Loss of parent compound as detected using LC-MS was measured to then infer chemical half-life and intrinsic clearance, as reported previously (Shibata *et al.*, 2002; Wetmore*, et al.*, 2015; Wetmore*, et al.*, 2012).

Application of toxicokinetic NAMs for internal exposure

Additional HTTK data collected for the chemicals in this case study, along with existing information in the library *httk* (v2.3.0), resulted in 151 chemicals with hepatic clearance data and 131 chemicals with both hepatic clearance and plasma protein binding empirical data. *In silico* prediction data included in *httk* were also loaded sequentially (httk::load_sipes2017(), httk::load_dawson2021, httk::load_pradeep2020) (Dawson *et al.*, 2021; Pradeep *et al.*, 2020b; Sipes *et al.*, 2017) and enabled AED estimation for most remaining chemicals in the case study. Hepatic clearance data and plasma protein binding data are required for the 3-compartment steady state model (referred to as 3compartmentss in *httk*) and the physiologically-based toxicokinetic (pbtk, as referenced by *httk*) model, but the 3compartmentss model permits use of a default 0.5% fraction unbound for plasma protein binding when empirical data are unavailable, whereas the pbtk model requires empirical information regarding fraction unbound for plasma protein binding. Since *in vitro* toxicokinetic measurements (plasma binding and hepatic clearance) are both presumed to have been attempted for each chemical, when protein binding data are unavailable but clearance data are available, a default assumption of 0.5% unbound can be used, given the assumption that high protein binding is the reason for the measurement to be missing (Rotroff *et al.*, 2010). The 0.5% unbound assumption is considered too imprecise for the pbtk model.

Both *httk* models (3compartmentss and pbtk) were used to perform IVIVE for as many chemicals as possible based on HTTK data availability. IVIVE was performed for each minimum *in vitro* bioactive concentration by assay (as described above in *Bioactivity NAMs*), based on the reverse dosimetry

assumption that the nominal minimum *in vitro* bioactive concentration (µM) is equivalent to a plasma concentration *in vivo*. AED_{50} values, corresponding to the median (50^{th} %ile) individual with respect to toxicokinetic variability, were estimated (*httk*::calc_mc_oral_equiv()) using assumptions of human physiology and restrictive clearance (using the median Css value) based on Monte Carlo simulation of human variability in physiological parameters including liver blood flow and the rate of kidney clearance and measurement uncertainty of *in vitro* toxicokinetic plasma protein binding and intrinsic hepatic clearance (Breen*, et al.*, 2021; Ring*, et al.*, 2017; Wetmore*, et al.*, 2015; Wetmore*, et al.*, 2012). Additionally, the default behavior of calc_mc_oral_equiv() in *httk* v2.3.0 was used, which now includes estimation of the fraction of chemical absorbed across the intestinal wall and the fraction of chemical absorbed from the gut to the portal vein, which when available, revises the fraction of chemical that is bioavailable for distribution and excretion. These estimates are based upon chemical-specific *in vitro* measurement of Caco-2 membrane permeability rate (Darwich *et al.*, 2010) or an *in silico* prediction of that rate. This improves upon previous versions of *httk* (prior to v2.3.0) in which the fraction bioavailable only accounted for the fraction not metabolized in first pass hepatic metabolism and assumed that intestinal and gut absorption were 100% (Honda et al., submitted). Where an AED_{50} was available using the pbtk model, it was used preferentially over the 3compartmentss model in derivation of a minimum AED_{50} by assay source. A minimum AED_{50}, based on the minimum *in vitro* bioactive concentration for the median individual based on a Monte Carlo simulation of human toxicokinetic variability (Ring*, et al.*, 2017) was calculated for each assay source. In Paul Friedman *et al.* (2020), the IVIVE methodology is similar, except that only the 3compartmentss model was run; multiple population quantiles (50^{th} and 95^{th}) were used in that analysis; and, previously estimates of intestinal and gut fraction absorbed were not included in applications of the *httk* library.

Comparison of POD_{NAM} to *in vivo* POD_{trad}

In vivo data were available for a subset of the substances in this case study from the Toxicity Value Database (ToxVal) version 9.4 (USEPA, 2023c). ToxVal was queried by DSSTox substance identifier (DTXSID) and repeated dose effect levels from oral exposures were retrieved for dog, rat, mouse, rabbit, and guinea pig studies, with effect levels including no effect level descriptions (in ToxVal, referred to as no effect level (NEL), no observable effect level (NOEL), no observable adverse effect level (NOAEL), highest no effect level (HNEL); lowest observed effect level descriptions (in ToxVal, referred to as LEL or LOAEL); and benchmark dose descriptions (in ToxVal, referred to as BMD, BMDL, BMDL10). Units were in or converted to mg/kg/day units, or the record was dropped from the dataset. Overall 5th, 10th, 15th, 20th, 25th, and 30th percentile summary values of a ToxVal-based POD were computed for any of these repeated dose data for 165 out of 201 chemicals in the case study (Supplemental File 2, Figure 2). Examination of the differences amongst 5-30th percentile summary values of ToxVal PODs suggested limited differences for this chemical set, and as such 5th percentile (lower bound estimate of a conservative POD value) and the 25th percentile (higher estimate of a conservative POD values) were used in subsequent analyses. The dataset was subset to only studies annotated as “repeated dose” or “subchronic” to generate a ToxVal subchronic POD for 160 chemicals in the case study. To further understand the impact of selecting different summarized *in vivo* PODs, an analysis of the differences between the ToxVal PODs (5th and 25th percentile) and the ToxVal subchronic-only POD values was performed (Supplemental File 2, Figure S2). This analysis suggested that subchronic to chronic POD values in ToxVal were linearly related with coefficient of variation (R^2) values of 0.8-0.9 and with root mean squared error (RMSE) values within 0.5 log_{10}-mg/kg/day, suggesting limited ability to see differences in the use of subchronic-only vs. ToxVal POD values that included all repeated dose study types.

Separate from the ToxVal information retrieval, a repeated dose POD was semi-manually curated via review of the ECHA International Uniform Chemical Information Database (IUCLID) for studies reported as OECD test guidelines 407, 408, and/or the systemic portion of the OECD test guideline 422. The minimum NOAEL or LOAEL value from this review was used as an estimate of a minimum systemic POD for 40 of the 201 substances in this case study (only 2 of these 40 lacked a POD value in ToxVal v9.4). When available, these ECHA IUCLID values were used as a repeated dose POD_{trad} in place of the 5th percentile or 25th percentile from ToxVal v9.4 to calculate POD ratios, as the ECHA IUCLID test guideline study values were considered a more definitive POD value for systemic toxicity in this case study rather than a composite value from many guideline and guideline-like study sources as is available from summary values from ToxVal. For most of the analyses herein, the 5th and 25th percentile summary values of ToxVal PODs (where possible supplemented with the minimum ECHA repeat dose POD value and referred to as POD_{trad}) were used as a comparator for the estimated POD_{NAM} values.

Summarizing AED_{50} values for defining the POD_{NAM}

Several summary values of the minimum AED_{50} value by assay source were attempted, including the minimum, median, multi-linear regression (MLR) model, and random forest (RF) models, using the 5th and 25th percentile ToxVal POD values as the benchmark values to be predicted. Equations 1a and 1b express the computation of the minimum and median, respectively, of the minimum AED_{50} value by assay and assay sets for different definitions of a POD_{NAM} in this work.

Equation 1a: $\min AED_{50} = \min\{\min(AED_{50,1}),\ldots, \min(AED_{50,i})\}$

Equation 1b: $\text{med } AED_{50} = \text{med}\{\min(AED_{50,1}),\ldots, \min(AED_{50,i})\}$,

Where $\min(AED_{50,i})$ are calculated by assay (Table 1) for a total of 12 min(AED50) values. As HTTr, HTPP, HIPPTox used multiple cell lines, a minimum AED_{50} was calculated for each assay-cell line combination. Min AED_{50} and med AED_{50} were also calculated for sets of assays. These assay sets

included all of the assays; only the targeted assays (ATG, BioMAP, NVS, and STM); only the broad profiling assays (HTPP U2-OS, HTTr HepaRG, HTTr MCF7, HTTr U2-OS); only ASTAR assays (ASTAR BEAS-2B, ASTAR HepG2, ASTAR HK-2); the "core" targeted assays (ATG, BioMAP, NVS); and, the core targeted assays plus the broad profiling assays (ATG, BioMAP, NVS, HTPP U2-OS, HTTr HepaRG, HTTr MCF7, HTTr U2-OS). Note that for all assays (including specific assay-cell line combinations) used herein, negative (inactive) results in the assay would result in a "missing" value, such that the assay would not contribute to the overall quantitative estimate of POD_{NAM} using a minimum or median summary value. Given the limited number of chemicals in this case study, and the related limitation of insufficient numbers of chemicals for training, testing, and external validation, as well as the very limited apparent performance gains in using MLR and RF with respect to root mean squared error (RMSE) in training, the emphasis in the analysis reported herein is for the minimum and median of the minimum AED_{50} values by assay and assay sets.

A limitation in a modeling approach to summarization of the AED_{50} values by assay is that not all chemicals are active in all assays. Missing (inactive) values were imputed as the median for the assay for both the MLR and the RF modeling attempts. A linear model (R function lm()) was used to find the coefficients needed to derive MLR models using the minimum AED_{50} in each assay as a covariate. The MLR models were constructed per the form in Equation 2.

Equation 2: $ToxVal_{P,k} = 0 + \text{beta coefficient}_1 * \min(AED_{50,1}) + \ldots \text{beta coefficient}_n * \min(AED_{50,n})$,

where $ToxVal_{P,k}$ is the 5^{th} or 25^{th} percentile from ToxVal values by chemical, predicted by the multi-linear regression model developed using each of the minimum AED_{50} values for the 12 assays in the bioactivity NAM battery (ATG, BioMAP, CCTE MEA, NVS, and STM; HTTr in MCF7, HepaRG, and U2-OS; HTPP in U2-OS; and HIPPTox models from ASTAR in BEAS-2B, HepG2, and HK2 cells). Similarly, RF models were constructed with R library caret (Kuhn, 2008) to predict the ToxVal 5^{th} and 25^{th} percentile

from the minimum AED_{50} values by assay, largely to understand the amount of variance in the ToxVal POD that could be explained by AED_{50} values. The RF models were trained with 12 input predictor values (minimum AED_{50} for each of the 12 bioactivity assays) using repeated cross-validation, with 10 folds and 3 repeats, with the number of variables to randomly sample as candidates at each split (mtry) tuned per model. In training, the MLR models for the 5th and 25th percentile ToxVal POD (RMSE = 1.24, 1.09; R^2 = 0.417, 0.654, respectively) and RF models (RMSE = 1.27, 1.02; R^2 = 0.17, 0.20; optimal mtry = 2, 7, respectively) had only small differences, if any, in RMSE from using a median rather than a model, whereas the median did not require inference of missing data. The R^2 values especially for the MLR models, but also for the RF models, are likely inflated, as the performance cannot be evaluated on an external test set. Given these early observations and limitations, including the limited ability to validate these models within the scope of this case study, further optimization of the MLR and RF models was not undertaken. Future work could be undertaken to employ additional modeling to obtain a POD_{NAM} from *in vitro* data using more chemicals.

POD ratios

POD ratios were calculated several ways to explore expectations on the difference between POD_{NAM} and POD_{trad} as well as to explore the impacts of decisions made in constructing the POD_{NAM} from different summarizations of the minimum AED_{50} by assay. In general, the POD ratio was calculated per Equation 3a.

Equation 3a: $\log10(POD_{trad}) - \log10(POD_{NAM})$,

where for this case study, POD_{NAM} was calculated a number of ways in Table 2 and for Figures 4, 5, 6, and 10. POD_{trad} is represented as indicated as the 5th, 10th, 15th, 20th, 25th, or 30th percentile of ToxVal PODs (unless minimum repeat dose ECHA values were available, in which case the minimum repeat dose ECHA value was used). In particular, the 5th and 25th percentile POD were used for many of

the comparative analyses (Figures 4, 5, and 10). Additional POD ratios were calculated to compare POD_{trad} to the threshold POD value assigned by TTC (POD_{TTC}), as well as the POD value calculated for only subchronic studies in ToxValDB (POD_{SUB}), per Equations 3b and 3c, respectively.

Equation 3b: POD_{TTC} ratio = $\log10(POD_{trad}) - \log10(POD_{TTC})$

Equation 3c: POD_{SUB} ratio = $\log10(POD_{trad}) - \log10(POD_{SUB})$

Bioactivity:exposure ratio (BER)

The upper limit on the credible interval (95th-percentile) for total population exposure was estimated using the consensus meta-model developed using the Systematic Empirical Evaluation of Models framework version 3 (SEEM3) (Ring*, et al.*, 2019). Log10-bioactivity:exposure ratios (BER) were calculated per Equation 4.

Equation 4: $BER = \log_{10}(\text{med } AED_{50}) - \log_{10}(SEEMU95)$,

Where med AED_{50} is defined by Equation 1b as-the median AED_{50} (median of minimums by *in vitro* assay) and SEEM3U95 represents the upper 95th percentile on the credible interval for prediction of median total US population exposure from SEEM3, all in log10-mg/kg/day units. It is important to note that the 95th percentile in this case reflects uncertainty in estimation of the median population value and does not take variability in human exposure into account.

Software and Supplemental File Descriptions

The code (produced with R version 4.4.1) and source data are all publicly available at EPA GitHub (https://github.com/USEPA/CompTox-APCRA-pro).

Supplemental File 1 is an Excel file that contains 10 tables, described in a README tab of the file in detail. The primary output from this study is provided in Supplemental File 1 Table S3 POD BER Flags Summary. Table S1 is a table of the AQC grades and flags for Tox21. Table S2 is the information used to define the applicability domain for chemical screening in *in vitro* aqueous-based assays. Table S4 is all of

the calculated POD ratios. Table S5 is the information used for the DART flag. Table S6 is the information used for the ER and AR flags. Table S7 is the information used for the developmental toxicity hazard flag. Table S8 is the information used for the target cell type flag. Table S9 is more detailed information from the BioMAP platform. Table S10 is more detailed information from the CCTE-MEA assay platform.

Supplemental File 2 contains all Supplemental Figures. Supplemental Figure 1 shows silicon-oxygen bond containing structures in the case study. Supplemental Figure 2 shows the distribution of all ToxVal POD values and the 5-30th percentile summary POD values by chemical from ToxVal version 9.4 for the chemicals in this case study. Supplemental Figure 3 shows a comparison of different summary values for *in vivo* POD in this case study. Supplemental Figure 4 shows a linear comparison of minimum *in vitro* bioactive concentrations from different *in vitro* NAMs to the 5th percentile ToxCast ACC value for the chemicals in this case study. Supplemental Figure 5 shows the size of the BER vs. ExpoCast SEEM3 exposure prediction credible interval size. Supplemental Figure 6 shows HepaRG potency relative to other estimates of dose.

Results

Chemicals evaluated

Initially, several considerations drove the selection of chemicals for this case study. Approximately half of the chemicals from the APCRA retrospective case study of data-rich chemicals (Paul Friedman*, et al.*, 2020) were carried over in order to evaluate the performance of the reduced bioactivity NAM battery proposed herein via ensuring that there would be *in vivo* POD information for comparison. For the other half of the chemicals included in this case study, importance to different regulatory authorities and presence in the ToxCast chemicals inventory but with relatively limited *in vivo* data coverage were the main criteria. Over the course of the case study, other efforts aimed at defining the applicability domain for *in vitro* screening matured, including available information on the AQC for chemicals in the ToxCast chemical library (Williams *et al*., in prep). In Figure 2A, the 24 (out of 201)

chemicals that did not fully pass AQC are shown. Twenty-two of these 24 chemicals are only in the prospective case study (did not overlap with the retrospective). Two nitrates (calcium and potassium nitrate) and cadmium chloride are inorganic salts for which OPERA physicochemical predictions cannot be generated simply based on descriptor coverage, and none of the applied analytical techniques for AQC would be applicable (i.e., liquid and gas chromatography with mass spectrometry and proton nuclear magnetic resonance would fail). Additionally, there are six silicon-containing chemicals within these 24 chemicals that did not fully pass AQC (eight silicon-containing substances in the case study overall), each containing multiple silicon-oxygen bonds, which are a category of chemicals noteworthy for discussion in terms of AQC (Supplement File 2, Supplemental Figure 1). These chemicals demonstrate a general pattern of degradation over time at room temperature in DMSO solvent based on AQC measurements. Silicon-oxygen-carbon bonds are known to be hydrolytically unstable, and hydrolysis and condensation reactions are common in this class of chemicals as exemplified by the instability of tetraethyl orthosilicate (Kaur *et al.*, 2022). Since DMSO is hygroscopic, and the samples contained water as evidenced by the large peak in the nuclear magnetic resonance spectra, instability based on hydrolytic attack on the silicon-oxygen bonds may be expected. This is borne out by the degradation observed through a combination of analytical techniques. Overall, chemicals containing silicon-oxygen-carbon bonds, when dissolved and stored in DMSO or applied to aqueous media, likely degrade via hydrolysis. The other 16 of the 22 substances have cautions associated with their AQC (caution definition was fairly permissive, as defined in the Methods – Cheminformatics section), indicating that resultant bioactivity might be due to the parent, one or more degradants, metabolites, or contaminants, and that the concentration of the parent chemical associated with any bioactivity has additional uncertainty. These results underscore the concept that some classes of chemicals once solvated may not remain as the parent chemical; however, this finding may be true once these chemicals are introduced to any aqueous environment (including *in vivo*), and as such, AQC may provide

context for understanding bioactivity results (or lack thereof) but may not always indicate that observed bioactivity should be disregarded and rather that the bioactivity should be qualified.

In general, most chemicals with cautions on the AQC data would not have been identified *a priori* based on molecular weight, logP, or vapor pressure, suggesting the importance of AQC verification of DMSO- or otherwise-solvated samples prior to screening. A Uniform Manifold Approximation and Projection (UMAP) to reduce the feature dimensionality of molecular weight and predicted logP, vapor pressure, and water solubility (Figure 2C) shows that chemicals with cautions on the AQC (as labeled in Figure 2C) distribute throughout the chemical space interrogated for the prospective case study (201 chemicals), retrospective case study (448 chemicals) and union of the case studies (both, 96 chemicals). The overlap suggests that a simple screening for physicochemical properties (molecular weight between 100 and 500 g/mol, vapor pressure < 100 mmHg, logP < 6.5, and measures of solubility or melting point) compatible with aqueous, cell-based assay bioavailability, would be insufficient to identify chemicals that might be unstable in DMSO or possess other properties that would result in transformation or loss of the chemical from the sample. The UMAP representation of these properties also failed to separate chemicals from the prospective and retrospective case studies, suggesting that the physicochemical property prediction space was similar between case studies. The results for evaluation of the applicability domain for chemicals in this case study suggest that future efforts should include not only physicochemical property and AQC amenability predictions, but also encoded structure alerts for structural moieties that may be related to transformation or degradation.

In an attempt to characterize the breadth of chemicals included in the prospective case study, the predicted exposure pathways used in a consensus model for total U.S. population exposure estimates (Ring*, et al.*, 2019) were used to indicate approximate function and exposure pathway. In the prospective case study, efforts were made to include chemicals with consumer and industrial uses. However, much of the overlap between the prospective and retrospective case studies comes from

chemicals with at least one predicted use related to pesticidal action (Figure 2B), resultant to selecting chemicals for this case study that were already in the ToxCast chemical library. Note that some chemicals had different combinations of predicted exposure pathways, and some chemicals had unknown exposure pathways or were not included in public releases from the model (designated as NA).

NAM battery results

In vitro potency for the case study chemicals generally spanned approximately 5 orders of magnitude (0.001-100 µM), with some outliers, across all of the assays employed (Figure 3A). All of the chemicals selected demonstrated some *in vitro* bioactivity, even those chemicals that suggest major loss over time or degradation of chemical sample in DMSO stock solution (Figure 3B). The relative sensitivity for *in vitro* PODs across different groups of bioactivity assays may inform selection of a panel of assays that could be applied for prospective chemical assessment; interestingly, the general potency distributions for the targeted NAM assay set and the Tier 1 HTTr and HTPP assay set were similar overall, and fairly similar to the 5th percentile ACC from all ToxCast assay endpoints as used in previously published work (Paul Friedman et al. 2020). However, on a chemical-specific basis, no one assay defined the lowest bioactive concentration for all chemicals (Figure 3C and 3D); i.e., no one assay could serve as a potency sentinel because no one assay contained all relevant biology and/or maximum sensitivity. Cell-free assays in the NVS suite and the acute MEA defined the lowest bioactive concentrations most frequently, but ATG, HTTr in U-2 OS cells, BioMAP (primary cell systems), STM, and HTPP in U-2 OS cells defined the minimum potency for some number of chemicals in this case study, with respective descending frequency (Figure 3D). The finding that NVS defined the minimum bioactive concentration (MBC) most frequently is not unexpected, as this assay suite covers many specific pharmacological targets that may not be present in other assays, and *in vitro* disposition of chemicals in these assays may be different from those assays that incorporate cells where diffusion or transport must be present for

chemicals to access the primary target in the assay. The potency values observed in the A*STAR HIPPTox assays tended to be higher for all chemicals with positive responses (10-100 µM, Figure 3A and Supplemental File 2 Supplemental Figure 4), in part because these assays were developed to provide broad bioactivity coverage (see Methods for HIPPTox). Within the HTTr assays, the U-2 OS cell line seemed to provide a higher frequency of positive responses relative to the other cell lines tested, but the signature analyses for all three cell lines (U-2 OS, MCF7, and HepaRG) appeared to result in minimum potencies typically between 1 and 100 µM, with very few substances resulting in potency at lower concentrations (0.01 to 1 µM) (Supplemental File 2, Supplemental Figure 6).

For the purposes of this case study, chemicals that were unlikely to be present due to degradation (as determined by analytical measurements) present a challenge to the domain of applicability for *in vitro* NAM screening. For samples with a parent chemical constituent that degrades, transforms, or otherwise lacks initial purity or correct molecular weight identity, it is unclear how to uniformly evaluate the bioactivity data. As illustrated in Figure 3B, chemicals with caution flags for their AQC data often still had bioactivity in several bioactivity NAMs, suggesting that the parent chemical or some degradants may be both bioactive and present in the bioassay wells (which were not directly sampled for AQC). For these 22 chemicals with cautions on the AQC that also only appear in the APCRA prospective case study, the *in vitro* results are likely less informative for identification of chemicals that could be further tested without some additional consideration (e.g., what is the most appropriate model system for the specific chemistry? What degradants may be generated that would be active?).

The concept of combining multiple assays to derive a protective POD_{NAM} is consistent with previous APCRA work (Paul Friedman*, et al.*, 2020) and other case studies for NBA (Baltazar*, et al.*, 2020; Middleton *et al.*, 2022; Thomas *et al.*, 2013a; Wetmore *et al.*, 2013). It was possible to derive AED_{50} values using empirical HTTK data for 131 chemicals using the pbtk model in R library *httk* and an additional 20 chemicals using the R library *httk* 3-compartment steady state model. Once *in silico*

models for hepatic clearance were loaded in library *httk*, 196 of the 201 substances in the case study list had sufficient data to compute AED_{50} values (bioactivity and HTTK data). The best and most practical means of combining these AED_{50} to form a chemical-level POD_{NAM} were investigated via several comparisons of the calculated POD_{NAM} (i.e., different summarizations of the min AED_{50} by assay) to a POD_{trad} from ToxVal (e.g., 5^{th} to 30^{th} percentile value for repeated dose data for each chemical; available for 164 chemicals). The objective of these comparisons was to understand maximal predictive performance and protective performance of different potential definitions of POD_{NAM}, meant to define the dose corresponding to the threshold concentration for bioactivity, relative to different potential definitions of POD_{trad}, meant to rapidly define the threshold dose for *in vivo* effects. In Table 2, a number of these comparisons of summarized AED_{50} and *in vivo* POD_{trad} definitions are provided, including: the minimum and median of the all assay MBC values (per Table 1 and Equations 1a and 1b); the minimum and median of only targeted assay MBC values (ATG, BioMAP, NVS and STM); the minimum and median of only broad profiling assay MBC values (HTPP U2-OS, HTTr HepaRG, HTTr MCF7, HTTr U2-OS); the MBC per individual assay; and, the results of MLR models trained using MBCs from all 12 assays and the indicated POD_{trad} percentile. Additionally, the predictive results for random forest models are provided (though these models likely demonstrate inflated performance due to the small training set and imputed missing values).

Table 2. Results of linear and direct comparisons of summary AED_{50} and summary ToxVal POD_{trad} values.

The summary AED_{50} value (min = minimum, med = median, mlr = multi-linear regression, rf = random forest) and the summary *in vivo* POD_{trad} percentile are provided for comparison, where each row is a separate comparison. The terms "min AED_{50}" and "med AED_{50}" express the computed values of the minimum and median, respectively, of the minimum AED50 value by assay and by assay sets for different definitions of a POD_{NAM} in this work. The column "Which assays?" indicates the assay minimum bioactive concentration (MBC) values used in the summary AED_{50}. In the section to evaluate whether the POD_{NAM} definitions were predictive, the listed summary AED50 value is linearly compared to the POD_{trad} percentile, with root mean square error RMSE and R^2 based on a linear model to relate these values. A direct comparison of the summary AED50 and the POD_{trad}, using a root mean squared difference (RMSD), is also provided. In the section designed to indicate whether the POD ratio would be protective (i.e., the POD_{NAM} less than or within a certain range of the POD_{trad}), the following values are tabulated: count = number of non-NA POD ratios for the listed summary AED50 and POD_{trad}; # Greater than 0 = number of POD ratios ≥ 0 on a log10-mg/kg/day scale; # Within ± 2 = number of POD ratios within ± 2 on a log10=mg/kg/day scale; # Greater than -2 = number of POD ratios ≥ -2 on a log10-mg/kg/day scale; % Greater than 0 = percent of POD ratios ≥ 0 on a log10-mg/kg/day scale; % Within ± 2 =% of POD ratios within ± 2 on a log10=mg/kg/day scale; % Greater than -2 = % of POD ratios ≥ -2 on a log10-mg/kg/day scale. The rows containing the med AED50 for all assays compared to the 5^{th} and 25^{th} of POD_{trad} values are in boldface font to indicate the POD_{NAM} vs. POD_{trad} that is used in much of the subsequent analysis in this work.

Which assays?	Summary AED50	POD_{trad} percentile	Predictive POD_{NAM}?			Protective POD ratio?						
			RMSE	R2	RMSD	Count	# Greater than 0	# Within ± 2	# Greater than -2	% Greater than 0	% Within ± 2	% Greater than -2
All	min AED50	5th	1.264	0.149	2.226	158	140	90	155	88.6	57	98.1
	min AED50	10th	1.201	0.151	2.326	158	140	83	155	88.6	52.5	98.1
	min AED50	15th	1.138	0.154	2.433	158	143	81	156	90.5	51.3	98.7
	min AED50	20th	1.091	0.154	2.529	158	146	77	158	92.4	48.7	100
	min AED50	25th	1.039	0.149	2.625	158	148	69	158	93.7	43.7	100
	min AED50	30th	0.998	0.153	2.689	158	150	66	158	94.9	41.8	100
	med AED50	**5th**	**1.278**	**0.13**	**1.782**	**158**	**84**	**134**	**146**	**53.2**	**84.8**	**92.4**
	med AED50	10th	1.207	0.142	1.803	158	94	135	148	59.5	85.4	93.7
	med AED50	15th	1.143	0.147	1.853	158	104	134	150	65.8	84.8	94.9
	med AED50	20th	1.089	0.157	1.9	158	108	133	152	68.4	84.2	96.2
	med AED50	**25th**	**1.035**	**0.156**	**1.962**	**158**	**110**	**131**	**153**	**69.6**	**82.9**	**96.8**
	med AED50	30th	0.994	0.16	2.006	158	112	133	156	70.9	84.2	98.7
Broad profiling	min AED50	5th	1.317	0.076	1.617	158	45	117	121	28.5	74.1	76.6
	min AED50	10th	1.248	0.083	1.622	158	49	122	126	31	77.2	79.7
	min AED50	15th	1.18	0.09	1.648	158	56	127	131	35.4	80.4	82.9
	min AED50	20th	1.124	0.102	1.679	158	62	131	136	39.2	82.9	86.1
	min AED50	25th	1.063	0.11	1.718	158	67	135	140	42.4	85.4	88.6
	min AED50	30th	1.02	0.116	1.753	158	70	135	140	44.3	85.4	88.6
Targeted	min AED50	5th	1.244	0.175	2.165	145	84	118	138	53.2	74.7	87.3
	min AED50	10th	1.184	0.174	2.263	145	91	116	139	57.6	73.4	88
	min AED50	15th	1.124	0.175	2.369	145	102	116	140	64.6	73.4	88.6
	min AED50	20th	1.079	0.174	2.463	145	106	116	141	67.1	73.4	89.2
	min AED50	25th	1.028	0.167	2.558	145	109	114	142	69	72.2	89.9
	min AED50	30th	0.988	0.17	2.622	145	112	114	143	70.9	72.2	90.5
Broad profiling	med AED50	5th	1.326	0.062	1.672	158	68	128	139	43	81	88
	med AED50	10th	1.255	0.073	1.577	158	77	130	142	48.7	82.3	89.9
	med AED50	15th	1.189	0.076	1.515	158	84	131	144	53.2	82.9	91.1
	med AED50	20th	1.138	0.08	1.479	158	90	132	147	57	83.5	93
	med AED50	25th	1.08	0.081	1.456	158	92	133	150	58.2	84.2	94.9
	med AED50	30th	1.038	0.085	1.445	158	94	132	151	59.5	83.5	95.6
Targeted	med AED50	5th	1.278	0.13	1.782	145	108	105	139	68.4	66.5	88
	med AED50	10th	1.207	0.142	1.803	145	113	101	140	71.5	63.9	88.6
	med AED50	15th	1.143	0.147	1.853	145	122	100	141	77.2	63.3	89.2
	med AED50	20th	1.089	0.157	1.9	145	125	98	142	79.1	62	89.9
	med AED50	25th	1.035	0.156	1.962	145	127	93	143	80.4	58.9	90.5
	med AED50	30th	0.994	0.16	2.006	145	127	89	144	80.4	56.3	91.1

Individual assays	ATG AED50	5th	1.278	0.13	1.782	123	87	93	119	55.1	58.9	75.3
	BSK AED50	5th	1.294	0.107	1.533	117	75	94	111	47.5	59.5	70.3
	NVS AED50	5th	1.269	0.142	2.024	101	81	62	99	51.3	39.2	62.7
	STM AED50	5th	1.296	0.105	1.524	30	19	26	30	12	16.5	19
	HTPP U2-OS AED50	5th	1.287	0.118	1.513	93	53	82	91	33.5	51.9	57.6
	HTTr HepaRG AED50	5th	1.304	0.094	1.88	156	48	115	120	30.4	72.8	75.9
	HTTr MCF7 AED50	5th	1.326	0.062	1.672	119	32	97	101	20.3	61.4	63.9
	HTTr U2-OS AED50	5th	1.336	0.049	1.64	156	100	119	144	63.3	75.3	91.1
	ASTAR BEAS-2B AED50	5th	1.319	0.073	1.565	61	17	44	50	10.8	27.8	31.6
	ASTAR HepG2 AED50	5th	1.336	0.048	1.615	48	14	40	42	8.9	25.3	26.6
	ASTAR HK-2 AED50	5th	1.326	0.063	1.549	51	9	37	39	5.7	23.4	24.7
Models	mlr AED50 trained to 5 (not pictured in Figure 4B)	5th	1.229	0.195	1.237	164	129	117	161	81.6	74.1	98.2
	mlr AED50 trained to 25	25th	1.04	0.148	1.087	164	145	103	164	91.8	65.2	100
	mlr AED50 trained to 25 eval 5 (pictured in Figure 4B)	5th	1.284	0.122	NA							
	rf AED50	5th	1.273	0.171								
	rf AED50	25th	1.031	0.206								

The results in Table 2 provided several learnings that informed further analysis in this case study. First, in terms of predictive performance, linear models constructed using different putative definitions of POD_{NAM} and POD_{trad} demonstrated RMSE that ranged 0.99 to 1.34 and coefficients of determination (R^2) that ranged from approximate 0.1 to 0.2. These results suggest POD_{NAM} values, regardless of how they were defined, explained only a small amount of variation in POD_{trad} values, and that the error on these linear models would place a majority of POD_{NAM} values within ± 1 to 1.3 log_{10}-mg/kg/day of the POD_{trad} value. Second, the RMSE values obtained appear to trend lower for higher percentiles of POD_{trad}, but with little difference based on how the AED_{50} may be summarized, suggesting that lower POD_{trad} values may represent noisier or more extreme values. In addition to calculating RMSE and R^2 for linear model comparisons, we also calculated a root mean squared difference (RMSD) as a means of directly comparing summary AED_{50} and summary POD_{trad} values (i.e., by calculating how far away from the unity line these values tend to be) (Table 2). When examining all assays in the set, the RMSD values suggest that the median AED_{50} values, especially when compared to the 5^{th} to 15^{th}

percentile POD_{trad} values, are a better direct approximation of POD_{trad} values than the minimum AED_{50} values, whereas the RMSD values for the minimum AED_{50} value tend to be larger, further suggesting the median AED_{50} value may provide a more plausible POD_{NAM} value than using the minimum AED_{50} for a set of heterogeneous assays. A similar trend was observed for the minimum and median AED_{50} values for targeted assays and broad profiling assays alone, where the RMSD values suggest that the median AED_{50} values were closer to the POD_{trad} values. The linear comparisons of all assays combined slightly outperformed subsets of the assays from a predictive perspective: for all assays combined, the RMSE ranged 0.99-1.28, the R^2 was 0.13-0.16, and the RMSD ranged 1.78 to 1.96. In contrast, the RMSE values ranged 1.04 to 1.33 and 1.04 to 1.28, the R^2 values ranged 0.062 to 0.085 and 0.14 to 0.16, and the RMSD values ranged 1.45 to 1.67 and 1.78 to 2, for the medians of the broad and targeted assays alone, respectively, with range dependent on the and the POD_{trad} used, where typically the higher POD_{trad} percentile corresponded to slightly improved linear performance. With respect to protective performance, here again performance was very similar but slightly improved when using all assays rather than subsets of assays: using the median AED_{50} for all assay values resulted in 92.4-98.7% of POD ratios greater than or equal to -2, indicating that nearly all POD_{NAM} were no more than 2 orders of magnitude greater than the POD_{trad}. For comparison, using the median of the broad profiling and targeted assays, 88-96% and 88-91% of POD ratios were greater than or equal to -2. The median AED_{50} from the targeted assays were slightly more conservative on average when compared to the median AED_{50} from the broad profiling assays, with 68 to 80% of the POD ratios greater than or equal to zero (meaning 68-80% of these POD_{NAM} would be equal or less than the POD_{trad}; in contrast, 43-60% of the POD ratios were greater than or equal to zero for the median AED_{50} of broad profiling assays alone. The median AED_{50} for all assays resulted in 53-71% POD ratios greater than or equal to zero. For all POD_{trad} percentiles, the percent of POD_{NAM} within ± 2 log10-mg/kg/day of the POD_{trad} for the median AED_{50} from all assays was 83-85%; the median AED_{50} of the broad profiling assays alone produced a similar result

(81-84%), but the median AED_{50} of targeted assays alone produced a more conservative result wherein only 56-64% of the POD_{NAM} produced where within ± 2 log10-mg/kg/day of the POD_{trad}. Balancing a desire for: (1) error in a more protective direction; (2) reduction in extreme POD_{NAM} values that might be produced by using a summary minimum AED_{50} value; (3) values that are largely within ± 2 log10-mg/kg/day of the POD_{trad} (as one type of benchmarking); (4) greater coverage of biology so as to mimic a repeated dose study; and, (5) the availability of the data generated for this case study, a median AED_{50} for the assay battery was used as a primary comparator in additional analyses along with 5^{th} and 25^{th} percentile POD_{trad} values.

The findings in Table 2 are further visualized in Figure 4. In Figure 4A, as expected, we confirmed that no single assay produced information equivalent to the POD_{trad} (5^{th} or 25^{th} percentile from ToxVal) for all chemicals. In Figure 4B, the minimum, median, and multi-linear regression (MLR) model prediction using all AED_{50} values by assay (see Equations 1a-b and 2 in Methods) are compared to the ToxVal 5^{th} and 25^{th} percentile POD_{trad}, where the minimum, median, and MLR model (based on all assays) demonstrate roughly equivalent coefficients of determination (R^2 values of 0.12-0.20) (see Table 2 for all). Further examination reveals that the median AED_{50} values span roughly 8 log10 orders of magnitude (-3.1 to 4.7 log10-mg/kg/day, or 0.0007 to 48,000 mg/kg/day). In comparison, the MLR model predicts AED_{50} values within a narrower range of 3 log10 order of magnitude (0 to 2.7 log10-mg/kg/day, or 1 to 500 mg/kg/day) to maximize performance. The MLR model training results suggest over-fitting, require inference of missing values, and have few chemicals with which to inform training and test set results, leading to de-emphasis in this work of the MLR model results. Consequently, these results support use of a simple median of the minimum AED_{50} by source as the POD_{NAM} for use in benchmarking quantitative POD_{NAM} performance at this time for this case study. In Figure 4C, the results for the minimum and median of the minimum AED_{50} values for broad profiling assays only (HTPP U2-OS, HTTr HepaRG, HTTr MCF7, HTTr U2-OS) and targeted assays only (ATG, BSK, NVS, and STM) are

visualized. Here we observed that the minimum of the targeted assay AED_{50} values was likely over-protective when compared to the median of the assay minimums. For the broad profiling assays, the choice of minimum and median of the minimum assay AED_{50} values made less of a difference on the RMSE, R^2, and RMSD. Given that the RMSE on any of these linear model comparisons between POD_{NAM} and POD_{trad} ranged from 1-1.3 log10-mg/kg/day, with a low R^2 typically less than 0.2, and an RMSD between about 1.5-2 log10-mg/kg/day, there are multiple choices in the derivation of a POD_{NAM} that have similar performance. A median of the minimum AED_{50} values from combined broad profiling and targeted assays appears to produce POD_{NAM} that are largely within ± 2 log10-mg/kg/day of the POD_{trad}, where values outside of this range are predominantly over-protective, and with a similar small amount of variance explained by a linear model comparing this POD_{NAM} to POD_{trad}. Another clear finding in this benchmarking exercise is a need for a consistent methodology to select an appropriate percentile from available traditional animal toxicity information and/or to rapidly develop or model a POD_{trad} value for benchmarking POD_{NAM} values. This might condense the many options available in benchmarking POD_{NAM} values to existing values from traditional animal models if a calibrated POD_{trad} could be used.

Expectations on PODs

As the NBA workflow presented herein is expected to be iteratively improved over time, with the possible addition and subtraction of different assays, we explored the impact of different groupings of AED_{50} in the derivation of the POD_{NAM} used to calculate the POD ratio, using the 5^{th} and 25^{th} ToxVal POD_{trad} values (Figure 5). Generally, the AED_{50} values from the HIPPTox platform were higher, resulting in a POD_{NAM} that was higher, and thus a log10-POD ratio that was lower (median approach -1 log10-mg/kg/day). The selected POD_{NAM} for this case study, the median of all minimum AED_{50} by assay, resulted in a median POD ratio of 0.14 log10-mg/kg/day. Similarly, the median POD ratio for POD_{NAM} based on the median of all broad profiling assays (HTTr in 3 cell lines and HTPP in one cell line) approached 0. In general, the inclusion of the core targeted NAMs (ATG, BioMAP, NVS) or all of the

targeted NAMs (ATG, BioMAP, CCTE MEA, NVS, and STM) resulted in more conservative median POD_{NAM} values, with log10-POD ratios that appear between 0 and 1 for the ToxVal 5th percentile POD_{trad}. Calculating the POD ratio using the ToxVal 25th percentile POD_{trad} resulted in slightly higher POD ratio values, as the POD_{NAM} appears slightly more conservative in comparison to the 25th percentile POD_{trad} value than the 5th percentile POD_{trad} value. For the purposes of further analysis, the median of the minimum AED_{50} values by source was used for further comparison in POD ratios and BER calculations.

We further compared the log10-POD ratio to other sources of potential POD ratios, including how well a POD from TTC (POD_{TTC}) and a POD from only *in vivo* subchronic studies (POD_{SUB}), might compare to the POD_{trad}. The log10-POD ratio using the median POD_{NAM} and the 5th percentile ToxVal POD_{trad} demonstrated a 10th percentile, 25th percentile, 50th percentile, 75th percentile, and 90th percentile of -1.7, -0.69, 0.14, 1.17, and 1.9 log10-mg/kg/day (calculated from the log10-POD ratio distribution, with distribution visualized in Figure 6). The distribution of this log10-POD ratio (log10-POD_{trad} minus the log10-POD_{NAM}, where POD_{NAM} was defined as in Equation 1b as the median of all assay minimum AED_{50} values) for the 158 chemicals with sufficient data to calculate this ratio demonstrates long tails, where only 10 substances demonstrate a POD ratio of ≤ -3 or ≥ 3 (Figure 6A and 6B). To put this into context, the SUB ratio was calculated as log10-POD_{trad} – log10-POD_{SUB}, using the 5th percentile values from ToxVal. Due to the nature of the chemicals included and the dataset available, much of the POD_{trad} was already based on a POD_{SUB}, resulting in a SUB ratio of zero for many chemicals in the case study. For those few substances where other non-SUB data informed the POD, the tails of the SUB ratio extend from approximately -2.5 to 1.5. Additionally, a TTC ratio was calculated as 5th percentile ToxVal POD_{trad} – TTC. This TTC ratio distribution also has long tails (approximately -1 to 7), with a median of 3.6 log10-mg/kg/day. As expected, even the 5th percentile POD_{trad} value tends to be much greater than the TTC value, and the median POD ratio for POD_{trad}:POD_{TTC} (3.6) is much greater than the median

POD_{trad}:POD_{NAM} ratio (0.14), indicating that the POD_{NAM} provided a value that was much less conservative than the POD_{TTC}.

For the 10 chemicals where the POD ratio was ≤ -3 or ≥ 3, there are some informative observations (Figure 6B): *in silico* models may help evaluate known structure-toxicity associations; some chemicals may not be amenable to *in vitro* methods; chemicals with large disparities between the range of *in vivo* POD values could be manually reviewed; and, limited *in vivo* data may indicate a chemical where it is not necessarily a good benchmark chemical for evaluating POD_{NAM} performance. Only 3 of these 10 chemicals have POD_{NAM} that are not conservative enough to be protective for a computed POD_{trad} value. Of these three, Aldicarb (DTXSID0039223) and dimethoate (DTXSID7020479) are well-characterized insecticides that based on chemical structure and indicated use would likely be managed using an *in silico* approach like TTC or read-across, in addition to existing *in vivo* data, as it has been previously reported that *in vitro* NAM-based POD values based on a broad battery fail to be conservative enough for some carbamate and organophosphate insecticides with studies specific to identifying reduced cholinesterase activity (Paul Friedman*, et al.*, 2020). Propylsilanetriyl triacetate (DTXSID0044608) is missing AQC information, but other similar siloxanes were determined during AQC to undergo chemical transformation in a DMSO sample via hydrolysis, and, as a result of this observation, it is unclear if the POD_{NAM} would be reliable prospectively for siloxanes. Siloxanes are considered corrosive substances, and the data used from the ECHA IUCLID dossier is largely for a read-across analogue (sodium acetate), and as such the anchoring POD_{trad} is not based on empirical data. Thus, for these 3 chemicals were POD_{NAM} was not conservative enough, two of the chemicals would be better handled by an *in silico* structure alert and 1 chemical may not have been amenable to *in vitro* screening and was associated with POD_{trad} data from read-across rather than empirical studies.

The remaining 7 substances with POD ratio ≥ |3| suggest that the POD_{NAM} was overly conservative when compared to our estimate of an animal-based POD_{trad}. For some substances, the

conservatism inherent in the POD_{NAM} may be due to a combination of the *in vitro* potency, the IVIVE approach taken, and data underlying the POD_{trad} estimate. Here, we take a closer look at some of the POD_{trad} values to try to understand potential limitations in available data and why POD_{NAM} may have been overly conservative. A few of these chemicals were well-characterized previously using traditional data. The 5^{th} percentile POD_{trad} for Di-n-octyl phthalate (DTXSID1021956) of 1.57 log10-mg/kg/day aligns with an available EPA Provisional Peer Reviewed Toxicity Value (PPRTV) of 1.57 log10-mg/kg/day (based on histopathological changes in the liver). And, *in vitro* ATG assay endpoints (see Methods – Table 1) related to peroxisome proliferator-activator receptor gamma and hypoxia-inducible factor 1-alpha were positive for di-n-octyl phthalate, but with doses estimated using IVIVE that were approximately 3 orders of magnitude lower than the POD_{trad}. Bifenthrin (DTXSID9020160) is a pyrethroid insecticide with potent *in vitro* activity for the acute MEA and transporters such as the dopamine transporter in NVS *in vitro*, with a POD_{NAM} of -3.03 log10-mg/kg/day, but with a POD_{trad} of 0.18 log10-mg/kg/day for this case study that aligns with the historical (now deprecated) IRIS NOEL value of 0.17 log10-mg/kg/day. There were also a few chemicals with less empirical information available. Bithionol (DTXSID9021342) is a soluble adenyl cyclase inhibitor, and thus has *in vitro* effects including cytotoxicity *in vitro*, and was withdrawn from use in topical drugs due to photosensitization. Curated information on bithionol is extremely limited; ECHA IUCLID lists a repeated dose POD_{trad} as 2.0 log10-mg/kg/day. 2-Butyloctan-1-ol (DTXSID0044818) has a POD_{trad} of 3.0 log10-mg/kg/day based on a subchronic study from ECHA IUCLID. (-)-Ambroxide (DTXSID0047113) has a POD_{trad} of 2.91 log10-mg/kg/day, based on the ECHA IUCLID values of 2.90 and 3 log10-mg/kg/day from one short-term and one subchronic study, respectively. Octrizole (DTXSID9027522) appears to be associated with some amount of nuclear and steroid hormone receptor activity across several ToxCast assays, but at concentrations that also approach cytotoxic concentration ranges *in vitro*; octrizole is associated with one repeated dose study in rats from ECHA IUCLID indicating a NOEL of 3.75 log10-mg/kg/day. For 4,4'-(9H-fluorene-9,9-diyl)diphenol

(DTXSID5037731), a potential bisphenol A substitute (McLaughlin *et al.*, 2023), ECHA IUCLID indicates a POD_{trad} of 3.0 log10-mg/kg/day, which is based on the reported NOAEL in a non-guideline repeated dose study with limited information on the study design and parameters evaluated. For octrizole and 4,4'-(9H-fluorene-9,9-diyl)diphenol, the POD_{NAM} may be based on mechanisms of toxicity that are not mechanistically evaluated in the *in vivo* study or evaluated with limited sensitivity. For some of these 7 chemicals, POD_{NAM} may have appeared overly conservative for other reasons as well, including assumptions in IVIVE.

Metrics for NBA

BER summary

The computed BERs are illustrated in Figure 7 and with numeric data available in Supplemental File 1 (see Supplemental Table 3). In Figure 7A, a red box surrounds the chemicals that have a BER < 4 (computed on a log10-mg/kg/day scale). Forty-three of the 194 chemicals in this case study with a POD_{NAM} satisfied the following criteria: included in the APCRA prospective case study only (not in the APCRA retrospective case study), passed AQC and are believed to be within the applicability domain for *in vitro* NAM-based screening, and demonstrated a BER < 4 on a log10-mg/kg/day scale, where the BER was based on the POD_{NAM} calculated as the median AED_{50} of the minimum AED_{50} by assay source. These 43 chemicals are displayed with all AED_{50} values by assay source in Figure 7B and then with only the overall POD_{NAM} in Figure 7C. Of these 43 chemicals, 15 of them might be defined as data-poor where no repeated dose *in vivo* POD information is available in the sources used for this case study. We noted that BER appeared to be inversely linearly related with the size of the credible interval to estimate total population exposure in SEEM3 (Supplemental File 2, Supplemental Figure 5), suggesting that support for refinements to exposure modeling may lead to greater certainty in exposure estimates and potentially fewer chemicals appearing to have BER < 4.

Hazard flag summary

A set of qualitative hazard flags for developmental and reproductive toxicity (DART) are illustrated in Figure 8 and available in Supplemental File 1. These hazard flags include indicators of developmental toxicity (from TEST *in silico* predictions and *in vitro* NAM data from the devTOX quickPredict assay) as well as combined *in silico* and *in vitro* predictors of ER and AR modulation. In Figure 8, DART hazard flags for the 43 chemicals with a BER < 4 that pass AQC and are in the APCRA prospective case study are shown, rank-ordered by BER. A number of BER-prioritized substances are associated with putative DART flags. For comparison, positive reference chemicals were selected for DART, including flutamide (DTXSID7032004), boric acid (DTXSID1020194), 5-fluorouracil (DTXSID2020634), diethylstilbestrol (DTXSID3020465), vinclozolin (DTXSID4022361), genistein (DTXSID5022308), hydroxyurea (DTXSID6025438), bisphenol A (DTXSID7020182), retinoic acid (DTXSID7021239), and thalidomide (DTXSID9022524).

To ensure possible relevance to regulatory toxicology, any NAM alternative to repeated dose toxicity studies should provide a quantitative POD and some putative indication of possible hazards. However, 90-day repeated dose toxicity tests alone may not provide enough information to conclusively determine effects on specific types of hazards such as developmental and reproductive toxicity or specific mechanistic effects in target tissues. As the hazard flags are putative indicators of hazard that could inform additional data gathering, a formal performance evaluation of these hazard flags has not been conducted. However, we do include a reference chemical panel in Figure 8 to demonstrate how known anti-androgenic (flutamide, diethylstilbestrol, vinclozolin), estrogenic (bisphenol A, diethylstilbestrol, genistein), and developmentally toxic (5-fluorouracil, hydroxyurea, retinoic acid, thalidomide) chemicals behaved with respect to the DART flag. Boric acid is one of the 42 benchmark chemicals used to evaluate the STM developmental toxicity model that is known to be a false negative (i.e., is developmentally toxic but not detected) in the STM assay (Zurlinden*, et al.*, 2020). Bisphenol A is

a known to be estrogenic *in vitro* (Judson*, et al.*, 2015) but is a true negative in the STM assay for developmental toxicity (Zurlinden*, et al.*, 2020). Genistein is estrogenic *in vitro* and was also found to be positive in the STM assay, but non-selectively at concentrations that overlapped with cytotoxicity (Zurlinden*, et al.*, 2020); genistein was also positive in the TEST DEV model. Diethylstilbestrol is a known developmental toxicant that is a known false negative in the STM assay (Zurlinden*, et al.*, 2020), but is positive in the TEST DEV model. Using the TEST DEV model and the STM assay results together may provide a highly conservative prediction of developmental toxicity, i.e. with limited specificity. As noted in the Methods, the TEST DEV model is based on a relatively small training set and biased toward positive predictions (69% of the training set compounds are positive for developmental toxicity).

A set of quantitative hazard flags for putative target tissue indications are illustrated in Figure 9 and available in Supplemental File 1. In Figure 9, these quantitative hazard flags for the same 43 chemicals with a BER < 4 that pass AQC and are in the APCRA prospective case study are shown, rank-ordered by BER (upper panel of Figure 9). Thirty-eight of these 43 chemicals shown in Figure 9 had screening data for the acute MEA assay; of these, a majority had some activity in the MEA, but only 4 chemicals shown would actually have the MEA neurotoxicity flag applied (2,6-Di-tert-butyl-4-[(dimethylamino)methyl]phenol (DTXSID0044997); 2-Butyloctan-1-ol (DTXSID0044818); 2,2’-Bisphenol F (DTXSID4022446); and Tetrabutylammonium bromide (DTXSID4044400), indicated by MEA in red bold left annotations), as only these 4 chemicals have the MEA potency as their minimum *in vitro* potency and >3 MEA assay endpoints in a single direction are positive. Chemicals in the case study, when active in the MEA, usually appeared active across multiple assay endpoints and multiple MEA activity types (firing, bursting, and connectivity), and tended to demonstrate greatest potency (lowest bioactive concentrations) in the MEA connectivity endpoints. Several chemicals were selected as positive reference chemicals for acute neuroactivity, as shown in the lower panel of Figure 9, including abamectin (DTXSID8023892), beta-cyfluthrin (DTXSID8032330), lindane (DTXSID2020686), and

tributyltin chloride (DTXSID3027403), which in general seems to be active across the activity types in the MEA (bursting, connectivity, firing) and at lower concentrations (increased potency) than chemicals in the case study. The BioMAP immunosuppression flags were observed for 10 of these 43 chemicals, and only 5 of these 10 were observed to be "selective" when compared to indicators of acute toxicity in the assay suite. Several immunosuppressive drugs in humans were selected as positive reference chemicals for the immunosuppressive activity, including azathioprine (DTXSID4020119), cyclosporin A (DTXSID0020365), dexamethasone sodium phosphate (DTXSID3047429), and methotrexate (DTXSID4020822), which appeared active in the BioMAP immunosuppression endpoints at sub-micromolar concentrations (see lower panel of Figure 9). The HIPPTox flags for lung, liver, and kidney were observed for many substances, but generally at higher *in vitro* concentrations than other assays. The hazard flag for target cell type may provide some limited information regarding cell types of interest on the basis of whether the chemical can affect these cell types at lower concentrations that approach their minimum bioactive concentration.

The fifteen chemicals of interest for further exploration due to BER < 4 and data-poorness are shown in Table 3 with their corresponding hazard flags (with full data available in Supplemental File 1). Additionally, the annotated harmonized functional use (Dionisio *et al.*, 2018; Handa *et al.*, submitted) from the ChemExpo Knowledgebase (USEPA, 2023a) was added to provide indications of potential commercial use.

Table 3. Fifteen chemicals of interest by BER prioritization, data-poorness, and flags.

These 15 chemicals had BER < 4 and were defined as "data poor, " i.e., chemicals with no associated repeat dose *in vivo* study information available. All numeric data are reported with log10-mg/kg/day units. These chemicals also had to pass AQC. Acute MEA = neuroactivity hazard flag; AR = androgen receptor hazard flag; ER = estrogen receptor hazard flag; DEV-TEST = positive in the TEST (Q)SAR for developmental toxicity; DEV = positive in the STM assay; DEV-S: selective positive in the STM assay; BioMAP flags for acute toxicity = acute; immunosupp = immunosuppression; immunosupp-S = selective immunosuppression hazard flag; HIPPTox target cell type flags for lung, liver, kidney. Harmonized functional use data indicates curated function category information obtained from the EPA's Chemicals and Products Database (Dionisio*, et al.*, 2018) (v4.0.0-alpha.2), accessed via the ChemExpo Knowledgebase (https://comptox.epa.gov/chemexpo/), which uses updated internationally harmonized function categories (OECD, 2017).

DSSTox substance id	Preferred name	BER (log10-mg/kg/day)	Median AED50 (log10-mg/kg/day)	SEEM3 U95 (log10-mg/kg/day	Flags	Harmonized Functional Use
DTXSID1025302	Octinoxate	-0.71	0.58	1.29	DEV, DEV-S, BioMAP immunosupp, immunosupp-S; HIPPTox liver and kidney	Fragrance, UV stabilizer
DTXSID9047592	9-Phenanthrol	1.18	-0.83	-2.01	DEV, DEV-S, AR, BioMAP acute, immunosupp, immunosupp-S	Not annotated
DTXSID6025301	2-Ethylhexyl glycidyl ether	1.21	0.23	-0.98	DEV, DEV-S, HIPPTox liver and kidney	Binder; chemical reaction regulator; heat stabilizer; thickening agent; solvent; viscosity modifier; wetting agent
DTXSID5038888	Basic Blue 7	1.34	-3.13	-4.47	DEV-TEST, DEV, BioMAP acute, immunosupp, immunosupp-S	Non-food use dye (toners used in printers, coolant or lubricants for metalworking industrial products)
DTXSID6024838	C.I. Solvent Red 80	2.25	-1.86	-4.11	DEV-TEST, DEV, DEV-S, ER, BioMAP acute, immunosupp, immunosupp-S	Dye
DTXSID8044836	2,4,4'-Trihydroxybenzophenone	2.37	0.47	-1.9	ER, AR	Not annotated
DTXSID0040707	4-Pentylaniline	2.42	0.11	-2.31	DEV, DEV-S, BioMAP acute, HIPPTox liver and kidney	Not annotated
DTXSID9040001	Monomethyl phthalate	2.52	0.46	-2.06		Not annotated
DTXSID0022436	Diphenolic acid	2.54	-0.48	-3.02	ER, DEV-TEST	Viscosity modifier
DTXSID1044354	N-Butyldiethanolamine	2.61	2.6	-0.01	Acute MEA, HIPPTox lung	pH regulating agent
DTXSID5022439	Phenolphthalin	2.75	0.23	-2.52	DEV-TEST, DEV, DEV-S, HIPPTox liver	Not annotated

DTXSID9047540	3-Hydroxyfluorene	2.74	0.4	-2.34	DEV-TEST, DEV, DEV-S, AR, BioMAP acute, HIPPTox lung, liver, kidney	Not annotated
DTXSID3022403	2,2'-Dihydroxy-4-methoxybenzophenone	2.89	0.75	-2.14	DEV-TEST, DEV, DEV-S, ER, AR, acute MEA	UV stabilizer
DTXSID9034361	Denatonium saccharide	3.52	0.44	-3.08		Not annotated; possible antimicrobial pesticide
DTXSID4022446	2,2'-Bisphenol F	3.76	1.48	-2.28	DEV-TEST, DEV, DEV-S, acute MEA	Not annotated

The intent of this case study was to provide an extensible, rapid approach for synthesizing NAM information to identify chemicals of potential interest. To evaluate if our NAM-based workflow results were reasonable given what is known from authoritative sources, we manually reviewed information for the chemicals identified in Table 3, as expert judgment is commonly used for evaluating single chemicals in regulatory contexts. Of these 15 chemicals of potential interest, some chemicals upon a closer manual inspection could be determined to be already well-characterized, such as C.I. Solvent Red 80, which is already listed as an International Agency for Research on Cancer (IARC) group 2B carcinogen, and it is banned from food use in the EU (used only for nonedible orange peels in the US (21CFR74.392)). Another important observation from this list of substances is that while one isomer may appear data-poor by the definition in this case study, it may be that related isomers can and have been used to make regulatory decisions. For example, 2,2’-bisphenol F (DTXSID4022446) is not registered in the EU, but an isomeric mixture of 4,4’-bisphenol F and 2,4’-bisphenol F and 2,2’-bisphenol F is registered (with endocrine activity noted for these isomers, (Punt *et al.*, 2019)), and 2,2’-bisphenol F is also in a proposed bisphenol A grouping in Canada (ECCC/HC, 2020) and was included in an IATA case study for evaluating the estrogenic potential of bisphenols (OECD, 2022). 2,2'-Dihydroxy-4-methoxybenzophenone, also known as benzophenone-8 and dioxybenzone, was indicated as having insufficient data to determine if

it was generally recognized as safe and effective (84 Federal Register (38) 6204, from 2019). Octinoxate does have information on systemic and reproductive effects, including data available from the National Toxicology Program as of 2022 (NTP, 2022) that was not included in ToxVal database v9.4. This indicates that further manual review of POD_{trad} values obtained from large, curated databases, or continued efforts to automate extraction and structuring of POD_{trad} information, could both enhance the results of preliminary screening for interesting chemicals. The approach taken herein is a case study for a baseline methodology to prospectively identify chemicals of interest for further data-gathering, such as 2,2'-Dihydroxy-4-methoxybenzophenone (DTXSID3022403), Phenolphthalin (DTXSID5022439), 3-Hydroxyfluorene (DTXSID9047540), N-Butyldiethanolamine (DTXSID1044354), Diphenolic acid (DTXSID0022436), Monomethyl phthalate (DTXSID9040001), 4-Pentylaniline (DTXSID0040707), 2,4,4'-trihydroxybenzophenone (DTXSID8044836), Basic Blue 7 (DTXSID5038888), 2-Ethylhexyl glycidyl ether (DTXSID6025301) [for which ECHA has requested systemic toxicity study information by August 2026], and 9-Phenanthrol (DTXSID9047592), for which little systemic toxicity information appear to be available.

In addition to NBA for developing POD values for safety assessment and prioritization of additional data to collect for hazard assessment, an important result is the demonstration of chemicals for which the NBA may suggest low priority, defined here as chemicals for which the POD_{NAM} is equal to or greater than 2 log10-mg/kg/day, log10-BER > 3, and log10-POD ratio greater than -0.5 log10-mg/kg/day (i.e., POD_{NAM} is within ± 0.5 log10-mg/kg/day or is greater than the POD_{trad}). Additionally, the chemical needed to pass AQC to provide greater confidence that the chemical was within the domain of screening. Of the 158 chemicals with sufficient information to calculate a POD ratio, 6 chemicals satisfied all of these criteria to demonstrate lower priority within this case study NBA (Table 4).

Table 4. Six chemicals demonstrated lower priority in this NBA

These six chemicals had some existing repeat dose information for comparison to the POD_{NAM} to understand priority within this NBA case study. These chemicals had POD_{NAM} > 2, BER > 3, and POD ratio > -0.5, and (all on a log10-mg/kg/day scale) and all passed AQC. Harmonized functional use data indicates curated function category information obtained from the EPA's Chemicals and Products Database (Dionisio*, et al.*, 2018) (v4.0.0-alpha.2), accessed via the ChemExpo Knowledgebase (https://comptox.epa.gov/chemexpo/), which uses updated internationally harmonized function categories (OECD, 2017).

DSSTox substance id	Preferred name	POD_{NAM} (log10-mg/kg/day)	POD_{trad} (log10-mg/kg/day)	POD ratio (log10-mg/kg/day)	BER (log10-mg/kg/day)	Flags	Harmonized functional use
DTXSID6025567	Methyl 2-aminobenzoate	3.19	2.7	-0.49	6.77	DEV, DEV-S, HIPPTox lung	Flavoring and nutrient; fragrance; deodorizer; solvent
DTXSID8034665	Imazapyr	2.01	2.94	0.93	5.78	DEV-TEST	Biocide
DTXSID8037750	(3Z)-Hex-3-en-1-yl salicylate	2.29	2.3	0.01	5.47	DEV-TEST, DEV, DEV-S, ER, BioMAP acute, BioMAP immunosupp, HIPPTox Lung	Flavoring and nutrient; fragrance
DTXSID1040245	Sucralose	2.03	2.18	0.15	5.26	DEV, DEV-S, HIPPTox liver	Flavoring and nutrient; fragrance; softener and conditioner
DTXSID9047201	Vanillin isobutyrate	3.05	3	-0.05	4.14	DEV-TEST, DEV, DEV-S	Flavoring and nutrient; fragrance
DTXSID4044791	Benzyl propanoate	2.81	2.7	-0.11	3.59	DEV-TEST, DEV, DEV-S	Flavoring and nutrient; fragrance

Of these 6 substances, all may have some dietary exposure based on estimates of exposure from pesticide residue, flavoring, or fragrance uses. Review of ECHA REACH information for methyl 2-aminobenzoate (DTXSID6025567) including a NOAEL of 500 mg/kg/day in a non-guideline 90-day study and review of the TTC and Cramer Class II designation for this substance (Api *et al.*, 2017) supports low priority identified within this NBA. Imazapyr (DTXSID8034665) also has a low-risk designation, based on

subchronic and chronic studies with NOAELs that range from greater than 286 mg/kg/day to in excess of 1000 mg/kg/day in REACH dossier information. Sucralose (DTXSID1040245) is generally regarded as nontoxic. Benzyl propanoate (DTXSID4044791) has been listed as low priority based on read-across assessment (Api *et al.*, 2023). Interestingly, of this short list, only (3Z)-Hex-3-en-1-yl salicylate (DTXSID8037750) has an associated regulatory action proposed under REACH, on the basis of reproductive hazard; however, the systemic toxicity risk is low based on NOAEL values of greater than 360 mg/kg/day, and the reported reproductive hazard NOAEL was 200 mg/kg/day. This substance did have hazard flags for developmental toxicity and ER activity, among others.

Evaluating NBA

Finally, we evaluated log10-POD ratios to understand how the POD_{NAM} might be used in practice following this NBA case study. In Figure 10, we examine the POD ratios using (A) the 5th percentile ToxVal POD_{trad} and (B) the 25th percentile ToxVal POD_{trad} and the POD_{NAM} defined as the median of all assay minimum AED_{50} values as defined in Equation 1b. Using the 5th percentile ToxVal POD_{trad}, 146 of 158 chemicals (92%) for which a log10 POD ratio could be calculated had a log10 POD ratio greater than -2 log10-mg/kg/day (Figure 10), using the median POD_{NAM}. Further, 134 of 158 (85%) chemicals with a log10 POD ratio using the 5th percentile POD_{trad} value had a POD ratio within ±2 log10-mg/kg/day. For alternative POD_{NAM} derivations, we observed similar results. Approximately 81% and 84% of the 158 chemicals had log10-POD ratios within ±2 log10-mg/kg/day using the broad profiling NAMs only or the broad and core targeted NAMs, respectively. Using the 25th percentile ToxVal POD_{trad} as a comparator, we observed similar results, where 83% of chemicals had a log10 POD ratio within ±2 log10-mg/kg/day. Using the higher 25th percentile ToxVal POD_{trad} as a comparator made POD_{NAM} appear slightly more conservative, but without a large numeric shift in the number of chemicals for which the POD_{NAM} is within 2 orders of magnitude of the POD_{trad}.

Discussion

Findings

In this research we demonstrate a NAM-based assessment (NBA) workflow for data-poor substances that focuses largely on application of a reduced *in vitro* NAM battery to develop quantitative POD estimates for systemic toxicity and adds hazard flags to indicate putative target toxicities, with the goal of providing enough NAM-based information to prioritize substances for further examination and/or possible data generation in models of greater biological complexity. More specifically, this case study expands the chemicals examined from the previous retrospective case study to include more industrial chemicals; combines broad profiling NAMs, including transcriptomics and Cell Painting, with a reduced set of targeted NAMs for deriving a POD_{NAM}; refines the toxicokinetic approach to utilize more complex generic toxicokinetic models when they can be parameterized; provides perspective on how decisions to summarize POD_{NAM} data may affect the predictive and protective performance of the POD_{NAM} derived; further demonstrates how different facets of this battery affect the POD ratio observed using traditional *in vivo* data; and, through this analysis, begins to inform expectations on how a POD_{NAM} and associated data may help identify chemicals and/or data that would be of interest for further consideration. This work expands upon previously published case study work from this consortium, as well as the work of colleagues in the NBA field, all with a similar central theme: characterization of a minimal *in vitro* NAM battery to obtain a suitably protective quantitative systemic toxicity POD estimate. Based on this work, and the work across the field, such an *in vitro* NAM battery should likely contain broad profiling screening assays in some number of cell models along with consideration of a suite of targeted assays that cover a number of key pharmacological targets as well as functional processes of interest relevant to the NBA decision to be made (Baltazar*, et al.*, 2020; Dent*, et al.*, 2021; Middleton*, et al.*, 2022). Herein, we attempted to define the performance of a putative minimal NAM battery intended to be useful as an alternative for a repeated dose toxicity test, such as a 90-day subchronic assay, and that would provide sufficient quantitative POD information and qualitative target toxicity

information to help prioritize substances for further testing (Gwinn*, et al.*, 2020; USEPA, 2023b). The 158 chemicals in this case study with both a POD_{NAM} and some existing repeated dose study information summarized as a 5th percentile POD_{trad} demonstrated a median log10-POD ratio of 0.14 log10-mg/kg/day; i.e., the median difference between the POD_{NAM} and the 5th percentile POD_{trad} approached zero (Figure 10). Eighty-five percent (134/158) of these chemicals demonstrated a POD_{NAM} within ±2 log10-mg/kg/day orders of magnitude from the 5th percentile POD_{trad}. Thus, within this NBA workflow, we find POD_{NAM} are typically within ± 2 log10-mg/kg/day of existing repeated dose POD data from animal models. Further, we applied specific assays or *in silico* tools to highlight specific hazard flags that might indicate the need for follow up, including developmental and reproductivity toxicity, neurotoxicity, immunosuppression, and target organ cell types, with the idea that these hazard flags might be informative of whether additional data would provide added value in an *in silico*, *in vitro*, or animal model with specific lifestages or endpoint measures. Together with the POD_{NAM} and BER, these hazard flags could be used to create customized strategies for additional modeling or data collection and/or generation (for hazard or exposure or both) to fulfill hazard and risk assessment needs.

A critical step of NBA should fundamentally include an assessment of whether the chemical will be amenable to the NAMs selected. In the context of this case study, it was necessary to understand if these chemicals when dissolved in DMSO solvent would likely be present when applied to aqueous cell-based and cell-free technologies. Herein, we explored the impact of constraining physicochemical properties of chemicals to those that are generally nonvolatile, of a molecular weight between 100 and 500 g/mol, and with a logP within a range suggesting aqueous availability and ability to cross cell membranes ($-0.4 < logP < 5.6$). Additionally, a large chemistry curation effort that ran in parallel to this work to compile and interpret multiple AQC readouts was leveraged to understand the presence and stability of chemicals in DMSO solution, and to align this understanding with *in vitro* bioactivity. Based on the initial design of this case study, most chemicals included were already nonvolatile, with a MW

more than 100 g/mol, and with a logP suggesting some aqueous availability. This case study highlights what has been observed previously in the US EPA ToxCast program: chemical samples that "fail" AQC because the parent chemical is not detected at sufficient purity or concentration most often still have bioactivity across a diverse panel of *in vitro* assays. This finding supports several actions for the future development of NBA workflows: (1) AQC measures of chemical samples applied to *in vitro* assays should be made to understand with what certainty the bioactivity observed should be ascribed to the parent chemical alone; (2) data from chemical samples with cautions on their AQC should not necessarily be discarded outright and should be examined for applicability to specific use cases for these data; (3) that the x-axis concentration units for bioactive samples may have additional uncertainty when a source chemical sample has cautions on the AQC, but that inactive samples that fail AQC pose greater problems in understanding whether the chemical is inactive at the targets screened or simply not present; (4) an *in silico* cheminformatics Tier 0 could be useful in predicting which chemicals will be more likely to be stable in solution; and, (5) *in silico* tools that can predict degradants and metabolites could be extremely useful within NBA to understand which chemicals might be present in a bioactive sample. All of these insights suggest the central role for an intensive cheminformatics analysis to precede any *in vitro* bioactivity screening. AQC directly identified specific classes of chemicals that were deemed hydrolytically unstable over time and, since DMSO is hygroscopic, hydrolysis is likely and bioactivity results for these chemicals may be based on parent, degradant(s), or a mixture of the two; such efforts can inform future cheminformatic alerts. Though beyond the scope of this initial work, additional structure alerts and a suite of QSAR models should be run for chemical inventories of interest, preferably in an automated way to enable ease of integration and reproducibility within NBA workflows. These concerns are not unique to NAMs; when test chemicals are administered to *in vivo* models or *in vitro* models with any metabolic capacity, test chemical will inevitably be metabolized or transformed, resulting in exposure to a mixture that may not be completely defined.

A novel aspect of this work when compared to our previous retrospective case study was the examination of different sets of assays informing the POD_{NAM}. As used in our case study, the term "prospective" is intended to convey that we generated much of the data for this case study as an attempted simulation of what it would be like to generate these data for a new chemical, rather than using voluminous existing data, which would be infeasible for new chemicals. Whereas previously we had used any available data in the ToxCast database, such that the number of assays tested per chemical varied up to over 1000, herein we constrained the set of assays to a battery and also included broad profiling assays, similar to the construction of assay batteries for screening cosmetics (Dent*, et al.*, 2021; Middleton*, et al.*, 2022). The set of 12 assays herein included targeted assays (ATG, BioMAP, CCTE-MEA, NVS, and STM), broad profiling assays (phenotypic profiling in one cell line and high-throughput transcriptomics in three cell lines), and three ASTAR HIPPTox assays. The assays were selected based on their ability to inform a threshold MBC value as well as the conceptual biological coverage they provided. ATG was selected to cover nuclear receptor and oxidative stress response in a HepG2 model with a small amount of metabolism. BioMAP was selected to screen a variety of primary cell models of human pathophysiology. CCTE-MEA was selected to provide a screen for neurotoxicity. NVS was selected to inform on cell-free protein interactions, including enzyme inhibition, nuclear receptor binding, and ion channel transport, similar to an *in vitro* pharmacology panel. STM was selected to provide an indicator of potential developmental toxicity. The ASTAR HIPPTox assays were designed to indicate target cell type effects for lung, kidney, and liver. And finally, the broad profiling assays were selected to directly inform the threshold MBC value using a range of cell types. Based on analysis described in Table 2, no one assay would be sufficient to define an optimally predictive and protective POD_{NAM}, but it may be that not all of these assays are necessary to develop POD_{NAM} of similar predictive and protective value. A minimal assay battery in the future could include multiple assays that are selected on the basis of the specific context of use for the POD_{NAM} value developed, which may vary by

geography, statute, and chemistry. In this case study, by examining the predictive and protective value of multiple configurations of assays, we can contribute to informing expectations on the linear performance and the rate of conservative POD_{NAM} values from any minimal set of assays for a POD_{NAM}.

Another new aspect of this workflow was the development of putative "hazard flags" as a means of indicating the types of additional hazard information that could be of interest to examine. The conceptual goal of these hazard flags was to provide preliminary information, similar to a repeated dose or subchronic study, on the types of target toxicities that might be of interest for the chemical. In part, the hazard flags helped to illustrate the biological learnings from the NAM data generated for this case study. However, these hazard flags represent a conceptual experiment that has not undergone a performance evaluation (though the underlying methods are all available and have been evaluated in peer-reviewed papers or in some cases by formal performance evaluation, such as the ToxCast ER and AR pathway models used to inform estrogen and androgen receptor activity within the DART flag). For instance, of the chemicals shown in Figure 9 with a BER < 4, it seems that many were active in the TEST DEV model and the STM assay, indicating that some model of developmental toxicity might be of interest (whether *in silico*, *in vitro*, or *in vivo*) if within the particular regulatory decision context there was a further need to evaluate this potential hazard. The development and evaluation of (Q)SARs and *in vitro* NAMs for DART could result in future improvements to this concept of a hazard flag for DART. In addition to alerts based on chemical structure or chemical category (Karamertzanis *et al.*, 2024; Patlewicz *et al.*, 2024), hazard flags from an initial application of *in vitro* NAMs could help inform what kinds of additional hazard information to develop if needed, noting the limitation that the specificity of such an approach has not been evaluated and may be limited.

Potential limitations

Limitations on metabolic competence have been a thematic concern for use of NBA in decisions, as conceptually metabolism of parent chemical could produce either less bioactive metabolites or in

some cases, bioactivated metabolites. An important iterative improvement to the NBA demonstrated herein would be more complete treatment of metabolism, both through *in silico* metabolite prediction (Boyce *et al.*, 2022) and *in vitro* methods for generation of metabolites within the *in vitro* test systems (DeGroot *et al.*, 2018; Deisenroth *et al.*, 2020; Hopperstad and Deisenroth, 2023). Herein, inclusion of HTTr in HepaRG as a broad profiling assay and ATG (using the HG19 subclone of HepG2 cells, (Medvedev*, et al.*, 2018)) as a targeted assay both provide some, albeit more limited than intact liver, metabolic competence for generation of metabolites (Gerets *et al.*, 2012; Hussain *et al.*, 2020; Jennen *et al.*, 2010; Stanley and Wolf, 2022). Inclusion of one or more *in vitro* assays with enhanced metabolic competence within an NBA panel, as we have done herein, may provide information for conservative POD_{NAM} derivation (based on potency). In comparing the AED_{50} derived from the HTTr HepaRG cell line to AED_{50} values from the other two cell lines used here, HepaRG HTTr AED_{50} were generally similar to other AED_{50} from HTTr in other cell lines, except for a small fraction for which the HepaRG HTTr AED_{50} was more than 0.5 log10-mg/kg/day higher. Similarly, the HepaRG HTTr AED_{50}, if different from the overall POD_{NAM}, tends to be higher than the overall POD_{NAM}. Without knowing the metabolic activity occurring in the time-course of the HTTr experiments with HepaRG, this preliminary view suggests that if anything putative metabolism in the HepaRG HTTr assay may serve to transform chemicals to a less potently bioactive form (Supplemental File 2, Supplemental Figure 6). Previous methods incorporating metabolism have more frequently observed changes in the chemical efficacy of assay wells with added Phase I metabolism, rather than significant shifts in potency, as captured by an area-under-the-fitted curve measure rather than a difference in potency resultant to assaying the mixture of parent and metabolite present (Deisenroth*, et al.*, 2020).

Use of a more diverse chemical space in this case study of 200 chemicals was intended to make the learnings from this case study as extensible as possible to “data-poor” chemicals. As with any study, given unlimited resources, the chemical coverage could always be larger and more extensive than it was

in order to increase this extensibility and reduce the risk of potential bias in findings. The chemical space was selected to include more industrial chemicals and chemicals with limited to no data available than in our previous retrospective case study, while still including a number of data-rich chemicals for anchoring our findings. Previous conclusions from the APCRA retrospective case study were made based on 448 chemicals that were extremely data-rich, with a majority of chemicals having at least one pesticidal use. Though more industrial and consumer chemicals were included in this APCRA prospective case study, the chemical space has inherent limitations in terms of extrapolating this case study to other chemicals, as this case study initially comprised 201 chemicals for prospective data generation that were already within the ToxCast chemical inventory. Despite the limitations on the size and chemical diversity of this case study, it is notable that when using the median of all the minimum AED_{50} values for all assays included, approximately 85% of chemicals (134/158) with a calculable log10 POD ratio (for which POD_{NAM} and POD_{trad} were available) were within ± 2 log10-mg/kg/day of each other. Only 12/158 chemicals had a log10-POD ratio less than -2, and only 3 chemicals had a log10-POD ratio less than -3. Many of these disparities could be explained by (1) existing known chemistries that are data-rich and/or would be subject to known structure alerts, such as organophosphate and carbamate insecticides; (2) additional review of available *in vivo* data; and, (3) potential incompatibility with aqueous based screening. Though beyond the scope of the work herein, ongoing work to develop and evaluate a rapid and standardized methodology for *in silico* cheminformatic analysis prior to screening (Patlewicz*, et al.*, 2024) would be useful in deciding if *in vitro* POD_{NAM} should be developed and what additional information might be needed to develop a POD. Further, and also beyond the scope of this case study, the derivation of calibrated toxicity values to estimate a POD_{trad} based on available repeated dose toxicity data (Aurisano *et al.*, 2023) could be useful in hazard data gap-filling and in benchmarking POD_{NAM} methods.

An additional limitation in this case study was the number of chemicals (approximately 200) for which prospective application of the NAM battery described herein could be applied. In an effort to increase screening of chemicals with limited *in vivo* information, not all of these chemicals were associated with publicly available *in vivo* repeated dose data. Further, not all chemicals in this case study were positive in all assays; thus, a modeling approach to POD_{NAM} is hindered by "missing data" across assays in addition to the relatively small number of chemicals for a machine learning exercise. Random forest modeling (results not shown) was attempted to describe the amount of variance in POD_{trad} that could be accounted for with a POD_{NAM}, but there were a number of limitations in such an approach, including imputation of values for "inactive" substances, in addition to the limited number of chemicals in the case study and relatedly the lack of a sizeable training, test, and validation sets. These considerations placed a machine learning exercise for POD_{NAM}, which could have involved inference of missing or "inactive" values, outside of the scope of this work and into future consideration for POD_{NAM} development, where a larger number of chemicals could be profiled and additional information, including methodology and results from an accepted POD_{NAM} approach, could be used to train POD_{NAM} values.

Additional refinements of the HTTK-based IVIVE approach taken herein should be examined for further implementation of NBA. Ongoing work is focused on application of *in vitro* disposition models to large bioactivity datasets to understand the overall trends and impacts of these models (Scherer *et al.*, in prep). One improvement within this work was leveraged by use of httk v2.3.0, which includes estimation of the amount of chemical that might be absorbed from the gastrointestinal tract (thereby affecting the percent of chemical that would ultimately be bioavailable), which in theory may result in less conservative POD_{NAM} (i.e., with reduced Css values given the same oral dose). Of the 12 chemicals for which the POD_{NAM} failed to be conservative enough (POD ratio < -2), it is unclear how toxicokinetic triage based on predicted plasma half-life or other physicochemical properties could have been used to

identify this group of chemicals. For these chemicals it seems likely that an expanded cheminformatics process to examine amenability and known structure-informed hazard and toxicokinetics would be helpful. A better POD modeling approach in the future could involve consensus of POD_{NAM} with QSAR predictions (Kvasnicka *et al.*, 2024; Pradeep *et al.*, 2020a) or other cheminformatic approaches to POD prediction.

Defining expectations for POD_{NAM}

In this case study we further examined the POD_{NAM} as a protective as well as a predictive value, finding unsurprisingly that POD_{NAM} fail to appear very predictive of an animal based POD_{trad}. This comparison is inherently limited as the types of effects measured in repeated dose toxicity tests differ from the measurements made using NAMs; for example, a change in overall body weight may not have a simple NAM corollary. There are several other important limitations to consider with respect to this relative lack of predictivity for an animal-based POD. First, it is important to take into account the potential error in the POD_{NAM} value as well as the potential variability in the animal based POD_{trad} value used for comparison. The error in the POD_{NAM} value (when predicting the POD_{trad}) may be resultant to lack of sufficient target coverage but is likely due, at least in part, to decisions in the generic IVIVE approach applied using *httk*. When examining chemicals with organ-specific hazard for liver and kidney and cell-based models of those organs, a similar POD ratio median and range for POD_{NAM} and POD_{trad} was observed for the median toxicokinetic individual (as represented by an AED_{50}) (Paul Friedman*, et al.*, 2023). Previously, HTTK methods for predicting human Css, a key toxicokinetic measure in determining an IVIVE-based AED, were shown to result in Css predictions within a factor of 10 of *in vivo* values for most chemicals (or an RMSE of 1 on a log10-scale) (Breen*, et al.*, 2021). For rodent HTTK predictions, the expectation on AED prediction of *in vivo* POD ranges demonstrated that reverse dosimetry based upon PBTK was more predictive than other approaches (Honda *et al.*, 2019) (RMSE was not reported). However, in these comparisons, an important consideration of note is the variability of *in vivo* POD

values themselves, which in repeated dose animal models may approach 0.5 log10-mg/kg/day (Pham*, et al.*, 2020). Some rationalization of previous work on the predictive accuracy of toxicokinetic parameters estimated from HTTK data and models and AED values for *in vivo* PODs is necessary to bring context to the RMSE values (~1 to 1.2 log10-mg/kg/day) in this work for AED values and *in vivo* animal-based PODs. Indeed, adding the RMSE values between the ability of *in vivo* repeated dose studies to predict themselves (RMSE ~ 0.5 log10-mg/kg/day) and the RMSE for prediction of Css for determination of the AED (RMSE ~ 1 log10-mg/kg/day) comes close to approximating the RMSE observed for POD_{NAM} and POD_{trad} in this study (~1.2 log10-mg/kg/day). In addition to some amount of uncertainty that may not be explained by the IVIVE approach currently employed, there is another consideration: animal based POD_{trad} are typically divided by large uncertainty factors to be protective of human health, signaling an assumption that for systemic toxicity evaluation in our current paradigm there is generally an expectation of protection rather than prediction of specific effects that are anticipated to occur in humans (Browne*, et al.*, 2024).

An important critique of NBA for POD_{NAM} determination and hazard in general has been the drive toward ensuring conservatism, resulting in no chemical ever appearing to be of low priority on the basis of such a workflow. This is a point well-taken, and in this case study work, some considerations become clear for managing this as an expectation. First is in the demonstration of chemicals for which BER is sufficiently large (log10-BER > 3) and the POD_{NAM} is sufficiently high (> 100 mg/kg/day); depending on the regulatory framework applied, these chemicals may be of lower interest for continued data gathering. Determination of the potential uses of these chemicals, and refined exposure modeling, could be part of data gathering. A second consideration may be in reframing how typical animal-based chemical safety assessment is performed. To some extent, the application of uncertainty factors and lack of positive predictive value for key hazards (Monticello*, et al.*, 2017) in humans suggest that current animal-based safety assessment paradigms are designed to be protective rather than predictive

(Browne*, et al.*, 2024). If a data-informed POD_{NAM} provides a value that is a conservative value in comparison to animal-based POD_{trad}, and the data used to develop the POD_{NAM} provides potential insights into mechanism or processes involved in the bioactivity of the chemical through hazard flags similar to those demonstrated herein, then a POD_{NAM} is providing similar value as a POD_{trad} from a repeated dose study (such as the 90-day repeated dose toxicity study). In this case study, we demonstrate an NBA workflow that is subject to iterative improvement in terms of addition of *in silico* and *in vitro* NAMs, but in general could be utilized with adjustment factors to provide reasonably protective systemic toxicity POD values and putative indications of hazard, as would be expected from a 90-day repeated dose study in animals.

Acknowledgements

The authors wish to thank Nisha Sipes, Kristin Isaacs, John Cowden, Sid Hunter, and Renee Beardslee of the US EPA, and Kristin Eccles and Marc Beal of Health Canada, for useful technical comments on a previous version of this manuscript, as well as the teams of many scientists who make data available publicly within the databases and tools used in the NBA used herein. We would also like to thank Oscar Fu and Carmen Kong from Bioinformatics Institute, A*STAR for helping to perform the HIPPTox assays.

Figure Legends

Figure 1. NAM-based assessment (NBA) workflow.

An overview of a NBA workflow that incorporates cheminformatics, broad and targeted bioactivity NAMs, via hazard flags, and exposure NAMs for internal and external exposures. The workflow culminates in a set of outputs for NBA, including hazard flags, POD_{NAM}, and BER estimates.

Figure 2. In vitro screening applicability domain

In A, the chemicals with caution on the analytical quality control (AQC) performed on DMSO solvated samples are shown, along with whether they are in the prospective case study only (Prosp = In); the predicted serum half-life (T1/2) is < 90 days (In); the molecular weight (MW) is between 100 and 500 g/mol (In); logP is > -0.4 and < 5.6 (In), and the log10-vapor pressure (logVP) is < 2 (In). In B, the exposure pathway predictions from the SEEM3 exposure model for each chemical in the prospective (pro), retrospective (ret), or both case studies are shown, where Pest. = Pesticide, Ind. = Industrial, Cons. = Consumer, Diet. = Dietary, All Four = Pest. + Ind. + Cons. + Diet, Unknown = not known in SEEM3; NA = not annotated in SEEM3. In C, a Uniform Manifold Approximation and Projection (UMAP) projection that reduced the feature dimensionality of molecular weight and predicted logP, vapor pressure, and water solubility failed to group chemicals which have AQC cautions (with chemical names labeled) or significantly distinguish the chemicals represented in the APCRA prospective and retrospective case studies.

Figure 3. In vitro assay battery

In A, the minimum log10-µM potency in each *in vitro* NAM source is illustrated with yellow hues indicating more potent bioactive concentrations and red to purple hues indicating less potent bioactivity. Two row annotations are provided to indicate AQC pass/caution (black = pass, white = caution) and chemical membership in the APCRA prospective case study only (blue = prospective only). In B, a detailed view of the chemicals which have cautions on their AQC is shown. In C, the distribution of the minimum *in vitro* potency (µM) is shown for the minimum of targeted NAMs, minimum of broad profiling NAMs, and the 5th percentile from all of ToxCast (ACC values) data available. In D, the count of each broad profiling NAM or targeted NAM underlying the minimum potency by chemical is displayed, indicating the potencies in NVS were most often the minimum potency by chemical, followed by CCTE-MEA potencies. *[CCTE-MEA = acute microelectrode array assay; STM = Stemina developmental toxicity assay; NVS = NovaScreen, cell-free data on protein and enzyme targets; ATG = Attagene transcription factor assay ; BioMAP = BioMAP Panel of 11 primary culture and co-culture models of human pathophysiology; http.u2os.pac.min = high-throughput phenotypic profiling phenotype altering concentration (PAC) from U-2 OS cells; httr.mcf7, u20s, and heparg: high-throughput transcriptomic data from MCF7, U-2 OS, and HepaRG screening for signature-based point of departure; Astar_BEAS2B, HK2, and HepG2: A*STAR high-throughput phenotypic profiling of BEAS2B, HK2, and HepG2 models of lung, kidney, and liver cells]*

Figure 4. POD_{NAM} benchmarking to POD_{trad}

In A, the minimum AED_{50} for each assay (abbreviated assay names under AED50 Source) demonstrates that no single assay produced a value with a strong linear relationship to the ToxVal-based POD_{trad} (5th-%ile or 25th-%ile). The solid black line represents unity. In B, reducing the population of AED_{50} values per chemical to the minimum (min), median (med), and multilinear regression (MLR) modeled-value from training to the ToxVal 25th %-ile (as defined in Equations 1a, 1b, and 2, respectively), relatively poor linear relationships are observed. The results of comparing the derived POD_{NAM} with the 5th and 25th-%ile ToxVal PODs are provided, along with coefficient of determination (R^2). The solid black line represents unity and the dashed lines represent ± 1 log10-mg/kg/day from unity. The purple or green lines represent the linear model, using the min, med, or MLR AED_{50} vs the ToxVal 5th or 25th percentile, respectively. In C, the graph elements are the same except that the POD_{NAM} definitions shown are the min and med AED_{50} values from the broad profiling and targeted assay subsets only. See Table 2 for all POD_{NAM} vs. POD_{trad} performance metrics.

Figure 5. POD ratios by different assays and ToxVal summary values

Boxplots of log10-POD ratio (log10-mg/kg/day) distributions using different derivations of the POD_{NAM} and two different ToxVal POD_{trad} summary values (5th and 25th %-ile). The median and upper and lower quartiles are described by each box. Each chemical point in the distribution is superimposed with jitter. Dashed horizontal red line indicates a log10-POD ratio of 0, i.e. POD_{trad} = POD_{NAM}. POD-Med HIPPTox = POD_{NAM} based on median of HIPPTox values; MED POD Ratio = POD_{NAM} used in this case study that is based on the median of all minimum AED_{50} values by assay; POD-Med Broad = POD_{NAM} based on the median of broad profiling AED_{50} values; POD-Med Broad + Core Targeted = POD_{NAM} based on median of AED_{50} values from broad and core targeted (ATG, BioMAP, NVS) NAMs; POD-Med Targeted = POD_{NAM} based on median of all targeted NAM AED_{50} values; POD-Core Targeted = POD_{NAM} based on core targeted NAM AED_{50} values.

Figure 6. In vitro, in silico, and in vivo POD comparisons.

In A, the frequency distributions of the POD ratio (purple) [5th %-ile POD_{trad} – Med AED_{50}], TTC ratio (teal) [5th %-ile POD_{trad} – TTC], and SUB ratio (yellow) [5th %-ile POD_{trad} – SUB]. Red dashed line = median POD ratio; blue dashed line = median TTC ratio. In B, the chemicals for which the POD ratio is > ± 3 log10-mg/kg/day are shown, where 5th %-ile POD_{trad} = *in vivo* POD based on the ECHA IUCLID POD or the 5th %-ile in ToxVal and Med AED50 = POD_{NAM} based on median of the minimum AED_{50} values by assay.

Figure 7. Bioactivity:exposure ratios (BERs) for prioritization.

BERs provide a simple prioritization for further examination of existing data and/or data generation. In (A) a zoomed out full view of all 201 chemicals in the case study is shown, sorted by the BER. The SEEM3 median and upper credible interval on the median estimate for human exposure, the 5th percentile POD_{trad} (5th%-ile POD All), and the AED_{50} for individual assays are all pictured. The red box indicates the chemicals with a BER <4. In (B), chemicals with a BER < 4, in the APCRA prospective case study only, and with passing AQC are shown, with all assay AED_{50} values depicted separated. In (C), complexity is reduced, showing chemicals with a BER < 4, in the APCRA prospective case study only, and with passing AQC, using the median AED_{50}

value alone rather than an AED_{50} value for each assay. The 5%-ile POD All (when available) and SEEM3 exposure estimates are included. All numeric values in panels A through C are reported in log10-mg/kg/day units.

Figure 8. Qualitative Hazard Flags for DART.

Illustrates a putative development and reproductive toxicity (DART) panel comprised of flags for ER and AR (integrated *in silico* and *in vitro* signals); developmental toxicity (DEV) and selective developmental toxicity (DEV-S); and TEST predictions of developmental toxicity (DEV-TEST). Only chemicals with log_{10}-BER < 4 are shown. The row annotation to the left indicates the assay that resulted in the minimum AED50. The upper panel illustrates chemicals included in this case study whereas the lower panel illustrates chemicals added as a reference to demonstrate how the hazard flag performs. White = negative, Gray = missing; Green = *in silico* flag; Purple = *in vitro* flag.

Figure 9. Quantitative Hazard Flags for Target Cell Types.

Illustrates target tissue concerns for HIPPTox lung, liver, and kidney; acute and immune suppression signals from the BioMAP panel; and acute neuroactivity in the microelectode array (MEA) assay. Only chemicals with log_{10}-BER < 4 are shown. The row annotation to the left indicates the assay that resulted in the minimum AED50; left annotation text in blue italics indicates that MEA data were not available for the chemical. Only where MEA underlies the minimum AED_{50}, and at least 3 MEA assay endpoints in a single direction are positive, is the neuroactivity flag applied (indicated by red bold MEA on the left annotation). The lower panel of this figure provides information on reference chemicals for this hazard flag.

Figure 10. Understanding POD ratios

In A, plots of the empirical cumulative distribution for the POD ratios by assay are shown. In B, plots of the empirical cumulative distribution for the POD ratios by summary values are shown. Red horizontal dashed line indicates 90% cumulative frequency and the vertical dashed red lines indicate ± 2 log10-mg/kg/day, between which 85 and 83% percent of POD ratios fall in A and B, respectively. Solid vertical line indicates a log10 POD ratio of 0. POD-Med HIPPTox = POD_{NAM} based on median of HIPPTox values; MED POD Ratio = POD_{NAM} used in this case study that is based on the median of all minimum AED_{50} values by assay; POD-Med Broad = POD_{NAM} based on the median of broad profiling AED_{50} values; POD-Med Broad + Core Targeted = POD_{NAM} based on median of AED_{50} values from broad and core targeted (ATG, BioMAP, NVS) NAMs; POD-Med Targeted = POD_{NAM} based on median of all targeted NAM AED_{50} values; POD-Core Targeted = POD_{NAM} based on core targeted NAM AED_{50} values.

References

Api, A. M., Belsito, D., Botelho, D., Browne, D., Bruze, M., Burton, G. A., Jr., Buschmann, J., Dagli, M. L., Date, M., Dekant, W*., et al.* (2017). RIFM fragrance ingredient safety assessment, methyl anthranilate, CAS Registry Number 134-20-3. *Food Chem Toxicol* **110 Suppl 1**, S290-S298.

Api, A. M., Belsito, D., Botelho, D., Bruze, M., Burton, G. A., Jr., Cancellieri, M. A., Chon, H., Dagli, M. L., Date, M., Dekant, W*., et al.* (2023). Update to RIFM fragrance ingredient safety assessment, benzyl propionate, CAS Registry Number 122-63-4. *Food Chem Toxicol* **182 Suppl 1**, 114237.

Aurisano, N., Jolliet, O., Chiu, W. A., Judson, R., Jang, S., Unnikrishnan, A., Kosnik, M. B., and Fantke, P. (2023). Probabilistic Points of Departure and Reference Doses for Characterizing Human Noncancer and Developmental/Reproductive Effects for 10,145 Chemicals. *Environ Health Perspect* **131**(3), 37016.

Baltazar, M. T., Cable, S., Carmichael, P. L., Cubberley, R., Cull, T., Delagrange, M., Dent, M. P., Hatherell, S., Houghton, J., Kukic, P*., et al.* (2020). A Next-Generation Risk Assessment Case Study for Coumarin in Cosmetic Products. *Toxicol Sci* **176**(1), 236-252.

Barton-Maclaren, T. S., Wade, M., Basu, N., Bayen, S., Grundy, J., Marlatt, V., Moore, R., Parent, L., Parrott, J., Grigorova, P.*, et al.* (2022). Innovation in regulatory approaches for endocrine disrupting chemicals: The journey to risk assessment modernization in Canada. *Environ Res* **204**(Pt C), 112225.

Basketter, D. A., Clewell, H., Kimber, I., Rossi, A., Blaauboer, B., Burrier, R., Daneshian, M., Eskes, C., Goldberg, A., Hasiwa, N.*, et al.* (2012). A roadmap for the development of alternative (non-animal) methods for systemic toxicity testing. *ALTEX* **29**(1), 3-91.

Beal, M. A., Audebert, M., Barton-Maclaren, T., Battaion, H., Bemis, J. C., Cao, X., Chen, C., Dertinger, S. D., Froetschl, R., Guo, X.*, et al.* (2023). Quantitative in vitro to in vivo extrapolation of genotoxicity data provides protective estimates of in vivo dose. *Environ Mol Mutagen* **64**(2), 105-122.

Beal, M. A., Gagne, M., Kulkarni, S. A., Patlewicz, G., Thomas, R. S., and Barton-Maclaren, T. S. (2022). Implementing in vitro bioactivity data to modernize priority setting of chemical inventories. *ALTEX* **39**(1), 123-139.

Betts, B. C., Bastian, D., Iamsawat, S., Nguyen, H., Heinrichs, J. L., Wu, Y., Daenthanasanmak, A., Veerapathran, A., O'Mahony, A., Walton, K.*, et al.* (2018). Targeting JAK2 reduces GVHD and xenograft rejection through regulation of T cell differentiation. *Proc Natl Acad Sci U S A* **115**(7), 1582-1587.

Bhuller, Y., Ramsingh, D., Beal, M., Kulkarni, S., Gagne, M., and Barton-Maclaren, T. S. (2021). Canadian Regulatory Perspective on Next Generation Risk Assessments for Pest Control Products and Industrial Chemicals. *Front Toxicol* **3**, 748406.

Boyce, M., Meyer, B., Grulke, C., Lizarraga, L., and Patlewicz, G. (2022). Comparing the performance and coverage of selected in silico (liver) metabolism tools relative to reported studies in the literature to inform analogue selection in read-across: A case study. *Comput Toxicol* **21**, 1-15.

Breen, M., Ring, C. L., Kreutz, A., Goldsmith, M. R., and Wambaugh, J. F. (2021). High-throughput PBTK models for in vitro to in vivo extrapolation. *Expert Opin Drug Metab Toxicol* doi: 10.1080/17425255.2021.1935867.

Browne, P., Paul Friedman, K., Boekelheide, K., and Thomas, R. S. (2024). Adverse effects in traditional and alternative toxicity tests. *Regul Toxicol Pharmacol* **148**, 105579.

Canada (2018). *New Substances Notification Regulations (Chemicals and Polymers), SOR/2005-247, Minister of Justice.* Available at: https://laws-lois.justice.gc.ca/eng/Regulations/SOR-2005-247/index.html. Accessed December 15, 2024.

Cassano, A., Manganaro, A., Martin, T., Young, D., Piclin, N., Pintore, M., Bigoni, D., and Benfenati, E. (2010). CAESAR models for developmental toxicity. *Chem Cent J* **4 Suppl 1**(Suppl 1), S4.

Clark, M., and Steger-Hartmann, T. (2018). A big data approach to the concordance of the toxicity of pharmaceuticals in animals and humans. *Regul Toxicol Pharmacol* **96**, 94-105.

Darwich, A. S., Neuhoff, S., Jamei, M., and Rostami-Hodjegan, A. (2010). Interplay of metabolism and transport in determining oral drug absorption and gut wall metabolism: a simulation assessment using the "Advanced Dissolution, Absorption, Metabolism (ADAM)" model. *Curr Drug Metab* **11**(9), 716-29.

Dawson, D. E., Ingle, B. L., Phillips, K. A., Nichols, J. W., Wambaugh, J. F., and Tornero-Velez, R. (2021). Designing QSARs for Parameters of High-Throughput Toxicokinetic Models Using Open-Source Descriptors. *Environ Sci Technol* **55**(9), 6505-6517.

DeGroot, D. E., Swank, A., Thomas, R. S., Strynar, M., Lee, M. Y., Carmichael, P. L., and Simmons, S. O. (2018). mRNA transfection retrofits cell-based assays with xenobiotic metabolism. *J Pharmacol Toxicol Methods* **92**, 77-94.

Deisenroth, C., DeGroot, D. E., Zurlinden, T., Eicher, A., McCord, J., Lee, M. Y., Carmichael, P., and Thomas, R. S. (2020). The Alginate Immobilization of Metabolic Enzymes Platform Retrofits an Estrogen Receptor Transactivation Assay With Metabolic Competence. *Toxicol Sci* **178**(2), 281-301.

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.*, et al.* (2021). Paving the way for application of next generation risk assessment to safety decision-making for cosmetic ingredients. *Regul Toxicol Pharmacol* **125**, 105026.

Dionisio, K. L., Phillips, K., Price, P. S., Grulke, C. M., Williams, A., Biryol, D., Hong, T., and Isaacs, K. K. (2018). The Chemical and Products Database, a resource for exposure-relevant data on chemicals in consumer products. *Sci Data* **5**, 180125.

ECCC/HC (2020). *Technical Consultation: Proposed Subgrouping of Bisphenol A (BPA) Structural Analogues and Functional Alternatives, Ottawa, ON, Canada.* Available at: https://www.canada.ca/en/environment-climate-change/services/evaluating-existing-substances/technical-consultation-proposed-subgrouping-bpa-structural-analogues-functional-alternatives.html.

ECHA (2023). The use of alternatives to testing on animals for the REACH Regulation. Fifth report under Article 117(3) of the REACH Regulation. ECHA-23-R-07-EN, doi: 10.2823/805454, Helsinki, Finland. https://echa.europa.eu/documents/10162/23919267/230530_117_3_alternatives_test_animals_2023_en.pdf.

EFSA (2012). Scientific Opinion on Exploring options for providing advice about possible human health risks based on the concept of Threshold of Toxicological Concern (TTC). In (doi: https://doi.org/10.2903/j.efsa.2012.2750). EFSA Scientific Committee.

EuropeanCommission (2007). *Registration, Evaluation, Authorisation and Restriction of Chemicals. In Regulation (EC) No. 1907/2006 of the European Parliament and of the Council. European Union.* https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32006R1907.

Filer, D. L., Kothiya, P., Setzer, R. W., Judson, R. S., and Martin, M. T. (2017). tcpl: the ToxCast pipeline for high-throughput screening data. *Bioinformatics* **33**(4), 618-620.

Gerets, H. H., Tilmant, K., Gerin, B., Chanteux, H., Depelchin, B. O., Dhalluin, S., and Atienzar, F. A. (2012). Characterization of primary human hepatocytes, HepG2 cells, and HepaRG cells at the mRNA level and CYP activity in response to inducers and their predictivity for the detection of human hepatotoxins. *Cell Biol Toxicol* **28**(2), 69-87.

Gilmour, N., Alepee, N., Hoffmann, S., Kern, P. S., Van Vliet, E., Bury, D., Miyazawa, M., Nishida, H., and Cosmetics, E. (2023). Applying a next generation risk assessment framework for skin sensitisation to inconsistent new approach methodology information. *ALTEX* doi: 10.14573/altex.2211161.

Gwinn, W. M., Auerbach, S. S., Parham, F., Stout, M. D., Waidyanatha, S., Mutlu, E., Collins, B., Paules, R. S., Merrick, B. A., Ferguson, S.*, et al.* (2020). Evaluation of 5-day In Vivo Rat Liver and Kidney With High-throughput Transcriptomics for Estimating Benchmark Doses of Apical Outcomes. *Toxicol Sci* **176**(2), 343-354.

Hagiwara, S., Paoli, G. M., Price, P. S., Gwinn, M. R., Guiseppi-Elie, A., Farrell, P. J., Hubbell, B. J., Krewski, D., and Thomas, R. S. (2023). A value of information framework for assessing the trade-offs associated with uncertainty, duration, and cost of chemical toxicity testing. *Risk Anal* **43**(3), 498-515.

Hammitzsch, A., Tallant, C., Fedorov, O., O'Mahony, A., Brennan, P. E., Hay, D. A., Martinez, F. O., Al-Mossawi, M. H., de Wit, J., Vecellio, M.*, et al.* (2015). CBP30, a selective CBP/p300 bromodomain inhibitor, suppresses human Th17 responses. *Proc Natl Acad Sci U S A* **112**(34), 10768-73.

Handa, S., Isaacs, K. K., Wall, J. T., Larger, A., Burns, S., Koval, L., Baron-Furuyama, K., Elonen, C. M., Lyons, D., Dionisio, K. L.*, et al.* (submitted). CPDat4.0, an updated database for chemical and product information supporting chemical evaluations. *Nature Scientific Data*.

Harrill, J. A., Everett, L. J., Haggard, D. E., Sheffield, T., Bundy, J. L., Willis, C. M., Thomas, R. S., Shah, I., and Judson, R. S. (2021). High-Throughput Transcriptomics Platform for Screening Environmental Chemicals. *Toxicol Sci* **181**(1), 68-89.

HC/ECCC (2023). *Notice of intent on the development of a strategy to guide the replacement, reduction, or refinement of vertebrate animal testing under the Canadian Environmental Protection Act (CEPA).* Available at: https://www.canada.ca/en/health-canada/programs/consultation-strategy-replace-reduce-refine-vertebrate-animal-testing/notice-intent.html. Accessed December 15, 2024.

HealthCanada (2016). Science Approach Document for the Threshold of Toxicological Concern (TTC)-based Approach for Certain Substances **150**(40).

Hisaki, T., Aiba Nee Kaneko, M., Yamaguchi, M., Sasa, H., and Kouzuki, H. (2015). Development of QSAR models using artificial neural network analysis for risk assessment of repeated-dose, reproductive, and developmental toxicities of cosmetic ingredients. *The Journal of toxicological sciences* **40**(2), 163-80.

Honda, G. S., Pearce, R. G., Pham, L. L., Setzer, R. W., Wetmore, B. A., Sipes, N. S., Gilbert, J., Franz, B., Thomas, R. S., and Wambaugh, J. F. (2019). Using the concordance of in vitro and in vivo data to evaluate extrapolation assumptions. *PLoS One* **14**(5), e0217564.

Hopperstad, K., and Deisenroth, C. (2023). Development of a bioprinter-based method for incorporating metabolic competence into high-throughput in vitro assays. *Front Toxicol* **5**, 1196245.

Houck, K. A., Friedman, K. P., Feshuk, M., Patlewicz, G., Smeltz, M., Clifton, M. S., Wetmore, B. A., Velichko, S., Berenyi, A., and Berg, E. L. (2023). Evaluation of 147 perfluoroalkyl substances for immunotoxic and other (patho)physiological activities through phenotypic screening of human primary cells. *ALTEX* **40**(2), 248-270.

Houck, K. A., Patlewicz, G., Richard, A. M., Williams, A. J., Shobair, M. A., Smeltz, M., Clifton, M. S., Wetmore, B., Medvedev, A., and Makarov, S. (2021). Bioactivity profiling of per- and polyfluoroalkyl substances (PFAS) identifies potential toxicity pathways related to molecular structure. *Toxicology* **457**, 152789.

Hussain, F., Basu, S., Heng, J. J. H., Loo, L. H., and Zink, D. (2020). Predicting direct hepatocyte toxicity in humans by combining high-throughput imaging of HepaRG cells and machine learning-based phenotypic profiling. *Arch Toxicol* **94**(8), 2749-2767.

Isaacs, K. K., Wall, J. T., Paul Friedman, K., Franzosa, J. A., Goeden, H., Williams, A. J., Dionisio, K. L., Lambert, J. C., Linnenbrink, M., Singh, A.*, et al.* (2023). Screening for drinking water contaminants of concern using an automated exposure-focused workflow. *J Expo Sci Environ Epidemiol* doi: 10.1038/s41370-023-00552-y.

Jennen, D. G., Magkoufopoulou, C., Ketelslegers, H. B., van Herwijnen, M. H., Kleinjans, J. C., and van Delft, J. H. (2010). Comparison of HepG2 and HepaRG by whole-genome gene expression analysis for the purpose of chemical hazard identification. *Toxicol Sci* **115**(1), 66-79.

Johnson, K. J., Auerbach, S. S., Stevens, T., Barton-Maclaren, T. S., Costa, E., Currie, R. A., Dalmas Wilk, D., Haq, S., Rager, J. E., Reardon, A. J. F.*, et al.* (2022). A Transformative Vision for an Omics-Based Regulatory Chemical Testing Paradigm. *Toxicol Sci* **190**(2), 127-132.

Judson, R. S., Magpantay, F. M., Chickarmane, V., Haskell, C., Tania, N., Taylor, J., Xia, M., Huang, R., Rotroff, D. M., Filer, D. L*., 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**(1), 137-54.

Karamertzanis, P. G., Patlewicz, G., Sannicola, M., Paul-Friedman, K., and Shah, I. (2024). Systematic Approaches for the Encoding of Chemical Groups: A Case Study. *Chem Res Toxicol* **37**(4), 600-619.

Kaur, H., Chaudhary, S., Kaur, H., Chaudhary, M., and Jena, K. C. (2022). Hydrolysis and Condensation of Tetraethyl Orthosilicate at the Air–Aqueous Interface: Implications for Silica Nanoparticle Formation. *ACS Applied Nano Materials* **5**, 411-422.

Kleinstreuer, N. C., Ceger, P., Watt, E. D., Martin, M., Houck, K., Browne, P., Thomas, R. S., Casey, W. M., Dix, D. J., Allen, D*., et al.* (2017). Development and Validation of a Computational Model for Androgen Receptor Activity. *Chem Res Toxicol* **30**(4), 946-964.

Knudsen, T. B., Houck, K. A., Sipes, N. S., Singh, A. V., Judson, R. S., Martin, M. T., Weissman, A., Kleinstreuer, N. C., Mortensen, H. M., Reif, D. M*., et al.* (2011). Activity profiles of 309 ToxCast chemicals evaluated across 292 biochemical targets. *Toxicology* **282**(1-2), 1-15.

Kosnik, M. B., Strickland, J. D., Marvel, S. W., Wallis, D. J., Wallace, K., Richard, A. M., Reif, D. M., and Shafer, T. J. (2020). Concentration-response evaluation of ToxCast compounds for multivariate activity patterns of neural network function. *Arch Toxicol* **94**(2), 469-484.

Kuhn, M. (2008). Building predictive models in R using the caret R package. *Journal of Statistical Software* **28**(5), 1-26.

Kulkarni, S. A., Benfenati, E., and Barton-Maclaren, T. S. (2016). Improving confidence in (Q)SAR predictions under Canada's Chemicals Management Plan - a chemical space approach. *SAR QSAR Environ Res* **27**(10), 851-863.

Kvasnicka, J., Aurisano, N., von Borries, K., Lu, E. H., Fantke, P., Jolliet, O., Wright, F. A., and Chiu, W. A. (2024). Two-Stage Machine Learning-Based Approach to Predict Points of Departure for Human Noncancer and Developmental/Reproductive Effects. *Environ Sci Technol* **58**(35), 15638-15649.

Laksameethanasan, D., Tan, R., Toh, G., and Loo, L. H. (2013). cellXpress: a fast and user-friendly software platform for profiling cellular phenotypes. *BMC Bioinformatics* **14 Suppl 16**(Suppl 16), S4.

Lautenberg, F. R. (2016). Frank R. Lautenberg Chemical Safety for the 21st Century Act. In (t. U. Congress, Ed.), pp. 114-182. Public Law.

Lee, J. J., Miller, J. A., Basu, S., Kee, T. V., and Loo, L. H. (2018). Building predictive in vitro pulmonary toxicity assays using high-throughput imaging and artificial intelligence. *Arch Toxicol* **92**(6), 2055-2075.

Loo, L. H., Wu, L. F., and Altschuler, S. J. (2007). Image-based multivariate profiling of drug responses from single cells. *Nat Methods* **4**(5), 445-53.

Mansouri, K., Abdelaziz, A., Rybacka, A., Roncaglioni, A., Tropsha, A., Varnek, A., Zakharov, A., Worth, A., Richard, A. M., Grulke, C. M*., et al.* (2016). CERAPP: Collaborative Estrogen Receptor Activity Prediction Project. *Environ Health Perspect* **124**(7), 1023-33.

Mansouri, K., Grulke, C. M., Judson, R. S., and Williams, A. J. (2018). OPERA models for predicting physicochemical properties and environmental fate endpoints. *J Cheminform* **10**(1), 10.

Mansouri, K., Kleinstreuer, N., Abdelaziz, A. M., Alberga, D., Alves, V. M., Andersson, P. L., Andrade, C. H., Bai, F., Balabin, I., Ballabio, D*., et al.* (2020). CoMPARA: Collaborative Modeling Project for Androgen Receptor Activity. *Environ Health Perspect* **128**(2), 27002.

Martin, M. M., Carpenter, A. F., Shafer, T. J., Paul Friedman, K., and Carstens, K. E. (2024). Chemical effects on neural network activity: Comparison of acute versus network formation exposure in microelectrode array assays. *Toxicology* **505**, 153842.

Martin, M. T., Dix, D. J., Judson, R. S., Kavlock, R. J., Reif, D. M., Richard, A. M., Rotroff, D. M., Romanov, S., Medvedev, A., Poltoratskaya, N*., et al.* (2010). Impact of environmental chemicals on key transcription regulators and correlation to toxicity end points within EPA's ToxCast program. *Chem Res Toxicol* **23**(3), 578-90.

McInnes, L., Healy, J., and Melville, J. (2018). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arxiv statML* **arXiv:1802.03426v3**.

McLaughlin, A. J., Kaniski, A. I., Matti, D. I., Monear, N. C., Tischler, J. L., and Xhabija, B. (2023). Fluorene-9-bisphenol affects the terminal differentiation of mouse embryonic bodies. *Curr Res Toxicol* **5**, 100133.

Medvedev, A., Moeser, M., Medvedeva, L., Martsen, E., Granick, A., Raines, L., Zeng, M., Makarov, S., Jr., Houck, K. A., and Makarov, S. S. (2018). Evaluating biological activity of compounds by transcription factor activity profiling. *Sci Adv* **4**(9), eaar4666.

Middleton, A. M., Reynolds, J., Cable, S., Baltazar, M. T., Li, H., Bevan, S., Carmichael, P. L., Dent, M. P., Hatherell, S., Houghton, J*., et al.* (2022). Are Non-animal Systemic Safety Assessments Protective? A Toolbox and Workflow. *Toxicol Sci* **189**(1), 124-147.

Miller, J. A., and Loo, L. H. (2020). Optimum concentration-response curve metrics for supervised selection of discriminative cellular phenotypic endpoints for chemical hazard assessment. *Arch Toxicol* **94**(9), 2951-2964.

Monticello, T. M., Jones, T. W., Dambach, D. M., Potter, D. M., Bolt, M. W., Liu, M., Keller, D. A., Hart, T. K., and Kadambi, V. J. (2017). Current nonclinical testing paradigm enables safe entry to First-In-Human clinical trials: The IQ consortium nonclinical to clinical translational database. *Toxicol Appl Pharmacol* **334**, 100-109.

Nicolas, C. I., Linakis, M. W., Minto, M. S., Mansouri, K., Clewell, R. A., Yoon, M., Wambaugh, J. F., Patlewicz, G., McMullen, P. D., Andersen, M. E*., et al.* (2022). Estimating provisional margins of exposure for data-poor chemicals using high-throughput computational methods. *Front Pharmacol* **13**, 980747.

NTP (2022). TP Developmental and Reproductive Toxicity Technical Report on the Modified One-Generation Study of 2-Ethylhexyl p-Methoxycinnamate (CASRN 5466-77-3) Administered in Feed to Sprague Dawley (Hsd:Sprague Dawley® SD®) Rats with Prenatal, Reproductive Performance, and Subchronic Assessments in F1 Offspring: DART Report 06 [Internet]. In (doi: 10.22427/NTP-DART-06.

Nyffeler, J., Haggard, D. E., Willis, C., Setzer, R. W., Judson, R., Paul-Friedman, K., Everett, L. J., and Harrill, J. A. (2021). Comparison of Approaches for Determining Bioactivity Hits from High-Dimensional Profiling Data. *SLAS Discov* **26**(2), 292-308.

O'Mahony, A., John, M. R., Cho, H., Hashizume, M., and Choy, E. H. (2018). Discriminating phenotypic signatures identified for tocilizumab, adalimumab, and tofacitinib monotherapy and their combinations with methotrexate. *J Transl Med* **16**(1), 156.

OECD (2022). Case Study on the use of Integrated Approaches to Testing and Assessment for potential Systemic Toxicity and Estrogen Receptor Activation of a Group of Bisphenols and Select Alternatives, Environment Directorate Chemicals & Biotechnology Committee Series on Testing and Assessment, Vol 373, ENV/CBC/MONO(2022)43, https://one.oecd.org/document/env/cbc/mono(2022)43/en/pdf. In (

OECD (2017), *Guidance on Grouping of Chemicals, Second Edition*, OECD Series on Testing and Assessment, No. 194, OECD Publishing, Paris, https://doi.org/10.1787/9789264274679-en.

OECD (2017). INTERNATIONALLY HARMONISED FUNCTIONAL, PRODUCT AND ARTICLE USE CATEGORIES, JOINT MEETING OF THE CHEMICALS COMMITTEE AND THE WORKING PARTY ON CHEMICALS, PESTICIDES & BIOTECHNOLOGY, ENVIRONMENT DIRECTORATE, ENV/JM/MONO(2017)14, https://one.oecd.org/document/ENV/JM/MONO(2017)14/en/pdf. In (

OECD (2023). (Q)SAR Assessment Framework: Guidance for the regulatory assessment of (Quantitative) Structure Activity Relationship models and predictions. In (doi: https://doi.org/10.1787/d96118f6-en). OECD Publishing, Paris, France.

Ouedraogo, G., Alexander-White, C., Bury, D., Clewell, H. J., 3rd, Cronin, M., Cull, T., Dent, M., Desprez, B., Detroyer, A., Ellison, C.*, et al.* (2022). Read-across and new approach methodologies applied in a 10-step framework for cosmetics safety assessment - A case study with parabens. *Regul Toxicol Pharmacol* **132**, 105161.

Paini, A., Cole, T., Meinero, M., Carpi, D., Deceuninck, P., Macko, P., Palosaari, T., Sund, J., Worth, A., and Whelan, M. (2020). EURL ECVAM in vitro hepatocyte clearance and blood plasma protein binding dataset for 77 chemicals. In (E. C. J. R. Centre, Ed.).

Palmer, J. A., Smith, A. M., Egnash, L. A., Conard, K. R., West, P. R., Burrier, R. E., Donley, E. L., and Kirchner, F. R. (2013). Establishment and assessment of a new human embryonic stem cell-based biomarker assay for developmental toxicity screening. *Birth Defects Res B Dev Reprod Toxicol* **98**(4), 343-63.

Patlewicz, G., Wambaugh, J. F., Felter, S. P., Simon, T. W., and Becker, R. A. (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.

Patlewicz, G., Williams, A., Adams, M., Shah, I., and Paul Friedman, K. (2024). A cheminformatics workflow to select representative TSCA chemicals for New Approach Methodology (NAM) screening. *Chem Res Toxicol* **Accepted**.

Paul Friedman, K., Foster, M. J., Pham, L. L., Feshuk, M., Watford, S. M., Wambaugh, J. F., Judson, R., Setzer, R. W., and Thomas, R. S. (2023). Reproducibility of organ-level effects in repeat dose animal studies. *Computational Toxicology* **28**.

Paul Friedman, K., Gagne, M., Loo, L. H., Karamertzanis, P., Netzeva, T., Sobanski, T., Franzosa, J. A., Richard, A. M., Lougee, R. R., Gissi, A.*, et al.* (2020). Utility of In Vitro Bioactivity as a Lower Bound Estimate of In Vivo Adverse Effect Levels and in Risk-Based Prioritization. *Toxicol Sci* **173**(1), 202-225.

Pearce, R. G., Setzer, R. W., Strope, C. L., Sipes, N. S., and Wambaugh, J. F. (2017). httk: R Package for High-Throughput Toxicokinetics. *2017* **79**(4), 26 (high-throughput; ToxCast; httk; toxicokinetics; pharmacokinetics).

Pham, L. L., Watford, S., Pradeep, P., Martin, M. T., Thomas, R., Judson, R., Setzer, R. W., and Paul Friedman, K. (2020). Variability in in vivo studies: Defining the upper limit of performance for predictions of systemic effect levels. *Comput Toxicol* **15**(August 2020), 1-100126.

Pognan, F., Beilmann, M., Boonen, H. C. M., Czich, A., Dear, G., Hewitt, P., Mow, T., Oinonen, T., Roth, A., Steger-Hartmann, T.*, et al.* (2023). The evolving role of investigative toxicology in the pharmaceutical industry. *Nat Rev Drug Discov* **22**(4), 317-335.

Pradeep, P., Friedman, K. P., and Judson, R. (2020a). Structure-based QSAR Models to Predict Repeat Dose Toxicity Points of Departure. *Comput Toxicol* **16**(November 2020).

Pradeep, P., Patlewicz, G., Pearce, R., Wambaugh, J., Wetmore, B., and Judson, R. (2020b). Using Chemical Structure Information to Develop Predictive Models for In Vitro Toxicokinetic Parameters to Inform High-throughput Risk-assessment. *Comput Toxicol* **16**.

Punt, A., Aartse, A., Bovee, T. F. H., Gerssen, A., van Leeuwen, S. P. J., Hoogenboom, R., and Peijnenburg, A. (2019). Quantitative in vitro-to-in vivo extrapolation (QIVIVE) of estrogenic and anti-androgenic potencies of BPA and BADGE analogues. *Arch Toxicol* **93**(7), 1941-1953.

Reynolds, G., Reynolds, J., Gilmour, N., Cubberley, R., Spriggs, S., Aptula, A., Przybylak, K., Windebank, S., Maxwell, G., and Baltazar, M. T. (2021). A hypothetical skin sensitisation next generation risk assessment for coumarin in cosmetic products. *Regul Toxicol Pharmacol* **127**, 105075.

Richard, A. M., Tao, D., LeClair, C. A., Leister, W., Tretyakov, K. V., White, E., Lewis, K. C., Sefler, A., Collins, B. J., Nguyen, D. T.*, et al.* (In review.). Chemical Quality Evaluation of the Tox21 Compound Library.

Ring, C. L., Arnot, J. A., Bennett, D. H., Egeghy, P. P., Fantke, P., Huang, L., Isaacs, K. K., Jolliet, O., Phillips, K. A., Price, P. S.*, et al.* (2019). Consensus Modeling of Median Chemical Intake for the U.S. Population Based on Predictions of Exposure Pathways. *Environ Sci Technol* **53**(2), 719-732.

Ring, C. L., Pearce, R. G., Setzer, R. W., Wetmore, B. A., and Wambaugh, J. F. (2017). Identifying populations sensitive to environmental chemicals by simulating toxicokinetic variability. *Environ Int* **106**, 105-118.

Rotroff, D. M., Wetmore, B. A., Dix, D. J., Ferguson, S. S., Clewell, H. J., Houck, K. A., Lecluyse, E. L., Andersen, M. E., Judson, R. S., Smith, C. M.*, et al.* (2010). Incorporating human dosimetry and exposure into high-throughput in vitro toxicity screening. *Toxicol Sci* **117**(2), 348-58.

Scherer, M., Feshuk, M., Paul Friedman, K., and Wambaugh, J. (in prep). Empirical Evaluation of In Vitro Disposition Models.

Shah, F., Stepan, A. F., O'Mahony, A., Velichko, S., Folias, A. E., Houle, C., Shaffer, C. L., Marcek, J., Whritenour, J., Stanton, R.*, et al.* (2017). Mechanisms of Skin Toxicity Associated with Metabotropic Glutamate Receptor 5 Negative Allosteric Modulators. *Cell Chem Biol* **24**(7), 858-869 e5.

Shibata, Y., Takahashi, H., Chiba, M., and Ishii, Y. (2002). Prediction of hepatic clearance and availability by cryopreserved human hepatocytes: an application of serum incubation method. *Drug Metab Dispos* **30**(8), 892-6.

Simms, L., Mason, E., Berg, E. L., Yu, F., Rudd, K., Czekala, L., Trelles Sticken, E., Brinster, O., Wieczorek, R., Stevenson, M.*, et al.* (2021). Use of a rapid human primary cell-based disease screening model, to compare next generation products to combustible cigarettes. *Curr Res Toxicol* **2**, 309-321.

Singer, J. W., Al-Fayoumi, S., Taylor, J., Velichko, S., and O'Mahony, A. (2019). Comparative phenotypic profiling of the JAK2 inhibitors ruxolitinib, fedratinib, momelotinib, and pacritinib reveals distinct mechanistic signatures. *PLOS ONE* **14**(9), e0222944.

Sipes, N. S., Martin, M. T., Kothiya, P., Reif, D. M., Judson, R. S., Richard, A. M., Houck, K. A., Dix, D. J., Kavlock, R. J., and Knudsen, T. B. (2013). Profiling 976 ToxCast chemicals across 331 enzymatic and receptor signaling assays. *Chem Res Toxicol* **26**(6), 878-95.

Sipes, N. S., Wambaugh, J. F., Pearce, R., Auerbach, S. S., Wetmore, B. A., Hsieh, J. H., Shapiro, A. J., Svoboda, D., DeVito, M. J., and Ferguson, S. S. (2017). An Intuitive Approach for Predicting Potential Human Health Risk with the Tox21 10k Library. *Environ Sci Technol* **51**(18), 10786-10796.

Stanley, L. A., and Wolf, C. R. (2022). Through a glass, darkly? HepaRG and HepG2 cells as models of human phase I drug metabolism. *Drug Metab Rev* **54**(1), 46-62.

Strickland, J. D., Martin, M. T., Richard, A. M., Houck, K. A., and Shafer, T. J. (2018). Screening the ToxCast phase II libraries for alterations in network function using cortical neurons grown on multi-well microelectrode array (mwMEA) plates. *Arch Toxicol* **92**(1), 487-500.

Su, R., Xiong, S., Zink, D., and Loo, L. H. (2016). High-throughput imaging-based nephrotoxicity prediction for xenobiotics with diverse chemical structures. *Arch Toxicol* **90**(11), 2793-2808.

Thomas, R. S., Bahadori, T., Buckley, T. J., Cowden, J., Deisenroth, C., Dionisio, K. L., Frithsen, J. B., Grulke, C. M., Gwinn, M. R., Harrill, J. A.*, et al.* (2019). The Next Generation Blueprint of Computational Toxicology at the U.S. Environmental Protection Agency. *Toxicol Sci* **169**(2), 317-332.

Thomas, R. S., Philbert, M. A., Auerbach, S. S., Wetmore, B. A., Devito, M. J., Cote, I., Rowlands, J. C., Whelan, M. P., Hays, S. M., Andersen, M. E.*, et al.* (2013a). Incorporating new technologies into toxicity testing and risk assessment: moving from 21st century vision to a data-driven framework. *Toxicol Sci* **136**(1), 4-18.

Thomas, R. S., Wesselkamper, S. C., Wang, N. C., Zhao, Q. J., Petersen, D. D., Lambert, J. C., Cote, I., Yang, L., Healy, E., Black, M. B.*, et al.* (2013b). Temporal concordance between apical and transcriptional points of departure for chemical risk assessment. *Toxicol Sci* **134**(1), 180-94.

USEPA (2023a). ChemExpo Knowledgebase, Harmonized Functional Use Data Bulk Download, doi: https://doi.org/10.23645/epacomptox.20422176. In (

USEPA (2022a). *The New Chemicals Collaborative Research Program: Modernizing the Process and Bringing Innovative Science to Evaluate New Chemicals Under TSCA, Office of Research and Development, Office of Chemical Safety and Pollution Prevention.* Available at: https://www.epa.gov/bosc/bosc-review-panel-meeting-october-2022. Accessed December 15, 2024.

USEPA (2022b). *Predictive Models and Tools for Assessing Chemicals Under TSCA.* Available at: https://www.epa.gov/tsca-screening-tools. Accessed May 31, 2023.

USEPA (2023b). *Scientific Studies Supporting Development of Transcriptomic Points of Departure for EPA Transcriptomic Assessment Products (ETAPs), EPA/600/X-23/084, Office of Research & Development, doi: 10.23645/epacomptox.25365550*.

USEPA (2022c). *ToxCast database: Invitrodb version 3.5, doi: https://doi.org/10.23645/epacomptox.6062623.v10.* Available at: https://clowder.edap-cluster.com/spaces/62bb560ee4b07abf29f88fef.

USEPA (2023c). *ToxVal 9.4, Office of Research & Development Center for Computational Toxicology and Exposure, doi: https://doi.org/10.23645/epacomptox.20394501.v5*.

USEPA (2020). *User's Guide for T.E.S.T. (version 5.1) (Toxicity Estimation Software Tool): A Program to Estimate Toxicity from Molecular Structure, Office of Research and Development Center for Computational Toxicology & Exposure, https://www.epa.gov/sites/default/files/2016-05/documents/600r16058.pdf*.

Valdivia, P., Martin, M., LeFew, W. R., Ross, J., Houck, K. A., and Shafer, T. J. (2014). Multi-well microelectrode array recordings detect neuroactivity of ToxCast compounds. *Neurotoxicology* **44**, 204-17.

Waters, N. J., Jones, R., Williams, G., and Sohal, B. (2008). Validation of a rapid equilibrium dialysis approach for the measurement of plasma protein binding. *J Pharm Sci* **97**(10), 4586-95.

Wetmore, B. A., Wambaugh, J. F., Allen, B., Ferguson, S. S., Sochaski, M. A., Setzer, R. W., Houck, K. A., Strope, C. L., Cantwell, K., Judson, R. S*., et al.* (2015). Incorporating High-Throughput Exposure Predictions With Dosimetry-Adjusted In Vitro Bioactivity to Inform Chemical Toxicity Testing. *Toxicol Sci* **148**(1), 121-36.

Wetmore, B. A., Wambaugh, J. F., Ferguson, S. S., Li, L., Clewell, H. J., 3rd, Judson, R. S., Freeman, K., Bao, W., Sochaski, M. A., Chu, T. M*., 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**(2), 327-46.

Wetmore, B. A., Wambaugh, J. F., Ferguson, S. S., Sochaski, M. A., Rotroff, D. M., Freeman, K., Clewell, H. J., 3rd, Dix, D. J., Andersen, M. E., Houck, K. A*., et al.* (2012). Integration of dosimetry, exposure, and high-throughput screening data in chemical toxicity assessment. *Toxicol Sci* **125**(1), 157-74.

Williams, A. J., Grulke, C. M., Edwards, J., McEachran, A. D., Mansouri, K., Baker, N. C., Patlewicz, G., Shah, I., Wambaugh, J. F., Judson, R. S*., et al.* (2017). The CompTox Chemistry Dashboard: a community data resource for environmental chemistry. *J Cheminform* **9**(1), 61.

Zurlinden, T. J., Saili, K. S., Rush, N., Kothiya, P., Judson, R. S., Houck, K. A., Hunter, E. S., Baker, N. C., Palmer, J. A., Thomas, R. S*., et al.* (2020). Profiling the ToxCast Library With a Pluripotent Human (H9) Stem Cell Line-Based Biomarker Assay for Developmental Toxicity. *Toxicol Sci* **174**(2), 189-209.

Zwickl, C. M., Graham, J., Jolly, R., Bassan, A., Ahlberg, E., Amberg, A., Anger, L. T., Barton-Maclaren, T., Beilke, L., Bellion, P*., et al.* (2022). Principles and Procedures for Assessment of Acute Toxicity Incorporating In Silico Methods. *Comput Toxicol* **24**.