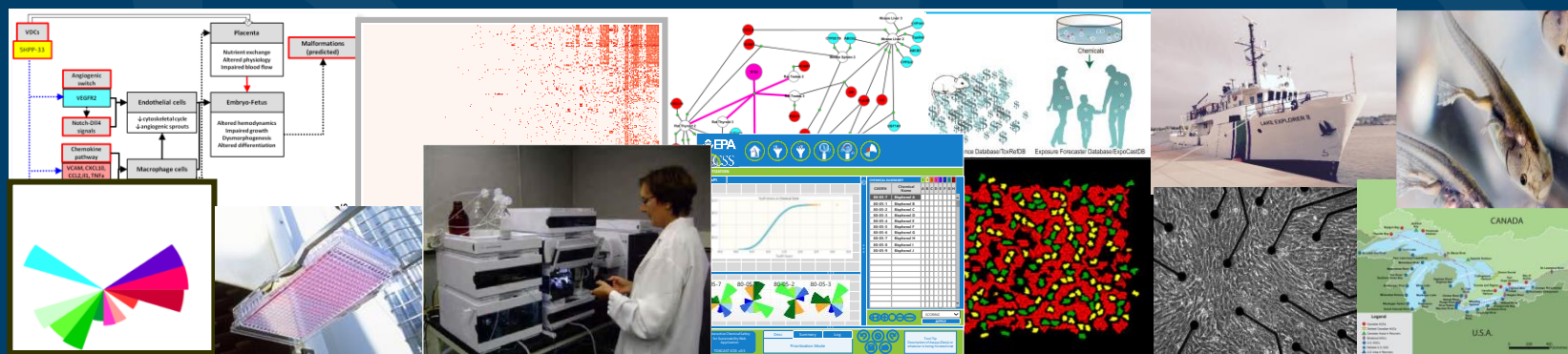


Overview of Data Science Activities & Progress within the Center for Computational Toxicology and Exposure

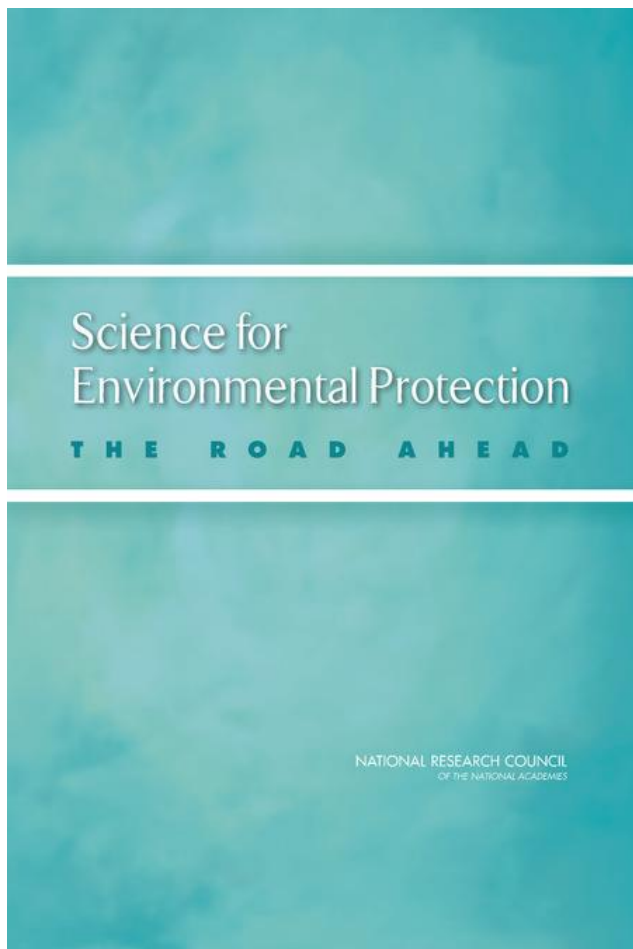


National Academies of Science: Data Science Session for *Anticipatory Research for EPA's Research and Development Enterprise to Inform Future Environmental Protection: The Road Ahead*
July 29, 2021

Reeder Sams
Deputy Director
Center for Computational Toxicology and Exposure

The views expressed in this presentation are those of the presenter and do not necessarily reflect the views or policies of the U.S. EPA

Software Tools, IT and IM are Integral to the Future of Environmental Protection



NAS Report 2012

Computer Science, Informatics, and Information Technology

The future needs for IT and informatics in support of science in EPA are subject to two principal influences: the future directions of EPA's mission and the underlying science in future directions taken by the IT industry. Science in EPA will increasingly depend on its capability in IT and informatics. IT is concerned with the acquisition, processing, storage, and dissemination of information with a combination of computing and telecommunication (Longley and Shain 1985). The term *informatics*, as used here, refers to the application of IT in the generation, repository, retrieval, processing, integration, analysis, and interpretation of data obtained in different media and across geographic and disciplinary boundaries that are related to the environment and ecosystem, community and human activities, and human health (see He 2003). Informatics is also concerned with the computational, cognitive, and social aspects of IT. One way in which IT can be used for data acquisition is through public engagement. Taking advantage of expertise outside of EPA (from academia, industry, and other agencies) and considering the general public as a source of new information is a way in which knowledge and resources can be combined in a cost-effective manner. Examples include taking advantage of social media and crowdsourcing. Appendix D provides additional background information on various important and rapidly changing tools and technologies in the field of information technology and informatics.

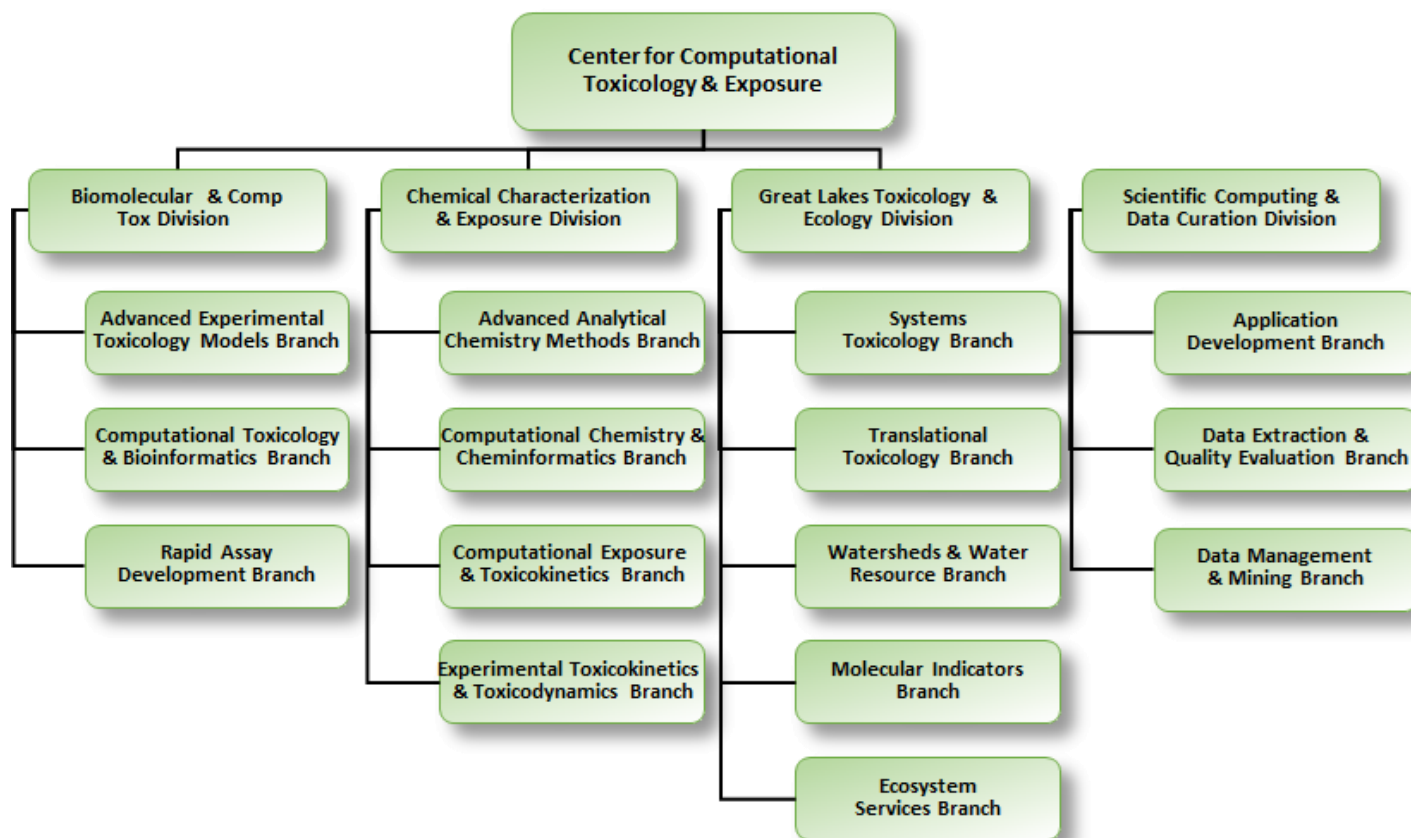
Excerpt from Report

Includes...

- 1) Data management
- 2) Data mining
- 3) Software and decision support tools
- 4) Curation and integration of information from other sources

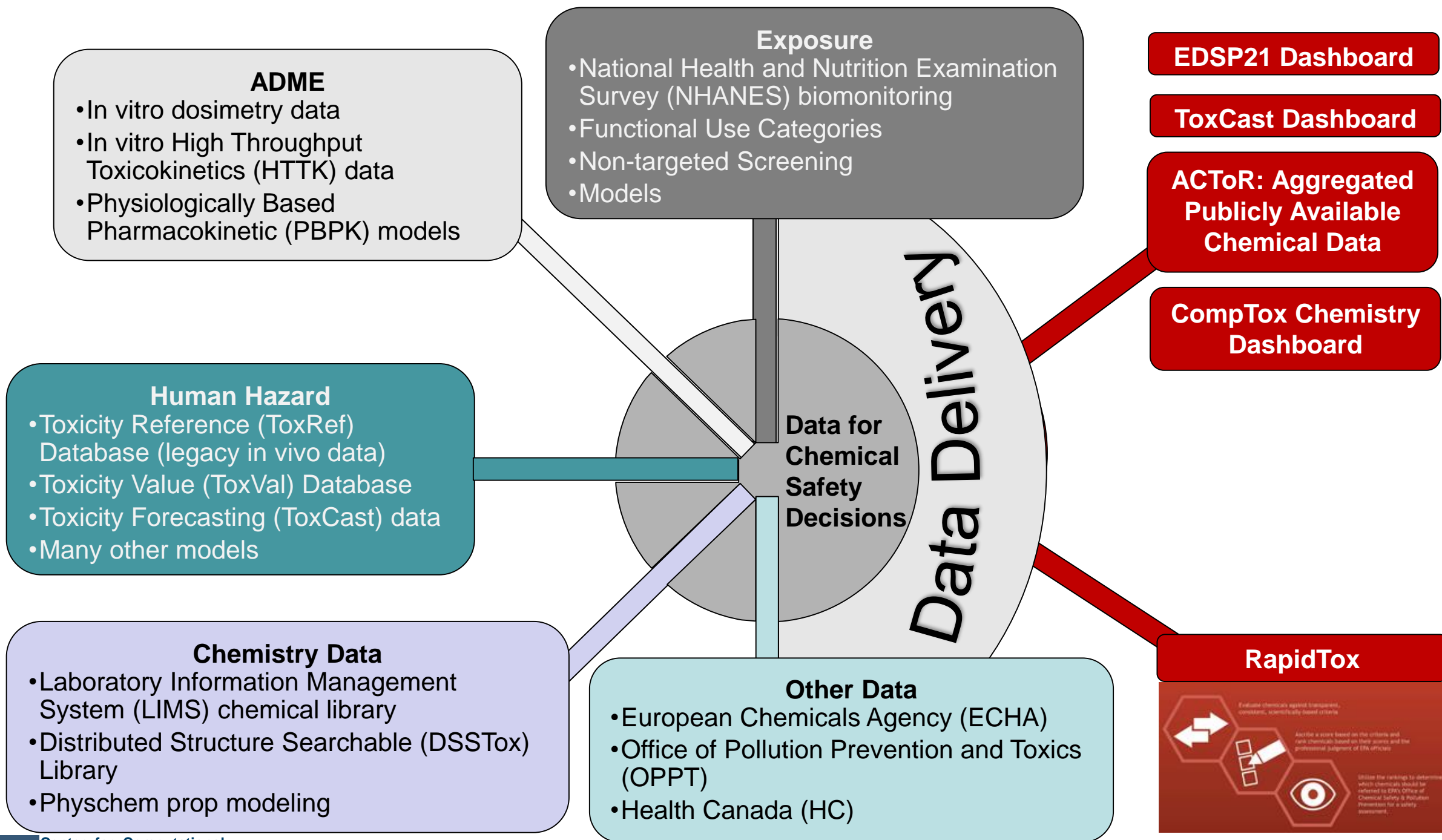
Reorganization Provided an Opportunity to Formally Integrate Scientific Tool Development and Data Management Capability (Informatics)

Organization chart for ORD's Center for Computational Toxicology & Exposure



- Sustainable support for development and maintenance of enterprise-level software tools that can inform chemical safety and ecological decision-making
- Centralized capability to collate and curate available chemistry, toxicity, exposure and ecological information from structured and unstructured sources to provide a comprehensive chemical safety knowledgebase.
- Integrated data management systems capable of efficiently storing and delivering a diverse range of chemistry, toxicity, exposure, and ecological data for internal and external needs.

Data Management and Curation

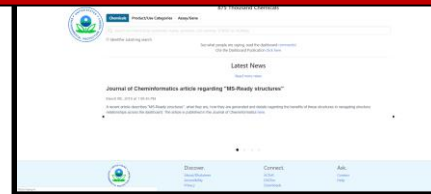


ORD/CCTE Data Management Overview

RapidTox



Dashboards, Scientific Tools & downloadable data



Production Application Layer

Virtual or Physical Application Databases
– Data Structured to efficiently support our external production applications

Data Mart / Presentation Layer (Fit-for-Purpose)

ORD Data Hub

Data
Analytics/
Learning

Generalized data model
providing integration
across the domains for
scientific exploration

Model +
Pred App

Raw data processing
and data curation,
quality tools

Chemistry

Chemistry

Human
Health
Hazard

Human
Health
Hazard

Exposure

ChemTrack

DSSTox

ChemProp

InVitroDB

ToxRef DB

CPDat

Models

Data Collection Layer (Transactional)

Data Curation Approach and Activities

Curation Process



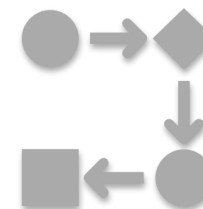
Catalog and
prioritize
relevant external
data sources



Design process
for extracting
and cleaning
data from each
source



Identify tools to
facilitate data
curation



Review data
curation process
regularly



Develop data
management
and quality
assurance plans

Current Data Focus

Chemistry

- Chemical Library
- DSSTox
- Phys/Chemical Prop

Exposure

- Biomonitoring
- CPDat
- Factotum

Human Health Hazard

- ToxVal
- CHD
- ToxRef

ADME

- CvT
- HTTK

Ecological Hazard

- ECOTOX

Ecological Integrity

Machine Learning

Example CCTE Projects Implementing Machine Learning

- Hazard

- Developmental Tox machine learning model-predictive model using Stemina and ToxCast Assays

(Zurlinden et al., 2020, <https://pubmed.ncbi.nlm.nih.gov/32073639/>)

- Toxicokinetics

- *In vitro* data can be used to predict toxicokinetics data through R package of HHTK

(Pierce et al., 2017, <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6134854/>)

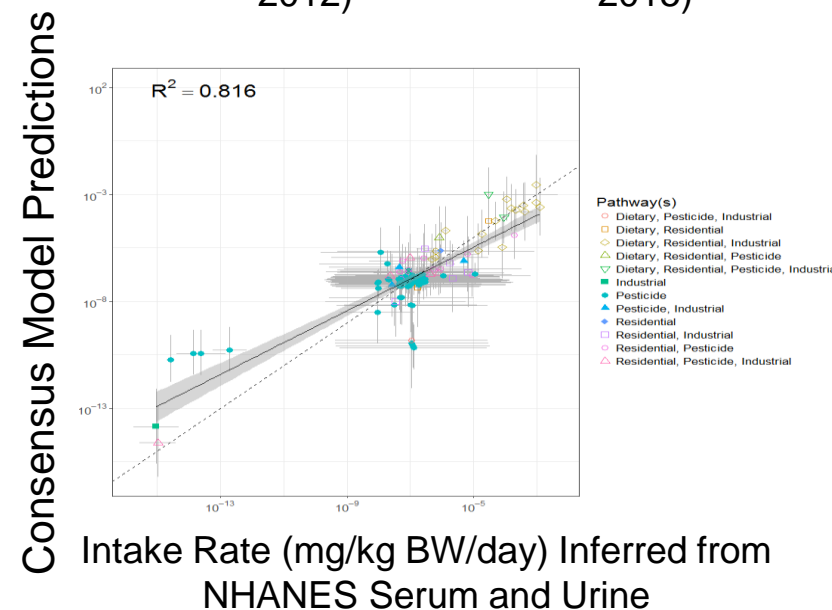
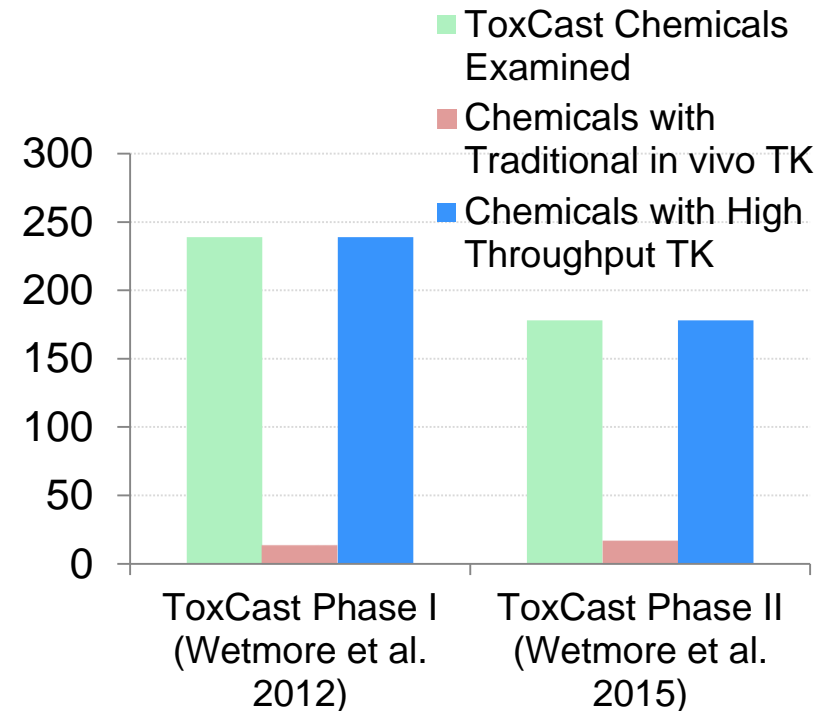
- Access through ORD's CompTox Chemicals Dashboard

(https://comptox.epa.gov/dashboard/chemical_lists/HTTKHUMAN/)

- Exposure

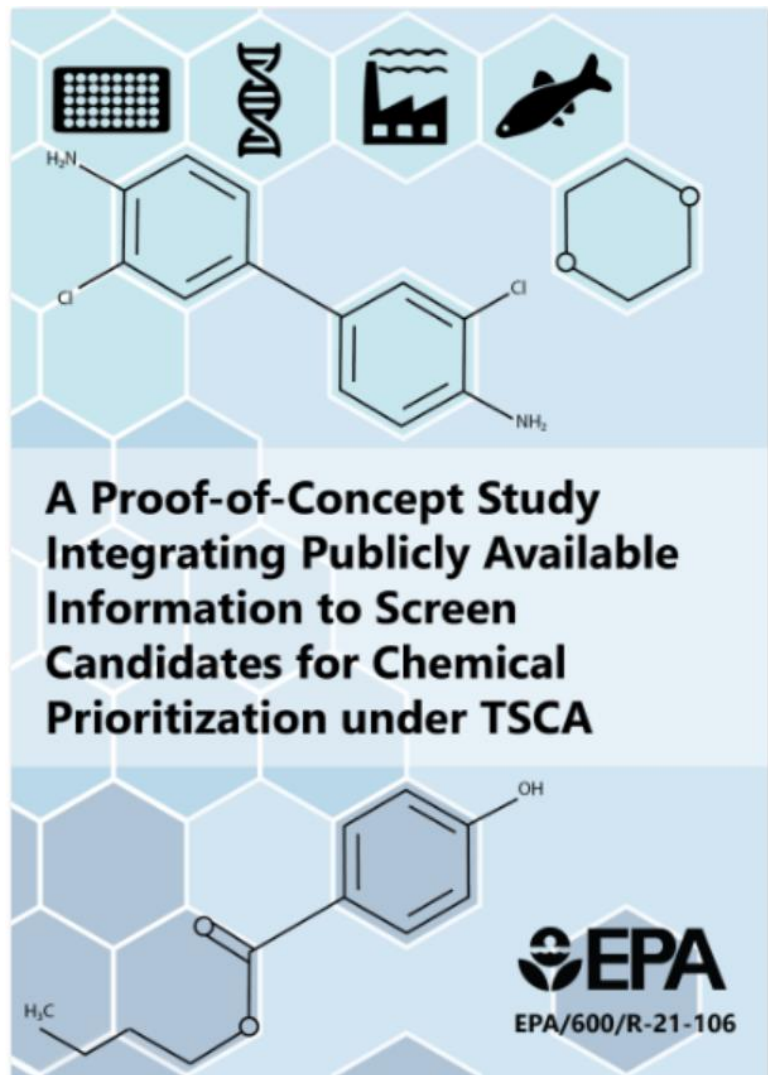
- Systematic Empirical Evaluation of Models (SEEM) Framework- used to predict exposure estimates for almost 500k chemicals

(Ring et al., 2019, <https://pubmed.ncbi.nlm.nih.gov/30516957/>)



Decision Support Tools

Public Information Curation and Synthesis (PICS) Approach



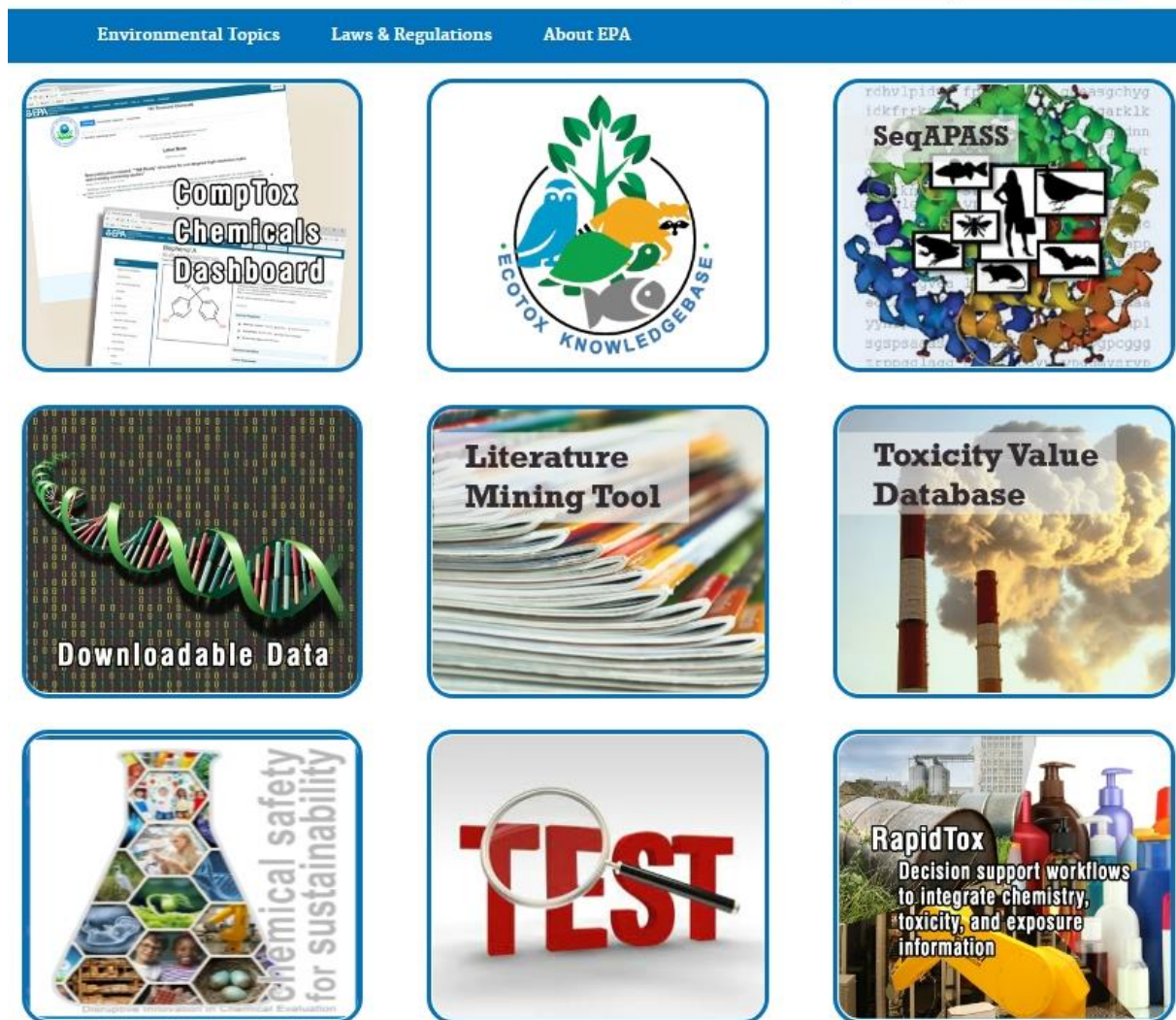
- GOAL: Develop a transparent and reproducible process for integrating available information, identifying potential information data gaps and understanding the landscape of publicly-available information for large chemical databases.
- Delivered: A proof-of-concept case study demonstrating an automated approach that extracts, stores, and integrates publicly available information from traditional and new approach methods in toxicology, exposure, and environmental fate-related studies.
 - Case study developed to reflect TSCA decision context and used a small subset of the TSCA active inventory.
 - PICS Approach includes a scientific domain metric and an information availability metric based on decision context specific analysis of information from seven scientific domains.
 - Results are presented in report and available on TSCA POC Dashboard at the link below.

<https://www.epa.gov/chemical-research/translation-and-knowledge-delivery>

Tool access through CCTE Portal

External Portal

(<https://comptox.epa.gov/index.html#/>)



Internal Portal



ORD's Comptox Chemicals Dashboard:

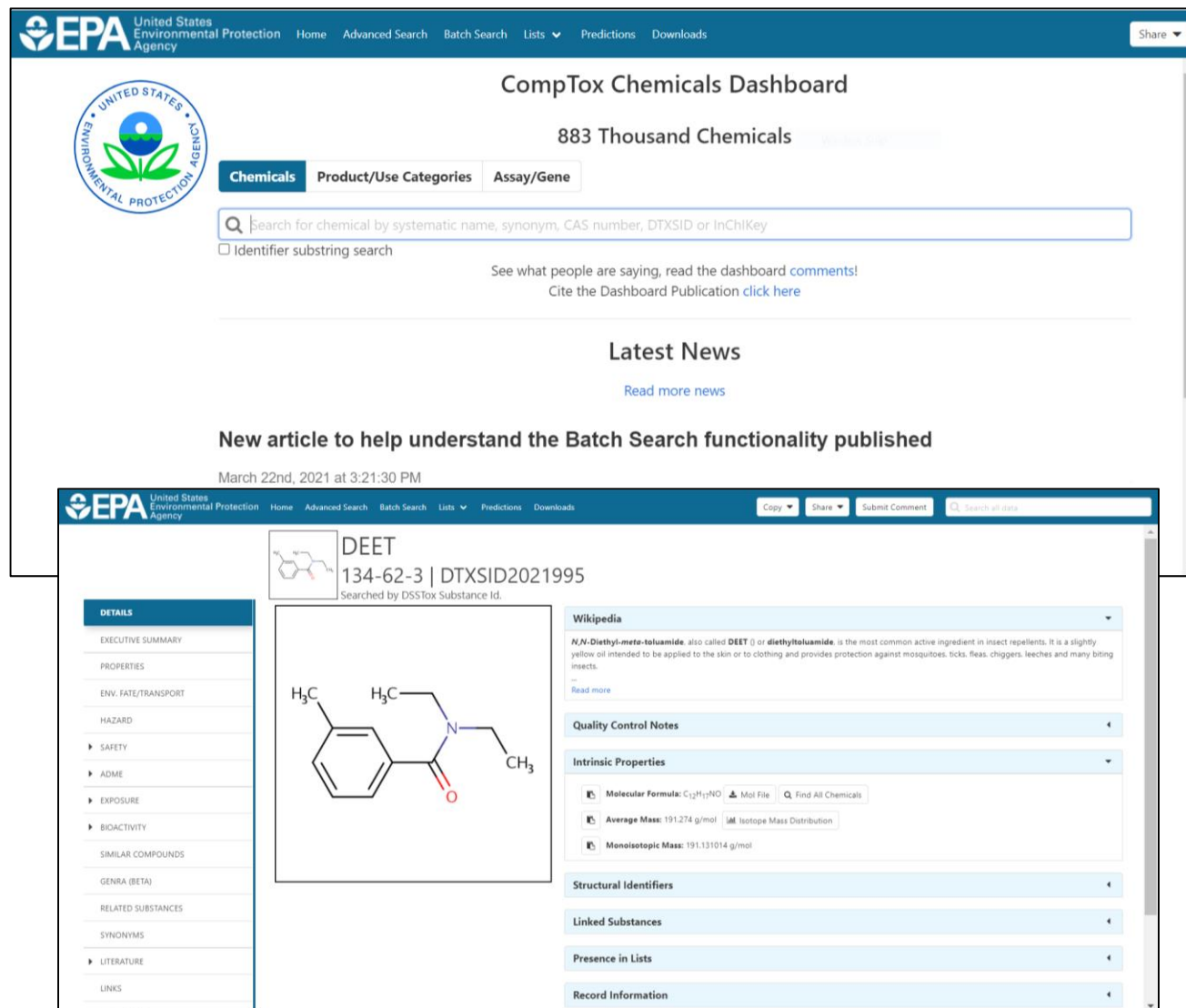
<https://comptox.epa.gov/dashboard>
(New Version Late 2020)

Data

- Soon with >900,000 chemicals
- New version of ToxVal v9 and invitrodb 3.4

User Experience

- Replacement of tables in the application with more flexible table handling for data
- The Abstract Sifter and GenRA will be presented as separated tools in a modular way that have been plugged into the new Tools menu of the dashboard.
- Performance will be much enhanced in terms of download speed and the batch search will be able to handle much larger input sets.



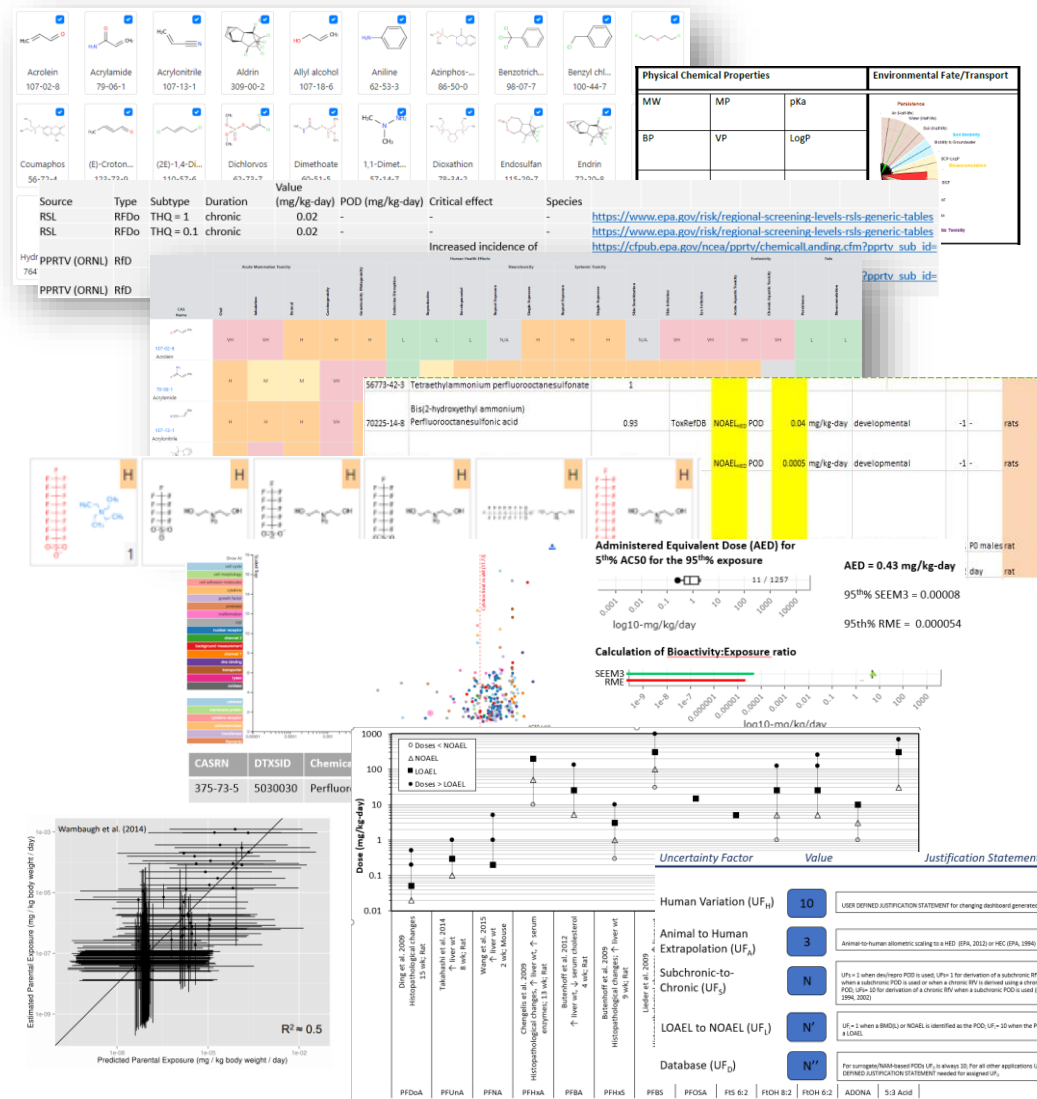
The screenshot displays the EPA Comptox Chemicals Dashboard. The top navigation bar includes the EPA logo, "United States Environmental Protection Agency", and links for Home, Advanced Search, Batch Search, Lists, Predictions, and Downloads. The main header reads "CompTox Chemicals Dashboard" and "883 Thousand Chemicals". Below this, there are tabs for "Chemicals", "Product/Use Categories", and "Assay/Gene". A search bar prompts users to "Search for chemical by systematic name, synonym, CAS number, DTXSID or InChIKey". A checkbox for "Identifier substring search" is present. Links to "See what people are saying, read the dashboard comments!" and "Cite the Dashboard Publication [click here](#)" are provided. The "Latest News" section features a headline: "New article to help understand the Batch Search functionality published" dated "March 22nd, 2021 at 3:21:30 PM". The bottom section shows a detailed view of a chemical, DEET (134-62-3 | DTXSID2021995). It includes a chemical structure diagram, a sidebar with a "DETAILS" menu (Executive Summary, Properties, Env. Fate/Transport, Hazard, Safety, ADME, Exposure, Bioactivity, Similar Compounds, GenRA (Beta), Related Substances, Synonyms, Literature, Links), and a main content area with Wikipedia information, Quality Control Notes, Intrinsic Properties (Molecular Formula: C₁₂H₁₇NO, Average Mass: 191.274 g/mol, Monoisotopic Mass: 191.131014 g/mol), Structural Identifiers, Linked Substances, Presence in Lists, and Record Information.

RapidTox Tool (in development)

- Data management challenges in chemical risk assessment necessitate the transition towards increased use of decision support tools
- RapidTox is a collection of decision-based workflows that integrate a range of information related to chemical properties, fate and transport, hazard, dose-response, and exposure
- Delivers information that facilitates development of quantitative screening-level assessments (with associated uncertainty) for hundreds to thousands of chemicals
- Benefits potentially include increased transparency, more readily updated datasets/models/assessments, expedited management review, harmonized data within and across regulatory agencies, and increased efficiency

Chemicals

Search for chemical by systematic name, synonym, CAS n



Workflow for Prioritizing Chemicals of Emerging Concern (CEC)-CRADA with MN DOH

Data Curation

MN-specific documents and other source documents extracted and curated into ORD's research databases via the Factotum curation application.

QA, document provenance, audit tracking

ORD's "Factotum" Curation Application

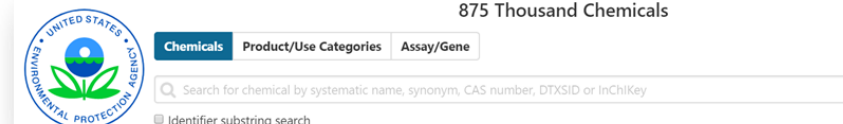
Chemicals
and
Products
Database
CPDat

Multimedia
Monitoring
Database
MMDB

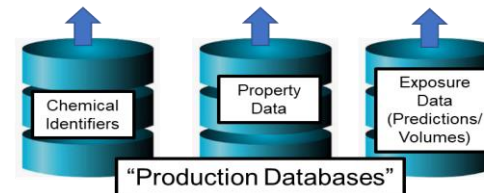
ORD "Research" Databases



Other public data streams, e.g.
USGS webservices or datasets
not yet incorporated into formal
ORD databases



CompTox Chemicals Dashboard



Data Hub

"Workflow-Specific Data Mart"



Main Scoring Criteria

Persistence and Fate
Release Potential
Occurrence

Unadjusted
Score

+

Scoring Adjustments (+/-)

Chemical Identity*
Exposure Potential
Detection Frequency

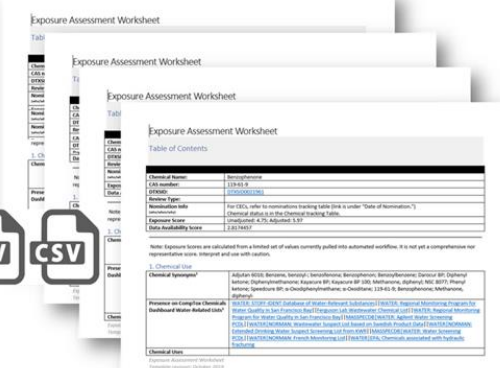
Score
Adjustments

=
Final Score

- Data retrieval and caching
- Chemical scoring
- Summary report and data table generation



Automated Reporting and Data Generation for In-Depth Assessment



Take Home Messages...

- ORD has made a sustained investment into the management of large datasets, data curation, machine learning, and decision support tools
- Dedicating an entire division within CCTE represents a commitment to informatics that will facilitate sustainable data management and tool development to meet the needs of our partners and clients
- ORD has been responsive to advice provided by the NAS on informatics and looks forward to the completion of the current effort.

Acknowledgements

EPA Colleagues:

CEMM
CPHEA
CESER
OCSP
OW
OLEM

Tox21 Colleagues:

NTP
FDA
NCATS

Collaborative Partners:

Unilever
A*STAR
ECHA
EFSA
Health Canada

Center for Computational Toxicology and Exposure (CCTE) Staff



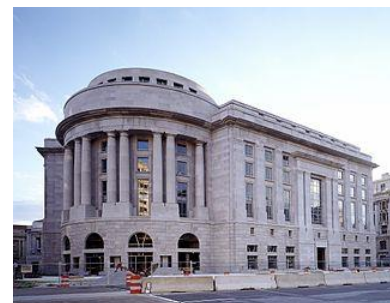
Research Triangle Park, NC



Cincinnati, OH



Duluth, MN



Washington, DC



Athens, GA



Gulf Breeze, FL