Chemical Research in Toxicology

ACS EDITORS' CHOICE

pubs.acs.org/crt Review

The Tox21 10K Compound Library: Collaborative Chemistry Advancing Toxicology

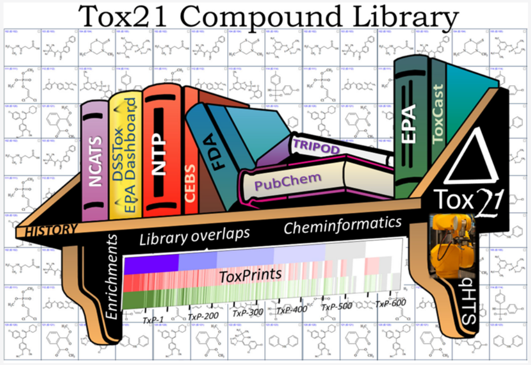
Ann M. Richard,* Ruili Huang, Suramya Waidyanatha, Paul Shinn, Bradley J. Collins, Inthirany Thillainadarajah, Christopher M. Grulke, Antony J. Williams, Ryan R. Lougee, Richard S. Judson, Keith A. Houck, Mahmoud Shobair, Chihae Yang, James F. Rathman, Adam Yasgar, Suzanne C. Fitzpatrick, Anton Simeonov, Russell S. Thomas, Kevin M. Crofton, Richard S. Paules, John R. Bucher, Christopher P. Austin, Robert J. Kavlock, and Raymond R. Tice

Cite This: *Chem. Res. Toxicol.* 2021, 34, 189−216

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Since 2009, the Tox21 project has screened ∼8500 chemicals in more than 70 high-throughput assays, generating upward of 100 million data points, with all data publicly available through partner websites at the United States Environmental Protection Agency (EPA), National Center for Advancing Translational Sciences (NCATS), and National Toxicology Program (NTP). Underpinning this public effort is the largest compound library ever constructed specifically for improving understanding of the chemical basis of toxicity across research and regulatory domains. Each Tox21 federal partner brought specialized resources and capabilities to the partnership, including three approximately equal-sized compound libraries. All Tox21 data generated to date have resulted from a confluence of ideas, technologies, and expertise used to design, screen, and analyze the Tox21 10K library. The different programmatic objectives of the partners led to three distinct, overlapping compound libraries that, when combined, not only covered a diversity of chemical structures, use-categories, and properties but also incorporated many types of compound replicates. The history of development of the Tox21 "10K" chemical library and data workflows implemented to ensure quality chemical annotations and allow for various reproducibility assessments are described. Cheminformatics profiling demonstrates how the three partner libraries complement one another to expand the reach of each individual library, as reflected in coverage of regulatory lists, predicted toxicity end points, and physicochemical properties. ToxPrint chemotypes (CTs) and enrichment approaches further demonstrate how the combined partner libraries amplify structure−activity patterns that would otherwise not be detected. Finally, CT enrichments are used to probe global patterns of activity in combined ToxCast and Tox21 activity data sets relative to test-set size and chemical versus biological end point diversity, illustrating the power of CT approaches to discern patterns in chemical−activity data sets. These results support a central premise of the Tox21 program: A collaborative merging of programmatically distinct compound libraries would yield greater rewards than could be achieved separately.



1. BACKGROUND

Prior to 2004, the construction and high-throughput screening (HTS) of compound libraries were primarily the domain of the pharmaceutical industry, directed toward the goal of identifying candidates for drug development that interact specifically, and with high potency, with a wide range of putative therapeutic targets. Drug libraries typically consist of large collections of small molecules (<500−1000 g/mol), ranging from tens of thousands to millions of chemicals and containing a diverse set of synthesized, designed, combinatorially produced, and/or commercially procured compounds along with their associated property data.[1,2] The handling, storage, plating, screening and analysis of large libraries, in turn, require modern robotics and HTS assay technologies as well as the bioinformatics and cheminformatics infrastructure and tools for storing and processing the large quantities of data generated. However, because pharmaceutical compound libraries represent valuable corporate investments, the contents and associated data are considered proprietary intellectual property and are not made publicly available.

Special Issue: Computational Toxicology

Received: July 2, 2020
Published: November 3, 2020

https://dx.doi.org/10.1021/acs.chemrestox.0c00264

In late 2004, the National Institutes of Health (NIH) announced the Molecular Libraries Initiative (MLI) as a component of the NIH Roadmap for Medical Research.[3] The MLI led to the creation of the NIH Molecular Libraries Small Molecule Repository (MLSMR), a compound library of >300 K substances procured from commercial sources. Recognizing the necessity of an informatics support infrastructure, the program also included the creation of PubChem (https://pubchem.ncbi.nlm.nih.gov/), a public, structure-centric, chem-informatics platform whose initial mandate was to serve as the public repository for all chemical structures and HTS data generated by the MLI program.[4,5] The MLSMR compound library underwent screening at multiple NIH-funded extra-mural sites as well as at the intramural, state-of-the-art robotics facility at NIH's Chemical Genomics Center (NCGC), later incorporated into the Division of Preclinical Innovation at the NIH National Center for Advancing Translational Sciences (NCATS). Thus, the MLSMR and PubChem, created with the aim of providing support for basic research in the medical sciences, represented the first major, publicly funded effort to generate and publish HTS bioassay results for a large compound library.

Not long after creation of the MLSMR and PubChem, the release of the National Research Council's "Toxicology in the 21st Century" report initiated a transformative paradigm shift in the field of toxicology toward development and application of higher throughput *in vitro* systems and computational modeling to replace costly and time-consuming *in vivo* animal testing.[6] The new research direction was aimed at addressing the failure of traditional toxicity testing methods to handle the increasingly large backlog of environmental chemicals lacking toxicity data and, additionally, providing more human and pathway-relevant data for informing chemical toxicity assessment by harnessing new computational and HTS technologies, many having been commercialized to service the pharmaceutical industry. Prior to the release of the NRC report, a major shift in research focus toward computational toxicology was already well underway within both the United States Environmental Protection Agency (EPA) and the National Toxicology Program (NTP), a Division of the National Institute of Environmental Health Sciences (NIEHS).[7,8] Institutionalizing this shift, EPA's National Center for Computational Toxicology (NCCT) was formed in 2005, followed by the launch of its signature Toxicity Forecaster (ToxCast) HTS program.[9] In addition, by 2006, both the EPA and NTP had begun creating collections of chemicals for proof-of-concept high-throughput testing in collaboration with NCGC, to be screened at the NCGC robotic testing facility. The Tox21 Program was formally launched in 2008 as a collaboration among these three federal partners—EPA, NTP, and NCGC/NCATS— with the United States Food and Drug Administration (FDA) joining the partnership in 2010.[10−12] (Note: For purposes of this report, we will henceforth use the label NCATS to encompass both organizational periods, that is, NCGC until 2012 and NCATS after 2012.) The stated goals of the program included: research, development, validation, and translation of innovative compound testing methods to characterize toxicity pathways, prioritizing compounds for more extensive toxicological evaluation, and developing predictive models for biological response in humans and the environment. To help achieve these goals, the Tox21 partners created a compound library of ∼10,000 agency-relevant chemical samples to be screened in a battery of quantitative HTS (qHTS) assays at the NCATS intramural robotics facility.[13] Testing initially focused on a broad set of assays associated with nuclear receptor activities and stress pathways. In contrast to the single concentration being typically tested during HTS in drug discovery, the qHTS approach screens compounds at multiple concentrations over 4 orders of magnitude, which is more suitable for detecting weakly active compounds, building full dose−response relationships for the active molecules, and more confidently designing inactive compounds. Given the Tox21 programmatic need to increase confidence in the results for individual chemicals, each chemical was run in triplicate in each assay, and analytical chemistry quality control (QC) testing was undertaken for each individual chemical. Finally, a signature feature of the program was that the compound library chemical identifiers and associated assay results were to be made publicly available.

Each of the Tox21 federal partners brought unique expertise, resources, and capabilities to the Tox21 partnership, including three separately sourced, partially overlapping, and approximately equal-sized compound libraries from the three original partners (i.e., EPA, NTP and NCATS). The consolidation of those partner libraries, in turn, came to comprise the full Tox21 screening library of ∼10,000 chemical samples, often referred to as the "Tox21 10K library". (Note: Henceforth, the term "Tox21 10K library" will be synonymous with "full Tox21 library", referring to the first and later iterations of the ∼10K chemical samples in the consolidated screening library.) Analogous to the large compound libraries that fuel drug discovery efforts, but toward different objectives and with a commitment to full public data release, the Tox21 10K library was designed to cover a structurally diverse chemical space, spanning the broad interests of the Tox21 partners, and to serve as a chemical probe set of a broad panel of *in vitro* bioactivities potentially informative of toxicity and adverse outcomes in humans. This library was orders of magnitude larger and more structurally diverse than any chemical library previously screened in traditional toxicity assays. Hence, the programmatic objectives of the Tox21 project were ambitiously broad - *to generate screening data across a wide diversity of chemicals and potential toxicity mechanisms* - as well as focused - *to generate HTS bioactivity profiles across a wide range of biochemical targets and cellular end points for individual chemicals of environmental toxicological interest*. Building the Tox21 10K library was the first step in this process.

The Tox21 library was largely completed in 2012, with some later additions and reprocurements, and was screened at NCATS from that point forward. Today, nearly 8 years after the start of full library screening, Tox21 qHTS data are still being generated and published. Tox21 qHTS results, to date, have yielded upward of 100 million chemical assay data points for over 70 distinct assays, corresponding to more than 200 separate assay end point read outs (http://www.ncbi.nlm.nih.gov/pcassay?term=tox21, accessed October 31, 2019). In addition, the Tox21 program has produced over 110 publications (https://tox21.gov/all-publications/) describing various aspects of the program and results, including details of the plating and robotic testing platform,[13] the assay data analysis pipeline,[14−16] analyses of assay results,[17−19] predictive modeling of assay results in relation to toxicity end points,[20−24] along with periodic updates and perspectives on the program,[11,25] and, more recently, a new strategic plan for future Tox21 collaborative projects.[26] Additionally, the commitment to the public release of all screening data within six months of the screen being completed has encouraged uptake and analysis of Tox21 chemical−activity data by the broader scientific community, resulting in hundreds more publications in areas ranging from predictive toxicity to exposure modeling and nontargeted screening.[20,27−30] Lastly a Tox21-sponsored challenge, specifically aimed at the machine-learning and predictive toxicity modeling communities, asked contributors to build models for nuclear receptor and stress response pathways.[31] This challenge yielded a series of models and articles from outside contributors (an e-book containing all author contributions is freely available for download at: https://www.frontiersin.org/research-topics/2954/tox21-challenge-to-build-predictive-models-of-nuclear-receptor-and-stress-response-pathways-as-media#articles).

Despite the broad use and application of Tox21 chemical and assay data, there has been no comprehensive account of the etiology and history of the Tox21 10K library construction to date. This history, which should inform use of the library today, includes descriptions of how the three federal partner libraries were built and combined and their respective relationships to the Phase I NTP HTS library,[32] NIH's NCGC Pharmaceutical Collection (presently the NCATS Pharmaceutical Collection (NPC)),[33] and EPA's ToxCast library.[34] This history also encompasses decisions pertaining to sample tracking and compound registration that were implemented to ensure the integrity of assay screening results as well as cheminformatics coordination challenges that impact downstream analysis and data sharing. These decisions as well as the overall chemical library construction and management undertaken by each Tox21 partner were largely the responsibility of the Tox21 Chemical Workgroup, comprised of a lead chemist from EPA, NTP, and NCATS. Unique to this public project, a decade-long effort to generate analytical QC results on the full Tox21 library was carried out in parallel to Tox21 screening efforts, with summary spectra and QC results for individual chemical samples made publicly available (https://tripod.nih.gov/tox21/samples, accessed February 10, 2020). However, analytical QC results linked to method details have not yet been published, nor have the QC data been analyzed to illuminate patterns associated with analytical QC method, molecular structure, or qHTS results. Finally, the detailed chemical−structural composition of the Tox21 compound library has received relatively little attention except in relation to EPA's Tox21 library contribution, which largely overlapped with EPA's ToxCast library.[34] A survey of the Tox21 compound library, in terms of chemical substance usage, features, and properties as well as the sample tracking and analysis steps taken to ensure sample integrity, extend beyond purely historical interest. Never before has a public HTS effort leveraged such a large, structurally and mechanistically diverse compound library toward the goal of enriching knowledge of the chemical−biological basis for toxicity. Hence, these details represent foundational aspects of the Tox21 program that directly influence how the data generated by the program, to date, are understood, analyzed, and interpreted moving forward.

This first paper in a planned series of Tox21 compound library publications will cover the history of the library construction and apply cheminformatics methods to profile the three federal partner contributions in relation to one another and the complete Tox21 library. Sections 2 and 3 of the present review will cover the initial phase of pilot testing and library construction, to subsequent consolidation of the three partner libraries in preparation for full Tox21 library screening (Phase I), to the means by which samples were tracked, annotated, and represented cheminformatically, and chemical assay results were publicly distributed (Phase II). Sections 4 and 5 will proceed to examine the Tox21 compound library through a variety of structural and property lenses, largely focusing on the overlaps, differences, and complementary nature of the three partner library contributions in relation to the whole. In particular, the cheminformatics profiling will endeavor to show how the three Tox21 partner libraries, developed independently and with different programmatic objectives, complement one another to expand the reach of each individual library, as reflected in coverage of regulatory lists, predicted toxicity end points, and physicochemical properties. ToxPrint chemotypes (CTs) and enrichment approaches will further demonstrate how combining the partner libraries successfully amplifies structure−activity patterns in the HTS results that would otherwise not be detectable in the separate libraries. Finally, we employ CT enrichments to probe global patterns of activity, illuminating influences of test-set size, and chemical versus biological space diversity on the enrichment results. Given that EPA's ToxCast compound library, largely comprising EPA's contribution to the Tox21 compound library, has been screened in a larger, more biologically varied set of *in vitro* HTS assays, the CT enrichment results across both Tox21 and ToxCast assay data sets provide a unique opportunity to examine the relative influence of chemical library size and diversity versus biological end point profile diversity on CT enrichment patterns.[35]

The present article is not intended as a review and evaluation of the overall Tox21 program and its progress and successes toward the stated goals of transforming predictive toxicology. Rather, the history, description, and cheminformatics profiling of the Tox21 compound library is intended to enrich public understanding of the Tox21 effort from a chemical perspective, encourage a more informed chemistry-based exploration of the Tox21 chemical−bioactivity data landscape, and lay the groundwork for the in-depth reporting and analysis of the analytical QC results. Subsequent papers will relate the history of the decade-long effort to generate, aggregate, and publish analytical chemistry QC results for the full Tox21 testing library, with the data revealing patterns of analytical QC failure (such as low purity, low concentration, and degradation over time) in relation to chemical structure features and properties, and ultimately assay results. Hence, this first in the series of Tox21 compound library papers is intended to inform and guide current application and analysis of Tox21 data as well as future compound screening, assay interpretation, and predictive modeling activities in toxicology.

2. TOX21 PHASE I

Construction of the full Tox21 compound library began as separate, agency-directed activities within the EPA, NTP, and NCATS, with each partner focused on selecting, procuring, solubilizing, and plating compounds within their areas of regulatory interest and to serve multiple agency-specific programmatic uses. For these and logistical reasons, there was little initial coordination of the library chemical selection among the partners. Phase I of the Tox21 program refers to the period from 2007 to 2011, covering library construction through to the start of screening of the full Tox21 library in early 2012, hereafter referred to as Phase II of the Tox21

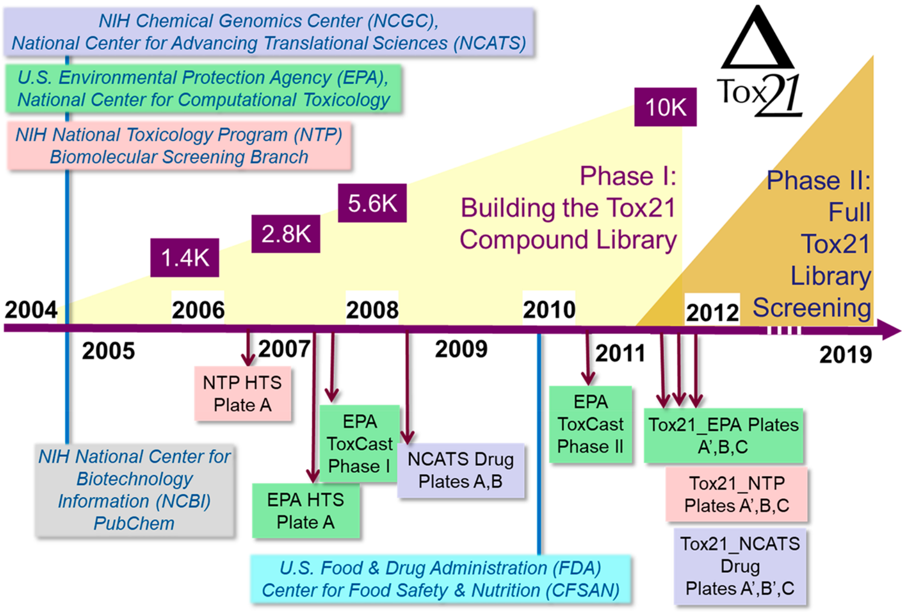


Figure 1. Approximate timeline for constructing the full Tox21 compound library. Plates A, B, and C (and revised/reprocured plates A′ and B′) refer to construction of 1536-well plates containing up to 1408 compounds each, for each of the three component partner libraries, Tox21_EPA, Tox21_NTP, and Tox21_NCATS.

program. An approximate timeline of the separate partner activities leading up to the start of screening of the full Tox21 library is provided in Figure 1.

Both EPA and NTP participated in early pilot projects in which an initial set of 1462 and 1408 compounds, respectively, from each partner (EPA HTS Plate A and NTP HTS Plate A), along with the available portion of the NCATS's drug library (NCATS Drug Plates A, B), were screened at NCATS. These initial compound sets, in each case, represented the chemical libraries that were available to each partner organization at that time. These samples were largely depleted during this initial screening effort, or shortly after Tox21 full library screening commenced (in the case of NCATS) and, thus, were reprocured for inclusion in the full Tox21 library; the reprocured plates are denoted A′ in the cases of EPA and NTP and A′ and B′ in the case of NCATS. For each organization, the process of compound selection during all testing phases was constrained by the ability to procure the chemical commercially (or to have it donated or custom synthesized in a few cases) and the extent to which a compound was soluble in dimethyl sulfoxide (DMSO), the solvent of choice for HTS studies, at a target concentration of 10−20 mM. Other programmatic considerations that weighed into compound selection for each of the three initial federal partners, leading to their final plate set contributions to the full Tox21 library, are summarized below. Whereas the FDA's inclusion in the Tox21 partnership did not entail a separate compound library contribution, it should be noted that the NCATS, NTP, and EPA libraries included many chemicals of interest to FDA researchers and safety assessors, including drugs and substances found in food-contact and personal care products.

2.1. NCATS Tox21 Compound Library. NCATS's contribution to the 10K library consisted of a total of 3764 unique compounds designated as "drugs"; this component library is referred to, henceforth, as "Tox21_NCATS". The initial contribution consisted of the subset of the full NPC library that was available as physical samples and deemed HTS amenable at the start of Phase II of the Tox21 program (https://tripod.nih.gov/npc/). The NPC was created first as a database and informatics resource to serve as a publicly accessible, definitive listing of drugs intended or approved for human use.[33] Second, the intent was to create a corresponding library of physical samples for qHTS that would be available to NCATS researchers and intramural collaborators. The NPC was designed to support research in drug repurposing, advanced drug development and chemical genomics, as well as to improve the mechanistic understanding of the toxicity of drugs through the Tox21 collaboration. For the purpose of compiling the NPC, a clear chemical lexicon for the term "drug" was required. Drug uniqueness was defined in association with a single molecule entity (ME), that is, encompassing neutral, salt, and hydrate forms, and a corresponding single active pharmaceutical ingredient (API), that is, a particular neutral, salt, or hydrate form. Although multiple APIs can represent a single ME and different APIs for the same ME can exhibit distinct absorption, distribution, metabolism, excretion (ADME) and toxicity properties that can impact *in vivo* activity, it was argued that API distinctions are less important for *in vitro* and *in silico* studies; hence, a single API form, that is, unique "drug" structure (i.e., parent, salt, or complex form), was chosen for inclusion in the NPC screening library. By 2011, the full NPC database listing reportedly contained 8969 unique MEs, of which 3526 physical samples were procured by the start of full Tox21 library screening.

Data sources for bioactive compounds and approved drugs (including unapproved substances tested in humans) used to construct the NPC included listings extracted from publicly available documents of the FDA, the European Medicines Agency, Health Canada, the United Kingdom's National Health Service, and the Japanese National Institute of Health Sciences Pharmacopeia (for a full listing of data sources, see Huang et al.).[33] The sources for compiling the NPC physical screening collection included commercial chemical suppliers, specialty collections, pharmacies, and limited custom synthesis. Due to compound cost and availability challenges and use of the library for non-Tox21 research projects, limited quantities of the initial NPC screening library were available as Phase II, full Tox21 library screening began; hence, efforts were undertaken to reprocure the library soon thereafter. In so doing, NCATS withdrew some drugs that were found to be problematic in earlier testing (e.g., volatile) and included several hundred drugs not previously included. The final count for the expanded Tox21_NCATS library incorporated into testing shortly after the start of Phase II was 3764.

2.2. NTP Tox21 Compound Library. NTP's contribution to the 10K library at the start of full Tox21 library screening consisted of a total of 3115 unique compounds spanning many areas of programmatic environmental and toxicological concern; this component library is referred to, henceforth, as "Tox21_NTP". In selecting compounds for the Tox21 10K library, NTP first considered compounds that had either been nominated for NTP testing or that had previously been included in the intramural NTP testing program, including compounds tested in the NTP rodent bioassay, screened for genotoxicity, or submitted for studies to evaluate ADME properties. The initial NTP Plate A set of 1408 compounds was largely drawn from these available, previously characterized chemicals, many of which were subsequently reprocured and/or replated for the NTP Plate A′ contribution to the full Tox21 library. [Note: The historical EPA DSSTox substance−structure list created for NTP Plate A, denoted NTPHTS (last updated on May 1, 2005), is available for download from PubChem at https://www.ncbi.nlm.nih.gov/pcsubstance/?term=NTPHTS]. For their expanded library, NTP selected chemicals from lists considered relevant to their toxicology research interests, including the NTP 11th Report on Carcinogens (https://ntp.niehs.nih.gov/whatwestudy/assessments/cancer/roc/), EPA's High Production Volume chemical list, and the Carcinogenic Potency Database (CPDB) list (obtained from EPA's DSSTox HPVCSI and CPDBAS chemical lists available at the time),[36] chemicals evaluated by NTP's Center for the Evaluation of Risks to Human Reproduction,[37] reference compounds from *in vivo* regulatory tests identified by the NTP Interagency Center for the Evaluation of Alternative Toxicological Methods/Interagency Coordinating Committee on the Validation of Alternative Methods (https://ntp.niehs.nih.gov/whatwestudy/niceatm/iccvam/), and compounds recommended by NTP external collaborators. Finally, NTP's library included a set of 118 formulated mixtures of known estrogenic and androgenic compounds, with varying potency and molar fraction ratios of mixture components (Parham et al., submitted). (Note: Historical DSSTox structure files for HPVCSI and CPDBAS are available for download from PubChem at: https://www.ncbi.nlm.nih.gov/pcsubstance/?term=HPVCSI and https://pubchem.ncbi.nlm.nih.gov/#query=CPDBAS, respectively. For an updated list of EPA's High Production Volume chemicals, see https://comptox.epa.gov/dashboard/chemical_lists/EPAHPV. An archived version of the CPDB is available for download from https://www.nlm.nih.gov/databases/download/cpdb.html.)

After consolidating the various lists and excluding duplicates, some chemicals were rejected for procurement if deemed likely to be a gas, explosive, nerve agent, controlled substance, or unstable in DMSO solution. NTP contractors used both internal stocks and commercial sources to procure compounds for Tox21 testing. During the reprocurement process, several issues were encountered, including: compounds were a different complex or salt than requested, with a different molecular weight and Chemical Abstracts Service Registry Number (CAS RN); the supplier certificate of analysis (CoA) did not match the CAS RN ordered or on the bottle label; the supplier CoA stated <90% purity although material ordered stated higher purity; the CoA purity stated as "passed" with no method or % purity mentioned; the material received was expired; and some compounds, especially isomers, dyes, and commercial products were not available at higher purities. Where possible, corrections to the compound identity were made, but in the end, some lower purity, technical grade compounds were included. Since DMSO solubility was not initially known, a literature search was used to estimate DMSO solubility for the most expensive chemicals prior to procurement, whereas for the majority of the chemicals, DMSO solubility was empirically determined after procurement.

2.3. EPA Tox21 Compound Library. EPA's contribution to the 10K library consisted of a total of 4078 unique compounds. This set comprised the complete set of procured ToxCast compounds available at that time that were deemed suitable for screening. These were either candidates for ToxCast testing or had already been included into ToxCast Phase I and Phase II testing.[34] This component library is referred to, henceforth, as "Tox21_EPA". A detailed review of the history of construction of EPA's ToxCast chemical inventory and its chemical structure−property usage landscape was published in 2016 and, at that time, the library maintained a 96% overlapping content with Tox21_EPA.[34] Hence, EPA's compound library was built for and was intended to serve dual purposes, that is, to supply chemicals for both the Tox21 federal partnership and the EPA ToxCast screening program. The reader is referred to the ToxCast Landscape review article for details of the Tox21_EPA chemical library construction and characteristics, which are briefly summarized below.[34]

For EPA's ToxCast program, chemicals were prioritized for testing in phases, referred to as ToxCast Phases I, II, and III. The ToxCast Phase I chemical inventory (ph1_v1), which predated Tox21 and initiated EPA's ToxCast screening program in 2007, was comprised of 310 unique compounds; the majority pesticide active ingredients that were associated with a rich complement of guideline *in vivo* animal toxicity study data for subchronic, chronic, reproductive, and developmental end points. ToxCast Phase I was intended to serve as a proof-of-principle of EPA's nascent screening program, largely relying upon contracts with a variety of commercial screening laboratories.[9] In 2009, with ToxCast Phase I chemical stocks depleted, the EPA set about reprocuring the majority of Phase I chemicals, along with a greatly expanded library of nominated chemicals that would serve as the EPA's contribution to the Tox21 program as well as fuel the expanding ToxCast testing program into Phases II and III.

Candidates for procurement moving into ToxCast Phase II and EPA's Tox21 library construction were compiled from a wide variety of ACToR and DSSTox lists available at the time[38,39] as well as nominations from EPA researchers and program offices, other government agencies (including FDA, which contributed lists of food-contact substances with available toxicity data), and outside collaborators. These included: lists pertaining to toxicity and bioactivity, such as rodent carcinogenicity, aquatic toxicity, genetic toxicity, developmental toxicity, and estrogen-receptor binding; regulatory lists pertaining to high production volume chemicals, disinfection byproducts, food-contact substances, and pesticides; and exposure-related use category lists, including fragrances, antimicrobials, and drugs. This combined candidate list of more than 19K substances was filtered to exclude most inorganics, complex mixtures, suspected volatiles (based on molecular weight and predicted vapor pressure), and highly lipophilic substances (based on predicted log octanol−water partition coefficient, log K_{ow}), resulting in a smaller set of approximately 7000 CAS RNs that were submitted to the EPA's ToxCast chemical contractor for sourcing and possible procurement. Of the 7000 candidates, approximately 4300 were commercially available and fell within cost limits; these were subsequently procured, and the majority (92%) were determined to be soluble in DMSO and became candidates for plating.[1] (Note: EPA's list of procured candidates for Tox21 testing that were determined by visual inspection to be insoluble in DMSO at concentrations <10 mM is available for download at https://comptox.epa.gov/dashboard/chemical_lists/CHEMINV_DMSOINSOLUBLES). In addition to commercially procured chemicals, approximately 150 chemicals were donated by non-EPA ToxCast collaborators, including 136 failed drugs donated by five major pharmaceutical companies, and a small set of "green" plasticizer alternatives and reference liver toxicants from the chemical industry and FDA's National Center for Toxicological Research, respectively.[34] Subsequent to the initial set of compounds plated for the 10K library (Plates A′ and B), the EPA added 352 more substances (Plate C), representing compounds most recently entered into the EPA's ToxCast Phase III testing program. This latter set expanded the coverage of chemicals of current regulatory concern to the EPA by including flame retardants and chemicals of interest to the EPA's Endocrine Disruption Screening Program (EDSP) for the 21st Century (EDSP21) (http://www2.epa.gov/endocrine-disruption/endocrine-disruptor-screening-program-21st-century-edsp21-workplan-summary).

3. TOX21 PHASE II

During the period of Tox21 partner library construction, plans were put in place to coordinate plating, chemical registration, library consolidation, analytical QC, and depositing of data into PubChem that would be necessary to track, publish, and analyze the Tox21 library and associated HTS data moving forward. The EPA's DSSTox program, already established and supporting the EPA's ToxCast program, was tasked to review and curate compounds in each of the partner libraries (further described in Section 3.1), ensuring that all unique samples and substances were registered with appropriate chemical structures and consistent identifiers according to established DSSTox quality review procedures.[40] Each of the three Tox21 partners, in turn, was responsible for procuring, solubilizing, and plating their unique sample libraries onto 384-well plates for shipment and further processing by NCATS. The one exception was in the selection and plating of 88 Tox21 replicate chemicals (further described in Section 3.2), which were to be procured and solubilized by the EPA from a single source and supplied to each of the three partners for inclusion on their respective plate sets. PubChem registration of the full set of Tox21 chemicals, along with deposits of soon-to-be-generated Tox21 qHTS assay data sets, would be managed by NCATS according to a similar protocol used to deposit qHTS data sets for the NIH MLI screening program into PubChem. Challenges moving forward would involve coordinating activities across the three federal agencies and taking into consideration partner capabilities and priorities in relation to overall Tox21 program objectives.

3.1. DSSTox Curation.

Both the EPA and NCATS had existing chemical library management systems to support their respective testing programs, that is, ToxCast and the MLI. NCATS had a structure-centric cheminformatics system that relied exclusively on supplier-provided information and was aligned with PubChem's data model, with NCGC sample IDs mapped to unique structures. In contrast, EPA's DSSTox data model for ToxCast sample registration considered the generic substance identifier (associated with a unique chemical name and CAS RN) as primary and the structure mapping as secondary.[40] Another way in which these programs differed was that most of the NCATS NPC drugs were procured in small quantities from commercial sources already solubilized (to 10 mM target concentration versus 20 mM target concentration for the EPA and NTP libraries) and with limited supporting documentation, that is, mostly without CoAs or safety data sheets. In contrast, both the EPA's ToxCast program and NTP's testing program procured chemicals in larger quantities, in "neat" (powder or pure liquid) form, with supporting documentation whenever possible. DSSTox registration of the EPA's ToxCast chemicals had implemented manual review of sample documentation since discrepancies across supplier compound identifiers, particularly structures, were commonly detected and required correction. Compared to final DSSTox-registered structures, this discrepancy rate for ToxCast chemicals was reported to exceed 20%, underscoring the need for manual documentation review.[34]

Hence, both NTP and NCATS provided the EPA with substance lists for their respective plated libraries, and these lists underwent independent DSSTox manual curation review. This process resulted in a unique DSSTox chemical listing for each of the three separately procured partner libraries, that is, Tox21_NCATS, Tox21_NTP, and Tox21_EPA, totaling over 10,000 samples differentiated at the substance-supplier/lot/stock solution level and comprising what was, henceforth, referred to as the Tox21 "10K" library. Substantial numbers of compound overlaps at the substance level (further discussed in Section 5.2) were the inevitable result of independent library construction efforts, but also resulted because the separate partner libraries were designed and intended to serve additional agency-specific programmatic needs beyond Tox21 (e.g., the EPA's ToxCast program).

3.2. Tox21 IDs.

The second informatics challenge was in determining how Tox21 chemical samples were to be tracked and reported with associated assay data over the course of Tox21 testing. To allow for differentiation of sample-level assay results in relation to Tox21 partner substance inventory (Tox21_NCATS, Tox21_NTP, or Tox21_EPA) as well as by supplier/lot/batch and solution properties (including date of solubility, concentration, and storage/thaw history), Tox21 partners agreed to track and publish assay screening results at the stock solution level. This would allow for assessment of variability in assay results due to supplier/lot/batch variation across the three partner libraries. This decision would also allow for the assessment and reporting of sample analytical QC results at the same stock solution level as was used for reporting assay results. Each stock solution, in turn, would have a documented provenance related to the original neat sample, supplier, date of solubilization, etc., separately tracked by each Tox21 partner.

Having agreed that the major avenue for public release of all Tox21 qHTS data would be through PubChem, the team also had to consider and work within the constraints of the PubChem data model. Up to that time, within each unique source assay data set deposited, PubChem required a strict 1:1 chemical substance-structure-assay (SID-CID-AID) data mapping. In the case of Tox21, a unique list of DSSTox chemical structures would be affiliated with the PubChem source "Tox21", and all Tox21 assays would be assigned PubChem assay identifiers (AIDs) associated with that single Tox21 source. Within the same assay data set, this data deposition model did not allow for the reporting of different assay outcomes for separately sourced samples and stock solutions from the separate partner libraries nor did it allow for reporting of results for replicates of the same stock solution (i.e., for the 88 intentional replicates). The solution proposed, and agreed upon by PubChem, was to register an intermediate level of chemical substance identifications at the stock solution level for the Tox21 project. The resulting "Tox21 ID", that is, PubChem SID, would allow Tox21 assay and analytical QC results for the different partner libraries to be reported at the differentiated stock solution level for the same compound, that is, same PubChem CID (and same DSSTox GSID and CID). (Note: Prior to 2013, DSSTox substance and structure identifiers were denoted as generic substance ID "GSID" and chemical (or structure) ID "CID" and were strictly numeric; with the migration of DSSTox content to DSSTox_v2, these numeric identifiers were incorporated into the current semantic web identifiers DTXSID and DTXCID.)[40] Hence, Tox21 IDs were assigned to differentiate samples across the three partner libraries: Tox21_1##### (e.g., Tox21_100034) for NCATS samples, Tox21_2##### for NTP samples and Tox21_3##### for the EPA samples. A fourth set of Tox21 samples, whose IDs took the form Tox21_4#####, were assigned to the set of 88 compounds that were to serve as stock solution replicates across all three partner library plate sets. These 88 compounds were selected on the basis of positive assay profiles obtained from the EPA's and NTP's Plate A HTS results run at NCATS as well as ToxCast's Phase I assay results. The compounds were sourced and solubilized by the EPA's chemical contractor, and an aliquot of the stock solution for each of the 88 compounds was shipped to NTP and NCATS prior to plating. A mapping file, listing Tox21 IDs mapped to NCGC IDs and PubChem IDs, is provided on the NCATS Tripod download page, https://tripod.nih.gov/tox21/assays/ (select "Download", tox21_10k_library_info.tsv.zip, accessed February 24, 2020). An extended mapping file additionally containing DSSTox identifiers and indicating partner library associations is provided in Supporting Information Table S1.

3.3. Tox21 Plating. Having assigned Tox21 IDs to each Tox21 plated substance and having mapped each substance to a DSSTox substance ID (GSID, later converted to DTXSID), the final step in partner library preparation was for the EPA and NTP to seal, freeze, and ship their respective sets of 384-well stock solution plates to NCATS for final plate preparation, storage, and analysis. Each EPA and NTP partner plate set consisted of four 384-well stock solution plates, labeled consecutively A, B, and C (or A′, B′, and C′ if reprocured), containing 80 μL of 10−20 mM DMSO solution. In addition, each partner randomly distributed two copies of the 88 replicate compounds (176 total) across each set of four 384-well plates and created 10 identical copies of each set, enough to contribute to the creation of 10 copies of the full Tox21 library for screening and analytical QC analysis (see Figure 2).

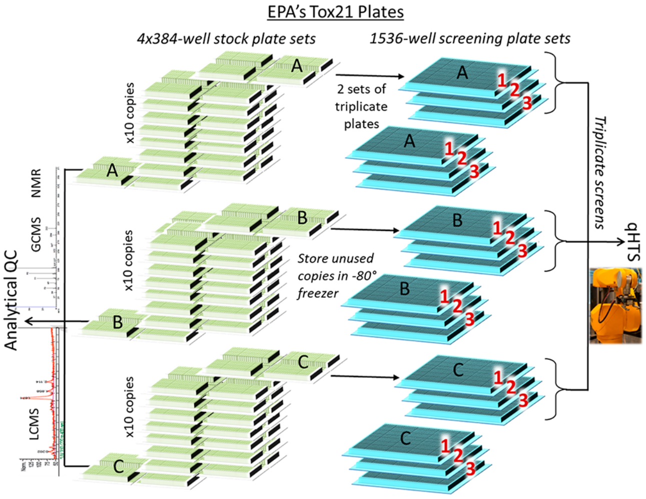


Figure 2. Schematic of the EPA's full Tox21 partner library plate set contribution consisting of 10 copies of three distinct 4 × 384-well stock plate sets (denoted A, B, and C), where each 384-well plate set (A, B, or C) was processed onto two triplicate sets of 1536-well plates (denoted 1−3, differing by shifted overlay pattern), and three sets of triplicate 1536-well plates (A, B, and C) comprised a full partner contribution to the active Tox21 screening library. All remaining plate copies were stored frozen at −80 °C until needed, and one full library plate set (A, B, and C) in 384-well format was reserved for analytical QC analyses.

Prior to screening, a complete partner library complement of the 384-well stock solution plate sets was transferred onto 1536-well plates. Two sets of triplicate copies of the 1526-well plates were created from one full set of 384-well stocks, with each copy within each triplicate set modified by a shifting overlay pattern each time the 384-well plates were combined (labeled 1−3 in Figure 2). In this way, each sample was located at a slightly different well position on the triplicate plates to control for edge and plate-well placement effects. Hence, for each Tox21 partner (e.g., the EPA), one full triplicate set (labeled A, B, C in Figure 2) of 1536-well plates was "actively" used for screening, whereas the second full triplicate set was stored at −80 °C until the first was depleted or had been exposed at room temperature for 4 months.[41] After depletion of both triplicate sets of 1536-well plates, the process was repeated with the next full set of 384-well stock solution plates. Hence, each copy of the full Tox21 library consisted of a total of nine sets of 1536-well plates, and each of the nine sets was screened in triplicate in each Tox21 qHTS assay run. In this manner, the EPA and NTP's original plate set contributions to the Tox21 library have supplied consistent samples for Tox21 screening from 2012 to the present. For further details of the instrumentation used for plating, the overlaying patterning used to add a degree of compound−well position variation to each of the full library plate sets, the 15 step dilution series, and plate storage and handling protocols used over the course of the Tox21 program, see Attene-Ramos et al.[13]

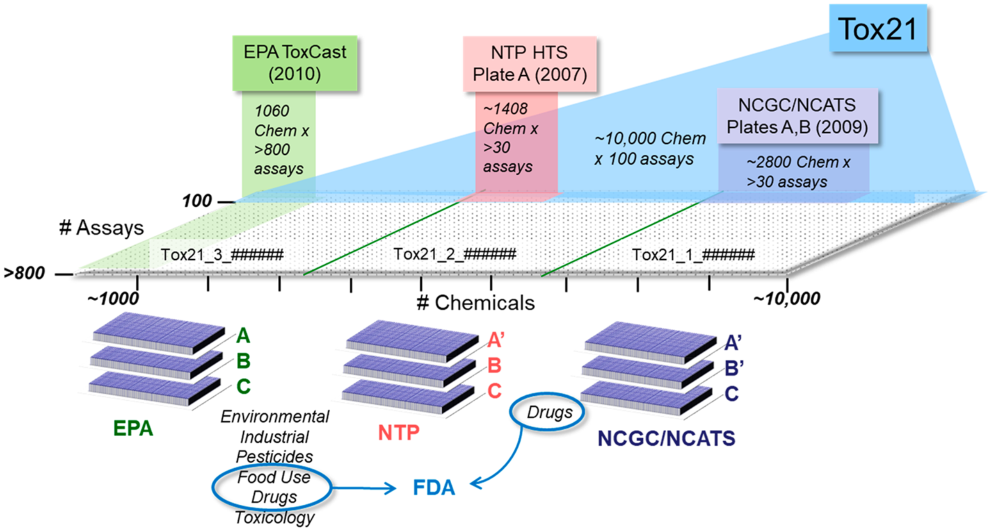


Figure 3. Tox21 partner plate set contributions to the full Tox21 library indicating the approximate chemical × assay overlap totals of the Tox21 Phase I NTP HTS Plate A (2007) and a portion of the EPA's ToxCast library (2009) available at the start of Tox21 Phase II screening.

As shown in Figure 2, one complete set of 384-well plate sets for the full Tox21 library was used to prepare plates for subsequent analytical QC analyses. These analyses would include both "time = 0 months" (T_0) plates, which would remain frozen until the time of analysis, and "time = 4 months" (T_{4mo}) plates, which would mirror the handling (freeze/thaw/room temperature storage) of a set of full library analysis plates for a period of 4 months, after which time the analysis plates would be retired and the T_{4mo} plates would be frozen until submitted for analysis. Analytical QC analysis of Tox21 plates performed over the course of the next several years would include liquid chromatography mass spectroscopy (LCMS), gas chromatography mass spectroscopy (GCMS), and nuclear magnetic resonance (NMR) spectroscopy on all or portions of the library. LCMS and NMR analytical analyses of Tox21 samples were carried out by a contract laboratory funded by the NTP (OpAns LLC, Durham, NC) and overseen by NCATS. GCMS analyses were carried out at the National Institutes of Standards and Technology. Raw results reported at the Tox21_ID stock solution level have been reviewed and summarized for publication by NCATS and are available for download from the NIH Tripod website (see https://tripod.nih.gov/tox21/samples, accessed February 8, 2020). (Note: Current Tox21 sample totals listed on the Tripod website are updated to include newly generated solutions from the same neat sample used to create the original library stock solutions. Tox21 IDs in these cases retain their original values with added extensions of the form _1, _2, etc., as in Tox21_113103_1.) Further details and analysis of the analytical methods and results will be presented in subsequent papers in this Tox21 Compound Library series (to be published).

Figure 3 illustrates Tox21 partner plate set contributions as well as the EPA's ToxCast assay coverage and NTP's chemical assay coverage at the start of Tox21 Phase II screening compared to future anticipated Tox21 screening coverage (~10,000 chemicals × 100 assays). The FDA contributed physical samples of a small set of drugs to the EPA Tox21 collection, whereas larger numbers of included chemicals on all three plate sets reflected content of regulatory interest to FDA, including thousands of food-use chemicals and drugs.

3.4. Publishing Tox21 Chemical Assay Results.

All Tox21 assay results produced thus far have been generated at the NCATS testing facility. Details of the robotic platform and Tox21 assays and results are provided elsewhere, in numerous publications (https://tox21.gov/all-publications/). As indicated previously, initial processing of Tox21 assay results was carried out by NCATS, and the data have been released to the public through PubChem in association with the original DSSTox-assigned structures and identifiers using the source library term "Tox21" (see, e.g., https://pubchem.ncbi.nlm.nih.gov/#query=tox21&tab=substance). These data have also been made available from NCATS's Tox21 Tripod data browser (https://tripod.nih.gov/tox21/index). In the same way in which the three Tox21 partner libraries began as separate endeavors to serve multiple programmatic objectives, Tox21 assay results have been separately processed by the three Tox21 partners by different means, in part to conform to their different in-house data tracking systems, assay data pipelines, and programmatic objectives. However, several data analysis (dose−response curve fitting) approaches were also developed to provide the best fits for a variety of anticipated and unanticipated dose−response patterns, and analysis methods were developed in parallel with data generation, resulting in varied approaches across the three agencies. Details of the data pipelining analysis used by NCATS have been described previously.[14,21] NTP has published some modified approaches (see, e.g., Hsieh et al. and Shockley)[15,42,43] and plans to release their processed Tox21 assay data through the Chemical Effects in Biological Systems (CEBS) website (ftp://anonftp.niehs.nih.gov/ntp-cebs/datatype/Tox21/, see also the NTP Tox21 Toolbox at https://ntp.niehs.nih.gov/results/tox21/tbox/index.html).[44] The EPA's ToxCast data analysis pipeline has been applied to both ToxCast and Tox21 HTS data sets.[16] The latter results are available through the EPA's ToxCast data downloads page (https://www.epa.gov/chemical-research/exploring-toxcast-data-downloadable-data) and from the EPA's CompTox Chemicals Dashboard (https://comptox.epa.gov/dashboard).[45] Consistent application of the same data pipeline processing has allowed the Tox21_EPA portion of the data set to be analyzed and combined with ToxCast assay profiles for all overlapping chemicals in the two libraries. The Tox21 data processing and publication workflows of the three Tox21 partners are schematically illustrated in Figure 4.

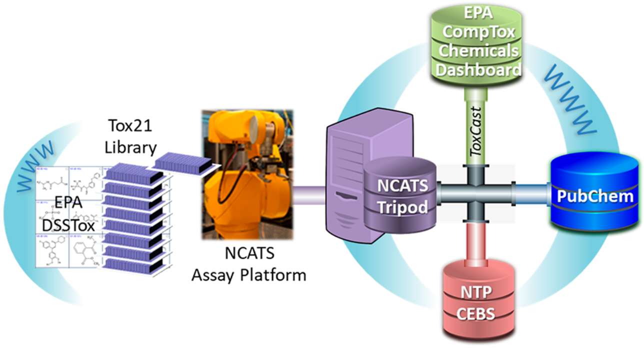


Figure 4. Tox21 chemical testing and data processing workflow, from initial DSSTox structure curation to plating and screening the Tox21 library at NCATS and to data distribution from the NCATS Tripod website (https://tripod.nih.gov/tox21) and PubChem (https://pubchem.ncbi.nlm.nih.gov/), with separate pipelined data analyses by both the EPA and NTP and data distribution through the EPA's ToxCast website (https://www.epa.gov/chemical-research/exploring-toxcast-data-downloadable-data), CompTox Chemicals Dashboard (https://comptox.epa.gov/dashboard), and NTP's CEBS website (ftp://anonftp.niehs.nih.gov/ntp-cebs/datatype/Tox21/).

In addition to divergence of assay analysis pipelines, a lack of a single centralized chemical database across the three Tox21 partners, encompassing different data models and levels of chemical curation, has inevitably led to some chemical identifier inconsistencies in different partner data releases in the public domain. At any given time, however, the publicly available DSSTox TOX21SL substance file will contain the most up-to-date Tox21 chemical identifiers (available at https://comptox.epa.gov/dashboard/chemical_lists/TOX21SL, accessed February 24, 2020; also available in Supporting Information Table S2). Despite lack of precise coordination of data releases, the various representations of the published Tox21 data serve to highlight the different ways in which qHTS assay data can be processed and interpreted to serve different regulatory and screening objectives.

4. METHODS

Methods used to generate the cheminformatics profiling results and figures presented in Section 5 are described below, accompanied by data available in Supporting Information Tables S1−S7.

4.1. Chemical Lists Cross-Referenced to Tox21 Partner Libraries.

The unique DSSTox substance list (TOX21SL) for the full Tox21 inventory used in the present study was downloaded from the EPA CompTox Chemicals Dashboard List Download page (https://comptox.epa.gov/dashboard/chemical_lists/TOX21SL, last updated 2017-02-23, 8947 chemicals). This file is provided in Supporting Information Table S2 and includes DTXSID, CAS RN, chemical name and, for the majority assigned to unique chemical structures, the DSSTox structure ID (DTXCID), molecular formula, molecular weight (MolWt), SMILES, InChI, and what have traditionally been termed "QSAR-ready SMILES" (i.e., SMILES for the desalted, neutralized parent with stereospecific information removed). A comparison of the various chemical representations is presented in Section 5.1.

Additional chemical lists of relevance to the Tox21 partners' interests were selected to evaluate corresponding coverage of the TOX21SL compound library (results in Section 5.2). The chemical lists, along with chemical totals, the EPA CompTox Chemicals Dashboard URL for accessing structure files, and source URL are provided in Table 1. Each of the chemical lists in Table 1 was cross-referenced by DSSTox substance identifier, DTXSID, to obtain overlaps with the TOX21SL list from the "Batch Search" page of the EPA CompTox Chemicals Dashboard. For the purpose of category and overlap comparisons, a "DRUG" category was created to include the following: chemicals listed in DRUGBANK; all Tox21 chemicals contained in the NCATS Tox21 NPC library; and the list of 134 donated failed pharmaceuticals included in the EPA's ToxCast and Tox21 libraries.[33,34,46] An additional "COSMOS+" category consisting of cosmetics, personal care products, and food-contact substances was created by combining the COSMOSDB and EFSAOFT lists. The publicly available REACH2017 list acts as a surrogate for chemicals covered under the EU's REACH chemicals program (https://ec.europa.eu/environment/chemicals/reach/reach_en.html), and the TSCAACTIVENONCONF list represents the publicly available portion of the EPA's Toxic Substances Control Act inventory (https://www.epa.gov/tsca-inventory/how-access-tsca-inventory). Finally, the OPPIN list represents pesticides tracked within the EPA's Substance Registry Service (https://ofmpub.epa.gov/sor_internet/registry/substreg/). A table providing binary overlap indicators (1,0) of the TOX21SL substance list with each of the above category lists is provided in Supporting Information Table S3.

4.2. QSAR Property and Toxicity Predictions.

For the purpose of discerning differential global property trends across the Tox21 partner libraries, we generated several physicochemical properties and predicted toxicities of the library compounds (results presented in Section 5.2). Physicochemical properties (e.g., vapor pressure, log K_{ow}, and bioconcentration factor) were calculated using the OPERA quantitative structure−activity relationship (QSAR) models, with precomputed results downloaded from the EPA's CompTox Chemicals Dashboard (https://comptox.epa.gov/dashboard, accessed March 1, 2020) using the Batch search query with TOX21SL DTXSID identifiers.[47]

A simple global measure of molecular complexity, denoted "Complexity" (computed based on paths, branching and number of atom types) was calculated using the commercially licensed CORINA Symphony software (Molecular Networks, GmbH, v1.0).[48]

The commercially licensed Derek Nexus v.6.0.1 for Windows application (Lhasa Ltd., Leeds, UK) was used to batch process the DSSTox SDF file for the TOX21SL inventory to generate predictions for rat carcinogenicity, denoted "RatCarc". Derek Nexus uses a combination of expert-judgment and structure-alerting features, modified in some cases by properties (such as log K_{ow}), along with algorithms for assessing confidence in predictions (e.g., labeled as Probable, Plausible, Equivocal). Binary predictions (1,0) were generated using the settings: Species:Mammals:Rat, selecting "Carcinogenicity" as the end point, and selecting the "Plausible" threshold. Predictions of Ames mutagenicity and developmental toxicity were generated using the EPA T.E.S.T. QSAR models (https://www.epa.gov/chemical-

Table 1. Lists Used to Compare Chemical Coverage of Tox21 Partner Libraries

list name	list description[c]	type	no. of chems	EPA dashboard URL	source URL
REACH2017	NORMAN: REACH chemicals list provided to NORMAN network (last updated 2019-05-19)	industrial, environmental	57,758	https://comptox.epa.gov/dashboard/chemical_lists/REACH2017	https://ec.europa.eu/environment/chemicals/reach/reach_en.htm
OPPIN	PESTICIDES|EPA: Office of Pesticide Programs Information Network (last updated 2019-10-26)	pesticides	3961	https://comptox.epa.gov/dashboard/chemical_lists/OPPIN	https://ofmpub.epa.gov/sor_internet/registry/substreg/searchandretrieve/searchbylist/search.do?search=&searchCriteria.substanceList=89&searchCriteria.substanceType=-1
DRUGBANK	DRUGS: DrugBank database from the University of Alberta (a partially mapped list, last updated 2017-07-30)	drugs	4632	https://comptox.epa.gov/dashboard/chemical_lists/DRUGBANK	https://www.drugbank.ca/
TSCAACTIVENONCONF[a]	EPA|TSCA: TSCA Inventory, active nonconfidential portion (last updated 2020-01-06)	industrial, environmental	31,460	https://comptox.epa.gov/dashboard/chemical_lists/TSCA_ACTIVE_NCTI_0320	https://www.epa.gov/tsca-inventory/how-access-tsca-inventory
COSMOSDB[b]	COSMOS DB v1 food additives and cosmetics ingredients list (a partially mapped list, last updated 2020-05-29)	food additives and cosmetics	7021	https://comptox.epa.gov/dashboard/chemical_lists/COSMOSDB	http://www.cosmostox.eu/what/COSMOSdb/
EFSAOFT	European Food Safety Authority's (EFSA) OpenFoodTox chemical hazards list (a partially mapped list, last updated 2020-05-29)	food use chemicals	3942	https://comptox.epa.gov/dashboard/chemical_lists/EFSAOFT	https://data.europa.eu/euodp/en/data/dataset/openfoodtox-efsa-s-chemical-hazards-database

[a] List name and contents modified in the EPA CompTox Chemicals Dashboard as of March 20, 2020 (current URL provided), 33,364 total chemicals. [b] Contains the COSMOS cosmetics inventory compiled from the EU Cosing database (the European Union's official database for cosmetic ingredients) and the U.S. Personal Care Products Council list (both COSMOS DB v1.0 content and Cosmetics Inventory files are available for download from http://www.cosmostox.eu/what/COSMOSdb/, accessed March 1, 2020). [c] Abbreviated list description from the EPA CompTox Chemicals Dashboard, with expanded descriptions found on the EPA Dashboard URL.

research/toxicity-estimation-software-tool-test), downloaded from the EPA CompTox Chemicals Dashboard (accessed March 3, 2020) using the Batch search query with Tox21 DTXSID identifiers. T.E.S.T. models are based on publicly available toxicity end point data sets, models, and expert rules, adhering to standard practices in QSAR modeling and reporting.[49] T.E.S.T. models for predicting Ames mutagenicity are based on a publicly available benchmark data set available for download from http://doc.ml.tu-berlin.de/toxbenchmark/.[50] T.E.S.T. models for predicting developmental toxicity are implementations of models initially developed within the CAESAR project (CAESAR: Computer assisted evaluation of industrial chemical substances according to regulations. EC project 022674 (SSPI); http://www.caesar-project.eu) and are based on a publicly available collection of developmental toxicity data for diverse chemical structures.[51,52] The best performing T.E.S.T. models for Ames mutagenicity and developmental toxicity, downloadable from the EPA's CompTox Chemicals Dashboard, were consensus models that combined the results from several types of models (hierarchical clustering, nearest neighbor, multiple linear regression, etc.). Model predictions are reported as quantitative scores, where a score below 0.5 is considered negative (0) and above 0.5 positive (1) for the end point. In the case of the developmental toxicity model, however, this cutoff heavily weights sensitivity, capturing all positives along with many false positives at the expense of specificity; hence, for the present global inventory comparisons, we used a cutoff of 0.8 to assign a chemical as positive for developmental toxicity (T. Martin, personal communication). Lastly, T.E.S.T. models for Rat Oral LD_{50} are based on data extracted from the National Library of Medicine's TOXNET database, accessed from within ChemID Plus (https://chem.nlm.nih.gov/chemidplus/).

Predicted OPERA, CORINA Symphony, Derek Nexus, and T.E.S.T. properties for Tox21 chemicals are provided in Supporting Information Table S4.

4.3. ToxPrints and Chemotype-Enrichment Analysis.

To profile and compare Tox21 partner structure inventories (results presented in Section 5.2), we employed the publicly available set of ToxPrint molecular fingerprints (https://toxprint.org/) developed by Altamira (Altamira, Columbus, OH) and Molecular Networks (Molecular Networks, Erlangen, GmbH) under a contract from the FDA. The ToxPrint set (V2.0_r711) used in the present study consists of 729 uniquely defined structural features, also referred to as "chemotypes" (CTs). ToxPrints are designed not only to capture known toxicity structural alerts used in the FDA's safety assessment workflows but also to provide broad feature coverage of structure inventories consisting of tens of thousands of environmental and industrial chemicals, including pesticides, high-production volume chemicals, cosmetics ingredients, food additives, drugs, and chemicals for which *in vitro* or *in vivo* toxicity data are available. ToxPrints are coded in an open, XML-based Chemical Subgraphs and Reactions Markup Language (CSRML) and can be downloaded from the ToxPrint website (https://toxprint.org/) and visualized and searched within the publicly available Chemotyper application (https://chemotyper.org/).[53] All Tox21 ToxPrint inventory comparisons and enrichment results presented here were produced from ToxPrint fingerprints generated using a command-line installation of CORINA Symphony (Molecular Networks, Erlangen, GmbH). The ToxPrint fingerprint file for the TOX21SL chemical list can be downloaded from the EPA's CompTox Chemicals Dashboard using the Batch Download feature and selecting "Enhanced Meta Datasheets: ToxPrint fingerprints (Chemotyper format−csv)"; these are provided in Supporting Information Table S5 (accessed March 3, 2020).

Chemotype enrichment analyses were performed using ToxPrint fingerprints and binarized activity hit calls for Tox21 and ToxCast HTS chemical assay data sets to identify structural features differentially enriched in Tox21 assay actives in each of the Tox21 partner libraries and to examine patterns of CT enrichment as a function of tested set size and Tox21 and ToxCast assay platforms (results presented in Sections 5.3 and 5.4). A standardized chemotype enrichment analysis workflow (CTEW) has been developed within the EPA to identify chemical substructural features significantly enriched within the active subset of an assay data set relative to the whole data set, that is, having a greater statistical association with activity than would be expected by chance. The CTEW was programmed using Python and directly interacts with the EPA's internal DSSTox MySQL database to access DSSTox substance-structure mappings. It is currently available within the EPA's intranet using a command-line interface and plans are to eventually publish it as a web application. Input to the CTEW consists of a ToxPrint fingerprint table (i.e., binary indicators for presence (1) or absence (0) of each ToxPrint in each molecule) for all tested chemicals that can be represented by a molfile structure (indexed by DTXSID or DTXCID), combined with binarized activity hit calls (1,0) for each chemical. The CTEW generates a confusion matrix for each CT in association with the corresponding set of binary assay hit-calls, where the following definitions apply: true positives (TP) are chemicals that contain the CT and had positive hit calls; true negatives (TN) had neither the CT nor positive hit calls; false positives (FP) contain the CT but had negative hit calls; and false negatives (FN) do not contain the CT but had positive hit calls. To determine if a CT is enriched, a standard set of statistics is calculated to filter out insignificant results while retaining both weak and strong enrichments for follow-up analysis. These filtering thresholds include: one-tailed Fisher's exact p-value ≤ 0.05, number of true positives ≥ 3, and odds ratio (OR) ≥ 3, where OR = $(TP \times TN)/(FN \times FP)$. CT enrichments results used in the present study are provided in Supporting Information Table S6.

4.4. Tox21 and ToxCast Assay Results.

A binarized Tox21 assay hit call matrix was generated from the current EPA ToxCast database invitroDBv3.2 (available at: https://www.epa.gov/chemical-research/exploring-toxcast-data-downloadable-data) and was used to examine variability of hit calls for stereoisomer variants and parent-salt pairs (results presented in Section 5.2). A binarized hit call matrix for both Tox21 and ToxCast assays was generated from an earlier version of the EPA's public ToxCast database, invitroDBv2 (archived version available at: https://epa.figshare.com/articles/Previously_Published_ToxCast_Data/6062551/2) and was used to generate all CT enrichment results presented in Sections 5.3 and 5.4. Binarized assay end points were extracted for "assay_component_end point_ID" (aeid) level results in each case, and CT enrichments were computed for each aeid according to the CTEW standardized protocol described in Section 4.3. Global enrichment results presented in the current study were aggregated at the assay platform level for the Tox21 and ToxCast-affiliated assays and, for purposes of the current comparisons, only included assay platforms with 15 or more distinct assay end points represented, and assay end points for which test results were reported for 450 or more chemicals. Aggregated assay platforms included in the global comparisons are listed below and described in more detail elsewhere; the total number of assay end points having one or more enriched CTs is indicated in parentheses.[35] (Note: All non-Tox21 assay vendor and collaborator platforms are affiliated with the EPA's ToxCast screening program.)

- Tox21 qHTS assays (108/221 total from invitroDBv2 having one or more enriched CTs), primarily focused on cytotoxicity, cell stress responses, and nuclear receptor activity; screened at the NCATS intramural screening facility
- Apredica (33 total), time-course, cell-based, high-content imaging assays for assessing aspects of cell damage, oxidative stress, apoptosis, mitochondrial misfunction, etc. (acquired by Cyprotex US in 2010, Watertown, MA, https://www.cyprotex.com/)
- Attagene (161 total), high-content cell-based assays spanning multiple biological pathways by transcription factor analysis (Morrisville, NC, http://www.attagene.com/)
- BioSeek (147 total), complex human primary cell-based assays probing multiple phenotypic end points (acquired by Eurofins DiscoverX in 2012, Fremont, CA, https://www.discoverx.com/home)
- CEETOX (19 total), steroidogenesis assay measuring multiple end points in a human adrenocortical carcinoma cell line (acquired by Cyprotex US in 2010, Watertown, MA, https://www.cyprotex.com/)
- OdysseyThera (17 total) high-content cell-based imaging assays probing critical signaling pathways (San Ramon, CA, https://www.linkedin.com/company/odyssey-thera/)
- Tanguay_ZebraFish (19 total), phenotypic zebrafish developmental toxicity end point screening platform[54]

Tox21 assay results used to examine assay hit reproducibility across stereoisomer variants and salt pairs are provided in Supporting Information Table S7. To view the EPA's current public listing of ToxCast and Tox21 assay end points, along with descriptions and links to results, see https://comptox.epa.gov/dashboard/assay_endpoints/. Results for a particular assay technology (e.g., Apredica, abbreviated APR) can be viewed by filtering the results, e.g., https://comptox.epa.gov/dashboard/assay_endpoints/?link==APR).

5. CHEMINFORMATICS PROFILE

In the remainder of this article, we profile the contents of the Tox21 partner libraries in relation to each other, through a variety of structural and CT enrichment lenses. Our goal is not to provide a detailed survey of the Tox21 library nor a thorough analysis of the Tox21 assay results. Rather, we take a high-level, global view of the full library from the perspective of its constituent partner libraries and assess their relative contributions to the whole. Our aim is to show how the consolidation of these distinct partner libraries significantly expands the structure and property characteristics of the full Tox21 screening library beyond what the individual partners could have achieved separately. Furthermore, we show how this broadened chemical coverage enables detection of chemotype−activity enrichments in assay space that would have been missed otherwise. Lastly, we use global chemotype−activity enrichment counts as a surrogate measure of structure−activity information content to probe the relative impact of chemical diversity (related to chemical test set size) versus biological diversity (related to the numbers and types of distinct assays within an assay platform or testing program).

5.1. Tox21 Library: Chemical Representations.

Figure 5 indicates the various types of chemical representations considered, along with relative totals in the full Tox21 library. Each level of sample or chemical representation, in turn, affords a different view of the library. As indicated previously, the Tox21 ID identifies samples at the unique stock solution level, thereby distinguishing the same DTXSID substance contributed by different Tox21 library partners. The Tox21 ID is also the level at which NCATS aggregates and publishes Tox21 qHTS assay results through the Tripod and PubChem websites (see Figure 4). The next aggregation layer in Figure 5 consolidates Tox21 IDs corresponding to different supplier/lot stock solutions to the same DTXSID generic chemical substance layer, typically associated with a unique CAS RN, chemical name, and, in most cases, chemical structure. The DTXCID level, in turn, is limited to substance records associated with a uniquely rendered chemical structure, thus excluding chemical mixtures and ill-defined substances of which there are relatively few (375) in the full Tox21 library. The next layer of aggregation collapses various salt, complex, or hydrate forms to the associated parent form, converting closely related forms of a compound to replicates. In addition, "QSAR-ready" processing typically removes inorganics and ionic compounds containing heavy metals (such as Hg, Cd, Zn, etc., where the metal may be responsible for the toxicity) from further analysis, yielding approximately 970 fewer structures than at the DTXCID level. Lastly, various types of molecular fingerprinting methods are publicly available to dimensionally reduce a compound library either at the unique DTXCID structure level or at the QSAR-ready structure level to a smaller set of predefined chemical fragments or features. In the current analysis, we employ the publicly available set of ToxPrint CTs (described in Section 4.3), 85% of which (623/729) are represented one or more times in the full Tox21 chemical structure library. ToxPrints are designed to capture chemically informative atom/ring/chain/bond features relevant to chemical safety assessment workflows and represented within chemical data sets of regulatory interest to both the FDA and EPA. They have proven useful for profiling and comparing distinct chemical libraries, for QSAR analyses yielding

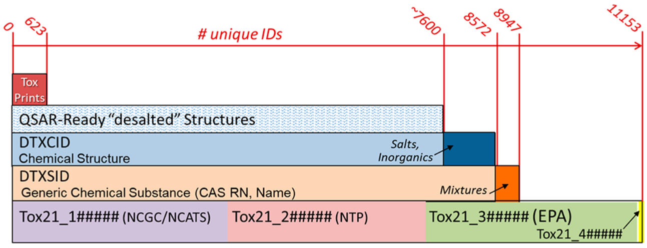


Figure 5. Various chemical representations of the full Tox21 compound library and the corresponding totals of unique identifiers at each level, ranging from Tox21_ID (stock solutions), to generic substances (DTXSID), structures (DTXCID), QSAR-ready structures, and the subset of ToxPrint chemical features (https://toxprint.org/) represented one or more times in the full Tox21 structure library (out of 729 total possible).

interpretable results, and for computing CT enrichments within and across chemical−activity space.[24,34,55−57]

5.2. Tox21 Partner Libraries: Totals, Overlaps, and Replicates.

Figure 6 indicates the number of unique substances, that is, DTXSIDs, independent of supplier, in each separate Tox21 partner library, as well as the total number of substance overlaps across the three partner libraries comprising the full Tox21 testing library. When DTXSID overlaps are not removed, there are a total of 10,957 substances comprising the three separate Tox21 partner libraries, approximately corresponding to the number of unique Tox21 IDs. (Note: This total is slightly less than the total number of unique Tox21 IDs listed in Figure 5 due to a small number of intralibrary DTXSID replicates assigned to distinct Tox21 IDs; in addition, the total non-overlapping unique DTXSIDs computed from Figure 6 (8944) is slightly smaller than shown in Figure 5 (8947) due to a small number of partner library overlaps with the 88 replicate set.) The largest overlapping content is between the EPA and NTP libraries (1254) due to their similar programmatic objectives in the realm of environmental toxicology. In contrast, the overlapping content of the NCATS library with either the EPA's (571) or NTP's (475) library is considerably less, a reflection of the exclusive drug focus of the NCATS library. Overlaps of the NCATS library with the EPA's and NTP's libraries, however, are indicative of compounds that are either multi-use or were included in the latter libraries due to availability of toxicity data or to serve as assay reference compounds. Including the 88 replicate set, there were a total of 287 substances shared by all three partner libraries, 199 of which were separately sourced and, thus, plated in triplicate across the full library.

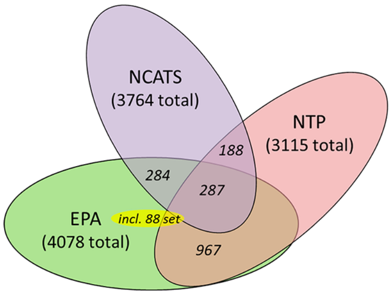


Figure 6. A representation of the unique and overlapping DSSTox substance content in the three Tox21 partner libraries, with library totals indicated in parentheses, and total overlapping content across each of the three libraries indicated. A single unique occurrence of the 88 replicate compound set is included in the EPA's Tox21 partner library totals for comparison purposes.

The large numbers and types of sample and substance replicates included in the full Tox21 library offer multiple opportunities to evaluate reproducibility of both assay and analytical QC results. At the sample level, the 88 replicate compound set, plated in duplicate and randomly located on each of the nine 1536-well plate sets (3 plates per partner), yielded 18 plate-well instances of each replicate compound across each copy of the full Tox21 testing library. The full Tox21 library, in turn, was screened in triplicate (Figure 2), and each plated compound was additionally tested in qHTS format at 15 concentrations for each assay. Reproducibility results for replicates at the plate-well level across an initial set of 30 cell-based Tox21 assays have been reported previously.[21]

The Tox21 library affords additional opportunities to evaluate the influence of variability in supplier/lot associated with over 1000 cross-partner replicates as well as across hundreds of pairs of salt/stereo-related forms on both the Tox21 assay results and associated analytical QC results. Here, we limit our examination to the variability of Tox21 assay profiles within the set of 882 replicate QSAR-ready structures in which salt and stereo information is removed to identify related "replicates". Closer examination of this set indicates that 258 structures are replicated due to removal of stereo information (e.g., *cis*/*trans* or *E*/*Z*, *R*/*S*, +/−, etc.), yielding 123 sets of chemicals differing only by one or more stereobonds. Approximately half of these sets correspond to drugs in the NCATS library, with the remainder in the EPA or NTP libraries, including pesticides, steroids, natural products, and some industrial chemicals. A view of the variation of Tox21 assay results (i.e., binarized activity hit calls) within stereoisomer sets is shown in Figure 7. The largest overall hit rate (hit%) within each stereofamily set is shown alongside the difference (diff) of hit percentages within the stereofamily, with the latter serving as a measure of hit rate variability within each stereogroup. A total of 56 stereosets had max hit% values of 3% or less across all Tox21 assays (hence, are not shown in plot), whereas 67 stereosets had differences exceeding 3%, averaging 50% difference in hit% (results plotted in Figure 7). These results indicate significant variation in assay hit rates within the majority of the stereochemical sets. Sample QC problems or assay artifacts could account for some of these variations or the results may be indicative of Tox21 assay end points being sensitive to stereochemistry variations in structures.

We also examined a subset of the 882 QSAR-ready replicates for which we had 84 pairs (168 total) of parent structures and their hydrochloride (HCl) acid counterpart. A total of 72

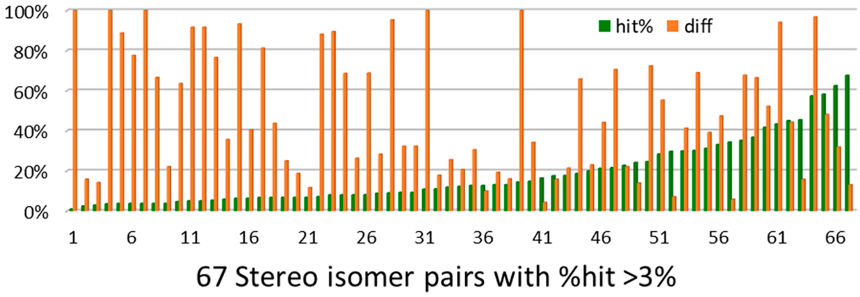


Figure 7. Maximum percent of hits (hit%) across Tox21 assays within a stereofamily (green bars) relative to the difference in hit% (diff) within the stereofamily (orange bars) for a total of 67 sets of stereoisomer chemicals.

replicates in this list were contained in the NCATS drug library, whereas the rest were from either the EPA or NTP libraries. Of the 58 pairs having >3% difference in overall assay hit rate (hit%) between the parent and HCl salt, a large majority (43) showed greater activity for the HCl salt, whereas only 15 showed greater activity for the parent (Figure 8). By considering the total number of assays in which both the salt and parent were tested, a Fisher's exact test indicated that more than a third, or 31 of the 84 pairs, had $p < 0.05$ significance. When the counts from all compounds were summed, the overall difference between salt hit rate and parent hit rate was also significant ($p < 1 \times 10^{-20}$). Thus, Tox21 assay results indicate a statistically significant enrichment of activity in HCl salts versus their corresponding parents.

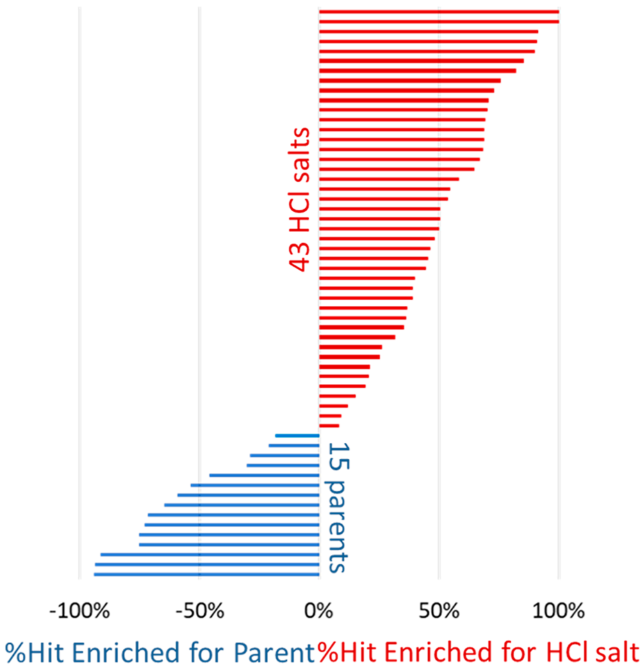


Figure 8. Plot of the largest enrichment of hits (difference in hit%) for either the HCl salt (red bars) or parent chemical (blue bars) relative to the other in the pair, that is, difference in % hits, across Tox21 assays for a total of 58 pairs of chemicals having >3% maximum hit rate.

A more detailed analysis of assay concordance and analytical QC results is needed to better understand the basis for stereoisomer and parent-salt variations in activity across Tox21 assays. Importantly, DSSTox substance annotation that appropriately captures stereoisomer and parent-salt specificity is a prerequisite to being able to further investigate such questions. Furthermore, the results of Figure 8 run counter to an assumption made in construction of the NPC library (see Section 2.1), where a single "active pharmaceutical ingredient" (or API) representative was included under the assumption that salt distinctions are less important for *in vitro* and *in silico* studies.[33]

5.3. Tox21 Partner Libraries: Relative Coverage of Lists, Toxicities, and Properties. In a previous study, the EPA's ToxCast chemical library, consisting of 4226 unique substances at the time, was surveyed from a variety of perspectives and structural lenses toward the goal of determining the fitness of the screening library for the predictive modeling objectives, that is, to provide a sufficient number, diversity, and property range of chemicals for representing the environmental chemical landscape of concern from both exposure and toxicity perspectives. The study concluded that there was substantial coverage of: (1) CAS RN lists used in the construction of the library and relevant to the ToxCast program objectives (such as lists pertaining to EPA and FDA regulatory programs, drugs, consumer products, exposure, and *in vivo* toxicity data lists); (2) toxicity and metabolism structure alerts; and (3) chemical class and structure similarity coverage in relation to larger potential EPA application inventories, such as the Endocrine Disruption Screening Program (EDSP) (https://www.epa.gov/endocrine-disruption) list as represented by the CERAPP collaborative modeling project.[58] Given that the EPA's ToxCast chemical library at that time fully encompassed the Tox21_EPA partner library, and the latter constituted 96% of the total ToxCast library, it is reasonable to assume that all results and conclusions of that study are applicable to the full Tox21 library; hence, the present analysis will build on that premise.

Figure 9 provides a high-level summary view of the relative coverage of a selection of chemical use-type and regulatory lists across the three exclusive partner libraries, where the Tox21_EPA partner library is considered in its entirety, and the remaining contributions for the non-overlapping portions of the NTP and NCATS libraries are shown for comparison. The lists considered include a broad collection of chemicals designated as drugs (DRUGS), an EPA pesticide inventory (OPPIN), industrial and commercial chemicals on regulatory lists in the US (TSCA) and EU (REACH), and a list that includes cosmetics, personal care products, and food-contact substances (COSMOS+); see Section 4.1 for details. These lists also align with the primary research and regulatory

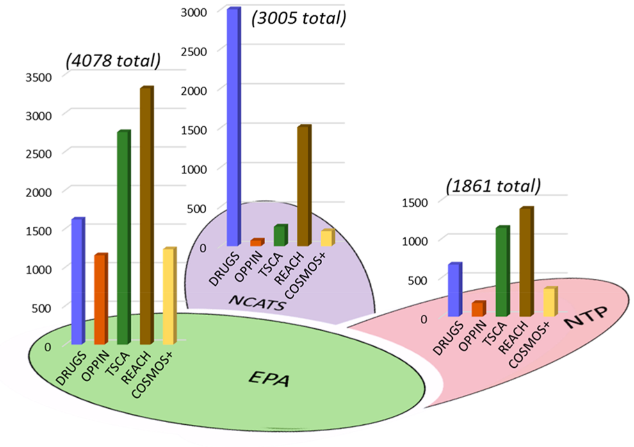


Figure 9. Comparison of chemical list coverages across the three non-overlapping Tox21 partner libraries: the full Tox21 EPA library (green, 4078 total), the portion of the Tox21 NTP library not overlapping with EPA (pink, 1861 total), and the portion of the Tox21 NCATS library not overlapping with either EPA or NTP (purple, 3005 total). Chemical lists include DRUGS (DrugBank, + Tox21_NCATS library + 134 EPA ToxCast donated pharma), OPPIN pesticides (EPA's Pesticide Program Information Network list), TSCA environmental and industrial chemicals (EPA's Toxic Substances Control Act list), REACH (NORMAN Network's list of REACH chemicals for use in suspect screening), and COSMOS+ (partially curated COSMOS DB, cosmetic ingredients and personal care products list, and European Food Safety Authority's (EFSA) OpenFoodTox list).

interests of the Tox21 partners, including NCATS (DRUGS), EPA (OPPIN, TSCA), NTP (varied), and FDA (DRUGS, COSMOS+).

Recognizing that the lists themselves have significant overlaps (see Table 2), Figure 9 clearly shows that each of the three non-overlapping libraries contributes substantial additional unique chemical list-associated content to the whole of the Tox21 library. The significant representation of chemicals in the DRUGS category in all three partner libraries, as well as the representation of non-DRUG lists in the NCATS library, particularly underscores the multi-use nature of many chemicals that also fall under different regulatory purviews.

Table 2. Overlap Totals for Chemical Lists Representing Different Regulatory Domains

LIST OVERLAPS[a]	DRUGS	OPPIN	TSCA	REACH	COSMOS+
DRUGS[b]	5293	630	1991	3478	962
OPPIN	630	1402	736	1087	416
TSCA	1991	736	4141	3880	1608
REACH	3478	1087	3880	6218	1631
COSMOS+	962	416	1608	1631	1783

[a] Numbers of overlapping DTXSID unique substances for five lists: DRUGS (DrugBank, + Tox21_NCATS library + 134 EPA ToxCast donated pharma), OPPIN pesticides (EPA's Pesticide Program Information Network list), TSCA environmental and industrial chemicals (EPA's Toxic Substances Control Act list), REACH (NORMAN Network's list of REACH chemicals for use in suspect screening), and COSMOS+ (partially curated COSMOS DB, cosmetic ingredients and personal care products list, and the European Food Safety Authority's (EFSA) OpenFoodTox list); details provided in Table 1. [b] Colors correspond to same as those used to represent lists in Figures 9 and 10.

Finally, Figure 10 presents the cumulative total of Tox21 library chemicals as a fraction of the full list for each of the five usage and regulatory lists represented in Figure 9. Broad coverage of these types of lists was an important objective for EPA's ToxCast program as well as for the Tox21 partners since the Tox21 screening library was intended to not only represent the chemical landscapes of research and regulatory interests of each of the Federal partner programs but also to serve as a probe of the chemical toxicity mechanism landscape toward the goal of building improved toxicity prediction models. The most complete list coverage is provided by Tox21 chemicals in the DRUGS category, which is not surprising given the more focused library objectives of NCATS.

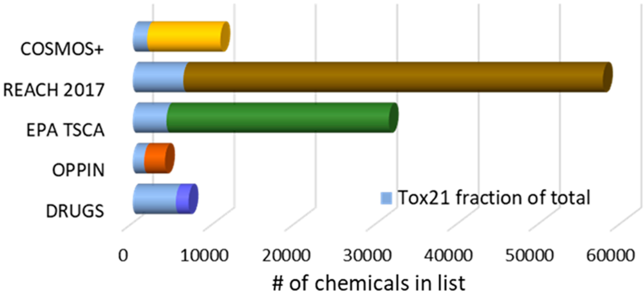


Figure 10. Total Tox21 library coverage of various lists where the bar color is unique to the list (and matches colors in Figure 9) and the total bar length indicates the total number of compounds in full list, whereas the light blue portion of bar indicates the number of list chemicals included in the Tox21 library. Lists include: DRUGS (DrugBank, + Tox21_NCATS library + 134 EPA ToxCast donated pharma), OPPIN pesticides (EPA's Pesticide Program Information Network list), TSCA environmental and industrial chemicals (EPA's Toxic Substances Control Act list), REACH (NORMAN Network's list of REACH chemicals for use in suspect screening), and COSMOS + (partially curated COSMOS DB, cosmetic ingredients and personal care products list, and the European Food Safety Authority's (EFSA) OpenFoodTox list).

Figure 11 presents a similar graphical view as in Figure 9 of the non-overlapping Tox21 partner libraries but from the perspective of relative coverage of a selection of predicted toxicity end points, including developmental toxicity (DevTox), mutagenicity (Mutag), and rat carcinogenicity (RatCarc); for descriptions of models used to predict these end points, see Section 4.2. (Note: Both NTP and EPA's compound libraries intentionally included known toxicants in all three categories; hence, a number of true positives are likely to be present within these end point predictions.) Once again, the goal is to highlight the significant number of unique contributions to each end point category from each of the exclusive partner libraries. Good coverage of potential toxicants spanning diverse chemical and property space across the Tox21 library is necessary from the standpoint of HTS screening, probing diverse biological mechanisms for toxicity, and improving toxicity prediction models. (Note: These plots do not attempt to convey "data-rich" chemicals, but rather chemicals that are predicted by the models to fall into one of the three toxicity categories. Similarly, chemicals not predicted to be positive should not be assumed to be predicted "negative", since the models may not be able to confidently predict portions of the remaining space.) Given the common interests and focus of EPA and NTP on the environmental toxicity landscape, as distinct from the NCATS's more exclusive interest in drugs, as well as the large overlapping

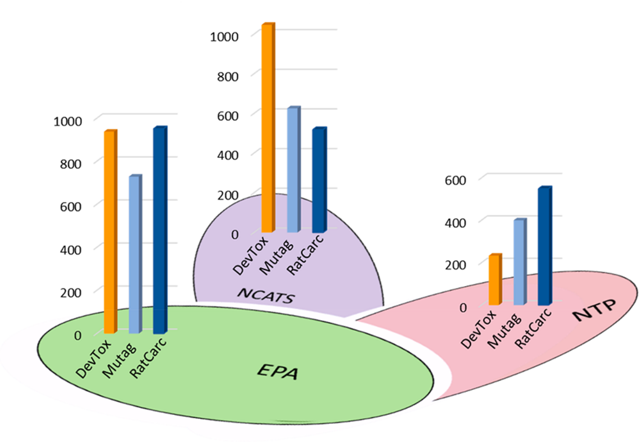


Figure 11. Comparison of numbers of predicted toxicants within the three non-overlapping Tox21 partner library sections: the full Tox21 EPA library (green, 4078 total), the portion of the Tox21 NTP library not overlapping with EPA (pink, 1861 total), and the portion of the Tox21 NCATS library not overlapping with either EPA or NTP (purple, 3005 total). DevTox (developmental toxicity) and Mutag (mutagenicity) were predicted from EPA T.E.S.T. models accessed from the EPA CompTox Chemicals Dashboard using 0.08 confidence threshold for DevTox and 0.5 forMutag. RatCarc (rat carcinogenicity) was predicted by the LHASA Derek Nexus software, v2.2.2 using the "Plausible" threshold (see Section 4.2 for details).

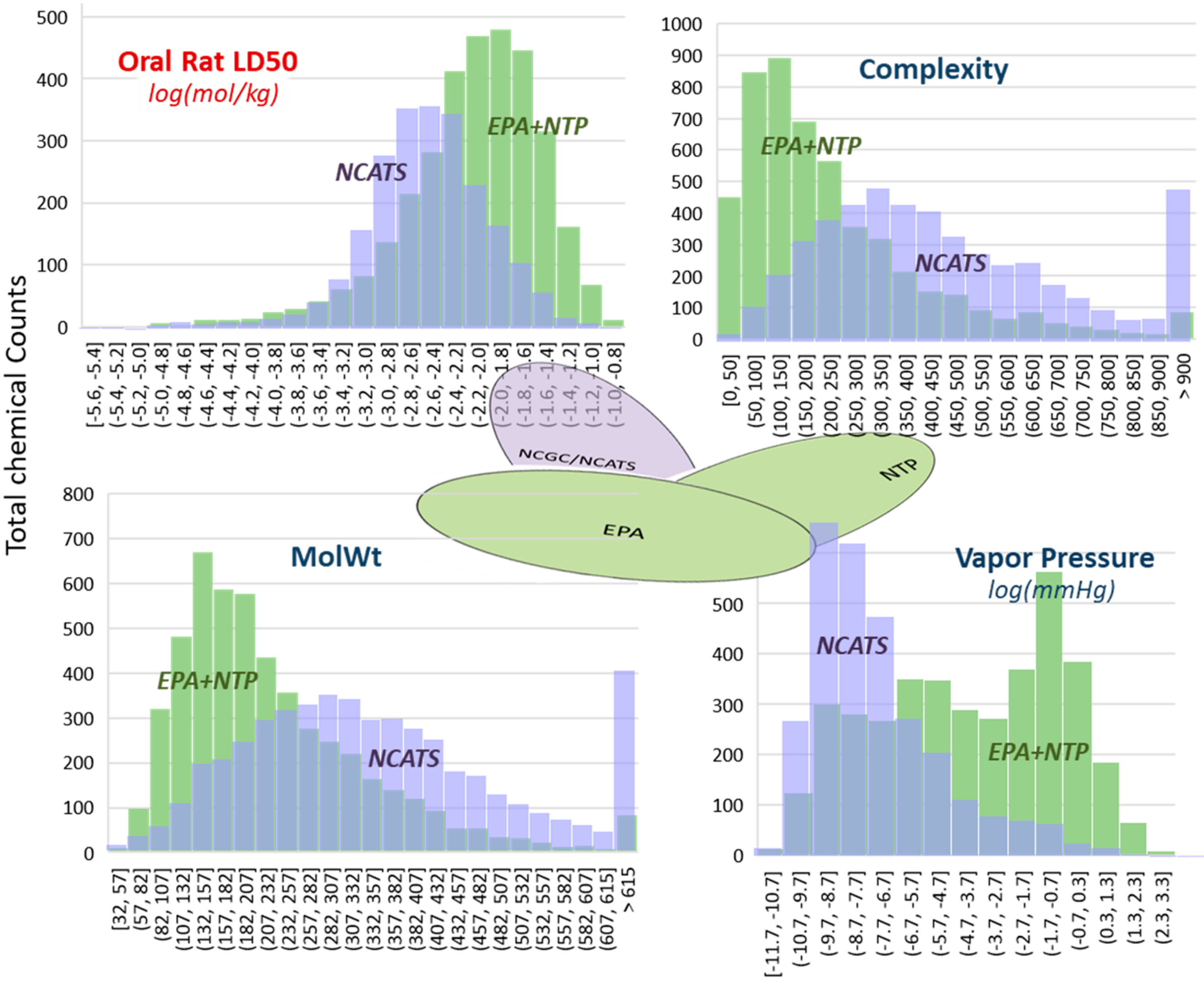


Figure 12. Histogram comparisons of computed properties (total chemical counts versus property range bins) for the combined EPA + NTP structure library versus the non-overlapping portion of the NCATS library (Venn diagram in center): Oral rat LD_{50} representing lethality to 50% of dosed rats predicted with the EPA T.E.S.T. QSAR model; complexity, based on atom, bond and path features, computed using the CORINA software (Molecular Networks GmbH); vapor pressure predicted using the OPERA QSAR models; and MolWt = molecular weight (see Section 4.2 for further details).

content of the EPA and NTP chemical libraries, the remaining comparisons presented will consolidate the EPA and NTP libraries into a single "EPA + NTP" library for the purpose of highlighting distinctions between inventories focused on environmental/industrial chemicals versus the NCATS library consisting exclusively of drugs. Note that a similar library split was made in a recent publication by Ngan et al. in order to examine differential Tox21 assay activity in the ENVR (EPA + NTP minus NCATS library overlap) versus DRUG (NCATS library) chemical landscapes.[24] The authors reported that Tox21 assay activity profiles were less able to discriminate between the ENVR and DRUG libraries than was chemical structure. The authors also found that DRUGs had slightly higher overall activity hit rates. Finally, the study reported higher structural diversity for the DRUG versus the ENVR library. We extend these observations in the following examples.

Figure 12 provides comparisons of four continuous properties predicted by structure-based models that were selected to highlight clear distinctions in the environmental toxicity landscape (EPA + NTP) versus the drug landscape (NCATS). Properties such as log K_{ow} and bioconcentration factor computed with the OPERA QSAR models showed little difference in their distributions between the EPA + NTP and NCATS libraries.[47] However, a QSAR-predicted quantitative toxicity end point (oral rat LD_{50}) and three chemical properties (molecular weight - MolWt, predicted vapor pressure, and complexity) each showed significantly shifted distributions between the two inventories. Interestingly, the peak of the EPA + NTP library distribution was biased toward greater oral rat LD_{50} values (indicating overall lower potency) than the NCATS/NCATS drug library. This is likely due to the designed bioactivity and higher overall potencies of drugs relative to commercial and industrial chemicals, consistent with the results of Ngan et al.[24] The remaining three plots in Figure 12 illustrate a substantial shift in the mean toward lower MolWt chemicals, with a related shift toward more volatile, higher vapor pressure chemicals in the EPA + NTP library

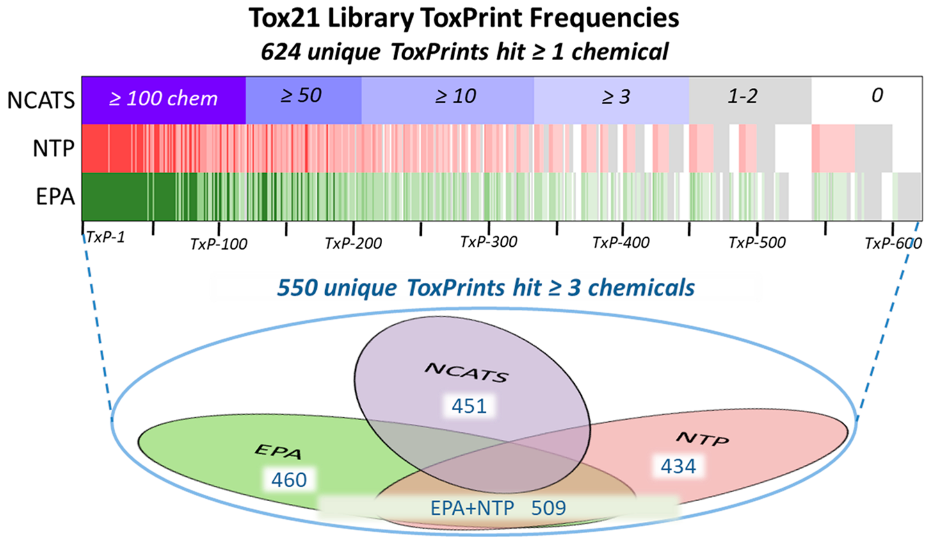


Figure 13. Frequency plot (heat map) of the number of chemicals containing each of 624 unique ToxPrint chemotypes sorted in the NCATS inventory from high (>100 chemicals per ToxPrint represented by the darkest shades of purple, red, and green), from 3 to 100 chemicals per ToxPrint represented by corresponding lighter color shades, and <3 chemicals per ToxPrint represented by light gray for all 3 inventories). The bottom portion of the figure indicates the numbers of unique ToxPrints present in three or more chemicals in each of the three total partner inventories as well as in the combined EPA + NTP inventory.

versus the NCATS drug library. A corresponding shift toward greater complexity is seen for the NCATS library, consistent with the report of greater chemical diversity (less interlibrary molecular similarity) within the NCATS drug library[24] and is likely also associated with higher mean MolWt for this library, as shown in Figure 12.

5.4. ToxPrint Fingerprint Profiles and Enrichments.

ToxPrint chemical structure fingerprints, or CTs, provide a simplified, generalized view of a chemical library by projecting it onto a smaller set of fixed chemical features. The total number of ToxPrint features represented within a compound library relative to the full set of ToxPrints (729) provides an estimate of chemical diversity and coverage, whereas comparison of ToxPrint feature profiles across libraries can highlight similarities and differences in local chemistry coverage.

A ToxPrint heat map projection of each of the Tox21 partner libraries is presented in Figure 13 along with the total number of unique ToxPrints (present in three or more chemicals) represented in each of the three partner libraries as well as in the combined EPA + NTP library and in the full Tox21 library. In total, 624 unique ToxPrints are represented across the full library in 1 or more chemicals, whereas 550 are represented in 3 or more chemicals. The missing ToxPrints mostly relate to individual metal atoms or metal-bond types; the ToxPrint set contains over 100 distinct chemotypes for such features. Each partner library section of the figure resembles a distinct ToxPrint "barcode", and the comparison emphasizes that there are areas of both commonality across partner libraries in feature space, enabling enrichment of chemotype−activity signals, and areas of significant differences in feature coverage across libraries, amplifying the importance of a single library contribution to that chemotype−activity subspace. Hence, the three libraries combined provide greater chemical coverage within local chemotype subspace and broader diversity across the ToxPrint landscape than each individually.

Another way in which the complementarity of each library and their respective contribution to the whole can be viewed is presented in Figure 14. Sets of ToxPrints that are "enriched" in each of the three partner libraries (and the EPA + NTP library) relative to the remaining libraries are shown, where a feature is labeled enriched if present in five or more chemicals in one library and in fewer than three chemicals in the remaining libraries. Since the libraries are projected onto a common set of ToxPrint features, one can more clearly see the cumulative enrichment in feature space when libraries are combined. Enrichment of feature space, in turn, provides greater opportunities to detect local structure−activity associations within assay results.

Of even greater interest is whether feature enrichments in the three libraries, samples of which are listed in Figure 14, translate into chemotype−activity enrichments that might not otherwise be detected if the Tox21 partner libraries were not combined. For this comparison, ToxPrint chemotype−activity enrichments were calculated for the binarized activity hit calls (1,0) of each Tox21 assay according to the EPA's standardized CTEW. Figure 15 presents results for a set of 23 ToxPrints that were found to be over-represented in the NCATS library relative to the EPA + NTP library and that were also found to be significantly enriched in the active space of one or more Tox21 assays according to the statistical thresholds of the CTEW. These include a variety of "ring:hetero_..." and "group:ligand_path_... tri- and bidentate" features more frequently represented in drugs. Given the relatively poor representation of these features in the EPA + NTP library, it is unlikely that their association with Tox21 activities would have been detected without the inclusion of the NCATS library chemicals.

Similarly, Figure 16 presents a set of 28 ToxPrints that were found to be over-represented in the EPA + NTP library relative to the NCATS library and that were also found to be significantly enriched in the active space of one or more Tox21 assays according to the statistical thresholds of the CTEW. These include several halide-containing features, polycyclic

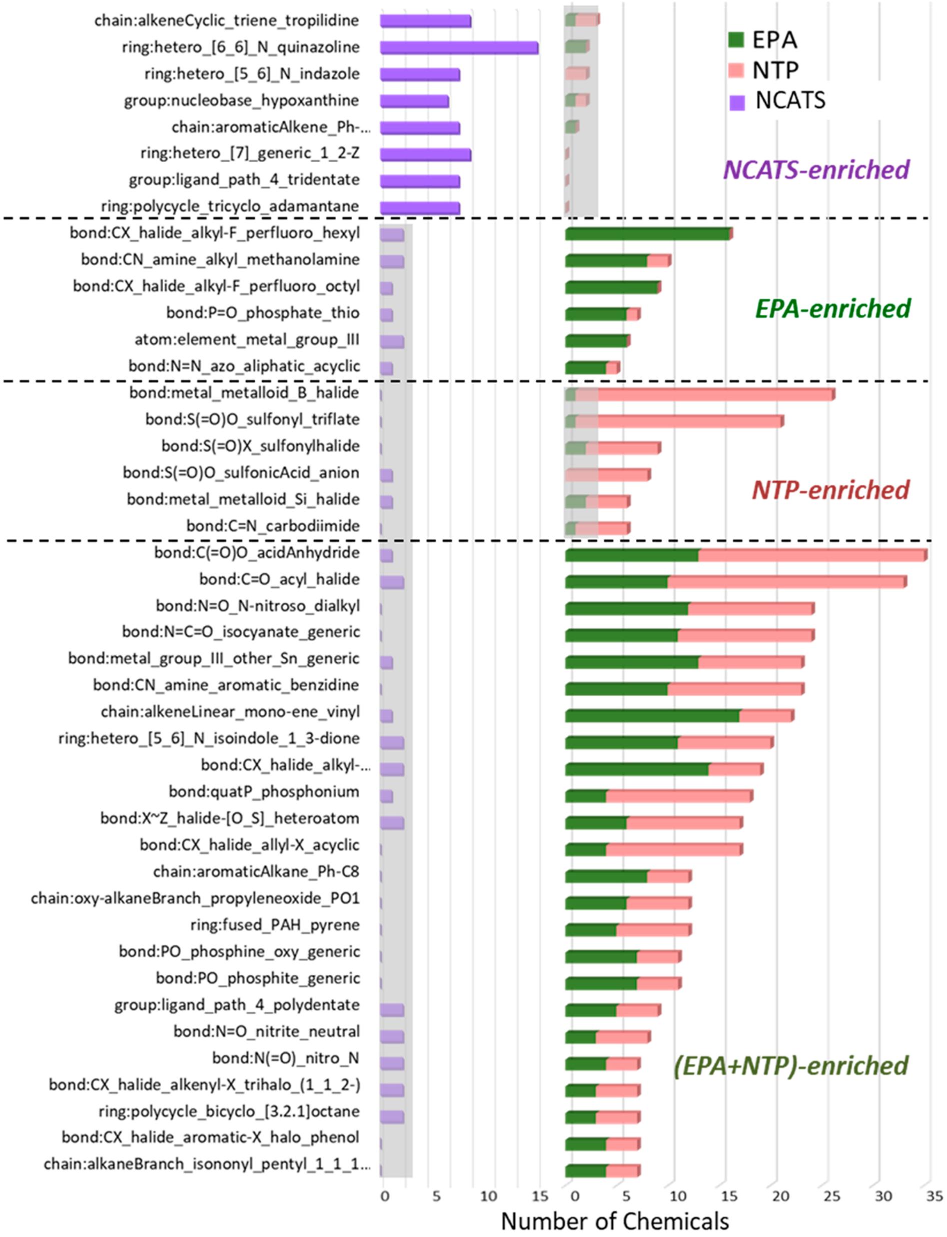


Figure 14. Subset of ToxPrint chemotypes showing the highest enrichment within a single partner inventory relative to the remaining inventories, comparing the incidence of each ToxPrint within the NCATS inventory (purple bars, far left) to the EPA (green) and NTP (pink) inventories (right panel), where enrichment is defined by ≥5 ToxPrint chemicals in the enriched library and ≤3 (grayed out) in each of the other partner libraries.

aromatic hydrocarbons, nitroso, azo, metal, and phosphorus-containing features, all less likely to be present in drugs. Given their poor representation in the NCATS drug library, it is unlikely that the association of these features with Tox21 activities would have been detected without the inclusion of the EPA + NTP library chemicals.

For the above two examples, one might argue that those who study environmental chemical toxicity are not interested in drugs (although drugs and their transformation products are found in the environment) and, likewise, those in the pharmaceutical industry are not interested in nondrug-like, environmental chemicals; hence, little is to be gained in combining libraries, as in the Tox21 effort, to provide greater coverage of chemical−activity space. Such divisions in chemical space are reflected in both the industrial and regulatory realms, but as we have pointed out previously (see Figure 9 and Table 2), many chemicals have multiple uses and cross over these use-type barriers. Additionally, when individual chemicals are viewed through the lens of more generalized chemical features, such as ToxPrint CTs, significant overlaps in CT space can be seen, presenting enhanced opportunities for detecting structure−activity patterns in assay results (see Figure 13). In addition, recognizing that the goal of the Tox21 project is to probe underlying biology to improve models to predict chemical toxicity, such chemical regulatory and use-type distinctions are counter-productive to improving understanding of underlying toxicity mechanisms that do not conform to such boundaries. In other words, the more examples of chemicals within a ToxPrint CT category that are contained in the full Tox21

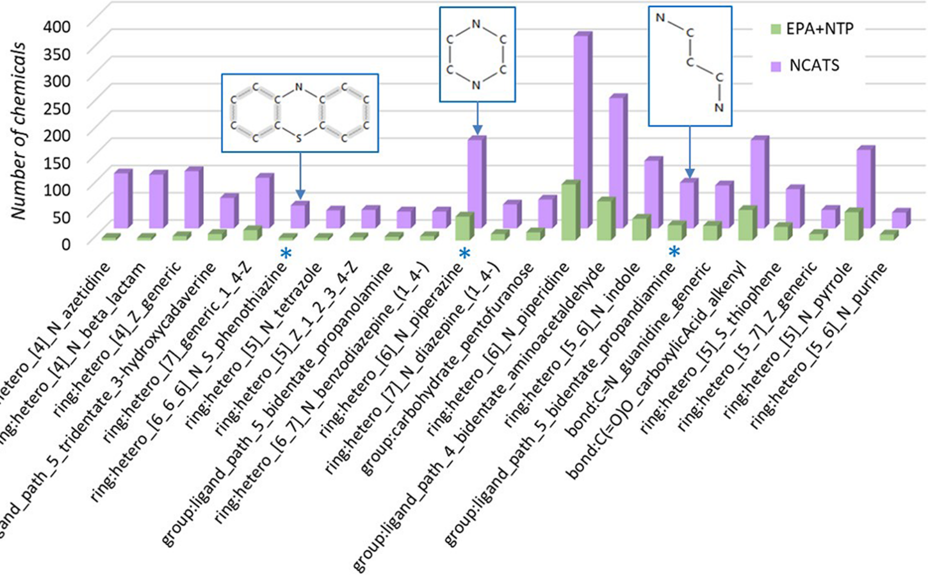


Figure 15. ToxPrint CTs with significantly greater representation in the NCTC/NCATS library (light purple bars) relative to the EPA + NTP library (light green bars), which were also found to be significantly enriched in the active chemical region of one or more Tox21 qHTS assays. Sample ToxPrint images are shown for starred names.

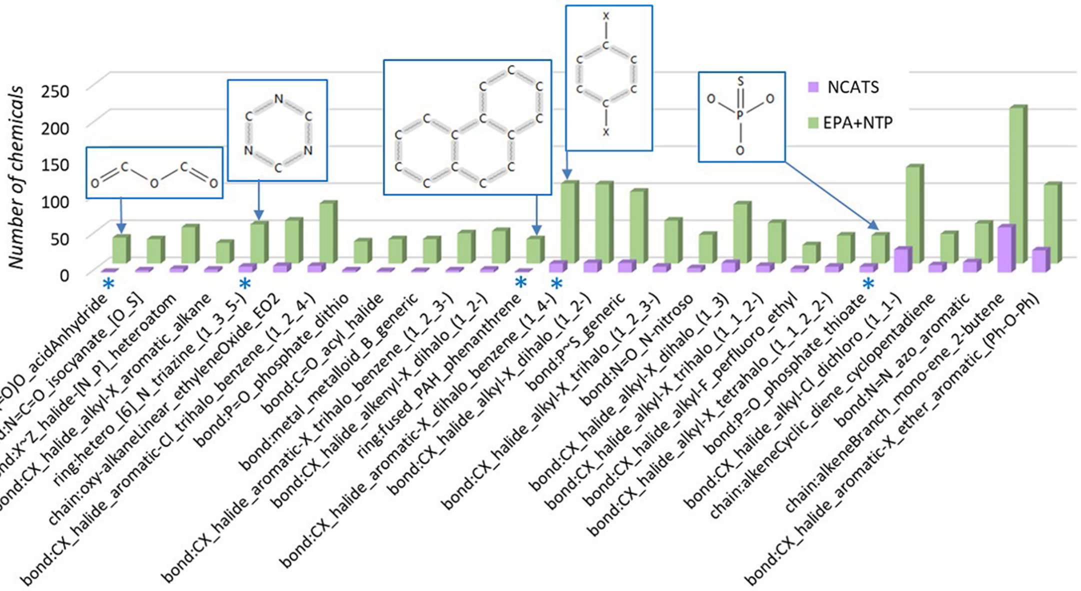


Figure 16. ToxPrint CTs with significantly greater representation in the EPA + NTP library (light green bars) relative to the NCATS library (light purple bars), which were also found to be significantly enriched in the active chemical region of one or more Tox21 qHTS assays. Sample ToxPrint images are shown for starred names.

screening library, the greater the chance of detecting patterns of activity in association with presence (or absence) of a CT, regardless of the intended use (e.g., pharmaceutical versus industrial).

5.5. ToxPrint Enrichments As a Function of Assay Platform and Test Set Size.

Up until now, we have considered the Tox21 compound library only in relation to Tox21 assays. However, all or part of EPA's Tox21 library also underwent screening within several other assay platforms (or technologies) as part of EPA's ToxCast program, significantly expanding the range of biological end points probed by this portion of the Tox21 library. ToxPrint CT enrichments have been computed not only for Tox21 assays but also for ToxCast assays, providing an opportunity for cross-comparisons between Tox21 and ToxCast assay results. Specifically, global patterns of CT enrichments provide a means for probing trends in assay results as a function of chemical test set size and diversity as well as biological diversity, where the latter is approximately represented by different assay platforms and end points.

Figure 17 presents the average number of ToxPrint CTs enriched per assay as a function of the average number of tested chemicals (i.e., the total number of screened chemicals for which binarized hit calls were reported) per assay within each distinct assay platform (e.g., CEETOX, Attagene) or testing program (i.e., All ToxCast), where the total number of

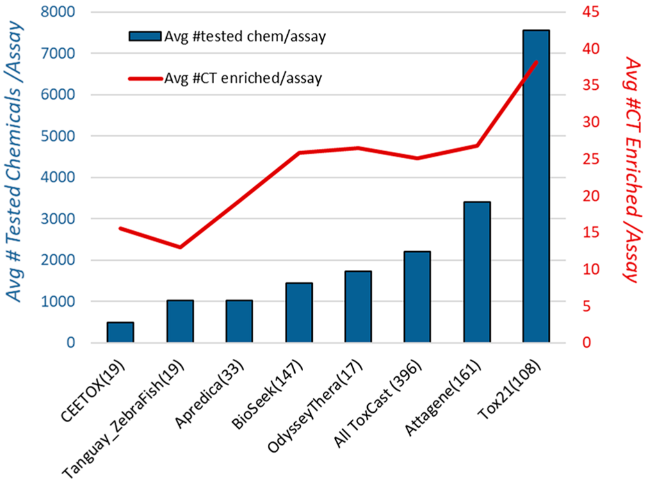


Figure 17. Plot of the average number of enriched CTs within each assay group (red line) superimposed on a bar plot of the average number of tested chemicals per assay (blue bars) within each assay group, where the number of assay end points per group is listed in parentheses beside the assay group name.

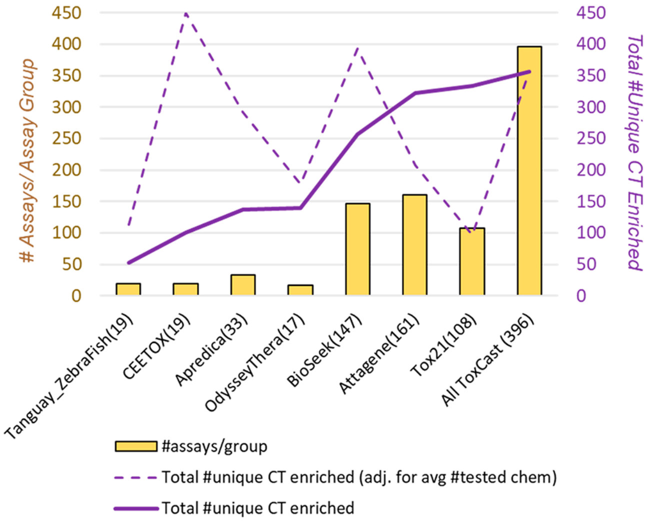


Figure 18. Plot of the total number of unique CTs enriched in assay actives (purple line) for all assays within the indicated assay platform (e.g., CEETOX) or testing program (All ToxCast), superimposed on a bar plot of the total number of assays contained within each assay platform or testing program (yellow bar), where the number of assay end points per assay platform is also listed in parentheses beside the platform name. The dashed purple line represents the total number of unique enriched CTs adjusted for the average total number of tested chemicals within the assay group (scaled to All ToxCast).

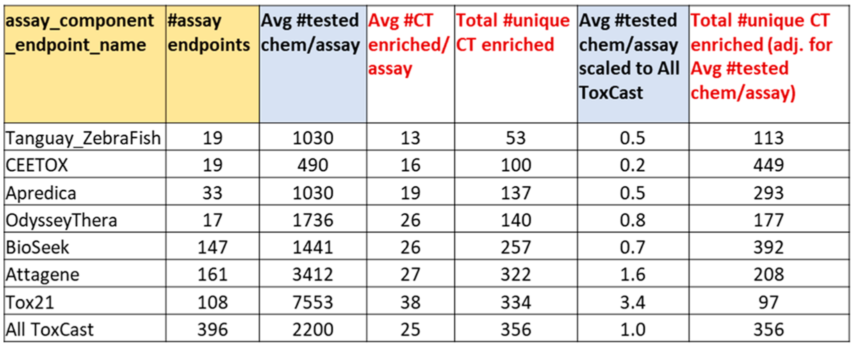
assays per group are indicated. (Note: Assay platforms with fewer than 15 end points and assay end points with test sets fewer than 450 chemicals were excluded from the comparisons.) The figure clearly shows a trend toward larger average numbers of enriched CTs per assay as the average number of tested chemicals per assay increases. This result is largely independent of assay group and reflects the increased ability to detect chemotype−activity enrichments in larger test sets that provide greater coverage of local CT chemical space (as seen in Figures 14−16).

In the same way that the number of enriched CTs is a measure of structure−activity information content within a single assay, we posit that the total number of unique enriched CTs detected across a battery of assays, such as within an assay platform or group (e.g., CEETOX), can provide an indication of biological end point diversity (and mechanism coverage) within the assay group. Figure 18 plots the total number of unique ToxPrint CTs enriched across all assays within each assay group as a function of the number of assays within the assay group. Note that the total numbers of unique enriched CTs across all assays within the assay groups are 4−15 times larger than the corresponding average numbers of enriched CTs for individual assays within a group (Figure 17 and Table 3). For instance, OdysseyThera averages 26 enriched CTs per assay but has 140 unique enriched CTs across the 17 total assays in the platform, suggesting a relatively high biological diversity of end points within the platform. Figure 18 orders the assay totals per group from low to high, and, although some trend is seen of increasing number of unique enriched CTs as a function of the number of assays within a group, there are clear exceptions to the rule. OdysseyThera, for instance, has the fewest number of assays (17), yet has 1.4 times as many unique enriched CTs detected in the assay actives as the similarly sized CEETOX assay group (140 CTs versus 100); however, its average test set size is also 3−4 times larger than CEETOX's (i.e., 1736 versus 490) (Table 3), which we have seen (Figure 17) has a strong influence on the number of enriched CTs per assay. To attempt to remove the influence of the test set size, the dashed line in Figure 18 shows the number of unique enriched CTs adjusted for the average test set size within the assay group. In this adjusted view, CEETOX, BioSeek and All ToxCast assay groups appear to provide similarly high levels of biodiversity information content per assay group (as measured by the adjusted number of unique CT enrichments). However, the influence of test set size cannot be so easily dismissed given that the larger absolute numbers of unique enriched CTs (i.e., 100 for CEETOX, 257 for BioSeek, 356 for All ToxCast) reflect increased coverage of local CT domains and associated chemical−activity mechanisms.

Table 3 provides a comparison of the information plotted separately in Figures 17 and 18, emphasizing the difference between average numbers of CTs enriched in a single assay versus the total number of unique CTs enriched across a battery of assays, where the latter count is 4−10 times larger in the above examples. The average number of CTs per assay is more heavily influenced by test set size (Figure 17), whereas the total number of unique CTs is influenced by both test set size and diversity of biological end points probed by the assays within a group (Figure 18). Attagene assays present an interesting case in that a larger proportion of the ToxCast chemical library was screened in this platform than for any other group of assays outside of Tox21. Furthermore, this platform detected over 90% of the unique CT enrichments represented in the "All ToxCast" group, which had four times as many assays. Again, however, the picture is clouded by the fact that Attagene tested, on average, 1.5 times as many chemicals per assay than the All ToxCast set (i.e., 3412 versus 2200).

Finally, a comparison of the unique CTs enriched in the Tox21 assays to those in the "All ToxCast" assays revealed a total of 90 enriched CTs in All ToxCast that were not detected in Tox21 assays and a total of 69 enriched CTs in Tox21 that were not found in ToxCast assays. Enriched CTs found only in Tox21 assay results are likely due to the Tox21-only screening

Table 3. Comparison of ToxCast and Tox21 Assay Platform and Chemotype−Activity Enrichment Characteristics As Plotted in Figures 17 and 18



assay_component_endpoint_name	#assay endpoints	Avg #tested chem/assay	Avg #CT enriched/assay	Total #unique CT enriched	Avg #tested chem/assay scaled to All ToxCast	Total #unique CT enriched (adj. for Avg #tested chem/assay)
Tanguay_ZebraFish	19	1030	13	53	0.5	113
CEETOX	19	490	16	100	0.2	449
Apredica	33	1030	19	137	0.5	293
OdysseyThera	17	1736	26	140	0.8	177
BioSeek	147	1441	26	257	0.7	392
Attagene	161	3412	27	322	1.6	208
Tox21	108	7553	38	334	3.4	97
All ToxCast	396	2200	25	356	1.0	356

of the large NCGC/NCAT drug library (minus the overlaps with EPA's ToxCast library), with these under-represented CTs including many of those shown in Figures 14 and 15. Enriched CTs found only in All ToxCast assays, in turn, are more likely due to the larger number and types of ToxCast assays (i.e., 396 for All ToxCast, 108 for Tox21) sampling more diverse biological space than the Tox21 assays. These results, although high-level and requiring follow-up, are intriguing and imply that sufficient test set size to cover diverse local CT domains as well as a diversity of assay platforms and end points are both necessary to capture the broadest range of chemical−activity associations. These observations can inform the timely argument related to whether a stated goal of the Tox21 program, to improve toxicity prediction models for all chemicals of interest, is better served by continued testing of the full Tox21 chemical library (or an expanded chemical space) or by expanding the scope of biological targets and screening fewer, strategically chosen chemical sets. The present results argue that a two-pronged approach involving strategic selection of chemicals and assays, balancing the influence of test set size (providing adequate diversity and local coverage of CT space) with sampling of biological end points and targets, informed by the results of CT-enrichment studies, can offer a practical and productive path forward.

6. TOX21 PRESENT AND FUTURE

The Tox21 multi-agency, federal partner project has now spanned over a decade and has helped to propel the field of toxicology into the 21st century by embracing new advances in quantitative high-throughput screening, by the application of modern cheminformatics and data analyses methods, and by committing to prompt, full public data release with accompanying tools for data viewing and analysis. This collaborative effort has succeeded in generating a wealth of chemical−bioactivity data and provides a multitude of opportunities for researchers to use and analyze these data in new ways. Underpinning and fueling this effort has been the largest compound library specifically designed for the purpose of gaining a better understanding the chemical basis of toxicology, spanning many areas of regulatory authority and applications (e.g., related to pharmaceuticals, environmental and industrial chemicals, cosmetics, food additives, and consumer products). The coming together of ideas and chemical libraries to create, manage, analyze, and screen the full Tox21 10K library was unprecedented when the Tox21 project launched in 2007 and full library screening commenced in 2012. Hence, many decisions had to be made in the course of library construction, prior to the commencement of screening of the full library, that would largely determine and constrain all future uses of Tox21 data. Whereas NCATS had prior experience with HTS screening for drug development, for EPA and NTP, the image of an "airplane being built while flying" has been used to convey the early years of the ToxCast and Tox21 programs in which these agencies were moving into new, uncharted territory and having to repurpose existing capabilities and contracts. A greater ability to coordinate databases and chemical library development early on might have allowed for a more rationally designed Tox21 compound library. However, during these early stages of the program, there were limited data relative to chemical−bioactivity determinants of toxicity for most environmental/industrial chemicals and drugs as to be insufficient for intelligently designing such a library, that is, we did not know what we did not know. Hence, each agency took a practical approach, focusing on procuring as many chemicals of programmatic relevance as possible that could be solubilized and screened. That this approach yielded a highly structurally diverse Tox21 library, spanning many use types and functionalities, broad property ranges, and many types of sample and compound replicates, is perhaps fortuitous yet has ultimately served the program well. On the other hand, the foresight to create a sufficient plate stock of library samples from each of the main partners (NCATS, NTP, and EPA) to last through to the present, to incorporate a range of quality control replicates and processes, and to institute high-quality chemical curation and sample tracking databases that would best support future analyses was deliberate and has also served the project well.

One aim of the present article was to document this history of the compound library construction and management to serve present and future aims of the Tox21 program and to potentially guide future testing programs. A second aim of the paper was to shine a light on the compound library contents, highlighting the chemical structure diversity and coverage of chemical use categories, regulatory interests, toxicity end points, and chemical features and properties. To provide a frame of reference, we chose to examine these issues by comparing the relative contributions of the three Tox21 partner libraries to the whole, showing by means of several examples how each partner's library contribution succeeded in expanding the diversity and scope of the full library, enabling enhancement of chemical−activity enrichment signals in the assay activity space that would otherwise not be detectable. The global CT enrichment analysis presented here provided a higher-level perspective of activity patterns, including a comparison to ToxCast library results made possible through use of the generalized ToxPrint fingerprints and CT enrichments. However, the devil is in the details, with these initial results requiring follow-up to explore the meaning of the various enrichment patterns. In particular, the role of assay artifacts, chemical promiscuity, and analytical QC on assay results and corresponding CT enrichments remains to be more fully described and understood.

Today, the full Tox21 full library is nearing the end of its screening and data generation run with the original library partner plate sets (Figure 2) and is moving into a new phase of strategically focused Tox21 partner projects.[26] To continue to serve the needs of these projects, the NCATS has replenished and expanded their drug library, and the EPA has continued to expand their screening library into new areas of chemistry (e.g., perfluorinated chemicals) and, in partnership with the NTP, into targeted bioactivity space (with expanded inclusion of reference chemicals). To continue to invest in and more efficiently support these collaborative projects, the NTP also recently reprocured a portion of its Tox21 library (excluding overlaps with EPA's existing compound library and some problematic chemicals), and both the EPA's and NTP's Tox21 libraries (plus EPA's expanded ToxCast library) have been consolidated under the EPA's chemical management to serve current and future collaborative Tox21 projects (for a list of current Tox21 partner projects, visit https://tox21.gov/projects/). The library continues to expand in areas of chemical diversity and bioactivity, but also retains limits imposed by practical constraints of DMSO solubility and volatility. However, continued screening of the full Tox21 library is not currently envisioned once the last original plate set is depleted. Rather, strategic selections of chemicals, informed by past Tox21 and ToxCast screening results and analyses, such as presented here, will move the Tox21 partner projects into new areas of chemical and biological screening, including high-throughput phenotypic profiling, high-throughput toxicokinetics, and toxicogenomics.

One final point to be made is relative to the future application of Tox21 chemical-bioassay results to the evaluation and screening of new chemicals. The generation of Tox21 bioactivity profiles, feeding into predictive models based wholly or in part on Tox21 HTS results, independent of structure–activity considerations, is an impractical path forward for evaluating the potential toxicity of new chemicals, particularly if the new chemicals are unsuitable for testing. Hence, a large "ask" of historical Tox21 data is to reveal structure–activity relationship insights and patterns that can potentially be combined with *in silico* modeling methods and domain knowledge to focus resources and testing into productive areas of inquiry. Here, an in-depth understanding of patterns of activity within local areas of chemical feature space, such as explored here with CT enrichments, and the ability to "look across" the bioactivity and chemical landscape using a fixed set of chemical features, such as ToxPrints, offer a promising path forward. In all of these future endeavors, the Tox21 compound library and its various cheminformatics representations, as described herein, will continue to play a pivotal role in fueling and defining the scope and ultimate success of Tox21 program projects.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.chemrestox.0c00264.

Table S1: Tox21 IDs mapped to NCGC IDs, PubChem IDs, and DSSTox IDs, and indicating NCATS, NTP and EPA partner library associations (date stamped February 24, 2020). Table S2: DSSTox TOX21SL list of substance IDs and structure formula, molecular weight, SMILES, InChI, and QSAR-ready SMILES (downloaded January 24, 2020). Table S3: DSSTox TOX21SL DTXSID overlaps with EPA CompTox Dashboard lists (downloaded January 24, 2020). Table S4: Predicted physicochemical properties and toxicities generated from OPERA, T.E.S.T, CORINA, and Derek Nexus models. Table S5: ToxPrint (V2.0_r711) fingerprint file for the TOX21SL chemical list. Table S6: Chemotype enrichment workflow results generated from binarized activity hit calls for ToxCast and Tox21 assay end points (aeids) obtained from EPA's public ToxCast database, invitroDBv2. Table S7: Tox21 binarized assay hit call matrix for stereo and salt pairs, extracted from EPA's public ToxCast database, invitroDBv3 (XLSX)

■ AUTHOR INFORMATION

Corresponding Author

Ann M. Richard − *Center for Computational Toxicology and Exposure, Office of Research and Development, United States Environmental Protection Agency, Research Triangle Park, North Carolina 27711, United States;* orcid.org/0000-0003-2116-2300; Phone: 919-541-3934; Email: richard.ann@epa.gov

Authors

Ruili Huang − *National Center for Advancing Translational Sciences, National Institutes of Health, Rockville, Maryland 20850, United States*

Suramya Waidyanatha − *Division of the National Toxicology Program, National Institute of Environmental Health Sciences, Research Triangle Park, North Carolina 27709, United States*

Paul Shinn − *National Center for Advancing Translational Sciences, National Institutes of Health, Rockville, Maryland 20850, United States*

Bradley J. Collins − *Division of the National Toxicology Program, National Institute of Environmental Health Sciences, Research Triangle Park, North Carolina 27709, United States*

Inthirany Thillainadarajah − *Senior Environmental Employment Program, United States Environmental Protection Agency, Research Triangle Park, North Carolina 27711, United States*

Christopher M. Grulke − *Center for Computational Toxicology and Exposure, Office of Research and Development, United States Environmental Protection Agency, Research Triangle Park, North Carolina 27711, United States*

Antony J. Williams − *Center for Computational Toxicology and Exposure, Office of Research and Development, United States Environmental Protection Agency, Research Triangle Park, North Carolina 27711, United States;* orcid.org/0000-0002-2668-4821

Ryan R. Lougee − *Center for Computational Toxicology and Exposure, Office of Research and Development, United States Environmental Protection Agency, Research Triangle Park,* North Carolina 27711, United States; Oak Ridge Institute for Science and Education, United States Department of Energy, Oak Ridge, Tennessee 37830, United States

Richard S. Judson − *Center for Computational Toxicology and Exposure, Office of Research and Development, United States Environmental Protection Agency, Research Triangle Park, North Carolina 27711, United States*

Keith A. Houck − *Center for Computational Toxicology and Exposure, Office of Research and Development, United States Environmental Protection Agency, Research Triangle Park, North Carolina 27711, United States;* orcid.org/0000-0002-0055-2249

Mahmoud Shobair − *Center for Computational Toxicology and Exposure, Office of Research and Development, United States Environmental Protection Agency, Research Triangle Park, North Carolina 27711, United States*

Chihae Yang − *Altamira, LLC, Columbus, Ohio 43235, United States; Molecular Networks, GmbH, Erlangen 90411, Germany;* orcid.org/0000-0003-2529-866X

James F. Rathman − *Altamira, LLC, Columbus, Ohio 43235, United States; Molecular Networks, GmbH, Erlangen 90411, Germany*

Adam Yasgar − *National Center for Advancing Translational Sciences, National Institutes of Health, Rockville, Maryland 20850, United States;* orcid.org/0000-0001-7350-1402

Suzanne C. Fitzpatrick − *Center for Food Safety and Applied Nutrition, United States Food and Drug Administration, College Park, Maryland 20740, United States*

Anton Simeonov − *National Center for Advancing Translational Sciences, National Institutes of Health, Rockville, Maryland 20850, United States*

Russell S. Thomas − *Center for Computational Toxicology and Exposure, Office of Research and Development, United States Environmental Protection Agency, Research Triangle Park, North Carolina 27711, United States;* orcid.org/0000-0002-2340-0301

Kevin M. Crofton − *Center for Computational Toxicology and Exposure, Office of Research and Development, United States Environmental Protection Agency, Research Triangle Park, North Carolina 27711, United States; R3Fellows, LLC, Durham, North Carolina 27701, United States*

Richard S. Paules − *Division of the National Toxicology Program and Division of the National Toxicology Program, National Institute of Environmental Health Sciences, Research Triangle Park, North Carolina 27709, United States*

John R. Bucher − *Division of the National Toxicology Program, National Institute of Environmental Health Sciences, Research Triangle Park, North Carolina 27709, United States*

Christopher P. Austin − *National Center for Advancing Translational Sciences, National Institutes of Health, Rockville, Maryland 20850, United States*

Robert J. Kavlock − *Center for Computational Toxicology and Exposure, Office of Research and Development, United States Environmental Protection Agency, Research Triangle Park, North Carolina 27711, United States; Kavlock Consulting, LLC, Washington, DC 20001, United States*

Raymond R. Tice − *Division of the National Toxicology Program, National Institute of Environmental Health Sciences, Research Triangle Park, North Carolina 27709, United States; RTice Consulting, Hillsborough, North Carolina 27278, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.chemrestox.0c00264

Author Contributions

A.M.R is lead author, heads the DSSTox and CTEW projects, and served as EPA's chemical contract manager for EPA's ToxCast and Tox21 library construction from 2007-2017. R.H. has served as co-lead on the Tox21 Chemical Library Team (with A.M.R. and S.W.) since the start of the Tox21 program, supervised construction of the NCATS Tox21 library and Tripod website, and heads the NCATS cheminformatics team. S.W. has served as co-lead of the Tox21 Chemical Library Team since 2013 and heads up the NTP's chemical contract management. P.S. led efforts within the NCATS to consolidate and plate the full Tox21 library for testing. B.J.C. led efforts to construct the NTP's Tox21 library and ship samples to NCATS and EPA. I.T. has been a lead DSSTox chemical curator of Tox21 content through to the present. C.M.G. is head cheminformatician supporting DSSTox chemical registration and chemical data linkages throughout the EPA's databases. A.J.W. is lead for the EPA's CompTox Chemicals Dashboard and property prediction (OPERA) projects and is responsible for hosting chemical lists used in the present analysis. R.R.L. developed the EPA's CTEW and generated the CT enrichment results used in the present study. R.S.J. led the ACToR project that provided the EPA's ToxCast and Tox21 chemical nominations and is the head bioinformatician leading efforts within the EPA to model ToxCast and Tox21 assay results. K.A.H. nominated chemicals for Tox21 inclusion and has been a HTS assay lead for both the ToxCast and Tox21 programs within the EPA. M.S. assisted in the Tox21 cheminformatics data analysis. C.Y. contributed FDA chemical lists to EPA Tox21 library nominations. C.Y. and J.R. led efforts to develop and publicly distribute the ChemoTyper and ToxPrints, providing an API version of the software for use on the public CompTox Chemicals Dashboard. A.Y. assisted in the plating of the initial NTP collection and played a major role in procurement of the NCATS drugs library. S.C.F. is the Tox21 lead for the FDA. A.S. is the Tox21 lead for the NCATS. R.S.T. is the Tox21 lead for the EPA. K.M.C. is a former Tox21 lead for the EPA. R.S.P. is a former Tox21 lead for the NTP. J.B. was a founding Tox21 lead for the NTP and has remained involved in the program to the present. C.P.A. was a founding Tox21 lead for the NCGC and later NCATS and has remained involved in the program to the present. R.J.K. was a founding Tox21 lead for the EPA. R.R.T. was a founding Tox21 lead for the NTP.

Funding

The work presented in this manuscript was solely supported by appropriated funds of the United States Environmental Protection Agency, National Institutes of Health, and United States Food and Drug Administration.

Notes

The authors declare no competing financial interest.

□K.M.C., R.S.P., R.J.K., and R.R.T. are retired from government service.

The views expressed in this paper are those of the authors and do not necessarily reflect the views or policies of the United States Environmental Protection Agency or the United States Food and Drug Administration. Mention of trade names or commercial products does not constitute endorsement or recommendation for use.

Biographies

Dr. Ann M. Richard received her Ph.D. (Theoretical Physical Chemistry) from the University of North Carolina at Chapel Hill and has been a Principal Investigator within the U.S. Environmental Protection Agency (EPA) since 1987. She joined the National Center for Computational Toxicology in 2005 and led the DSSTox project and chemical management efforts in support of the EPA's ToxCast and Tox21 programs through to 2019. Her research interests lie in creating a knowledge-informed, quality cheminformatics interface between the chemical landscape and the *in vitro* and *in vivo* data landscapes that can be used to guide modeling into productive areas of mechanistic inquiry.

Dr. Ruili Huang is the informatics group leader on the toxicity profiling team at the National Center for Advancing Translational Sciences (NCATS) and co-chair of the Tox21 chemical library working group. She came to NCATS from the National Cancer Institute in 2006. As an informatics team, her group contributes to qHTS data processing and interpretation and integrates biological pathway information and assay data to support interpretation of results. Additionally, her group analyzes compound toxicity profiling data and develops software tools and algorithms to build predictive models for *in vivo* toxicity. She received her Ph.D. in Chemistry from Iowa State University.

Dr. Suramya Waidyanatha leads the Chemistry and ADME Resources Group at the National Toxicology Program (NTP) and manages contracts for ADME, toxicokinetics, and analytical chemistry in support of NTP toxicology studies. Her research interests include metabolism and disposition of xenobiotics and mechanism(s) of interactions of xenobiotics and/or their metabolites with cellular macromolecules resulting in toxicity. She received her M.S. in Biochemistry from University of Illinois, Chicago and her Ph.D. in Analytical Chemistry from University of Maine. She joined NTP in 2008 as a staff scientist.

Paul Shinn is the compound management group leader at the National Center for Advancing Translational Sciences (NCATS) and joined the NIH in 2006. His group maintains the compound libraries that support high-throughput screening campaigns performed on various automation platforms. His primary focus is to create new methods or technologies that will enable researchers to advance their research. Prior to joining NIH, he worked at the Salk Institute for Biological Studies in La Jolla, California. He received his B.A. in Biochemistry with a minor in Chemistry from the University of Pennsylvania.

Bradley J. Collins serves as a project officer in the Program Operations Branch for the National Toxicology Program (NTP). He manages contracted chemistry efforts in support of NTP toxicology studies and has served on study design teams and interagency working groups to provide technical assistance in the areas of chemistry and toxicokinetics. He helps to design toxicokinetic studies and provides cheminformatics support to the NTP high-throughput toxicology screening program. He received his B.S. in Science and Environmental Change from the University of Wisconsin-Green Bay and his MSPH in Environmental Chemistry from the University of North Carolina at Chapel Hill.

Inthirany Thillainadarajah received her B.S. degree in Chemistry and Chemical Technology from University of Manchester, Manchester, UK and M.S. in Chemical Education from University of Arizona, Tucson, Arizona, USA. She has worked in R&D departments in the pharmaceutical industry and preclinical research in medical facilities. She has many years of research experience in molecular/cell biology, biochemistry, and pharmaceutical sciences. Presently, she is a lead curator/quality control analyst for the DSSTox database supporting ToxCast and Tox21 programs at the U.S. Environmental Protection Agency's Center for Computational Toxicology and Exposure.

Dr. Christopher M. Grulke received a B.S.E. from the University of Michigan in Chemical Engineering (2003) and a Ph.D. in Pharmaceutical Science, Medicinal Chemistry, and Biophysics from the University of North Carolina at Chapel Hill (2011). He is currently employed as a Cheminformatician Scientist at the U.S. Environmental Protection Agency's Center for Computational Toxicology and Exposure. He is applying advanced database and software development skills to build a cheminformatics infrastructure for integrating chemical and biological data to support the development of predictive models pertaining to exposure, pharmacokinetics, and toxicity.

Dr. Antony J. Williams received a Ph.D. in analytical chemistry (NMR) from the University of London, UK in 1988. He joined ACD/Labs as Chief Science Officer focusing on structure representation, nomenclature, and analytical data management. He was a co-founder of the ChemSpider database, later acquired by the Royal Society of Chemistry. In 2015, he joined the U.S. Environmental Protection Agency and is a project lead for the CompTox Chemicals Dashboard in the Center for Computational Toxicology and Exposure. His research is focused on development of web-based applications to access chemistry data and structure identification using combined mass spectrometry and informatics approaches.

Ryan R. Lougee received his B.S. in Biochemistry from the University of New Hampshire in 2012 and M.S. in Computational Toxicology from North Carolina State University in 2016. In 2017, he received a Student Trainee grant under the Oak Ridge Institute for Science and Education (ORISE) program to work at the U.S. Environmental Protection Agency's Center for Computational Toxicology and Exposure under the mentorship of Dr. Ann Richard. He is involved in several projects applying computational chemistry, cheminformatics, and programming skills to problems in environmental toxicology.

Dr. Richard S. Judson is with the U.S. Environmental Protection Agency where he develops computer models and databases to help predict toxicological effects of environmental chemicals. A past focus was on the development of models of chemicals interacting with the endocrine system. He has published in areas including computational biology and chemistry, bioinformatics, genomics, human genetics, toxicology, and applied mathematics. Prior to joining the EPA, he held positions in biotechnology and with the Department of Energy laboratories. Dr. Judson has a B.A. in Chemistry and Chemical Physics from Rice University and an M.A. and Ph.D. in Chemistry from Princeton University.

Dr. Keith A. Houck received an M.S. (Chemistry) from the University of North Carolina at Chapel Hill in 1985 and a Ph.D. (Toxicology and Pathology) from Duke University in 1989. Following a Postdoctoral Fellowship at Genentech, Inc., he worked in the biotechnology and pharmaceutical fields for 13 years at Sphinx Pharmaceuticals and Eli Lilly & Co. He joined the U.S. Environmental Protection Agency as a Principal Investigator in the National Center for Computational Toxicology in 2005 supporting the ToxCast project and Tox21 program. His research interests focus on understanding the complex interactions of chemical and biological targets underlying mechanisms of toxicity.

Dr. Mahmoud Shobair is a Post-Doctoral Fellow at the U.S. Environmental Protection Agency's Center for Computational Toxicology and Exposure, working with Drs. Ann Richard and Christopher Grulke. His research interests focus on combining computational chemistry and experimental techniques to investigate structure−function relationships and biochemical mechanisms of toxicity. He is currently modeling the molecular initiating events along the hypothalamic-pituitary-thyroid axis. He received a B.S. in physics from the University of Florida in Gainesville and a Ph.D. in Biochemistry and Biophysics with a graduate certificate in Bioinformatics and Computational Biology from the University of North Carolina at Chapel Hill.

Dr. Chihae Yang is Managing Director and CEO of MN-AM (Molecular Networks GmbH and Altamira LLC) and an adjunct professor at Ohio State University. She was previously a visiting fellow in the computational toxicology program at the U.S. Food and Drug Administration (2008-2011) and Chief Scientific Officer at Leadscope (2000-2008). She has developed the ToxML database standard, chemoinformatics-based data mining techniques, and a structural feature-based computational modeling system. Dr. Yang was also formerly a tenured professor in the Department of Chemistry at Otterbein College and a Senior Scientist at The Clorox Company.

Dr. James F. Rathman is a Professor of Chemical and Biomolecular Engineering at Ohio State University (OSU) and Managing Director of Altamira, LLC. Before coming to OSU in 1991, he spent seven years in industrial R&D with Conoco Inc. and The Clorox Company. His research interests include molecular informatics, interfacial and colloidal phenomena, computational modeling and simulation, machine learning, and statistical analysis and experimental design. His current research efforts focus on computational risk assessment of complex chemical systems, with emphasis on pharmaceuticals and cosmetics.

Adam Yasgar is a staff scientist at NCATS, where he works the miniaturization of biochemical and cell-based assays for high-throughput screening campaigns. He received a B.S. in chemistry at George Washington University. He worked in Pfizer's Pharmacokinetics, Dynamics, and Metabolism preclinical group as a research associate, specializing in the CNS therapeutic area, and later joined the National Human Genome Research Institute, where he helped build a team of automation and compound management experts before joining the biology team led by Dr. Anton Simeonov. He has worked on several screening campaigns using the NCATS' robotic screening system.

Dr. Suzanne C. Fitzpatrick is a Senior Science Advisor for Toxicology in FDA's Center for Food Safety and Applied Nutrition. She is the FDA lead for Tox21, chair of the FDA Predictive Toxicology Roadmap Committee, and played a pivotal role in helping to launch the organs-on-a-chip tool being evaluated by FDA. She is an Adjunct Professor at Johns Hopkins University, the FDA representative to the Johns Hopkins Center for Alternatives to Animal Testing Board, and past president of the American College of Toxicology. She received her B.A. from the University of California at San Diego and Ph.D. from Georgetown University.

Dr. Anton Simeonov is the scientific director at the National Center for Advancing Translational Sciences. His background ranges from bioorganic chemistry and molecular biology to clinical diagnostic research and development. He received a Ph.D. in Bioorganic Chemistry from the University of Southern California and a B.A. in Chemistry from Concordia College. Prior to joining NIH in November 2004, Simeonov was a senior scientist at Caliper Life Sciences, a leading developer of microfluidic technologies, where he was responsible for basic research on novel assay methodologies and development of microfluidic products for research and clinical diagnostics.

Dr. Russell Thomas is the Director of the Center for Computational Toxicology and Exposure at the U.S. Environmental Protection Agency. The Center is performing research to rapidly evaluate the potential human health and environmental risks due to exposures to environmental chemicals and ensure the integrity of the freshwater environment and its capacity to support human well-being. He received his B.A. in Chemistry from Tabor College and M.S. in Radiation Ecology and Health Physics and Ph.D. in Toxicology from Colorado State. He performed postdoctoral research in molecular biology and genomics at the McArdle Cancer Research Laboratory at the University of Wisconsin.

Dr. Kevin M. Crofton served as the Deputy Director of the National Center for Computational Toxicology at the U.S. Environmental Protection Agency prior to his retirement in late 2017. He received a M.S. in Zoology from Miami University, and a Ph.D. in Toxicology from the University of North Carolina at Chapel Hill. His research interests include developmental neurotoxicity, with an emphasis on the development of adverse outcome pathways and *in vitro* alternative testing methods, particularly as they relate to the impact of endocrine disruptors and the cumulative risk of thyroid disruptors and pesticides.

Dr. Richard S. Paules served as Chief of the Biomolecular Screening Branch in the National Toxicology Program and NTP lead for Tox21 until his retirement in early 2020. The BSB develops and carries out programs in medium and high-throughput screening of substances for detection of biological activities of significance to toxicology and administers programs to implement the NTP vision for Tox21. Until his retirement, he also held adjunct appointments at University of North Carolina at Chapel Hill (UNC-Chapel Hill) as Professor in the Department of Pathology and Member in the Lineberger Comprehensive Cancer Center. Dr. Paules received his Ph.D. from the Department of Pathology at UNC-Chapel Hill.

Dr. John R. Bucher is the past Director of the Division of the National Toxicology Program (NTP) at the National Institute of Environmental Health Sciences and Associate Director of the National Toxicology Program. He is a Diplomate of the American Board of Toxicology and a member of the Collegium Ramazzini. He received a M.S. in Biochemistry from the University of North Carolina at Chapel Hill and a Ph.D. in Pharmacology from the University of Iowa. He is the founding NTP Tox21 project lead.

Dr. Christopher P. Austin is director of the National Center for Advancing Translational Sciences (NCATS) at the National Institutes of Health. He leads the center's work translating observations in the laboratory, clinic, and community into interventions that benefit patients. Under his direction, NCATS researchers and collaborators are developing new technologies, resources, and collaborative research models, demonstrating their usefulness, and disseminating data, analysis, and methodologies for use by the worldwide research community. Prior to joining NCATS, he worked at the National Human Genome Research Institute. He earned an M.D. from Harvard Medical School and an A.B. in biology from Princeton University.

Dr. Robert J. Kavlock retired from the U.S. Environmental Protection Agency after 39 years of service in 2017. His last position was that of Agency Science Advisor and previous positions included Deputy Assistant Administrator for Science, the inaugural Director of the National Center for Computational Toxicology, and the Director of the Reproductive Toxicology Division. His expertise includes hazard and risk assessment of environmental chemicals, especially for noncancer health outcomes, and he was the original EPA lead of the

Tox21 project. He received a Ph.D. in Biology from the University of Miami and is currently the principal in Kavlock Consulting, LLC.

Dr. Raymond R. Tice has a M.S. in Biology from San Diego State University, and a Ph.D. in Biology from Johns Hopkins University. He joined the National Toxicology Program in 2005 as the first Deputy Director of the NTP Interagency Center for the Evaluation of Alternative Toxicological Methods and in 2008 became the first Branch Chief of the Division of the NTP's Biomolecular Screening Branch until his retirement in 2015. He played a key role in establishing the Tox21 program as an innovative way to advance the field of toxicology testing.

■ ACKNOWLEDGMENTS

The authors wish to acknowledge the following past and present Tox21 leads and working group team members for: (1) leadership of the Tox21 program - past (David Dix and Reeder Sams) and present (Warren Casey, Monica Linnenbrink, and Donna Mendrick); (2) the compound library (Katherine Coutros, present manager of the EPA's chemical contract, and William Leister, who oversaw analytical QC analysis of the library); (3) generation and analysis of screening results (NCATS: Menghang Xia, Matias Attene-Ramos, and Noel Southall; NTP: Nicole Kleinstreuer, Nisha Sipes, Keith Shockley, Kristine Witt, Fred Parham, Scott Auerbach, and Alex Merrick; EPA: Katie Paul-Friedman, Thomas Knudsen, and Michael DeVito). The authors acknowledge the invaluable contributions of EPA's chemical contractor (Evotec (US), Branford, CT) and NTP's chemical contractor (MRIGlobal, Kansas City, MO) in helping to procure and manage their respective chemical libraries. Finally, the authors thank Grace Patlewicz and Scott Auerbach for their expert review and constructive suggestions for improving the manuscript.

■ DEDICATION

The authors wish to dedicate this manuscript to Cynthia Smith (1950−2017), who led NTP's Tox21 compound library efforts through to her retirement in 2012, and Maritja (Marty) Wolf (1946−2014), who established high standards for DSSTox curation and sample registration of both EPA's ToxCast and Tox21 libraries.

■ ABBREVIATIONS

ADME, absorption, distribution, metabolism, excretion; API, active pharmaceutical ingredient; CAS RN, Chemical Abstracts Service Registry Number; CEBS, Chemical Effects in Biological Systems database; CoA, certificates of analysis; COSMOS+, combined overlap of EPA Dashboard COSMOSDB and EFSAOFT lists with TOX21SL list; CPDB, Carcinogenic Potency Database; CSRML, chemical subgraphs and reactions markup language; DMSO, dimethyl sulfoxide; DSSTox, Distributed Structure-Searchable Toxicity Data Network; EFSA, European Food Safety Authority; EPA, United States Environmental Protection Agency; EU, European Union; FDA, Food and Drug Administration; GCMS, gas chromatography mass spectroscopy; HPV, high-production volume; HTS, high-throughput screening; LCMS, liquid chromatography mass spectroscopy; log K_{ow}, log octanol/water partition coefficient; ME, molecular entity; MLI, NIH Molecular Libraries Initiative; MLSMR, NIH Molecular Libraries Small Molecule Repository; MolWt, molecular weight; NCATS, NIH National Center for Advancing Translational Sciences, NCGC, NIH Chemical Genomics Center; NMR, nuclear magnetic resonance; NIEHS, National Institute of Environmental Health Sciences; NIH, National Institutes of Health; NPC, NIH NCGC (later NCATS) Pharmaceuticals Collection; NTP, National Toxicology Program; OPPIN, EPA's Office of Pesticides Programs Information Network; qHTS, quantitative high-throughput screening; QC, quality control; QSAR, quantitative structure−activity relationship; SDF, structure-data file format; SMILES, simplified molecular-input line-entry system; TSCA, EPA's Toxic Substances Control Act

■ REFERENCES

(1) Eldridge, G. R., Vervoort, H. C., Lee, C. M., Cremin, P. A., Williams, C. T., Hart, S. M. S.M., Goering, M. G., O'Neil-Johnso, M., and Zeng, L. (2002) High-throughput method for the production and analysis of large natural product libraries for drug discovery. *Anal. Chem. 74* (16), 3963−3971.

(2) Geysen, H. M., Schoenen, F., Wagner, D., and Wagner, R. (2003) Combinatorial compound libraries for drug discovery: an ongoing challenge. *Nat. Rev. Drug Discovery 2* (3), 222−230.

(3) Austin, C. P., Brady, L. S., Insel, T. R., and Collins, F. S. (2004) NIH molecular libraries initiative. *Science 306*, 1138−1139.

(4) Bolton, E. E., Wang, Y., Thiessen, P. A., and Bryant, S. H. (2008) PubChem: integrated platform of small molecules and biological activities. In *Annual reports in computational chemistry*, Vol. *4*, pp 217−241, Elsevier, Amsterdam.

(5) Kim, S., Thiessen, P. A., Bolton, E. E., Chen, J., Fu, G., Gindulyte, A., Han, L., He, J., He, S., Shoemaker, B. A., Wang, J., Yu, B., Zhang, J., and Bryant, S. H. (2016) PubChem substance and compound databases. *Nucleic Acids Res. 44* (D1), D1202−D1213.

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

(7) Kavlock, R. J., Ankley, G. T., Blancato, J. N., Collette, T. W., Francis, E. Z., Gray, L. E., Jr., Hammerstrom, K., Swartout, J., Tilson, H. A., Toth, G. P., Veith, G. D., Weber, E. J., Wolf, D. C., and Young, D. M. (2003) *A Framework for a Computational Toxicology Research Program in ORD*, U.S. Environmental Protection Agency, Washington, DC, EPA 600/R-03/065 (NTIS PB2005-105438).

(8) National Toxicology Program Vision (2004) Toxicology in the 21st Century: The Role of the National Toxicology Program, https://ntp.niehs.nih.gov/ntp/main_pages/ntpvision.pdf (accessed 2020-09-14).

(9) Dix, D. J., Houck, K. A., Martin, M. T., Richard, A. M., Setzer, R. W., and Kavlock, R. J. (2007) The ToxCast program for prioritizing toxicity testing of environmental chemicals. *Toxicol. Sci. 95* (1), 5−12.

(10) Collins, F. S., Gray, G. M., and Bucher, J. R. (2008) Transforming environmental health protection. *Science (New York, NY, U.S.) 319* (5865), 906−907.

(11) Tice, R. R., Austin, C. P., Kavlock, R. J., and Bucher, J. R. (2013) Improving the human hazard characterization of chemicals: a Tox21 update. *Environ. Health Perspect. 121* (7), 756−765.

(12) *FDA's Predictive Toxicology Roadmap* (2017) U.S. Food and Drug Administration, Silver Spring, MD. https://www.fda.gov/media/109634/download (accessed 2020-09-14).

(13) Attene-Ramos, M. S., Miller, N., Huang, R., Michael, S., Itkin, M., Kavlock, R J., Austin, C. P., Shinn, P., Simeonov, A., Tice, R. R., and Xia, M. (2013) The Tox21 robotic platform for the assessment of environmental chemicals-from vision to reality. *Drug Discovery Today 18* (15−16), 716−723.

(14) Parham, F., Austin, C., Southall, N., Huang, R., Tice, R., and Portier, C. (2009) Dose-response modeling of high-throughput screening data. *J. Biomol. Screening 14* (10), 1216−1227.

(15) Hsieh, J. H., Sedykh, A., Huang, R., Xia, M., and Tice, R. R. (2015) A data analysis pipeline accounting for artifacts in Tox21 quantitative high-throughput screening assays. *J. Biomol. Screening 20* (7), 887−897.

(16) Filer, D. L., Kothiya, P., Setzer, R. W., Judson, R. S., and Martin, M. T. (2016) tcpl: the ToxCast pipeline for high-throughput screening data. *Bioinformatics 33* (4), 618−620.

(17) Attene-Ramos, M. S., Huang, R., Michael, S., Witt, K. L., Richard, A., Tice, R. R., Simeonov, A., Austin, C. P., and Xia, M. (2015) Profiling of the Tox21 chemical collection for mitochondrial function to identify compounds that acutely decrease mitochondrial membrane potential. *Environ. Health Perspect. 123* (1), 49−56.

(18) Chen, S., Hsieh, J. H., Huang, R., Sakamuru, S., Hsin, L. Y., Xia, M., Shockley, K. R., Auerbach, S., Kanaya, N., Lu, H., Svoboda, D., et al. (2015) Cell-based high-throughput screening for aromatase inhibitors in the Tox21 10K library. *Toxicol. Sci. 147* (2), 446−457.

(19) Huang, R., Sakamuru, S., Martin, M. T., Reif, D. M., Judson, R. S., Houck, K. A., Casey, W., Hsieh, J. H., Shockley, K. R., Ceger, P., Fostel, J., et al. (2015) Profiling of the Tox21 10K compound library for agonists and antagonists of the estrogen receptor alpha signaling pathway. *Sci. Rep. 4*, 5664.

(20) Abdelaziz, A., Spahn-Langguth, H., Schramm, K. W., and Tetko, I. V. (2016) Consensus modeling for HTS assays using in silico descriptors calculates the best balanced accuracy in Tox21 challenge. *Front. Environ. Sci. 4*, 2.

(21) Huang, R., Xia, M., Sakamuru, S., Zhao, J., Shahane, S. A., Attene-Ramos, M., Zhao, T., Austin, C. A., and Simeonov, A. (2016) Modelling the Tox21 10 K chemical profiles for in vivo toxicity prediction and mechanism characterization. *Nat. Commun. 7* (1), 1−10.

(22) Huang, R., Xia, M., Cho, M. H., Sakamuru, S., Shinn, P., Houck, K. A., Dix, D. J., Judson, R. S., Witt, K. L., Kavlock, R. J., Tice, R. R., et al. (2011) Chemical genomics profiling of environmental chemical modulation of human nuclear receptors. *Environ. Health Perspect. 119* (8), 1142−1148.

(23) Huang, R., Xia, M., Sakamuru, S., Zhao, J., Lynch, C., Zhao, T., Zhu, H., Austin, C. A., and Simeonov, A. (2018) Expanding biological space coverage enhances the prediction of drug adverse effects in human using in vitro activity profiles. *Sci. Rep. 8* (1), 1−12.

(24) Ngan, D. K., Ye, L., Wu, L., Xia, M., Rossoshek, A., Simeonov, A., and Huang, R. (2019) Bioactivity Signatures of Drugs vs. Environmental Chemicals Revealed by Tox21 High-Throughput Screening Assays. *Front. Big Data 2*, 50.

(25) Merrick, B. A., Paules, R. S., and Tice, R. R. (2015) Intersection of toxicogenomics and high throughput screening in the Tox21 program: an NIEHS perspective. *International journal of biotechnology 14* (1), 7.

(26) Thomas, R. S., Paules, R. S., Simeonov, A., Fitzpatrick, S. C., Crofton, K. M., Casey, W. M., and Mendrick, D. L. (2018) The US Federal Tox21 Program: A strategic and operational plan for continued leadership. *ALTEX-Alternatives to animal experimentation 35* (2), 163−168.

(27) Zhu, H., Zhang, J., Kim, M. T., Boison, A., Sedykh, A., and Moran, K. (2014) Big data in chemical toxicity research: the use of high-throughput screening assays to identify potential toxicants. *Chem. Res. Toxicol. 27* (10), 1643−1651.

(28) Fischer, F. C., Henneberger, L., König, M., Bittermann, K., Linden, L., Goss, K. U., and Escher, B. I. (2017) Modeling exposure in the Tox21 in vitro bioassays. *Chem. Res. Toxicol. 30* (5), 1197−1208.

(29) Tilley, S. K., Reif, D. M., and Fry, R. C. (2017) Incorporating ToxCast and Tox21 datasets to rank biological activity of chemicals at Superfund sites in North Carolina. *Environ. Int. 101*, 19−26.

(30) Newton, S. R., McMahen, R. L., Sobus, J. R., Mansouri, K., Williams, A. J., McEachran, A. D., and Strynar, M. J. (2018) Suspect screening and non-targeted analysis of drinking water using point-of-use filters. *Environ. Pollut. 234*, 297−306.

(31) Huang, R., and Xia, M. (2017) Editorial: Tox21 challenge to build predictive models of nuclear receptor and stress response pathways as mediated by exposure to environmental toxicants and drugs. *Front. Environ. Sci. 5* (3), 5.

(32) Xia, M., Huang, R., Witt, K. L., Southall, N., Fostel, J., Cho, M. H., Jadhav, A., Smith, C. S., Inglese, J., Portier, C. J., Tice, R. R., et al. (2008) Compound cytotoxicity profiling using quantitative high-throughput screening. *Environ. Health Perspect. 116* (3), 284−291.

(33) Huang, R., Southall, N., Wang, Y., Yasgar, A., Shinn, P., Jadhav, A., Nguyen, D. T., and Austin, C. P. (2011) The NCGC pharmaceutical collection: a comprehensive resource of clinically approved drugs enabling repurposing and chemical genomics. *Sci. Transl. Med. 3* (80), No. 80ps16.

(34) Richard, A. M., Judson, R. S., Houck, K. A., Grulke, C. M., Volarath, P., Thillainadarajah, I., et al. (2016) ToxCast chemical landscape: paving the road to 21st century toxicology. *Chem. Res. Toxicol. 29* (8), 1225−1251.

(35) Kavlock, R., Chandler, K., Houck, K., Hunter, S., Judson, R., Kleinstreuer, N., Knudsen, T., Martin, M., Padilla, S., Reif, D., Richard, A., et al. (2012) Update on EPA's ToxCast program: providing high throughput decision support tools for chemical risk management. *Chem. Res. Toxicol. 25* (7), 1287−1302.

(36) Gold, L. S., Slone, T. H., Manley, N. B., Garfinkel, G. B., Hudes, E. S., Rohrbach, L., and Ames, B. N. (1991) The Carcinogenic Potency Database: analyses of 4000 chronic animal cancer experiments published in the general literature and by the US National Cancer Institute/National Toxicology Program. *Environ. Health Perspect. 96*, 11−15.

(37) McKee, R. H., Butala, J. H., David, R. M., and Gans, G. (2004) NTP center for the evaluation of risks to human reproduction reports on phthalates: addressing the data gaps. *Reprod. Toxicol. 18* (1), 1−22.

(38) Judson, R., Richard, A., Dix, D., Houck, K., Elloumi, F., Martin, M., Cathey, T., Transue, T., Spencer, R., and Wolf, M. (2008) ACToR - Aggregated Computational Toxicology Resource. *Toxicol. Appl. Pharmacol. 233* (1), 7−13.

(39) Richard, A. M. (2004) DSSTox web site launch: Improving public access to databases for building structure-toxicity prediction models. *Preclinica 2* (2), 103−108.

(40) Grulke, C. M., Williams, A. J., Thillanadarajah, I., and Richard, A. M. (2019) EPA's DSSTox database: History of development of a curated chemistry resource supporting computational toxicology research. *Computational Toxicology 12*, 100096.

(41) MacArthur, R., Leister, W., Veith, H., Shinn, P., Southall, N., Austin, C. P., Inglese, J., and Auld, D. S. (2009) Monitoring compound integrity with cytochrome P450 assays and qHTS. *J. Biomol. Screening 14* (5), 538−546.

(42) Shockley, K. R. (2012) A three-stage algorithm to make toxicologically relevant activity calls from quantitative high throughput screening data. *Environ. Health Perspect. 120* (8), 1107−1115.

(43) Shockley, K. R. (2014) Using weighted entropy to rank chemicals in quantitative high throughput screening experiments. *J. Biomol. Screening 19*, 344−353.

(44) Lea, I. A., Gong, H., Paleja, A., Rashid, A., and Fostel, J. (2017) CEBS: a comprehensive annotated database of toxicological data. *Nucleic Acids Res. 45* (D1), D964−D971.

(45) 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., and Richard, A. M. (2017) The CompTox Chemistry Dashboard: a community data resource for environmental chemistry. *J. Cheminf. 9* (1), 61.

(46) Wishart, D. S., Knox, C., Guo, A. C., Cheng, D., Shrivastava, S., Tzur, D., Gautam, B., and Hassanali, M. (2008) DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res. 36*, D901−D906.

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

(48) Hendrickson, J. B., Huang, P., and Toczko, A. G. (1987) Molecular Complexity: A Simplified Formula Adapted to Individual Atoms. *J. Chem. Inf. Model. 27*, 63−67.

(49) (2016) *User's Guide for T.E.S.T. (version 4.2) (Toxicity Estimation Software Tool): A Program to Estimate Toxicity from Molecular Structure*, EPA, Washington, DC. https://www.epa.gov/

sites/production/files/2016-05/documents/600r16058.pdf (accessed 2020-09-14).

(50) Hansen, K., Mika, S., Schroeter, T., Sutter, A., Ter Laak, A., Steger-Hartmann, T., Heinrich, N., and Müller, K. R. (2009) Benchmark data set for in silico prediction of Ames mutagenicity. *J. Chem. Inf. Model. 49* (9), 2077−2081.

(51) 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* (S1), S4.

(52) Arena, V. C., Sussman, N. B., Mazumdar, S., Yu, S., and Macina, O. T. (2004) The utility of structure−activity relationship (SAR) models for prediction and covariate selection in developmental toxicity: comparative analysis of logistic regression and decision tree models. *SAR QSAR Environ. Res. 15* (1), 1−18.

(53) Yang, C., Tarkhov, A., Marusczyk, J., Bienfait, B., Gasteiger, J., Kleinoeder, T., Magdziarz, T., Sacher, O., Schwab, C. H., Schwoebel, J., Terfloth, L., et al. (2015) New publicly available chemical query language, CSRML, to support chemotype representations for application to data mining and modeling. *J. Chem. Inf. Model. 55* (3), 510−528.

(54) Truong, L., Reif, D. M., St Mary, L., Geier, M. C., Truong, H. D., and Tanguay, R. L. (2014) Multidimensional in vivo hazard assessment using zebrafish. *Toxicol. Sci. 137* (1), 212−233.

(55) 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.

(56) Wang, J., Hallinger, D. R., Murr, A. S., Buckalew, A. R., Lougee, R. R., Richard, A. M., Laws, S. C., and Stoker, T. E. (2019) High-throughput screening and chemotype-enrichment analysis of ToxCast phase II chemicals evaluated for human sodium-iodide symporter (NIS) inhibition. *Environ. Int. 126*, 377−386.

(57) Nelms, M. D., Pradeep, P., and Patlewicz, G. (2019) Evaluating potential refinements to existing Threshold of Toxicological Concern (TTC) values for environmentally-relevant compounds. *Regul. Toxicol. Pharmacol. 109*, 104505.

(58) Mansouri, K., Abdelaziz, A., Rybacka, A., Roncaglioni, A., Tropsha, A., Varnek, A., Zakharov, A., Worth, A., Richard, A. M., Grulke, C. M., Trisciuzzi, D., et al. (2016) CERAPP: collaborative estrogen receptor activity prediction project. *Environ. Health Perspect. 124* (7), 1023−1033.