ToxCast: Chemical List(s), Chemical Space, and Chemotype Information

Ann Richard
U.S. EPA, National Center for Computational Toxicology
Office of Research and Development

The views expressed in this presentation are those of the author and do not necessarily reflect the views or policies of the U.S. EPA
ToxCast Phase II Data Release
http://www.epa.gov/ncct/toxcast/

- ToxCast Assay Summary Activity Files
- ToxCast Assay Annotation Files
- ToxCast Chemical Library & Structure Files (DSSTox)
- ToxCast Concentration Response Data Files
- ToxRefDB Effect & Endpoint Data Files
ToxCast Data: Chemical Files

http://www.epa.gov/ncct/toxcast/data.html

Chemical IDs
Structures
Chemotypes

ToxCast Data

All available ToxCast Phase II data (including assay summary activity files, assay description files, effect and endpoint data files from animal toxicity studies, concentration response data files & chemical library and structure files) can be downloaded below.

ToxCast Phase II Data & Other Files

- Chemical list & Annotation Files (DSS Tox Information)
  - Contains the most current, official public listing of all of the ToxCast 1,800 and Tox21 8,307 chemicals in both an Excel and SD file format. It also contains files used to analyze chemical structure (fingerprinting & SMILES). More information is provided in the README file available below and at: http://www.epa.gov/ncct/dsstox

- High-Throughput Chemical Screening Data from ToxCast & Tox21
  - Contains high-throughput in vitro data on 1,800 chemicals from more than 800 high-throughput screening assays. It also contains concentration response files and curve fits as well as detailed descriptions of all the assays. More information is provided in the README file available below and at: http://actor.epa.gov/actor/faces/CSSDashboardLaunch.jsp

Download
(10 MB, ZIP)

Download
(623 MB, ZIP)
ToxCast Data: Chemical Files

http://www.epa.gov/ncct/toxcast/data.html

Contents
Introduction

How were chemicals selected?
How are chemicals identified?
How are chemicals being QC’d?
What assays were run and where were they run?
What does the assay data look like?
How was the assay data processed?
How is the assay data being QC’d?
What information is provided for *in vivo* animal tests in ToxRe
What other information are we providing in this data release?

Appendix 1: ToxCast *In Vitro* Data Processing Pipeline
Appendix 2: Assay Description files
Appendix 3: Chemical Information
Appendix 4: Chemical Selection, Processing and QC

ToxCast Phase II Data & Other Files

This file is intended to be a brief guide to help navigate and analyze all the ToxCast Phase II data.
Release README File

Download (481 KB, PDF)
DSSTox Chemical Structure Files & Browser
http://www.epa.gov/ncct/dsstox/

- Source of high quality structure files and chemical IDs for Tox21 (TOX21S) and ToxCast (TOXCST) programs

- Is my chemical in ToxCast or Tox21?
- Are “similar” chemicals being tested?

- What data are available across EPA and Internet resources for my chemical or its analogs?
NCCT Public Data  [http://www.epa.gov/ncct/](http://www.epa.gov/ncct/)

**ACToR**

**Data Integration**

**Chemical – Assay Linkage**

**In vivo data Linkage**

**Analysis tools**

[http://actor.epa.gov/dashboard/](http://actor.epa.gov/dashboard/)
ToxCast Inventories

293  ToxCast Phase I (293 unique cmpds)
- *EPA pesticidal actives w/ rich in vivo data*
- *PFOAs, BPA, approx 12 metabolite/parent pairs*

1060 ToxCast Phase II (767 unique new cmpds)
- *EPA pesticides, high interest EPA and stakeholder inventories, data rich chemicals (EDSP, OPPT, antimicrobials, inerts, green alternatives, fragrances, water …)*
- *FDA CFSAN data rich, NCTR LTKB Priority 1 drugs*
- *Toxicity reference chemicals, data-rich chemicals, NTP immunotox*
- *135 Donated pharma cmpds -- failed drugs w/ pre-clinical or clinical tox data*

1860 ToxCast E1K (800 unique new cmpds)
- *Endocrine active reference cmpds, SAR predicted ER-active/inactives, EDSP cmpds*

3727 EPA’s Tox21 library (3727 unique cmpds out of current 8599 total)
- *Complete on-hand EPA sample library used to build ToxCast inventories*
EPA ToxCast Inventories

ACToR – overlapping high-interest EPA inventories (xCAS)
Phys-chem filters (MW, LogP, VP, BP)
Commercially procurable
DMSO soluble (not reactive or volatile)

ToxCast Phase I
(mostly pesticides)
293

ToxCast Phase II
135

EPA ToxCast Inventories
consists of all cmpds tested in ToxCast and/or Tox21

EPA’s “ToxCast” inventory
(8599 unique substances)

TOX21S Inventory
(TOX21S)
NCPG (3816)
NTP (3312)
EPA (3726)

Donated failed pharma
(Merck, Pfizer, GSK, Astellas, Sanofi, Roche)

Donated “green” alternatives

4.4K
9K
17K
ToxCast & Tox21: Chemicals, Data and Release Timelines

<table>
<thead>
<tr>
<th>Set</th>
<th>Chemicals</th>
<th>Assays</th>
<th>Endpoints</th>
<th>Completion</th>
<th>Available</th>
</tr>
</thead>
<tbody>
<tr>
<td>ToxCast Phase I</td>
<td>293</td>
<td>~600</td>
<td>~700</td>
<td>2011</td>
<td>Now</td>
</tr>
<tr>
<td>ToxCast Phase II</td>
<td>767</td>
<td>~600</td>
<td>~700</td>
<td>03/2013</td>
<td>Now</td>
</tr>
<tr>
<td>ToxCast Phase III</td>
<td>~900</td>
<td>~100</td>
<td>~100</td>
<td>Just starting</td>
<td>2014</td>
</tr>
<tr>
<td>E1K (endocrine)</td>
<td>800</td>
<td>~50</td>
<td>~120</td>
<td>03/2013</td>
<td>Now</td>
</tr>
<tr>
<td>Tox21</td>
<td>8,599</td>
<td>~25</td>
<td>~50</td>
<td>Ongoing</td>
<td>Ongoing</td>
</tr>
</tbody>
</table>

Pesticides, antimicrobials, food additives, green alternatives, HPV, MPV, endocrine reference cmpds, tox reference cmpds, NTP in vivo, FDA GRAS, EDSP, water contaminants, exposure data, industrial, failed drugs, marketed drugs, fragrances, etc.
ToxCast PhI&PhII 1060: # Compounds per Inventory

- **AIR**: 90
- **GRAS**: 26
- **EDSP**: 130
- **IRIS**: 240
- **TRI**: 216
- **MPV**: 83
- **HPV**: 232
- **Green Chemistry**: 85
- **Antimicrobials**: 91
- **Consumer**: 210
- **Water**: 217
- **PesticideInerts**: 243
- **PesticideActives**: 329
- **Donated Pharmaceuticals**: 135
- **NTP In Vivo**: 202
- **FDA CFSAN**: 94
- **Total In vivo**: 580

- **Excellent coverage of multiple high-interest inventories**
- **Broad diversity of chemical-use categories**
- **Large overlap with data-rich *in vivo* inventories**
ToxCast Phase II Pharmaceuticals: Multipurposing

- 275 PhII cmpds classified as drugs based on presence on FDAMDD, NCGC, NCTR LKB, Donated pharma lists
- 150 drugs appear on additional lists (i.e., multipurposing)
- Caffeine appears on 18 lists
Integrated Chemical-Assay Data Management

**DSSTox**
- Chemical structure - CID
- Substance details - SID
- Project inventory record - RID

**ACToR/ToxMiner**
- Assay name
- Assay details
- Assay outcome

**Structures**
- DSSTox RID

**Assay Results**
- Solution ID
- Plate ID
- Plate Address ID

**Test Sample**
- DSSTox RID
- Bottle ID (→ COA ID)
- Solution ID (→ QC ID)

**Tox21 Sample Tracking Database**
Chemical & Data Quality Issues

DSSTox: Quality substance & structure annotations of test data

ToxCast & Tox21 Sample Database

Valid Structures
Accurate substance annotation:
CAS – Name - Structure

Compound Libraries

Solutions
Bottles
Supplier/Lot/Batch

Supplier-provided info (or lack thereof)
Reported (actual) purity
Limited DMSO solubility
Volatility (missing sample, stench)
Reactivity, explosive …
A copy of each parent Tox21 assay plate is subjected to analytical QC for assessing purity, identity, concentration, stability.

- **PASS =** Confirm parent ion peak and >90% purity
- **Fail, inconclusive or analytical method inappropriate**
- **Retest at later time point under assay conditions for stability**

Publish QC summary results in association with assay data.
Tox21 Analytical QC

- Analytical QC performed on entire Tox21 chemical library at soln-well-level (i.e., test plate copy x time)
- Generate summary calls at soln and cmpd level
- Open discussion on how best to release and communicate QC results
- Recommend using to inform interpretation of assay results
Tox21/ToxCast Analytical QC

- “inconclusives” require additional follow-up & method development
- EPA’s library contains higher proportion of inconclusives (e.g., low MW cmpds, dyes, metals)
- Overall very low “fail” rate

Tox21 Library (8288 cmpds)

- Inconclusive, 3060 (37%)
- Purity >90%, 4417 (53%)

EPA Tox21/ToxCast (3019 cmpds)

- Inconclusive, 1488 (49%)
- Purity >90%, 1433 (48%)

- Mixture, 58
- Decomposition, <50%, 487
- Isomers found, 64
- <50%, 81
- >50%, 49

- Isomers found, 13
- Failure, 14
- Purity <50%, 30
- >75%, 24
- >50%, 17

Office of Research and Development
National Center for Computational Toxicology
Chemical Elements to Data Integration: Chemical representations → Uses

Chemical Name, CASRN

Features, Properties

Structure

Generic Substance

Test Sample

Chemotypes, fingerprints, phys-chem properties, ...

SMILES, InChI

Chemical analogs, Read-across, SAR modeling

Structure searching & modeling

Public toxicity datasets

Experimental Endpoint Data

Supplier, Lot/Batch, physical description
### DSSTox TOXCST (PhI,II,E1K) Structure File

<table>
<thead>
<tr>
<th>Field</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>DSSTox_RID</td>
<td></td>
</tr>
<tr>
<td>DSSTox_GSID</td>
<td></td>
</tr>
<tr>
<td>DSSTox_CID</td>
<td></td>
</tr>
<tr>
<td>DSSTox_FileID</td>
<td></td>
</tr>
<tr>
<td>TestSubstance_ChemicalName</td>
<td></td>
</tr>
<tr>
<td>TestSubstance_CASRN</td>
<td></td>
</tr>
<tr>
<td>TestSubstance_Description</td>
<td></td>
</tr>
<tr>
<td>ChemicalNote</td>
<td></td>
</tr>
<tr>
<td>STRUCTURE_Shown</td>
<td></td>
</tr>
<tr>
<td>STRUCTURE_Formula</td>
<td></td>
</tr>
<tr>
<td>STRUCTURE_MolecularWeight</td>
<td></td>
</tr>
<tr>
<td>STRUCTURE_ChemicalType</td>
<td></td>
</tr>
<tr>
<td>STRUCTURE_TestedForm_DEFINEDOrganic</td>
<td></td>
</tr>
<tr>
<td>STRUCTURE_ChemicalName_IUPAC</td>
<td></td>
</tr>
<tr>
<td>STRUCTURE_SMILES</td>
<td></td>
</tr>
<tr>
<td>STRUCTURE_Parent_SMILES</td>
<td></td>
</tr>
<tr>
<td>STRUCTURE_InChIS</td>
<td></td>
</tr>
<tr>
<td>STRUCTURE_InChIKey</td>
<td></td>
</tr>
<tr>
<td>Substance_modify_yyyymmdd</td>
<td></td>
</tr>
<tr>
<td>ToxCast_TestingStatus</td>
<td></td>
</tr>
<tr>
<td>ToxCast_PhI_v1</td>
<td></td>
</tr>
<tr>
<td>ToxCast_PhI_v2</td>
<td></td>
</tr>
<tr>
<td>ToxCast_PhII</td>
<td></td>
</tr>
<tr>
<td>ToxCast_PhII_DonatedPharma</td>
<td></td>
</tr>
<tr>
<td>ToxCast_e1k</td>
<td></td>
</tr>
<tr>
<td>ToxCast_Tox21</td>
<td></td>
</tr>
<tr>
<td>AnalyticalQC_Problems</td>
<td></td>
</tr>
<tr>
<td>Note_TOXCST</td>
<td></td>
</tr>
</tbody>
</table>

#### Total 1860 cmpds:
- 10 macromolecules
- 96 mixture or formulation
- 17 inorganics
- 43 organometallics
- 84 complexes
- 136 salts

#### MW ranges from 30 (Formalin) to 1700 (Tannic acid)

- Alkylbenzenes
- Chlorobenzenes
- Methyl phenols
- Nitrobenzenes
- Aromatic amines
- Phthalates
- Perfluorinates...
LOG P = Octanol/Water partition coefficient
TPSA = log (Total Polar Surface Area)
Complexity = log (complexity based on paths, branching, atoms)

ToxCast & Tox21 Property Space

LOGP

TPSA

COMPLEXITY

ToxCast Phasel (293)
ToxCast Phasell (767)
Donated Pharma (135)
ToxCast e1k (+800)
Tox21 (7324 unique)

Chemical properties computed using “Adrianna” software by Molecular Networks (P. Volarath)
Estimating Toxicity Mechanism Coverage: DEREK (LHASA) Predictions for ToxCast PhII (1060)

- 328/450 unique DEREK alerts fired across entire dataset
- 128 alerts fired 5 or more times across dataset

DEREK predicts 1 or more toxicity endpoints for 80% of chemicals
DEREK predicts 3 or more endpoints for 40% chemicals

Categories:
- Aromatic nitro compound
- 1,2-Ethleneglycol or derivative
- Organophosphorus di- or tri-ester
- HERG Pharmacophore I
- Di- to poly-halogenated alkane or cycloalkane
- Simple aniline or precursor
- Hydrazine or precursor
- Alkylphenol
- Alkyl aldehyde or precursor
- Polyhalogenated benzene
- beta-O/S-Substituted carboxylic acid or...
- 1,2-Dihalogenated hydrocarbon
- Aromatic primary or secondary amine
- Substituted pyrimidine or purine
- Alkyl ester of phosphoric or phosphonic acid
- Organophosphorus ester
- Phenol or precursor
- Alkylationg agent
- Polyhalogenated aromatic
- Halogenated benzene
New Publicly Available Resources: Chemotyper & ToxPrint Chemotypes

Chemotyper: [http://www.chemotyper.org](http://www.chemotyper.org)

- MS Windows application allows for searching and highlighting of chemical chemotypes (chemical substructures or subgraphs) in structure files.
- Developed by *Molecular Networks GmbH (MN)* under contract from FDA CFSAN to house the “ToxPrint” chemotypes

ToxPrint Chemotypes: [http://www.toxprint.org](http://www.toxprint.org)

- Developed by *Altamira LLC* for FDA CFSAN’s CERES project
- Designed to provide coverage of EPA & FDA inventories and capture chemical features important for chemical safety assessment workflow
- Contains three subsets:
  - *generic structural fragments* (729 total)
  - *Ashby-Tennant genotoxic carcinogen rules* *(Ashby, J; Tennant, RW, 1988)*
  - *cancer TTC categories* *(Kroes, R. et. al. 2004)*
ToxCast ToxPrint Chemotype “Fingerprints”
"toxprint_v2_vs_TOX21S_v4a_8599_03Dec2013.csv"

**Excellent Coverage (#chem w/chemotypes):**
Tox21: 8599 chemicals x 729 chemotypes
- all 8454 structures have ≥ 1 chemotype
- 95% have ≥ 5 chemotypes each
- 65% have ≥ 10 chemotypes each

**Diversity (#chemotypes present)**
ToxCast (1860) \(\rightarrow\) Tox21(8599)
500/729 (68%) \(\rightarrow\) 627/729 (86%)
Chemotyper Application: ToxCast x ToxPrint

ToxCast PhII (1060)

ToxPrint (729)
1. Search “estradiol”
2. Identify common chemotype (steroid backbone)
3. Find all chemicals containing chemotype (238/8599 hits)

E.g. Tox21 Estradiol chemotype search
e.g. ToxCast (1860) “Bisphenol A” chemotype search (25 hits)

Can export list of chemotypes for selected chemicals
Can export structures containing chemotypes

→ Use in iCSS Dashboard to explore ToxCast HTS results
Chemistry: What’s needed

- Accurate chemical annotations of testing libraries (e.g., ToxCast & Tox21), transparency, & reporting of error sources
- Cheminformatics foundation to enable structure modeling

Public data release: ability of chemists & non-chemists (biologists, statisticians) to access & utilize chemical information

Use all available data (HTS+chemistry) to form hypotheses, guide & inform analog selection, and improve prediction models

Incorporate chemical information into usable tools for chemical prioritization & safety assessments
Questions?