EPA

www.epa.gov

# ScrubChem: Building Bioactivity Datasets from PubChem Bioassay Data

**Jason Bret Harris[1], * and Richard Judson[2]**

1 .Oak Ridge Institute of Science and Education (ORISE), Oak Ridge, TN
2. National Center for Computational Toxicology (NCCT), U.S. EPA, RTP, NC

*This poster does not necessarily reflect U.S. EPA policy.*

Jason Bret Harris Harris.Jason@epa.gov ORCID: 0000-0002-7371-0463

## Abstract

ScrubChem

PubChem

ChEMBL Toxcast Tox21
BindingDB MLSP etc.

The PubChem Bioassay database is a non-curated public repository with data from 64 sources, including: ChEMBL, BindingDb, DrugBank, EPA Tox21, NIH Molecular Libraries Screening Program, and various other academic, government, and industrial contributors. Methods for extracting this public data into quality datasets, useable for analytical research, presents several big-data challenges for which we have designed manageable solutions. According to our preliminary work, there are approximately 549 million bioactivity values and related meta-data within PubChem that can be mapped to over 10,000 biological targets. However, this data is not ready for use in data-driven research, mainly due to lack of structured annotations.

**We used a pragmatic approach that provides increasing access to bioactivity values in the PubChem Bioassay database. This included restructuring of individual PubChem Bioassay files into a relational database (ScrubChem). ScrubChem contains all primary PubChem Bioassay data that was: reparsed; error-corrected (when applicable); enriched with additional data links from other NCBI databases; and improved by adding key biological and assay annotations derived from logic-based language processing rules. The utility of ScrubChem and the curation process were illustrated using an example bioactivity dataset for the androgen receptor protein.** This initial work serves as a trial ground for establishing the technical framework for accessing, integrating, curating, analyzing, and making use of such massive bioactivity data.

## Introduction

**Rational:**

- PubChem Bioassay has 1.2M experimental assay records (2.1M chemicals & 10K protein targets).
- Bioactivity datasets built on PubChem data can empower researchers skilled at using data for discovery.

**Problem:**

- Flexibility of the assay record structure (allowing for ease of data submission) coupled with a lack of minimal biological information has accumulated into a huge big-data problem when comparing bioactivity results for chemicals across different assay records.
- For example, hit calls for active and inactive results are only meaningful relative to the assay design.

**Solution:**

- Parse PubChem and related data sources into a custom relational database (**ScrubChem**).
- Programmatically query and iteratively build logic to correct submission errors and add missing annotations about assay design.

**Specific Goals:**

- Improve identification of target data, biological system, testing technology, time, action mode, outcome justification, dose concentration, control, average, and replicate information.

**Expected Results:**

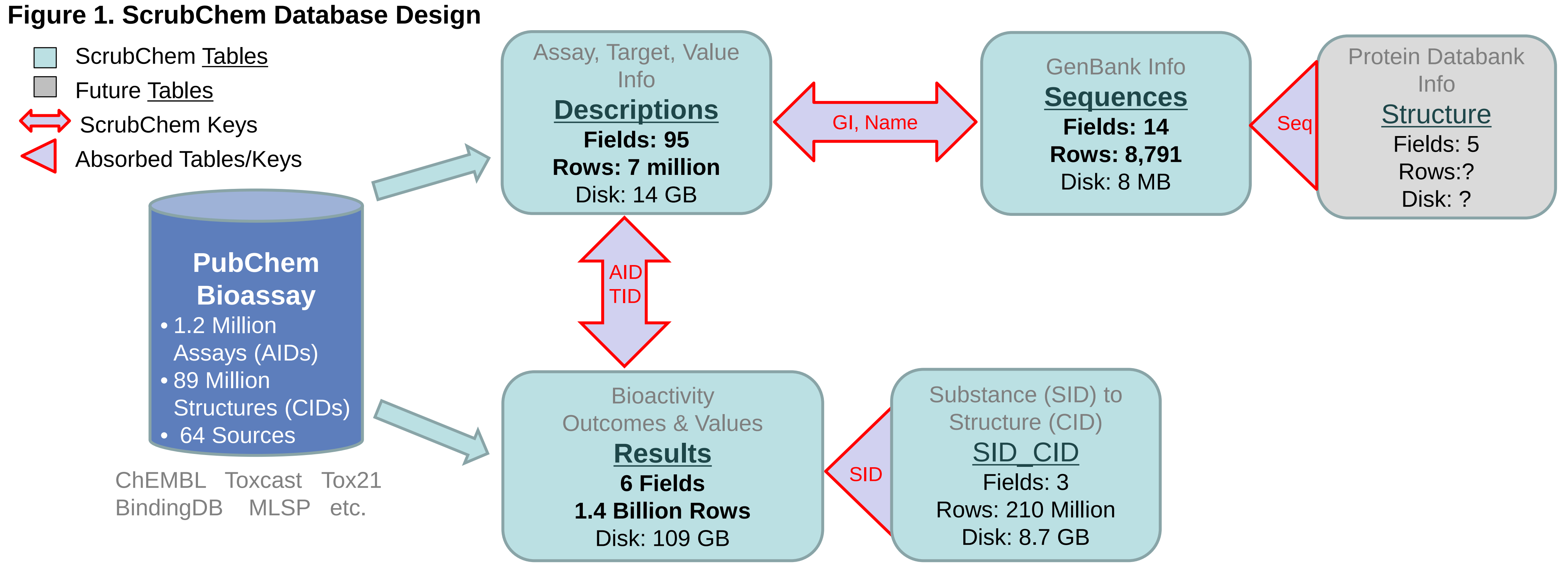
- Present a case study of the human androgen receptor to illustrate the large size potential of extracting target-specific data from PubChem.
- Display the need to capture minimal information to analytically compare chemical activities between different assay records.

**Overview of Terminology:**

1. Assay ID (AID) – An individual assay record.
2. Panel Member ID (MID) – A sub assay record within an AID for a separate target
3. Test ID (TID) – A depositor defined field to hold an assay readout or description (relates descriptive data between result data).
4. Substance ID (SID) – A depositor assigned ID for each substance they test (only static for depositor).
5. Compound ID (CID) – A PubChem assigned ID for each unique chemical structure derived from an SID.
6. GenBank ID (GI) – Accession number for a protein sequence record (records may point to multiple IDs).
7. Outcome – 0 = Inactive, 1 = Active, 2 = Inconclusive, 3 = Unspecified
8. Active Concentration (AC) – Flag for a TID value (e.g., EC50) used as justification for the outcome across an entire assay.
9. Test Concentration (TC) – Flag for a TID value specifying a single dose concentration being tested.

## Methodology

**Figure 1. ScrubChem Database Design**



**CLEANING PROCEDURES**

Assay Annotations (sourced from):

1. **Target** (2 structured target description fields & 4 unstructured comment fields),
2. **System** (unstructured assay/panel name),
3. **Technology** (unstructured assay/panel name)
4. **Time** (unstructured assay/panel name and test name),
5. **Action Mode** - e.g., agonist, antagonist, activator, inhibitor, etc. – (unstructured assay/panel name and test name/description)
6. **Justification** - e.g., AC, TC, No_AC/TC, AC_FIX, PubVal – (3 structured fields and unstructured TID name/description)
7. **Dose Response Concentrations -** e.g., curve data - min/mid/max – sourced from structured dose field and unstructured TID name/description)
8. **Controls, Averages, Replicates** (unstructured assay/panel name and TID name/description)

## Results & Discussion

**Interesting observations & anecdotes of issues identified and fixed using ScrubChem.**

General

- 616,484 result descriptions (AID_TID) have Protein Target (GI) information in unexpected fields (comment fields).
- 1,190,617 result descriptions (AID_TID) have Bioactivities not marked as a value (AC or TC). (i.e. "published values")
- Result outcomes may contain no result value (requires parsing every result).
- Cases of assays in a panel format not using a panel flag.
- Cases of Summary assays marked as Confirmatory assays. (duplicate data)

Spelling

- AID 588719 has "cytoxicity" in assay name.
- AID 48346 and 272704 have "antagonsim" in assay name .
- AID 446013 has "antagonsit" in assay name.

Missing Key Annotations

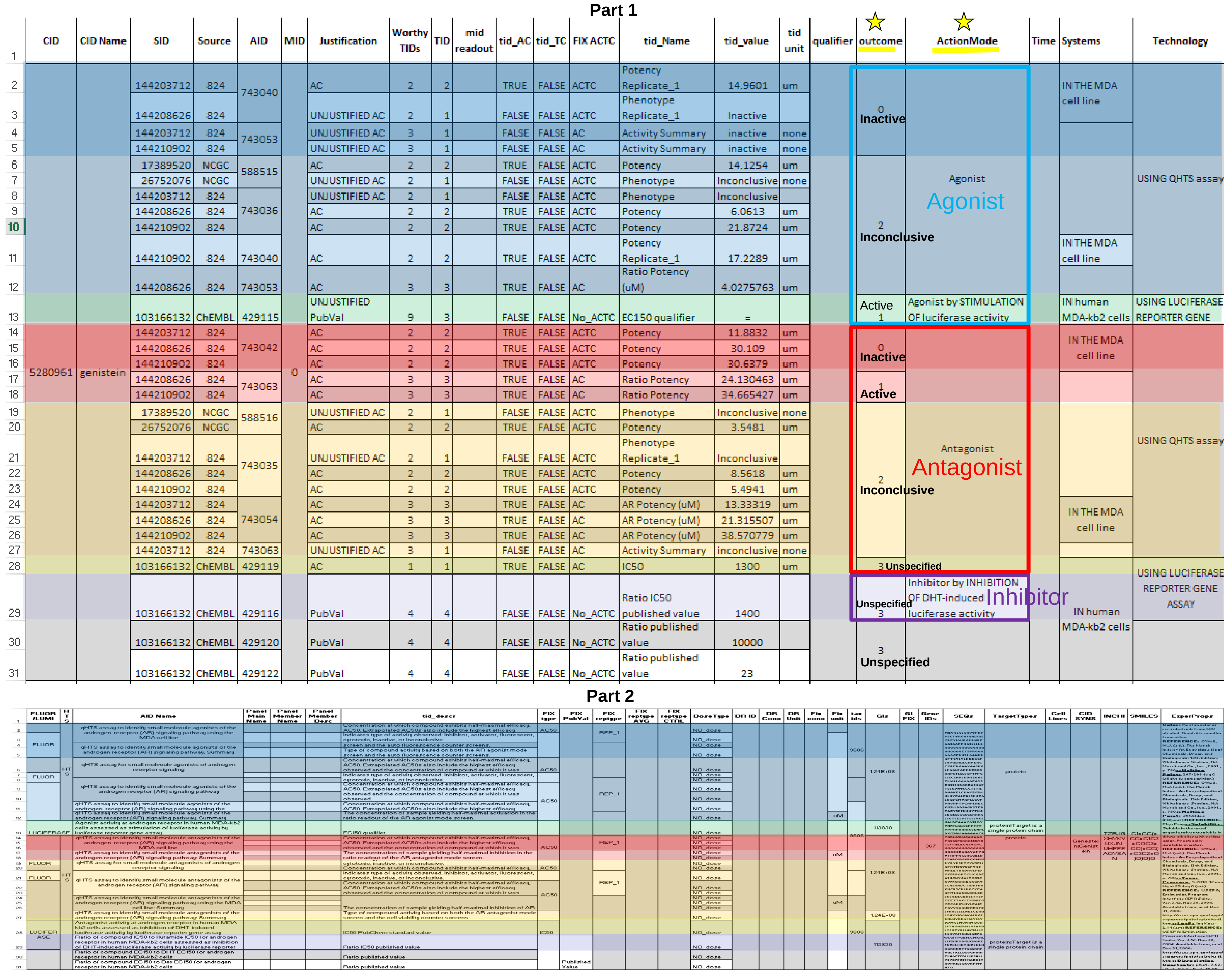
- AID 1000 has an IC50 value not marked as AC TRUE. ***FIXED***
  - Also, does not appear in pre-computed IC50 list.
  - Also, missing its units.
- AID 1208 (Toxcast) without outcomes fields for at least 2 SIDs (e.g., 48413336) *PubChem requirement*

## Results & Discussion

Selecting for **human androgen receptor** (GENE ID: 367) returns **85,030 results**, representing **10,795 unique chemical structures**. As a comparison**, ChEMBL** data only contains **7030 results** and **3142 Compounds**.

**Figure 2. Genestein Results for Androgen Receptor -** grouped by Action Mode & Outcome

- Grouping allows for resolving hit calls across different assays.
- Additional bioassay annotations (e.g., technology, system, time, dose) aid in understanding outcome variance.
- Other fields are output to better describe a chemical or target (e.g., Chemical Props, Seq, Syns).



Part 1

| 1 | CID | CID Name | SID | Source | AID | MID | Justification | Worthy TIDs | TID | mid readout | tid_AC | tid_TC | FIX ACTC | tid_Name | tid_value | tid unit | qualifier | ☆ outcome | ☆ ActionMode | Time | Systems | Technology |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2 | 5280961 | genistein | 144203712 | 824 | 743040 | 0 | AC | 2 | 2 | | TRUE | FALSE | ACTC | Potency Replicate_1 | 14.9601 | um | | 0 Inactive | Agonist Agonist | | IN THE MDA cell line | USING QHTS assay |
| 3 | | | 144208626 | 824 | | | UNJUSTIFIED AC | 2 | 1 | | FALSE | FALSE | ACTC | Phenotype Replicate_1 | Inactive | | | | | | | |
| 4 | | | 144203712 | 824 | 743053 | | UNJUSTIFIED AC | 3 | 1 | | FALSE | FALSE | AC | Activity Summary | inactive | none | | | | | | |
| 5 | | | 144210902 | 824 | | | UNJUSTIFIED AC | 3 | 1 | | FALSE | FALSE | AC | Activity Summary | inactive | none | | | | | | |
| 6 | | | 17389520 | NCGC | 588515 | | AC | 2 | 2 | | TRUE | FALSE | ACTC | Potency | 14.1254 | um | | 2 Inconclusive | | | | |
| 7 | | | 26752076 | NCGC | | | UNJUSTIFIED AC | 2 | 1 | | FALSE | FALSE | ACTC | Phenotype | Inconclusive | none | | | | | | |
| 8 | | | 144203712 | 824 | 743036 | | UNJUSTIFIED AC | 2 | 1 | | FALSE | FALSE | ACTC | Phenotype | Inconclusive | | | | | | | |
| 9 | | | 144208626 | 824 | | | AC | 2 | 2 | | TRUE | FALSE | ACTC | Potency | 6.0613 | um | | | | | | |
| 10 | | | 144210902 | 824 | | | AC | 2 | 2 | | TRUE | FALSE | ACTC | Potency | 21.8724 | um | | | | | | |
| 11 | | | 144210902 | 824 | 743040 | | AC | 2 | 2 | | TRUE | FALSE | ACTC | Potency Replicate_1 | 17.2289 | um | | | | | IN THE MDA cell line | |
| 12 | | | 144208626 | 824 | 743053 | | AC | 3 | 3 | | TRUE | FALSE | AC | Ratio Potency (uM) | 4.0275763 | um | | | | | | |
| 13 | | | 103166132 | ChEMBL | 429115 | | UNJUSTIFIED PubVal | 9 | 3 | | FALSE | FALSE | No_ACTC | EC150 qualifier | = | | | Active 1 | Agonist by STIMULATION OF luciferase activity | | IN human MDA-kb2 cells | USING LUCIFERASE REPORTER GENE |
| 14 | | | 144203712 | 824 | 743042 | | AC | 2 | 2 | | TRUE | FALSE | ACTC | Potency | 11.8832 | um | | 0 Inactive | Antagonist Antagonist | | IN THE MDA cell line | USING QHTS assay |
| 15 | | | 144208626 | 824 | | | AC | 2 | 2 | | TRUE | FALSE | ACTC | Potency | 30.109 | um | | | | | | |
| 16 | | | 144210902 | 824 | | | AC | 2 | 2 | | TRUE | FALSE | ACTC | Potency | 30.6379 | um | | | | | | |
| 17 | | | 144208626 | 824 | 743063 | | AC | 3 | 3 | | TRUE | FALSE | AC | Ratio Potency | 24.130463 | um | | 1 Active | | | | |
| 18 | | | 144210902 | 824 | | | AC | 3 | 3 | | TRUE | FALSE | AC | Ratio Potency | 34.665427 | um | | | | | | |
| 19 | | | 17389520 | NCGC | 588516 | | UNJUSTIFIED AC | 2 | 1 | | FALSE | FALSE | ACTC | Phenotype | Inconclusive | none | | 2 Inconclusive | | | | |
| 20 | | | 26752076 | NCGC | | | AC | 2 | 2 | | TRUE | FALSE | ACTC | Potency | 3.5481 | um | | | | | | |
| 21 | | | 144203712 | 824 | 743035 | | UNJUSTIFIED AC | 2 | 1 | | FALSE | FALSE | ACTC | Phenotype Replicate_1 | Inconclusive | | | | | | | |
| 22 | | | 144208626 | 824 | | | AC | 2 | 2 | | TRUE | FALSE | ACTC | Potency | 8.5618 | um | | | | | | |
| 23 | | | 144210902 | 824 | | | AC | 2 | 2 | | TRUE | FALSE | ACTC | Potency | 5.4941 | um | | | | | | |
| 24 | | | 144203712 | 824 | 743054 | | AC | 3 | 3 | | TRUE | FALSE | AC | AR Potency (uM) | 13.33319 | um | | | | | IN THE MDA cell line | |
| 25 | | | 144208626 | 824 | | | AC | 3 | 3 | | TRUE | FALSE | AC | AR Potency (uM) | 21.315507 | um | | | | | | |
| 26 | | | 144210902 | 824 | | | AC | 3 | 3 | | TRUE | FALSE | AC | AR Potency (uM) | 38.570779 | um | | | | | | |
| 27 | | | 144203712 | 824 | 743063 | | UNJUSTIFIED AC | 3 | 1 | | FALSE | FALSE | AC | Activity Summary | inconclusive | none | | | | | | |
| 28 | | | 103166132 | ChEMBL | 429119 | | AC | 1 | 1 | | TRUE | FALSE | AC | IC50 | 1300 | um | | 3 Unspecified | | | | USING LUCIFERASE REPORTER GENE ASSAY |
| 29 | | | 103166132 | ChEMBL | 429116 | | PubVal | 4 | 4 | | FALSE | FALSE | No_ACTC | Ratio IC50 published value | 1400 | | | Unspecified 3 | Inhibitor by INHIBITION OF DHT-induced luciferase activity Inhibitor | | IN human MDA-kb2 cells | |
| 30 | | | 103166132 | ChEMBL | 429120 | | PubVal | 4 | 4 | | FALSE | FALSE | No_ACTC | Ratio published value | 10000 | | | 3 Unspecified | | | | |
| 31 | | | 103166132 | ChEMBL | 429122 | | PubVal | 4 | 4 | | FALSE | FALSE | No_ACTC | Ratio published value | 23 | | | | | | | |

Part 2

## Conclusions and Future Directions

- ScrubChem incorporates programmatic methods to learn vocabulary and logic to parse PubChem data.
- Systemic and anecdotal submission errors have been corrected (or handled) in order to increase the underlying data available for query and dataset generation.
- Critical assay annotations have been added to coalesce hit calls for chemicals tested in different assays.
- Consensus rules still need to be derived for accepting hit calls.

U.S. Environmental Protection Agency
Office of Research and Development