National Institute of Environmental Health Sciences
Division of Translational Toxicology

Harmonizing Tox Data to Enhance Confidence and Utility

Aswani Unnikrishnan[1], Alexandre Borrel[1*], Kimberly To[1*], Victoria Hull[1], Amber Daniel[1], Emily Reinke[1], Nicole Kleinstreuer[2]

[1]Inotiv, Research Triangle Park, NC; [2]NIH/NIEHS/DTT/NICEATM, Research Triangle Park, NC

Background

- Toxicologically relevant data encompass information essential for assessing the safety and risk of chemical substances to human health and the environment.
- Examples of such data include:
  - Dose-response data to assess the relationship between an exposure to a chemical and its related effects.
  - Absorption, distribution, metabolism, and excretion (ADME) data to understand toxicokinetics and tissue-specific chemical behaviors.
  - Acute toxicity study results measuring the short-term effects of a substance.
- Data harmonization is important for chemical risk characterization and the development and validation of new approach methodologies. Without data harmonization, inconsistencies and discrepancies can lead to conflicting conclusions and complicate the development of effective safety measures.
- Here we present a comprehensive approach to clean and aggregate toxicologically relevant data from disparate sources for inclusion in the NTP Interagency Center for the Evaluation of Alternative Toxicological Methods (NICEATM) Integrated Chemical Environment (ICE: https://ice.ntp.niehs.nih.gov/).

Integrated Chemical Environment

Data Collection and Pre-Processing

- The data in ICE are gathered from multiple sources. Automated processes and expert-driven methods are employed to harmonize, standardize, and format the data to adhere to FAIR (findability, accessibility, interoperability, and reusability) principles.
- ICE curation efforts include expert-driven manual curation and harmonization (Daniel et al., 2022). To support the growing number of data sets being added to ICE, a computational approach was developed to automate data pre-processing, including harmonization of chemical identifiers and error corrections.

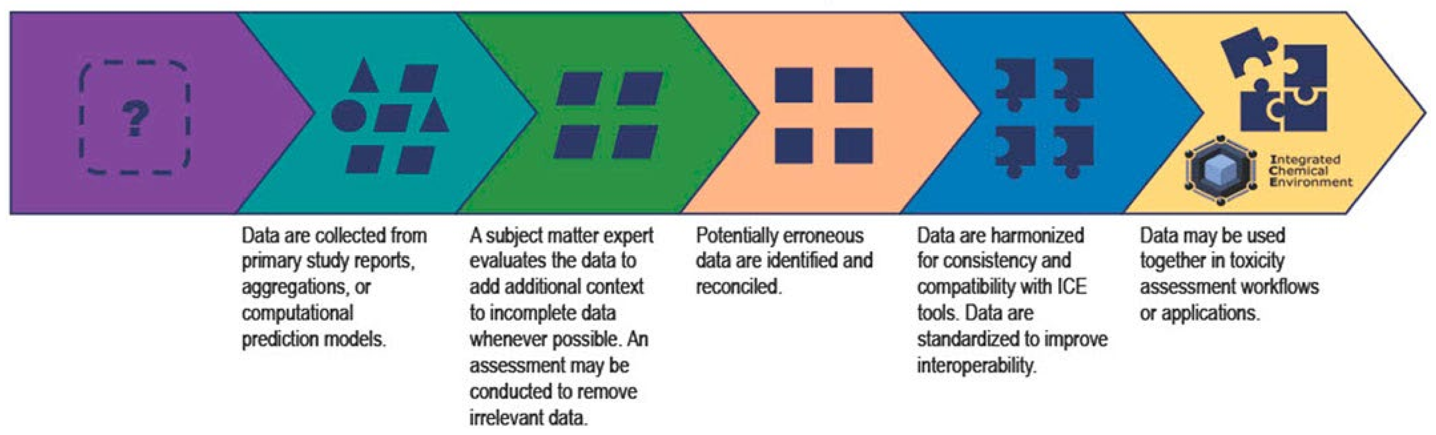

Figure 1:The ICE workflow begins with collection of raw data from various sources and formats and applies expert assessment and curation to produce pre-processed data sets. Diagram from Daniel et al., 2022.

Acute Systemic Toxicity Endpoint

Name	No: of Chemicals	Data Type	Source
Dermal	275	In Vivo	Submissions to EPA for pesticide registration
Inhalation	1781	In Vivo	Submissions to EPA for pesticide registration
			EPA Acute Exposure Guideline Levels for Airborne Chemicals
			ECHA Registration, Evaluation, Authorisation, and Restriction of Chemicals
			DoD study reports
			ChemIDplus
			NIOSH Pocket Guide to Chemical Hazards
Oral	9110	In Vivo	Submissions to EPA for pesticide registration
			ICCVAM validation reports
		In Vitro	EPA ToxCast and Tox21 (invitrodb v3.5)
		In Silico	CATMoS predictions

Abbreviations: CATMoS: Collaborative Acute Toxicity Modeling Suite; DoD: U.S. Department of Defense; ECHA: European Chemicals Agency; EPA: U.S. Environmental Protection Agency; ICCVAM: U.S. Interagency Coordinating Committee on the Validation of Alternative Methods; NIOSH: National Institute for Occupational Safety and Health.

Figure 2: Example of diverse sources of acute toxicity data in ICE.

Data	Flag: Curve Fit	Flag: Chemical QC	Flag: Technological Interference	Annotate
Retrieve data from the U.S. EPA's invitrodb v3.5 (Feshuk 2022).	Review curve fits. Integrate flags and custom rules for flagging less robust fits.	Integrate chemical QC information released via the Tox21 Tripod site.	Review assay technology and chemical structures to identify potential false signals.	Leverage EPA's technological assay annotations and create annotations for biological interpretation.

Figure 3: Represents the ICE cHTS data curation pipeline. It integrates HTS concentration-response curve fit information, chemical QC data, and technological interference flags to increase confidence in hit calls and also provides annotations for biological interpretation of the data.

Data Quality Control (QC) Pipeline

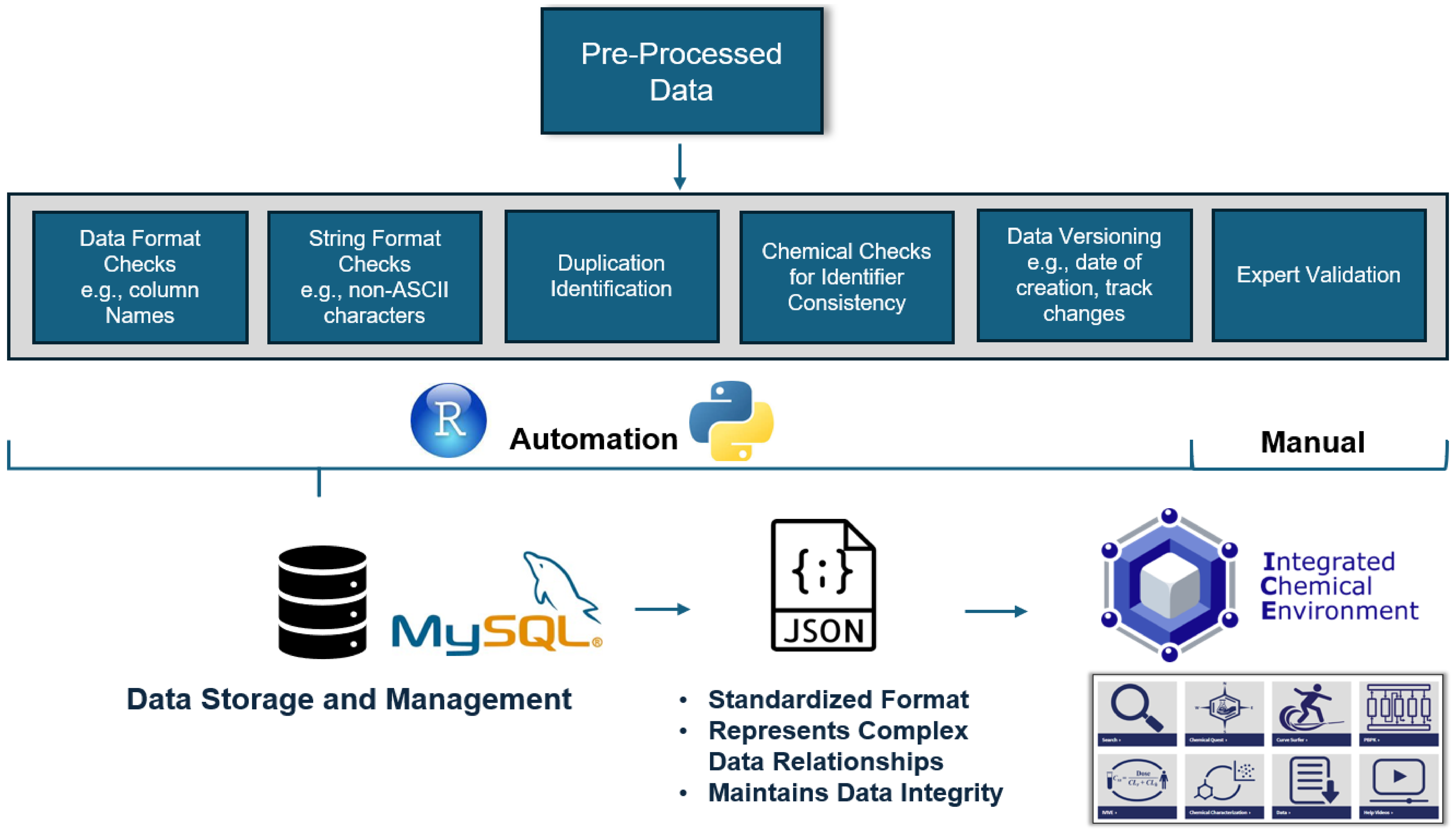

Figure 4: Represents ICE Data QC pipeline where pre-processed data from different toxicologically relevant data sets are subjected to a series of automated processing scripts using R and Python programming languages and subject matter expert validation.

The processed data are stored in a MySQL database, supporting versioning and allowing for efficient storage, retrieval, and management through SQL queries.

The pipeline also includes guidance documents such as SOPs for expert curation, ensuring that every step of the QC process is carried out consistently, minimizing errors and discrepancies.

The data is retrieved in JSON format, which includes substance and assay endpoint information. This data is integrated into ICE and used within its exploratory computational tools for data interpretation and analysis.

Data Representation and Availability

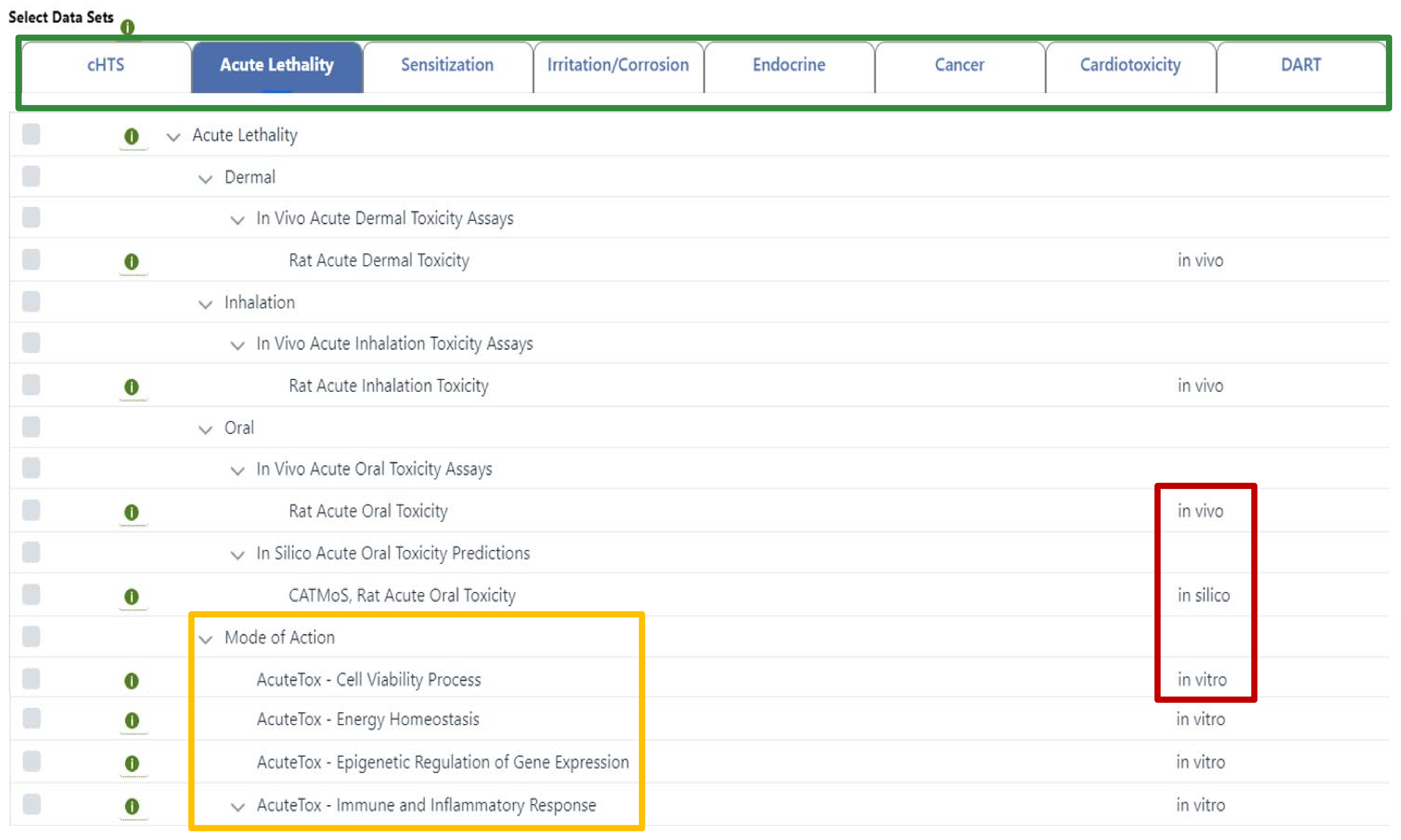

Figure 5: Data in the ICE Search tool is organized according to data type and toxicological endpoints of regulatory interest (green highlight at top). Consolidating in vivo, in vitro, and in silico data for acute oral toxicity endpoints in one location facilitates comprehensive assessment and effective cross-validation (red highlight). Mode of action annotations focusing on endpoints related to toxicological pathways providing increase in interoperability with other databases and harmonized reporting templates (yellow highlight).

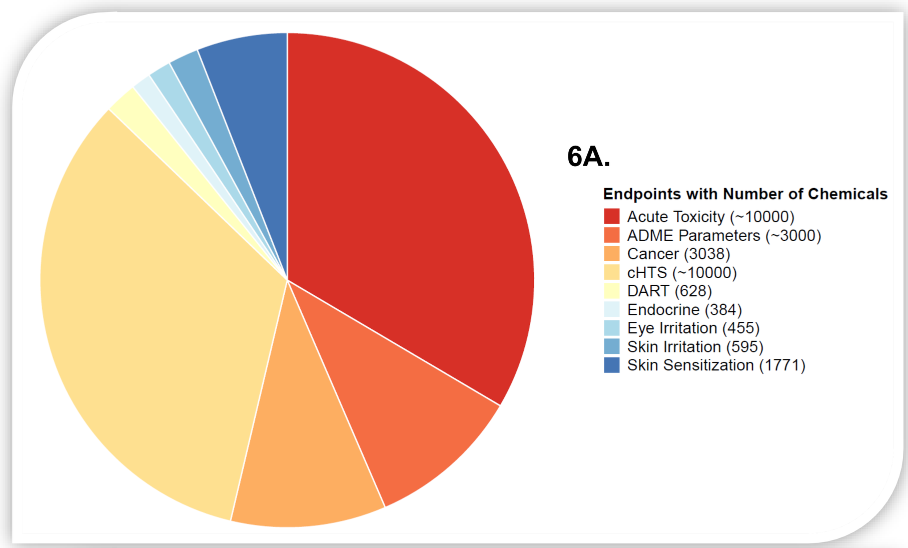

Figure 6A: Number of chemicals in data sets with in vivo or in vitro endpoints.

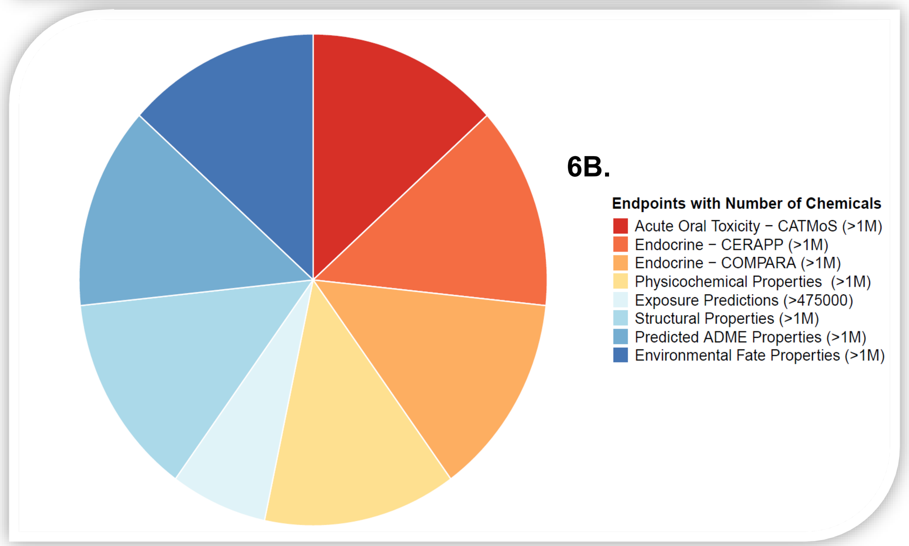

Figure 6B: Number of chemicals in data sets with in silico endpoints. These data sets can provide estimates for data-poor chemicals and help fill in data gaps. The data can be explored through ICE tools, directly downloaded from the site, or retrieved via the ICE REST API.

Conclusion

- The ICE infrastructure allows integration of disparate data from different sources, based on toxicity endpoints of regulatory interest, to yield a more comprehensive understanding of potential hazards associated with chemicals.
- Challenges such as inconsistencies in data formatting and terminology complicate combining information from various sources.
- This comprehensive approach to clean and consolidate toxicologically relevant data from disparate sources incorporates subject matter expertise with computational techniques like programmatic databases and processing scripts to streamline data harmonization.
- The processing pipeline was applied across millions of data points via a series of quality control steps to identify potential errors or inconsistencies (e.g., duplicate entries, missing values) and appropriately rectify them. Pipelined data were organized using standardized terminology to facilitate comparisons across datasets, derive meaningful insights, and uphold FAIR principles.
- This approach highlights data transparency and curation for ensuring reliability and integrity through metadata annotations of data provenance, assumptions, and quality assurance practices.
- The iterative nature of data cleanup and aggregation processes emphasizes the need for ongoing collaborations across stakeholder groups to continually refine, curate, and validate data.

References

- Daniel et al. 2022. Data curation to support toxicity assessments using the Integrated Chemical Environment. Front Toxicol. 4:987848. doi:10.3389/ftox.2022.987848
- Feshuk et al. 2022. Invitrodb version 3.5 release. EPA, Washington, DC. doi:10.23645/epacomptox.6062623.v8

Acknowledgments and More Information

- Learn more about ICE at ASCCT 2024 Poster No. 47, Reisfeld et al., and ICE annotations at ASCCT 2024 Poster No. 40, Hill et al.
- This project was funded with federal funds from NIEHS, NIH under Contract No. HHSN273201500010C. The views expressed above do not necessarily represent the official positions of any federal agency.
- We thank Catherine Sprankle and Elizabeth Farley-Dawson, Inotiv, for editorial input.
- Kimberly To is currently affiliated with ICF, Reston, VA. Alexandre Borrel is currently affiliated with Sciome, Research Triangle Park, NC.
- To subscribe to the NICEATM News email list visit: https://list.nih.gov/cgi-bin/wa.exe?SUBED1=niceatm-l&A=1.

Subscribe to NICEATM News