ScrubChem: Cleaning of PubChem Bioassay Data

to Create Diverse and Massive Bioactivity Datasets for Use in Modeling Applications

Jason B. Harris[1, 3*], Joshua C. Harris[2], Olexandr Isayev[4], Alex Tropsha[4], and Richard Judson[3]

1 .Oak Ridge Institute of Science and Education (ORISE), Oak Ridge, TN **2.** Independent Scholar, Oak Ridge, TN **3.** National Center for Computational Toxicology (NCCT), U.S. EPA, RTP, NC
4. Laboratory for Molecular Modeling, UNC Eshelman School of Pharmacy, University of North Carolina, Chapel Hill, NC.

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

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

Abstract

The PubChem Bioassay database is a non-curated public repository with bioactivity data from many sources, including: ChEMBL, BindingDb, DrugBank, Tox21, NIH Molecular Libraries Screening Program, and various academic, government, and industrial contributors. However, this data is difficult to use in data-driven research, mainly due to lack of interoperability and standardization among its **1.2 million assay records**. Methods for extracting this public data into high-quality, computable datasets, useable for predictive and analytical research, presents several big-data challenges for which ScrubChem is being developed as a manageable solution.

Our approach was to use logic-based text and language processing rules in order to digitally curate and correct the many issues related to the flexible deposition structure of PubChem (e.g., variable placement of biological target information, variable endpoint terminologies, result readouts not distinguished from non-result readouts, improper use of the null vs zero concept, incorrect spellings). Currently, ScrubChem contains approximately **694 million bioactivity values** and related meta-data within PubChem and maps this data to **9 thousand biological targets** and **2.3 million chemical structures**. This work presents case issues identified and resolved through ScrubChem and provides an example dataset for the human androgen receptor with over **85 thousand reference bioactivities** to further illustrate the results of the cleaning process**. Data and updates are available at** ScrubChem.org

Introduction

Rational:

- PubChem Bioassay has 1.2M bioassay submissions (231M outcomes, 2.3M chemicals & 10K protein targets).
- Bioactivity datasets built on PubChem data can be the most diverse and large for use in data-driven discovery & development.

Problems:

- Varying use of PubChem's assay record format, for ease of data submission, coupled with a lack of required minimal biological information has resulted in interoperability and standardization issues in using the data.
- Format of data does not allow for machine-readable retrieval of minimal information required to combine data (make hit calls).
- Data submission procedures have allowed for gaps or errors in some information.

Solution:

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

Specific Goals:

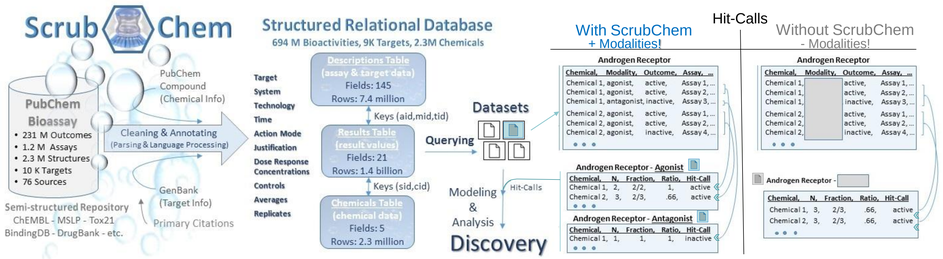
- Improve identification of target data, biological system, testing technology, time, modality (action mode), outcome justification, dose concentration, control, average, and replicate information.
- Use new information to compare chemical activities between different assay records and develop hit call methods.
- Develop a case study of the human androgen receptor to illustrate the large size of datasets available from ScrubChem.

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.
10. Justification – Any measurement(s) used as evidence to justify the reported activity outcome of an experiment.

Methodology

Figure 1. ScrubChem Design



Results

Figure 2. Summary of Data & Online Access: ScrubChem.org

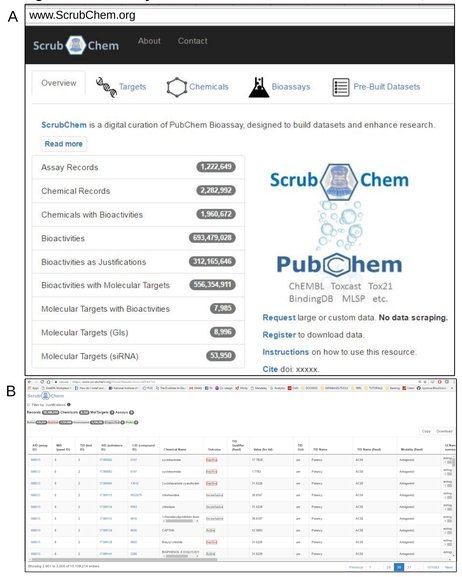


Figure 3. Example Issues & Solutions

A

Issue	Impact	Solution	Affected
[1]1. Target Identifier (Protein GI) not in a designated target field (found instead in comment sections).	Loss of detected targets	- Pattern parse description comment sections. - Re-annotate target fields.	- **Assays** 85,804 7.02% - **Bioactivities** 1,145,061 0.17% - **Chemicals** 248,222 10.87% - **Molecular Targets** 3,499 38.90% --- e.g., AID 2900, MID 0, TID 4, SID 103263999, CID 32817, GI 550543404, TARGET 15-LOX

B

Issue	Impact	Solution	Affected
[2]8. Modality of Action not defined.	Loss of comparable data between assays (ambiguous data)	- Sample diversity of modality types. - Standardize vocabulary for each modality. - Parse data fields (hierarchically) to identify most reliable modality descriptions.	- **Assays:** 691,802 56.58% - **CIDs:** 2,094,323 91.74% - **Bioactivities:** [4]large - **Molecular Targets:** 8,103 90.07% --- e.g., AID 3181 "Inhibitor" e.g., AID 3353 "Inhibitor| by INHIBITION OF PDGF-dependent autophosphorylation"

C

Issue	Impact	Solution	Affected
[3]5. Min-, Max-, Mid-, Single-, and No-dose concentrations are not annotated.	Loss of detected bioactivities	- Create library of values and dose types (min-, max-, mid-, single-, and no-dose) for each AID. - Annotate each TID in an AID with values corresponding to the dose types. - When no other justification options are available: - TC is used as a justification when tested at a min-dose and the result outcome is active (single-dose actives are questionable). - TC is used as a justification when tested at a max- or single-dose and the result outcome is inactive.	- **Assays:** 7,798 0.64 % - **CIDs** 696,206 30.50% - **Bioactivities:** 564,969,544 81.47% (...Active) 12,669,383 1.83% (...Inactive) 513,513,427 74.05% (TC_MIN) 15,753,029 2.27% (...ACTIVE) 421,810 0.06% (...Justification) 52,303 0.01% (...INACTIVE) 14,689,205 2.12% (TC_MID) 311,922,207 44.98% (...ACTIVE) 9,524,101 1.37% (...INACTIVE) 267,243,364 38.54% (TC_MAX) 33,280,753 4.80% (...ACTIVE) 673,504 0.10% (...INACTIVE) 30,098,046 4.34% (...Justification) 1,529,465 0.22% (TC_SINGLE) 204,013,323 29.42% (...ACTIVE) 2,049,968 0.30% (...INACTIVE) 201,482,812 29.05% (...Justification) 181,747,109 26.21% --- e.g., AID 743088, MID 2, TID 4, SID 164348278 "max-dose and inactive" • e.g., AID 731, MID 0, TID 1, SID 842121 "single-dose & inactive"

Fig 3.a. A target identifier does not always follow an expected format. One case is shown, affecting 7% of assays, where protein identifiers are found in a comment field instead of the target identifier field. **Fig 3.b**. Modalities (e.g., agonist) are not structured in at least 56% of assays. **Fig 3.c**. Many assays test a single or a max/min concentration for activity and results (>26% of all bioactivities) are conclusive (used as 'Justifications') without calculating a dose response curve.

Results (continued)

Figures 4-6. Example Data for Human Androgen Receptor

Figure 4. Data for hAR

Fig 4. Selecting for **human androgen receptor** (GENE ID: 367) returns **85,580 results**, representing **10,934 unique chemical structures**. As a comparison, **ChEMBL** data only contains **7,030 results** and **3,142 Compounds** and no modalities for deriving hit calls.

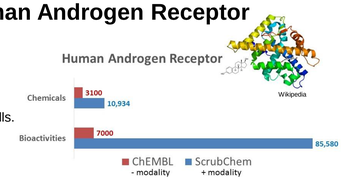


Figure 5. Making Hit Calls

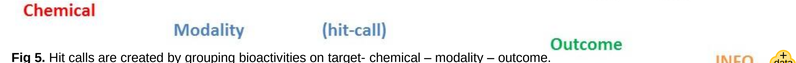
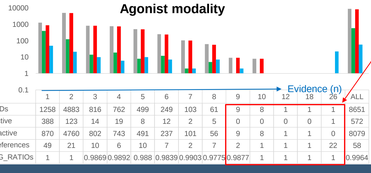


Fig 5. Hit calls are created by grouping bioactivities on target- chemical – modality – outcome.

CID_TYPE	cid_count	cid	NAME	MOD_TYPE	mod_count	modality	hitcall	ratio	fraction	n	sources	refcount	references	OUTCOME_TYPE	out_count	outcome	#REPORTS	INFO_1/info_c
ROOT CID:	1140 of 10934	5757 estradiol																
SUB CID	1140 of 10934	5757 estradiol	ROOT MODALI	1 of 4	Agonist	Active		1	1	6	824	1	21543282					
SUB CID	1140 of 10934	5757 estradiol	SUB MODALI	1 of 4	Agonist	Active		1	1	6	824	1	21543282	ROOT OUTCOME:	1 of 2	1	6	
SUB CID	1140 of 10934	5757 estradiol	SUB MODALI	1 of 4	Agonist	Active		1	1	6	824	1	21543282	SUB OUTCOME:	1 of 2	1	6	INFO: 1 of 6
SUB CID	1140 of 10934	5757 estradiol	SUB MODALI	1 of 4	Agonist	Active		1	1	6	824	1	21543282	SUB OUTCOME:	1 of 2	1	6	INFO: 2 of 6
SUB CID	1140 of 10934	5757 estradiol	SUB MODALI	1 of 4	Agonist	Active		1	1	6	824	1	21543282	SUB OUTCOME:	1 of 2	1	6	INFO: 3 of 6
SUB CID	1140 of 10934	5757 estradiol	SUB MODALI	1 of 4	Agonist	Active		1	1	6	824	1	21543282	SUB OUTCOME:	1 of 2	1	6	INFO: 4 of 6
SUB CID	1140 of 10934	5757 estradiol	SUB MODALI	1 of 4	Agonist	Active		1	1	6	824	1	21543282	SUB OUTCOME:	1 of 2	1	6	INFO: 5 of 6
SUB CID	1140 of 10934	5757 estradiol	SUB MODALI	1 of 4	Agonist	Active		1	1	6	824	1	21543282	SUB OUTCOME:	1 of 2	1	6	INFO: 6 of 6
SUB CID	1140 of 10934	5757 estradiol	SUB MODALI	1 of 4	Agonist	Active		1	1	6	824	1	21543282	ROOT OUTCOME:	2 of 2	2	1	
SUB CID	1140 of 10934	5757 estradiol	SUB MODALI	1 of 4	Agonist	Active		1	1	6	824	1	21543282	SUB OUTCOME:	2 of 2	2	1	INFO: 1 of 1
ROOT CID:	1140 of 10934	5757 estradiol																
SUB CID	1140 of 10934	5757 estradiol	ROOT MODALI	2 of 4	Antagonist	Active		0.75	3/4	4	824	1	21543282					
SUB CID	1140 of 10934	5757 estradiol	SUB MODALI	2 of 4	Antagonist	Active		0.75	3/4	4	824	1	21543282	ROOT OUTCOME:	1 of 3	0	1	
SUB CID	1140 of 10934	5757 estradiol	SUB MODALI	2 of 4	Antagonist	Active		0.75	3/4	4	824	1	21543282	SUB OUTCOME:	1 of 3	0	1	INFO: 1 of 1
SUB CID	1140 of 10934	5757 estradiol	SUB MODALI	2 of 4	Antagonist	Active		0.75	3/4	4	824	1	21543282	ROOT OUTCOME:	2 of 3	1	3	
SUB CID	1140 of 10934	5757 estradiol	SUB MODALI	2 of 4	Antagonist	Active		0.75	3/4	4	824	1	21543282	SUB OUTCOME:	2 of 3	1	3	INFO: 1 of 3
SUB CID	1140 of 10934	5757 estradiol	SUB MODALI	2 of 4	Antagonist	Active		0.75	3/4	4	824	1	21543282	SUB OUTCOME:	2 of 3	1	3	INFO: 2 of 3
SUB CID	1140 of 10934	5757 estradiol	SUB MODALI	2 of 4	Antagonist	Active		0.75	3/4	4	824	1	21543282	SUB OUTCOME:	2 of 3	1	3	INFO: 3 of 3
SUB CID	1140 of 10934	5757 estradiol	SUB MODALI	2 of 4	Antagonist	Active		0.75	3/4	4	824	1	21543282	ROOT OUTCOME:	3 of 3	2	3	
SUB CID	1140 of 10934	5757 estradiol	SUB MODALI	2 of 4	Antagonist	Active		0.75	3/4	4	824	1	21543282	SUB OUTCOME:	3 of 3	2	3	INFO: 1 of 3
SUB CID	1140 of 10934	5757 estradiol	SUB MODALI	2 of 4	Antagonist	Active		0.75	3/4	4	824	1	21543282	SUB OUTCOME:	3 of 3	2	3	INFO: 2 of 3
SUB CID	1140 of 10934	5757 estradiol	SUB MODALI	2 of 4	Antagonist	Active		0.75	3/4	4	824	1	21543282	SUB OUTCOME:	3 of 3	2	3	INFO: 3 of 3
SUB CID	1140 of 10934	5757 estradiol	ROOT MODALI	3 of 4	Inhibitor	Active		1	1	1	ChEMBL	1	16309907					
SUB CID	1140 of 10934	5757 estradiol	SUB MODALI	3 of 4	Inhibitor	Active		1	1	1	ChEMBL	1	16309907	ROOT OUTCOME:	1 of 1	1	1	
SUB CID	1140 of 10934	5757 estradiol	SUB MODALI	3 of 4	Inhibitor	Active		1	1	1	ChEMBL	1	16309907	SUB OUTCOME:	1 of 1	1	1	INFO: 1 of 1
SUB CID	1140 of 10934	5757 estradiol	ROOT MODALI	4 of 4	Affinity	Active		1	1	2	ChEMBL	2	17448656 | 17	90084				
SUB CID	1140 of 10934	5757 estradiol	SUB MODALI	4 of 4	Affinity	Active		1	1	2	ChEMBL	2	17448656 | 17	SUB OUTCOME:	1 of 2	1	2	
SUB CID	1140 of 10934	5757 estradiol	SUB MODALI	4 of 4	Affinity	Active		1	1	2	ChEMBL	2	17448656 | 17	SUB OUTCOME:	1 of 2	1	2	INFO: 1 of 2
SUB CID	1140 of 10934	5757 estradiol	SUB MODALI	4 of 4	Affinity	Active		1	1	2	ChEMBL	2	17448656 | 17	SUB OUTCOME:	1 of 2	1	2	INFO: 2 of 2
SUB CID	1140 of 10934	5757 estradiol	SUB MODALI	4 of 4	Affinity	Active		1	1	2	ChEMBL	2	17448656 | 17	SUB OUTCOME:	2 of 2	2	1	
SUB CID	1140 of 10934	5757 estradiol	SUB MODALI	4 of 4	Affinity	Active		1	1	2	ChEMBL	2	17448656 | 17	SUB OUTCOME:	2 of 2	2	1	INFO: 1 of 1

Figure 6. Hit Call Summary

Fig 6. Hit calls can be sorted by modalities and confidence metrics, such as the number of results included in each hit call (**n**) and **Ratio of agreement**.



e.g., There are 8,651 chemicals tested for the agonist modality. A selection grid of n > 9 yields 20 chemicals (1 active, 19 inactive). This means that each hit call derived from this selection have at least 9 separate evidences (**n**). The lowest **AVG Ratio of agreement** for a bin in this selection is 98.77% (n=9).

Summary

- ScrubChem digitally curates 1.2M assays that can be used to build datasets which would otherwise be unbuildable.
- ScrubChem database has 694 million+ measurements.
- Developed hit-call methods to build datasets.
- Developed hit-call evaluation metrics.
- Data & updates available online at ScrubChem.org