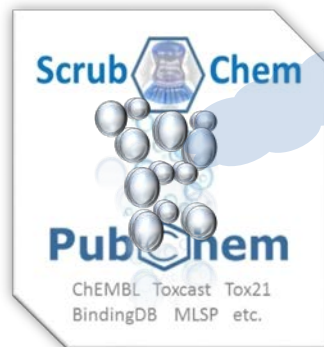


ScrubChem



Cleaning PubChem Bioassay Data

Jason Bret Harris

Jason.B.Harris@gmail.com Harris.Jason@epa.gov Jason@scrubchem.org

Postdoc Research Participant

Oak Ridge Institute for Science and Education (ORISE)
& U.S. Environmental Protection Agency (EPA), Research Triangle Park, NC
National Center for Computational Toxicology (NCCT)

Visiting Research Scholar

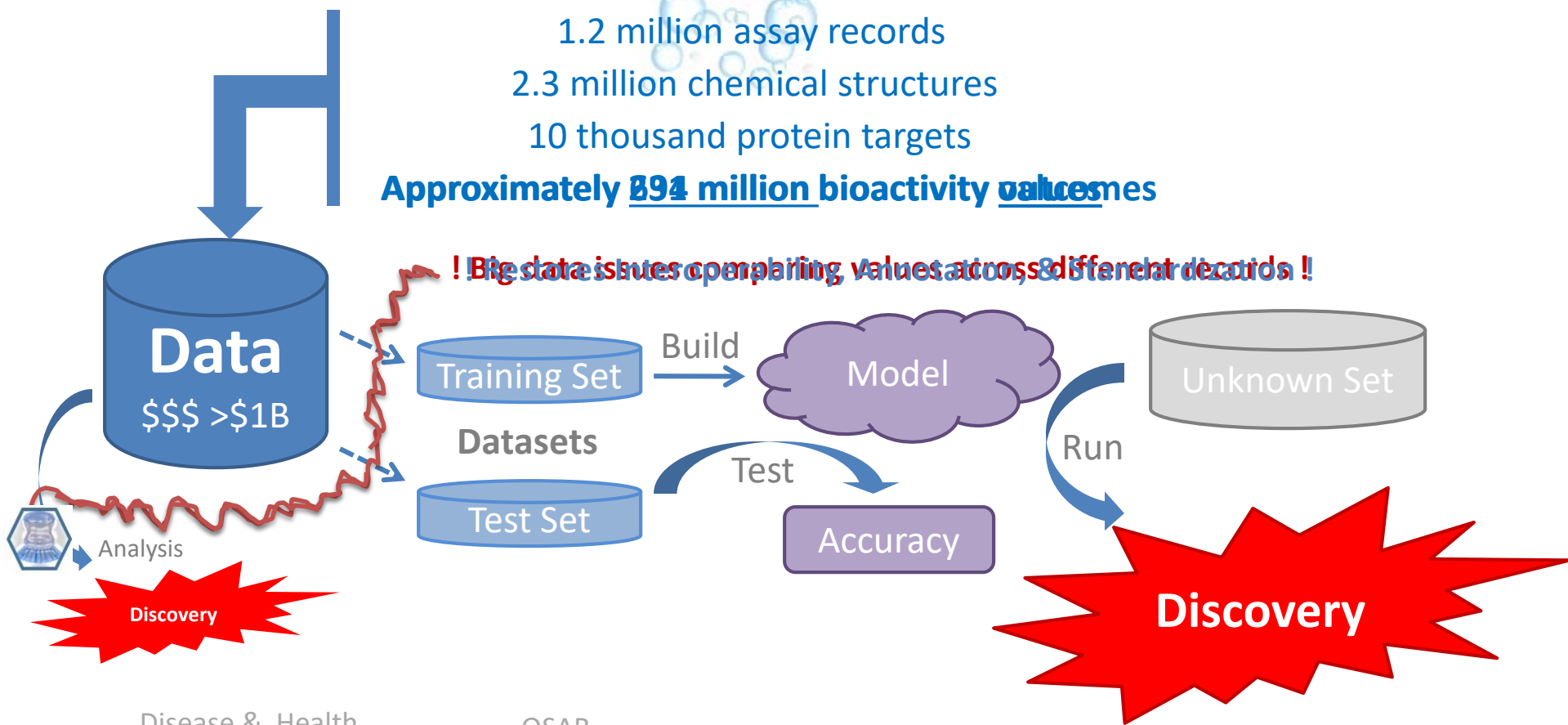
University of North Carolina Chapel Hill
Eshelman School of Pharmacy, Molecular Modeling Laboratory



ChEMBL DrugBank Tox21
BindingDB PDBind MLSP etc.

1.2 million assay records
2.3 million chemical structures
10 thousand protein targets

Approximately 694 million bioactivity values



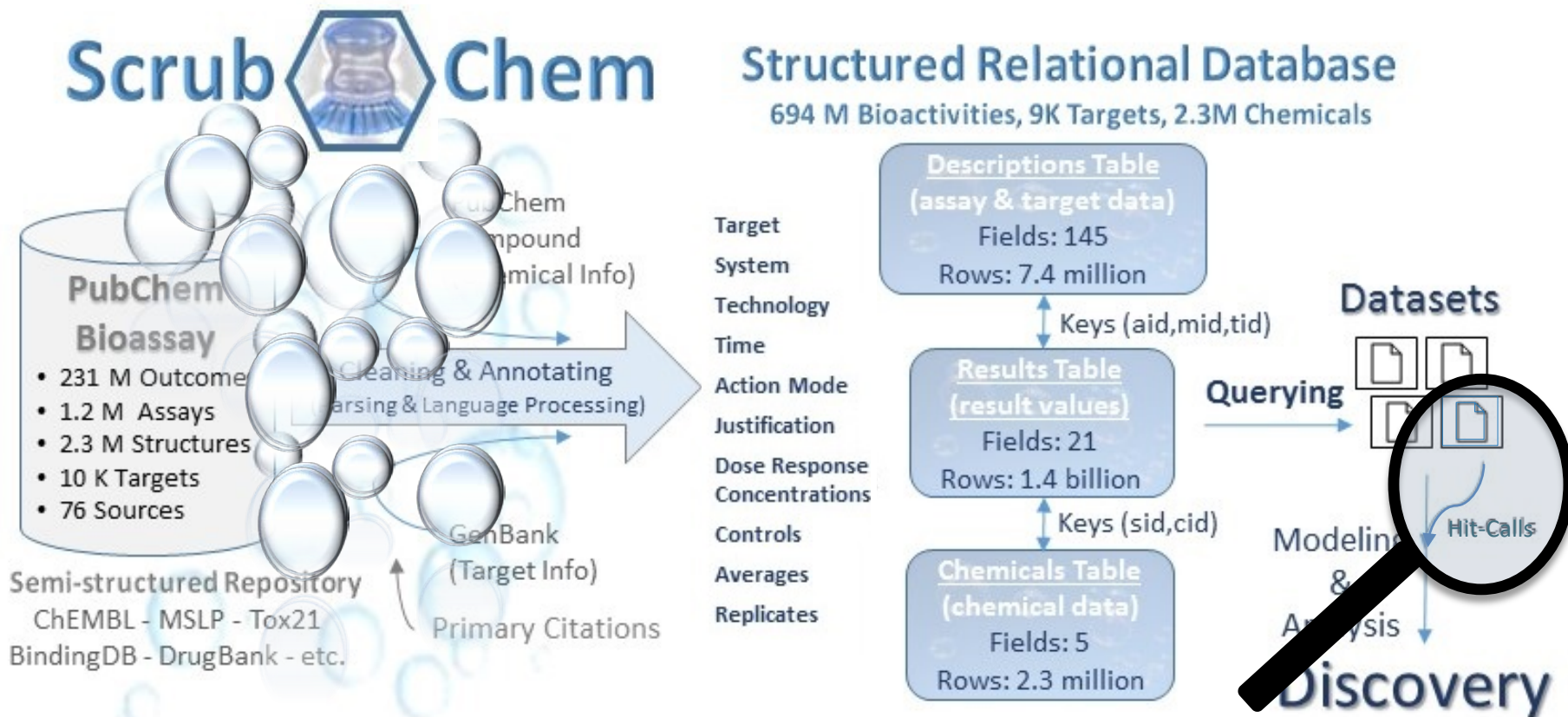
Disease & Health Toxicity QSAR
Drug Discovery Assay Development Docking
Reproducibility Metabolism Machine Learning
Deep Learning ADMET Hazard & Risk

Overview (in words)

- **ScrubChem** is an effort to programmatically **identify and correct** for many big data issues related to the interoperability and standardization of biochemical data across **millions** of bioassay records in PubChem.
- **ScrubChem** concepts (e.g., modalities & justifications) establish the way forward for organizing disparate data so that it can be reliably combined into hit-calls and datasets.
- A direct result of the **ScrubChem** effort is a **database** which can be used to build many different kinds of datasets for various data-driven use cases (e.g., docking, QSAR, machine-learning, regulatory decision-making, assay design).
- **ScrubChem** datasets **expand the number of targets and chemicals** with robust and reliable data.

ScrubChem Framework

Cleaning of this data increases the number of results usable for building datasets.



“Hit-Calls”

With ScrubChem (+) Modality

Generic Datasets



(~9,000 Molecular Targets)

Androgen Receptor

Chemical	Modality	Outcome	Assay, ...
Chemical 1	agonist	active	Assay 1, ...
Chemical 1	antagonist	inactive	Assay 2, ...
Chemical 1	antagonist	inactive	Assay 3, ...
Chemical 2	agonist	active	Assay 1, ...
Chemical 2	agonist	active	Assay 4, ...
Chemical 2	agonist	inactive	Assay 5, ...

Androgen Receptor - Agonist ✓

Chemical	N	Fraction	Ratio	Hit-Call
Chemical 1	1	1/1	1	active
Chemical 2	3	2/3	.66	active

Androgen Receptor - Antagonist ✓

Chemical	N	Fraction	Ratio	Hit-Call
Chemical 1	2	2/2	1	inactive

active in any modality = active for protein

[Androgen Receptor] ✓

[Hit-Calls]

(combined modalities)

Chemical	[M]	[N]	Hit-Call
Chemical 1	2	3	active
Chemical 2	1	3	active

Without ScrubChem (-) Modality

Androgen Receptor

Chemical	Modality	Outcome	Assay, ...
Chemical 1	agonist	active	Assay 1, ...
Chemical 1	antagonist	inactive	Assay 2, ...
Chemical 1	antagonist	inactive	Assay 3, ...
Chemical 2	agonist	active	Assay 1, ...
Chemical 2	agonist	active	Assay 4, ...
Chemical 2	agonist	inactive	Assay 5, ...

ISSUE:

Without modality information chemical 1 appears as active 1x and inactive 2x.

Androgen Receptor - Agonist ✗

Chemical	N	Fraction	Ratio	Hit-Call
Chemical 1	3	2/3	.66	inactive
Chemical 2	3	2/3	.66	active

Case Issue 1

Issue	Impact	Solution	Affected
^a 1. Target Identifier (Protein GI) not in a designated target field (found instead in comment sections).	Loss of detected targets	<ul style="list-style-type: none"> - Pattern parse description comment sections. - Re-annotate target fields. 	<ul style="list-style-type: none"> - Assays 85,804 7.02% - Bioactivities 1,145,061 0.17% - Chemicals 248,222 10.87% - Molecular Targets 3,499 38.90% <p>-----</p> <p>e.g., AID 2900, MID 0, TID 4, SID 103263999, CID 32817, GI 550544304, TARGET 15-LOX</p>

Assay ID (AID)

Assay Description:

Assay Name	Assay Description	Assay Comments	Assay Protocol	Assay Targets (ids, comments)	External References (xrefs)	Panel (T/F)
				Protein GIs (target IDs)		

Result Descriptions:

TID 1 (name, descr, unit, DR conc, AC/TC, mid #)
TID 2 (name, descr, unit, DR conc, AC/TC, mid #)
TID ... (name, descr, unit, DR conc, AC/TC, mid #)

Panel Descriptions (if panel is true):

MID 1 (Panel Name, Descr, Comments, Protocol, Target (ids, comments))
MID 2 (Panel Name, Descr, Comments, Protocol, Target (ids, comments))
MID 3 (Panel Name, Descr, Comments, Protocol, Target (ids, comments))

Assay ID (AID)

Results:

Substance ID (SID)	Substance ID (SID)	...
Outcome	Outcome	
TID 1 - (value)	TID 1 - (value)	
TID 2 - (value)	TID 2 - (value)	
TID 3 - (value)	TID 3 - (value)	
...	...	

Case Issue 8 (2-7 in paper)

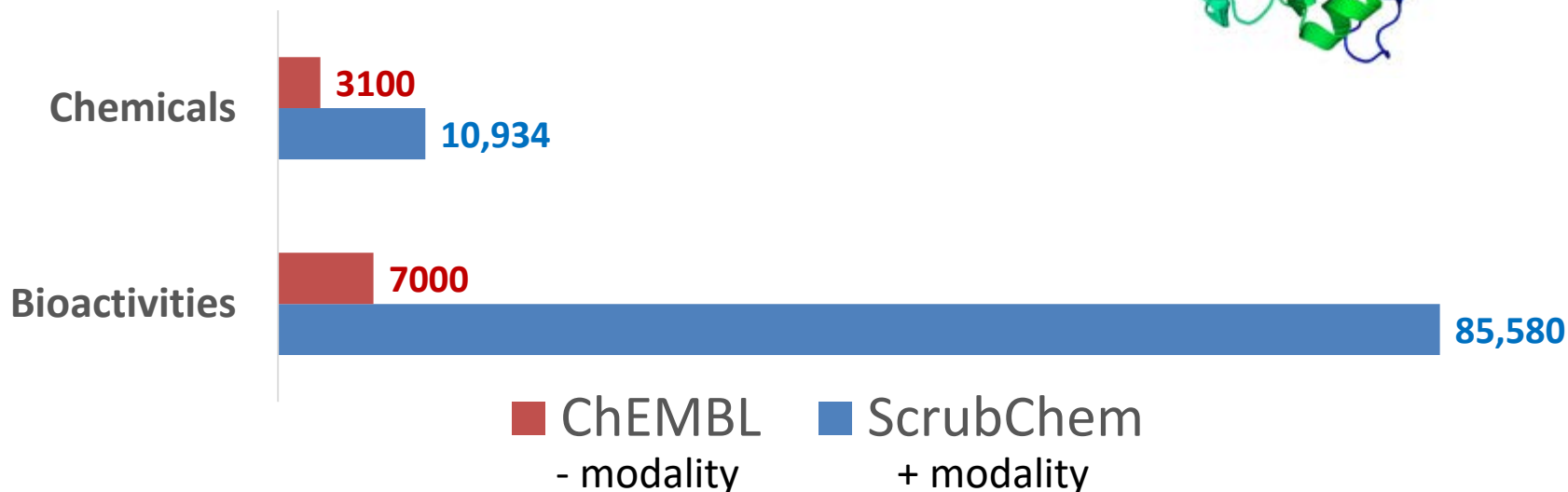
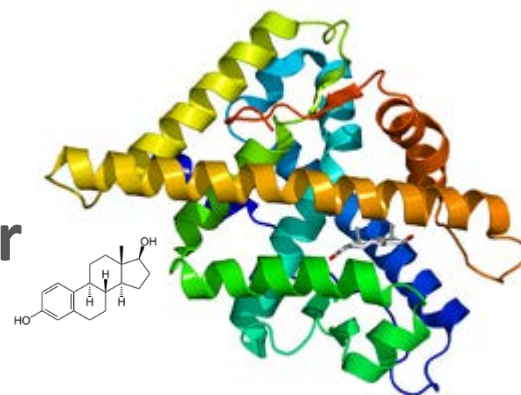
Issue	Impact	Solution	Affected												
^{cl} 8. Modality of Action not defined.	Loss of comparable data between assays (ambiguous data)	<ul style="list-style-type: none">- Sample diversity of modality types.- Standardize vocabulary for each modality.- Parse data fields (hierarchically) to identify most reliable modality descriptions.	<table><tr><td>- Assays:</td><td>691,802</td><td>56.58%</td></tr><tr><td>- CIDs:</td><td>2,094,323</td><td>91.74%</td></tr><tr><td>- Bioactivities:</td><td>#large</td><td></td></tr><tr><td>- Molecular Targets:</td><td>8,103</td><td>90.07%</td></tr></table> <div><div></div><div>e.g., AID 3181 "Inhibitor"</div><div>e.g., AID 3353 "Inhibitor by INHIBITION OF PDGF-dependent autophosphorylation"</div><div>e.g., AID 3586 "Inhibitor by DISPLACEMENT OF [3H]-5-HT"</div><div>e.g., AID 3187 "Inhibition Constant (Ki)"</div><div>e.g., AID 3183 "Inhibition Constant (Ki) by DISPLACEMENT OF binding"</div><div>e.g., AID 3191 "Inhibition Constant (Ki) by DISPLACEMENT OF [125I]DOI"</div><div>e.g., AID 3179 "Affinity"</div><div>e.g., AID 3174 "Affinity (Km)"</div><div>e.g., 243422 "Affinity Ratio"</div><div>e.g., AID 3534 "Agonist"</div><div>e.g., AID 3169 "Antagonist "</div><div>e.g., AID 3189 "Activity"</div><div>e.g., AID 3385 "Activity (EC) by stimulate glucose uptake"</div><div>e.g., AID 3192 "Effective Dose (ED)"</div></div>	- Assays:	691,802	56.58%	- CIDs:	2,094,323	91.74%	- Bioactivities:	#large		- Molecular Targets:	8,103	90.07%
- Assays:	691,802	56.58%													
- CIDs:	2,094,323	91.74%													
- Bioactivities:	#large														
- Molecular Targets:	8,103	90.07%													

Case Issues 9-14

Issue	Impact	Solution	Affected
^a 9. Missing digital identifiers for literature data (ChEMBL)	Loss of source citation information	- Use ChEMBL IDs to link back to ChEMBL database and extract primary citation data.	- Assays: 46,556 3.81% ----- e.g., AID 1207595 doi: 10.6019/CHEM e.g., AID2901 doi:10.1016/S096
Spelling - anecdotes			
^a 10. Cytotoxicity as "Cytotoxicity" in AID description's name.	Loss of retrievable data	- Awareness of issue Fix during pattern parse.	e.g., AID 588719
^a 11. Antagonism as "antagonism" in AID description's name.	Loss of retrievable data	- Awareness of issue - Fix during pattern parse.	e.g., AID 48346 e.g., AID 272704
^a 12. Antagonist as "antagonist" in AID description's name.	Loss of retrievable data	- Awareness of issue - Fix during pattern parse.	e.g., AID 446013
Missing Key Annotations – anecdotes			
^a 13. Missing result units	Loss of retrievable data	- Build library of unit types and phrases. - Pattern parse description sections for unit data.	e.g., AID 588523, MID 0, TID 1, SID 99494248
^a 14. AIDs with no outcome field (required field)	Slower parsing	- Requires null checks for every data field (even required fields) Build library of unit types and phrases.	e.g., AID 1208, SID 48413336 e.g., AID 555, SID 843951

Example Datasets

Human Androgen Receptor



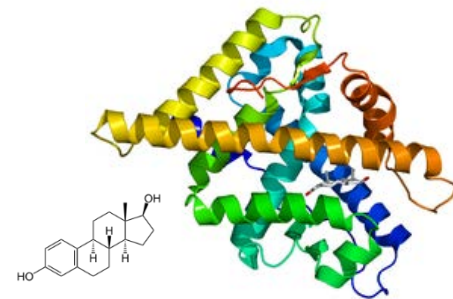
Selecting data for the **human androgen receptor** (GENE ID: 367) returns: **85,580 results** and **10,934 Chemicals** with modalities.

As a comparison, **ChEMBL** data only contains:

7,030 results and **3,142 Chemicals** and **no modalities** for deriving hit calls.

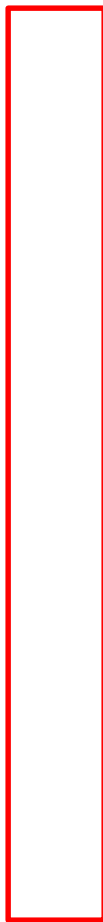
Bioactivities for hAR

Showing 1 Chemical (estradiol) out of 11,000



1 Chemical: Estradiol

Group by
Modality & Outcome



Chemical Hit-Calls for hAR

Tree structure for flat file filtering

Showing 1 Chemical (estradiol) out of 11,000

Chemical

Modality

(hit-call)

Outcome

INFO...

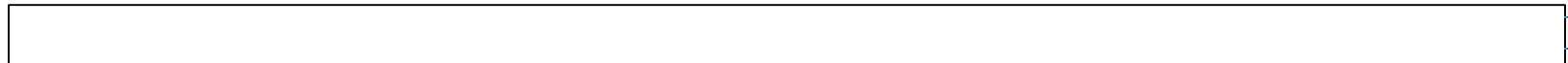
CID_TYPE	cid_count	cid	NAME	MOD_TYPE	count	modality	hitcall	ratio	fraction	n	sources	#references	references	OUTCOME_TYPE	count	outcome	#REPORTS	INFO_1	info_2
ROOT CID:	1140 of 1093	5757	estradiol																
SUB CID	1140 of 1093	5757	estradiol	ROOT MODALITY:	1 of 4	Agonist	Active	1	1	6	824	1	21543282						
SUB CID	1140 of 1093	5757	estradiol	SUB MODALITY:	1 of 4	Agonist	Active	1	1	6	824	1	21543282	ROOT OUTCOME:	1 of 2	1	6		
SUB CID	1140 of 1093	5757	estradiol	SUB MODALITY:	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 1093	5757	estradiol	SUB MODALITY:	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 1093	5757	estradiol	SUB MODALITY:	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 1093	5757	estradiol	SUB MODALITY:	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 1093	5757	estradiol	SUB MODALITY:	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 1093	5757	estradiol	SUB MODALITY:	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 1093	5757	estradiol	SUB MODALITY:	1 of 4	Agonist	Active	1	1	6	824	1	21543282	ROOT OUTCOME:	2 of 2	2	1		
SUB CID	1140 of 1093	5757	estradiol	SUB MODALITY:	1 of 4	Agonist	Active	1	1	6	824	1	21543282	SUB OUTCOME:	2 of 2	2	1	INFO:	1 of 1
SUB CID	1140 of 1093	5757	estradiol	ROOT MODALITY:	2 of 4	Antagonist	Active	0.8	3/4	4	824	1	21543282						
SUB CID	1140 of 1093	5757	estradiol	SUB MODALITY:	2 of 4	Antagonist	Active	0.8	3/4	4	824	1	21543282	ROOT OUTCOME:	1 of 3	0	1		
SUB CID	1140 of 1093	5757	estradiol	SUB MODALITY:	2 of 4	Antagonist	Active	0.8	3/4	4	824	1	21543282	SUB OUTCOME:	1 of 3	0	1	INFO:	1 of 1
SUB CID	1140 of 1093	5757	estradiol	SUB MODALITY:	2 of 4	Antagonist	Active	0.8	3/4	4	824	1	21543282	ROOT OUTCOME:	2 of 3	1	3		
SUB CID	1140 of 1093	5757	estradiol	SUB MODALITY:	2 of 4	Antagonist	Active	0.8	3/4	4	824	1	21543282	SUB OUTCOME:	2 of 3	1	3	INFO:	1 of 3
SUB CID	1140 of 1093	5757	estradiol	SUB MODALITY:	2 of 4	Antagonist	Active	0.8	3/4	4	824	1	21543282	SUB OUTCOME:	2 of 3	1	3	INFO:	2 of 3
SUB CID	1140 of 1093	5757	estradiol	SUB MODALITY:	2 of 4	Antagonist	Active	0.8	3/4	4	824	1	21543282	SUB OUTCOME:	2 of 3	1	3	INFO:	3 of 3
SUB CID	1140 of 1093	5757	estradiol	SUB MODALITY:	2 of 4	Antagonist	Active	0.8	3/4	4	824	1	21543282	ROOT OUTCOME:	3 of 3	2	3		
SUB CID	1140 of 1093	5757	estradiol	SUB MODALITY:	2 of 4	Antagonist	Active	0.8	3/4	4	824	1	21543282	SUB OUTCOME:	3 of 3	2	3	INFO:	1 of 3
SUB CID	1140 of 1093	5757	estradiol	SUB MODALITY:	2 of 4	Antagonist	Active	0.8	3/4	4	824	1	21543282	SUB OUTCOME:	3 of 3	2	3	INFO:	2 of 3
SUB CID	1140 of 1093	5757	estradiol	SUB MODALITY:	2 of 4	Antagonist	Active	0.8	3/4	4	824	1	21543282	SUB OUTCOME:	3 of 3	2	3	INFO:	3 of 3
SUB CID	1140 of 1093	5757	estradiol	ROOT MODALITY:	3 of 4	Inhibitor	Active	1	1	1	ChEMBL	1	16309907						
SUB CID	1140 of 1093	5757	estradiol	SUB MODALITY:	3 of 4	Inhibitor	Active	1	1	1	ChEMBL	1	16309907	ROOT OUTCOME:	1 of 1	1	1		
SUB CID	1140 of 1093	5757	estradiol	SUB MODALITY:	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 1093	5757	estradiol	ROOT MODALITY:	4 of 4	Affinity	Active	1	1	2	ChEMBL	2	17448656	17890084					
SUB CID	1140 of 1093	5757	estradiol	SUB MODALITY:	4 of 4	Affinity	Active	1	1	2	ChEMBL	2	17448656	ROOT OUTCOME:	1 of 2	1	2		
SUB CID	1140 of 1093	5757	estradiol	SUB MODALITY:	4 of 4	Affinity	Active	1	1	2	ChEMBL	2	17448656	SUB OUTCOME:	1 of 2	1	2	INFO:	1 of 2
SUB CID	1140 of 1093	5757	estradiol	SUB MODALITY:	4 of 4	Affinity	Active	1	1	2	ChEMBL	2	17448656	SUB OUTCOME:	1 of 2	1	2	INFO:	2 of 2
SUB CID	1140 of 1093	5757	estradiol	SUB MODALITY:	4 of 4	Affinity	Active	1	1	2	ChEMBL	2	17448656	ROOT OUTCOME:	2 of 2	2	1		
SUB CID	1140 of 1093	5757	estradiol	SUB MODALITY:	4 of 4	Affinity	Active	1	1	2	ChEMBL	2	17448656	SUB OUTCOME:	2 of 2	2	1	INFO:	1 of 1

Grouping Format: flat file filtering

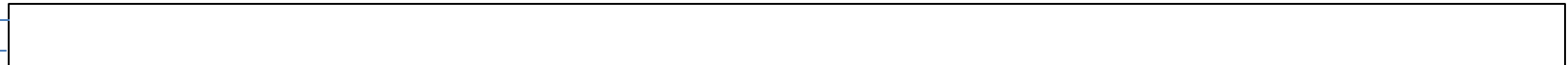
Main Grouping Scheme



Filtering by **INFO** Lines (view results and retains hit-calls)



Filtering by **Root Modality** (view hit calls and hides results)

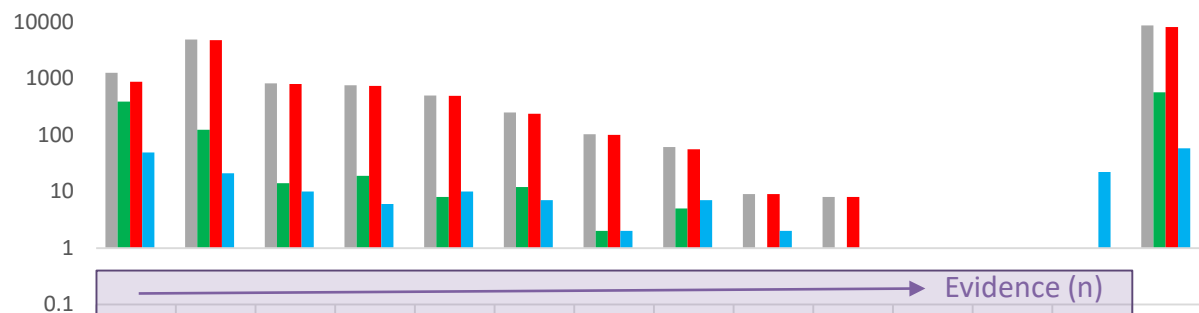


Hit-Call Summary:

Choosing Confidence Thresholds

Hit-calls can be divided by the **modality** and binned by **confidence/reproducibility metrics**, such as the number (**n**) of results (**evidences**) used in each hit-call or **% agreement**.

8,651 chemical hit-calls for hAR-agonist modality binned by Evidence (n).



e.g.,

There are 8,651 total chemicals tested for the agonist modality with an (**n**) of at least 1.

0.1		Evidence (n)													ALL
		1	2	3	4	5	6	7	8	9	10	12	18	26	
#CIDs	1258	4883	816	762	499	249	103	61	9	8	1	1	1	8651	
#Active	388	123	14	19	8	12	2	5	0	0	0	0	1	572	
#Inactive	870	4760	802	743	491	237	101	56	9	8	1	1	0	8079	
#References	49	21	10	6	10	7	2	7	2	1	1	1	22	58	
AVG_RATIOS	1	1	0.986930	0.989170	0.987980	0.983940	0.990290	0.977460	0.98765	1	1	1	1	0.99637	

A selection grid of $n \geq 9$ yields:

20 chemicals out of 8,651

1 active out of 572

19 inactive out of 8,079.

This means that each hit call derived from this selection will have at least 9 separate evidences (**n**). The lowest **AVG Ratio of agreement** for a bin in this selection is 98.77% ($n=9$).

Protein Hit-Calls for a Chemical

Chemical Summary for hit-calls from combining modalities ← Summary for protein and modality hit-calls →

Targets:

Hit-Calls:

(target-modality)

Single
modality hit-call
for protein

Agreement of evidences
of evidences

active : inactive

Combined
modality hit-calls
for protein

Combined # of evidences
Combined # of modalities

658 more target-modalities
376 more protein targets

Protein Hit-Calls for 155 TSCA Chemicals

← Combined Modality Hit-calls for 578 Protein Targets → Bisphenol A
[Active] = 15 : [Inactive] = 289
summary

155

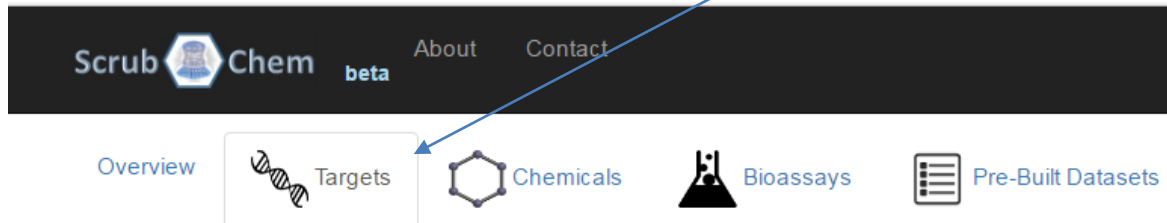
Chemicals

ScrubChem.org: Home Page

www.ScrubChem.org

Searching Data: Input

1. Select a **Search Tab** Category



Search Targets

Input

Select an Identifier:

- Gene Symbol
- Accession Number
- Gene Symbol
- Gene ID
- Gene Name
- Full Gene Name

Search

2. Select an **Identifier**



Search Targets

Input

Select an Identifier:

Gene Symbol

Input ID:

esr

- ESR1
- ESR1|ESR2
- ESR2
- ESR2|ESR1
- ESRRA
- ESRRB
- ESRRG


3. Type input and select from the autocomplete

Searching Data: Results

Search Targets

[Input](#)

[Results](#)

☐ Filter by 'Justifications' 

Records **11,035,823** Chemicals **98,944** MolTargets **8** Assays **887**

Active 676,645 **Inactive** 8,622,638 **Inconclusive** 1,720,513 **Unspecified** 16,016 **Probe** 7

[Copy](#)

[Download](#)

AID (assay ID)	MID (panel ID)	TID (test ID)	SID (substance ID)	CID (compound ID)	Chemical Name	Outcome	TID Qualifier (fixed)	Value (for tid)	TID Unit
636765	0	1	136930240	56969898	CHEMBL1928181	Unspecified	<		
636765	0	1	136930241	56969899	CHEMBL1928182	Unspecified	<		
636765	0	1	136936880	56969904	CHEMBL1928179	Unspecified	<		
636765	0	1	136936881	57398877	CHEMBL1928188	Unspecified	<		
636765	0	1	136940061	56969901	CHEMBL1928176	Unspecified	<		
636765	0	1	136940062	56969902	CHEMBL1928185	Unspecified	<		
636765	0	1	136940063	56969903	CHEMBL1928186	Unspecified	<		

Showing 1 to 100 of 11,035,823 entries

[Previous](#)

[Next](#)

Filtering Results: Justifications

Input Results View only records that are used to 'best' justify each outcome.

☐ Filter by 'Justifications' ⓘ

Records 11,035,823 Chemicals 98,944 MolTargets 8 Assays 887

Active 676,649 Inactive 8,622,638 Inconclusive 1,720,513 Unspecified 16,016 Probe 7

Copy Download

Chemical Name	Outcome	TID Qualifier (fixed)	Value (for tid)	TID Unit	TID Name	TID Name (fixed)	Modality (fixed)
CHEMBL1928181	Unspecified	<			RBA activity comment		Inhibitor by DISPLACEME ◀ ▶
CHEMBL1928182	Unspecified	<			RBA activity comment		Inhibitor by DISPLACEME ◀ ▶
CHEMBL1928179	Unspecified	<			RBA activity comment		Inhibitor by DISPLACEME ◀ ▶

Input Results

Filters results from 11 M records to 262 K.

☒ Filter by 'Justifications' ⓘ

Records 262,694 Chemicals 98,944 MolTargets 8 Assays 887

Active 9,823 Inactive 234,602 Inconclusive 12,541 Unspecified 5,724 Probe 4

Copy Download

Chemical Name	Outcome	TID Qualifier (fixed)	Value (for tid)	TID Unit	TID Name	TID Name (fixed)	Modality (fixed)
CHEMBL1928181	Unspecified	<	0.1		RBA published value	Relative Binding Affinity	Inhibitor by DISPLACEME ◀ ▶
CHEMBL1928182	Unspecified	<	0.1		RBA published value	Relative Binding Affinity	Inhibitor by DISPLACEME ◀ ▶
CHEMBL1928179	Unspecified	<	0.1		RBA published value	Relative Binding Affinity	Inhibitor by DISPLACEME ◀ ▶

Embed ScrubChem.org Tables

iframe of <https://www.scrubchem.org/Home/Results?CIDs=2244>

EPA's CompTox Dashboard (to be added)

← → ↻ 🏠 **Secure** | <https://comptox.epa.gov/dashboard/dsstoxdb/results?utf8=✓&search=aspirin> 🔍 ⚙️ ☆ ⋮

EPA United States Environmental Protection Agency Home Advanced Search Lists Search Chemistry Dashboard Submit Comment Share Copy Aa Aa Aa

Chemistry Dashboard

Aspirin

50-78-2 | DTXSID6020108

🔍 Searched by Approved Name: Found 1 result for 'aspirin':

Wikipedia

Aspirin, also known as acetylsalicylic acid (ASA), is a medication used to treat pain, fever, and inflammation. Specific inflammatory conditions in which it is used include Kawasaki disease, pericarditis, and rheumatic fever. Aspirin given shortly after a heart attack decreases the risk of death. Aspirin is also used long-term to help prevent heart attacks, strokes, and blood clots, in people at high risk. Aspirin may also decrease the risk of certain types of cancer, particularly... Read more

Intrinsic Properties

Structural Identifiers

Related Compounds (Beta)

Presence in Lists

Record Information

Chemical Properties Env. Fate/Transport Synonyms External Links Toxicity Values (Beta) Exposure Bioassays Similar Molecules (Beta) Literature Comments

ToxCast

ScrubChem

PubChem

BioAssay Results

ScrubChem

Filter by 'Justifications' ⓘ

Records 70,136 Chemicals 1 MolTargets 566 Assays 3,730

Active 763 Inactive 66,674 Inconclusive 4,071 Unspecified 5,228 Probe 0

AID (assay ID)	MID (panel ID)	TID (test ID)	SID (substance ID)	CID (compound ID)	Chemical Name	Outcome	TID Qualifier (fixed)	Value (for tid)	TID Unit	TID Name
634118	0	1	103164874	2244	aspirin	Active	=	3.12	um	IC50
634118	0	2	103164874	2244	aspirin	Active	=			IC50 activit
634118	0	4	103164874	2244	aspirin	Active		=		IC50 qualifi
634118	0	5	103164874	2244	aspirin	Active	=	3.12		IC50 publi
634118	0	6	103164874	2244	aspirin	Active	=	3120	nm	IC50 stand
637947	0	1	103164874	2244	aspirin	Active	=	6.07	um	IC50

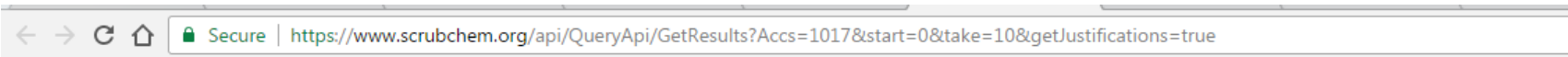
Showing 1 to 100 of 70,136 entries

Previous Next

About Contact Privacy Accessibility Help Downloads

API Access to Database

<https://www.scrubchem.org/api/QueryApi/GetResults?Accs=1017&start=0&take=10&getJustifications=true>



This XML file does not appear to have any style information associated with it. The document tree is shown below.

```
<ArrayOfQueryService.DescriptionsResultsCidInfo2 xmlns:i="http://www.w3.org/2001/XMLSchema-instance"
xmlns="http://schemas.datacontract.org/2004/07/ScrubChem.Services">
  <QueryService.DescriptionsResultsCidInfo2>
    <ACTC i:nil="true"/>
    <ChEMBL_Bioactivities i:nil="true"/>
    <ChEMBL_Cell_Type i:nil="true"/>
    <ChEMBL_Description i:nil="true"/>
    <ChEMBL_DocID i:nil="true"/>
    <ChEMBL_Doi i:nil="true"/>
    <ChEMBL_Format i:nil="true"/>
    <ChEMBL_ID i:nil="true"/>
    <ChEMBL_Journal i:nil="true"/>
    <ChEMBL_Organism i:nil="true"/>
    <ChEMBL_Pmid i:nil="true"/>
    <ChEMBL_Strain i:nil="true"/>
    <ChEMBL_Subcell i:nil="true"/>
    <ChEMBL_TargetID i:nil="true"/>
    <ChEMBL_Tissue i:nil="true"/>
    <ChEMBL_Type i:nil="true"/>
    <DoseType i:nil="true"/>
    <GI_Fix i:nil="true"/>
    <HTSWords i:nil="true"/>
    <accs>1017</accs>
    <action_mode>|</action_mode>
    <aid>493040</aid>
    <aid_name>Navigating the Kinome</aid_name>
    <aid_version>0</aid_version>
```


LIVE DEMOS

[ScrubChem.org](https://www.scrubchem.org)

<https://www.scrubchem.org>

[ScrubChem API](https://www.scrubchem.org/api/QueryApi/GetResults?Accs=1017&start=0&take=10&getJustifications=true)

<https://www.scrubchem.org/api/QueryApi/GetResults?Accs=1017&start=0&take=10&getJustifications=true>

[ScrubChem Embed](https://www.scrubchem.org/Home/Results?Accs=1017)

<https://www.scrubchem.org/Home/Results?Accs=1017>

[EPA CompTox Dashboard](https://comptox.epa.gov/dashboard/dsstoxdb/results?search=aspirin) (to be added)

<https://comptox.epa.gov/dashboard/dsstoxdb/results?search=aspirin>

Conclusions

ScrubChem Accomplishments:

- Accesses millions of assay records.
- Programmatically identifies & corrects data issues (systemic and anecdotal)
 - Adds critical annotations (e.g., modality & justifications) and implements concepts (hit calls) needed for aggregating data from different assays into datasets.
- Provides online access to the database and datasets.

In Progress:

- Improving assay annotation and code (iterative cleaning).
- Expanding vocabularies and ontology needed to better standardize endpoints, targets, and system information (grow scope of database).
- Reviewing hit calls (improving the methodology and scope).
- Model/Analysis/Tools (implement datasets)!
- Enhance web access (open to ideas!).
- Publish & Continued Support!

Acknowledgements

*Contributed code

Mentors:

Joshua C. Harris*

Independent Scholar, Oak Ridge, TN

Richard Judson

Environmental Protection Agency, Research Triangle Park, NC
National Center for Computational Toxicology

Alexander Tropsha & Olexandr (Oles) Isayev

University of North Carolina- Chapel Hill
Eshelman School of Pharmacy

Funded in part through a research participation program administered by **ORISE** (Oak Ridge Institute for Science & Education) and carried out at the **EPA** (Environmental Protection Agency).

