Advancing the accuracy of open chemical information with CAS Common Chemistry

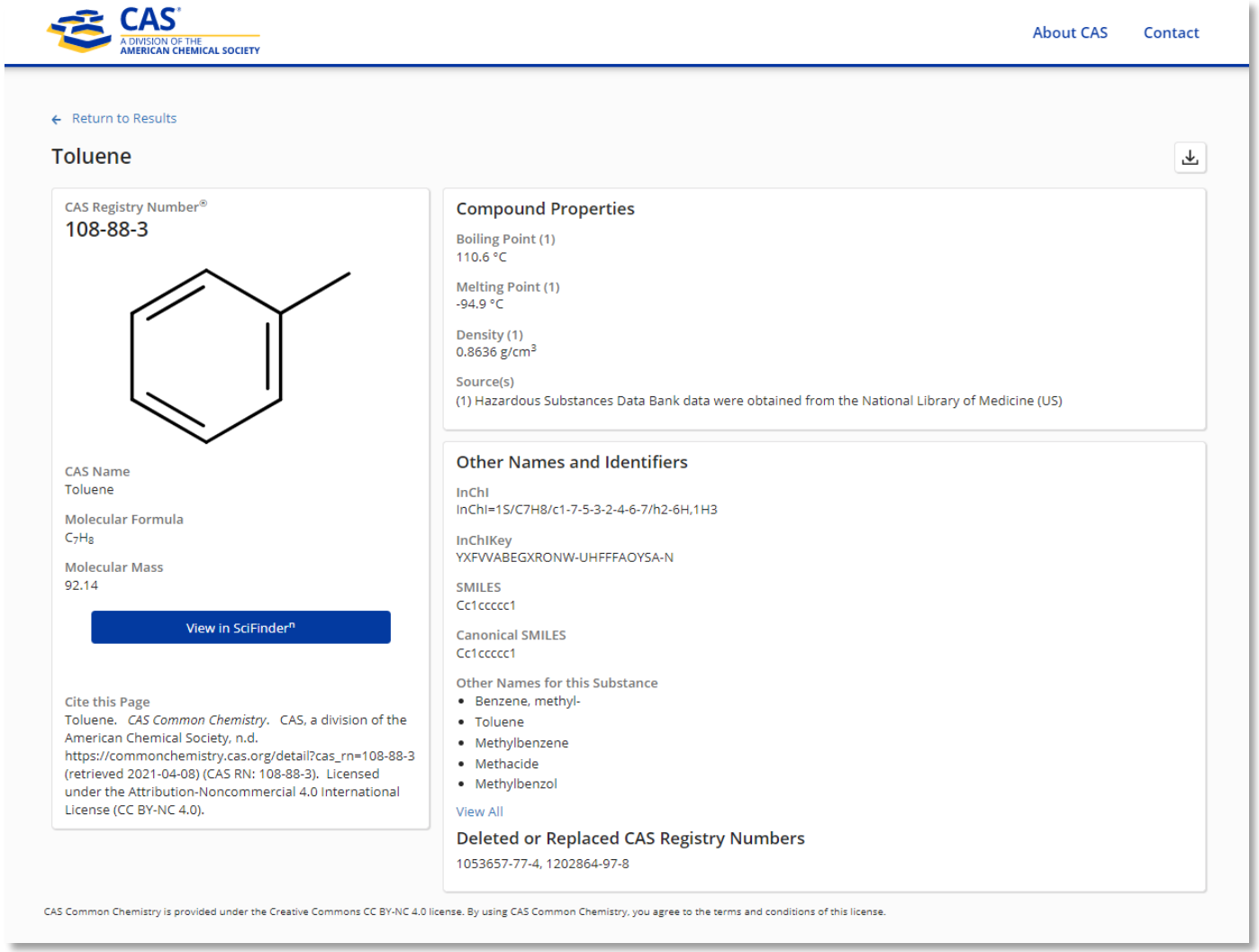
Egon L. Willighagen (Maastricht Uni/NL), @egonwillighagen, https://orcid.org/0000-0001-7542-0286
Antony J. Williams (US-EPA), @ChemConnector, https://orcid.org/0000-0002-2668-4821
Christopher Grulke (US-EPA)
Ann Richard (US-EPA)
Andrea Jacobs (CAS, a division of the American Chemical Society)

The views expressed in this presentation are those of the authors and do not necessarily reflect the views or policies of the US EPA.

What is CAS Common Chemistry?

https://commonchemistry.cas.org

- An open community resource from CAS, a division of the American Chemical Society
- First launched in 2009
- Trusted information from CAS REGISTRY®, including CAS RNs, chemical names, and structures
- Significant re-launch in March 2021, in collaboration with community contributors
- Now ~500K substances, accessible via API



Join Us Wednesday to Learn More

CAS Common Chemistry and the value of community collaboration for chemical informatics

Wednesday, April 14

2:50pm – 3:10pm PDT

Topic for Today: Comparing CAS Common Chemistry with Open Databases

- What is the chemistry the database is representing?
- When are two entries representing the same entity?

- PubChem
- EPA CompTox Dashboard
- Wikidata / Wikipedia

Database entries and matching

- CAS Common Chemistry
  - Substance (salts, solutions, …)
- PubChem
  - Compound (unique InChIKey)
- EPA CompTox Dashboard
  - Compound (CASRN)
- Wikidata
  - Neutral compounds (unique InChIKey)
- Wikipedia
  - It's complicated

- Entry equivalence
  - Structure normalization
  - InChIKey match

Hähnke, V. D., Kim, S. & Bolton, E. E. PubChem chemical structure standardization. *J Cheminform* **10**, 36 (2018).

Williams, A. J. *et al.* The CompTox Chemistry Dashboard: a community data resource for environmental chemistry. *Journal of Cheminformatics* **9**, (2017).

PubChem

https://pubchem.ncbi.nlm.nih.gov/

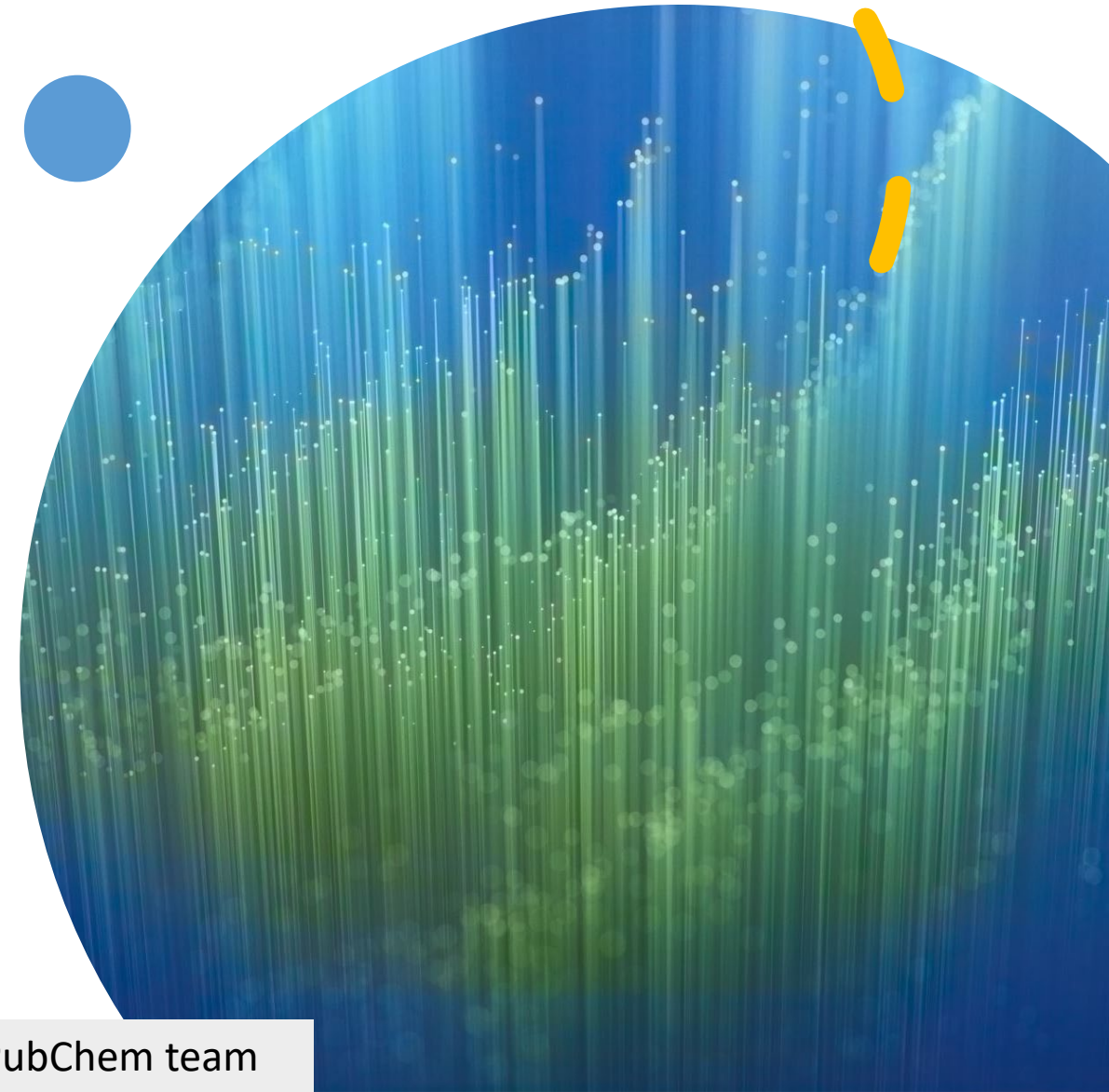
Comparing CAS Common Chemistry to PubChem

53.8% of CAS RN® found in PubChem

82.4% of SMILES found in PubChem

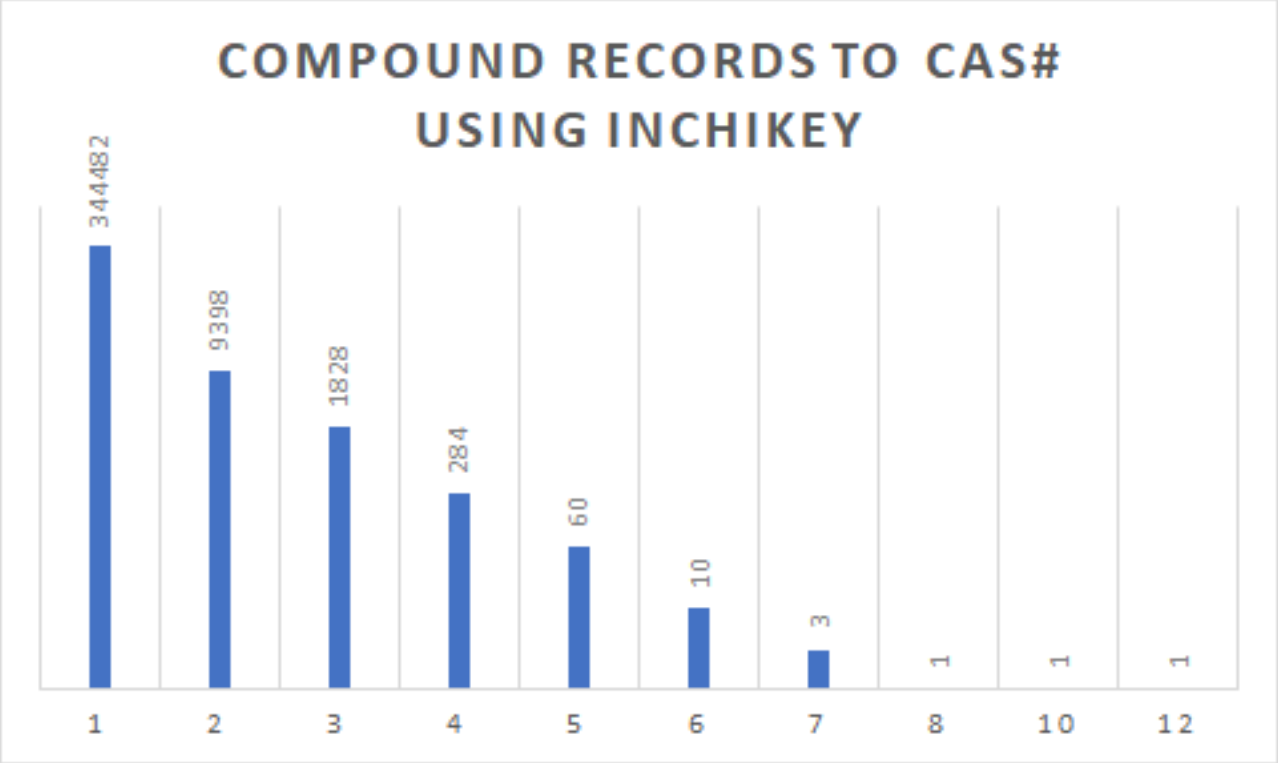
24.2% of Names found in PubChem

83.1% of InChIKey found in PubChem



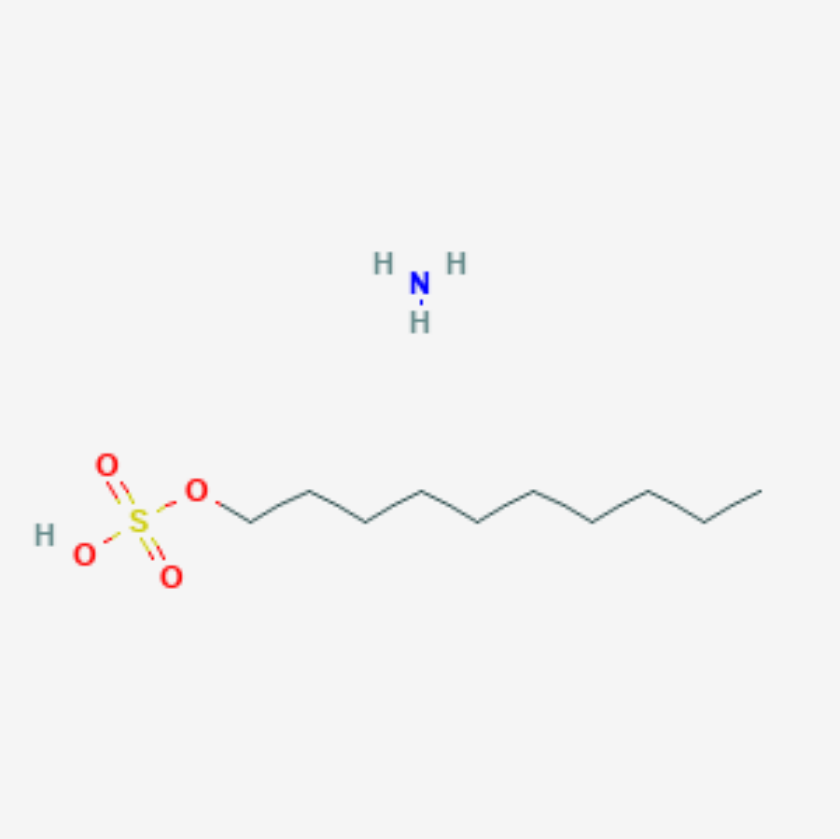
Slide credit: Evan Bolton, Jeff Zhang, and Paul Thiessen from the PubChem team

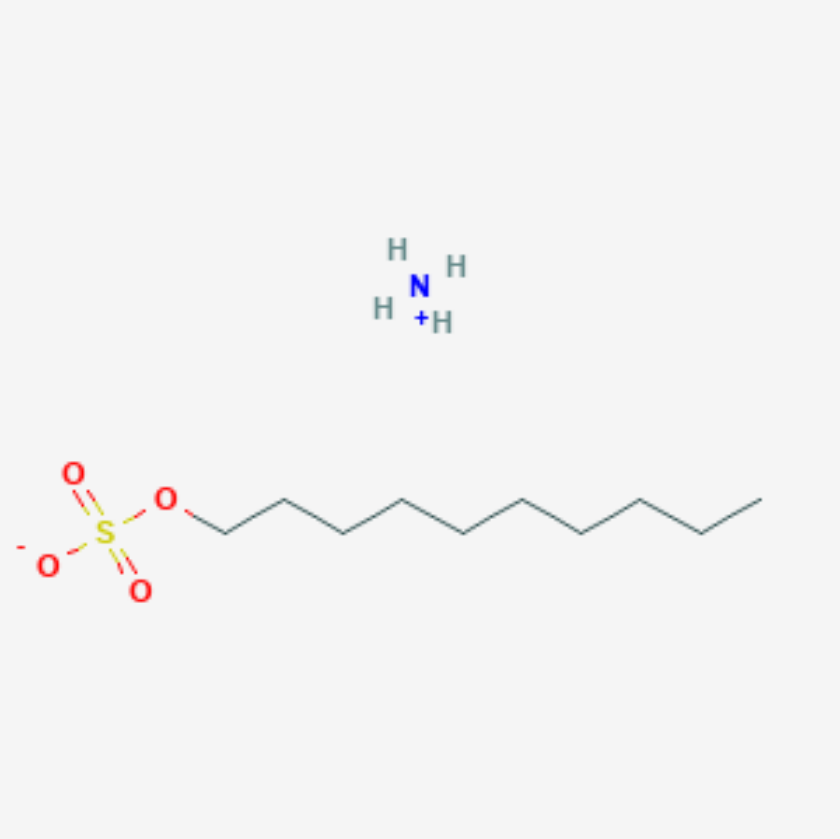
CAS Common Chemistry PubChem Comparison



- Many CAS RN may correspond to same InChI / SMILES
- Each CAS RN may have SMILES, Name, and InChIKey
- Each may be found in PubChem and matched to a chemical record (CID)
- InChIKey-based matching is many to many, due to chemical structure normaliza tion differences, but with the majority (96.7%) one CAS RN to one PubChem CID
- Highlights challenges with chemical structure-based data mapping between resources

Slide credit: Evan Bolton, Jeff Zhang, and Paul Thiessen from the PubChem team





Many PubChem InChIKey-based "many CID to one CAS RN" cases are due to structure drawing differences and PubChem normalization conservativism

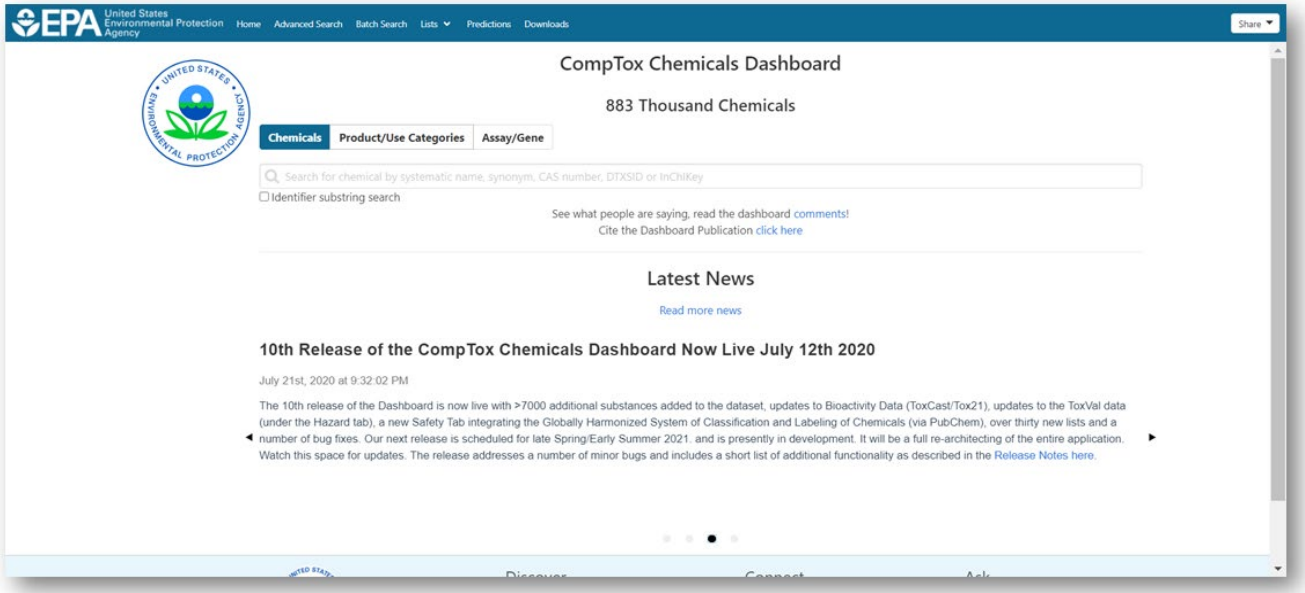
EPA CompTox Chemicals Dashboard

https://comptox.epa.gov/dashboard

The views expressed in this presentation are those of the authors and do not necessarily reflect the views or policies of the US EPA.

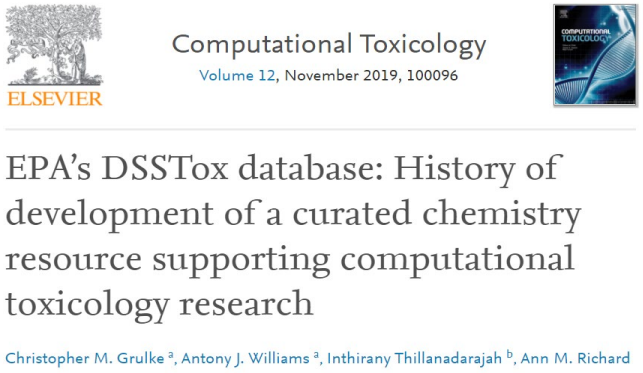
EPA CompTox Chemicals Dashboard https://comptox.epa.gov/dashboard

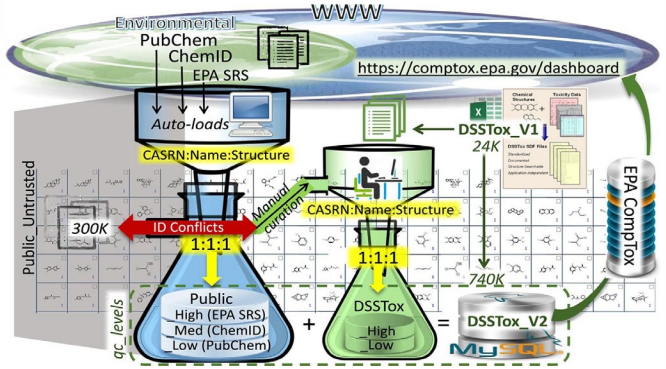
- Data for ~900,000 chemical substances of interest to the agency
- Includes structures and "UVCB chemicals": **U**nknown or **V**ariable Composition, **C**omplex Reaction Products and **B**iological Materials
- Data include experimental and predicted data:
  - Physicochemical properties
  - *In vivo* and *in vitro* hazard data
  - Exposure data – consumer uses
  - Mapped relationships

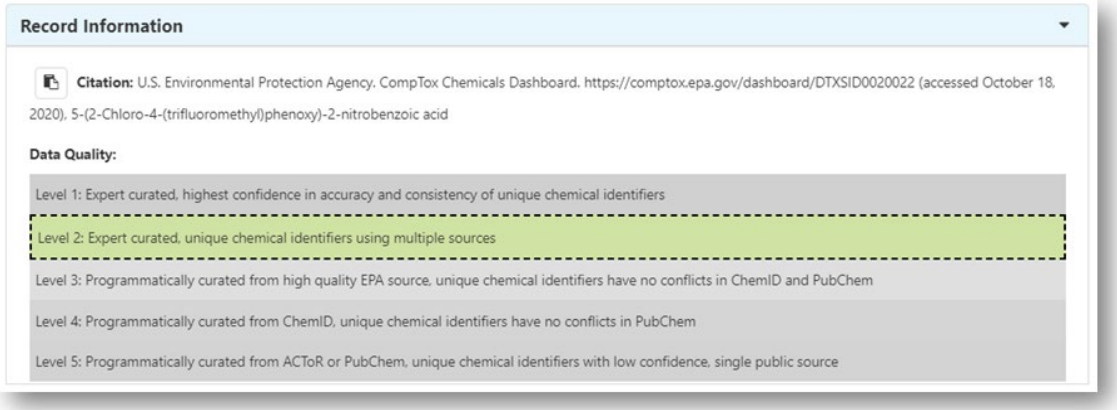


Chemical Registration System

- Dashboard is underpinned by the ChemReg registration system
- Load procedures check for collisions in names, CAS RNs and structures
- Manual curators check relationships and mappings between data
- Relationship mappings include:
  - Monomer to polymer
  - Parent to transformation product (e.g. degradant, metabolite)







Data Observations and Challenges

- Stoichiometry

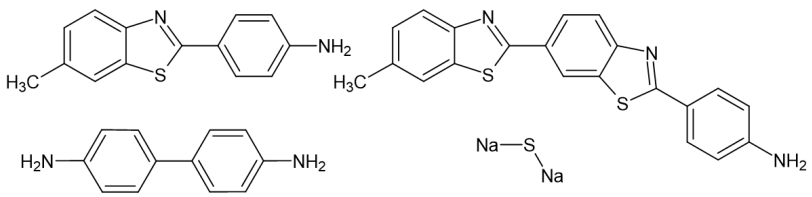
CAS RN	Chemical Name	Isomeric SMILES	InChIKey
142473-50-5	Water-d2, trimer	[2H]O[2H]	XLYOFNOQVPJJNP-ZSJDYOACSA-N
26352-74-9	Water-d2, dimer	[2H]O[2H]	XLYOFNOQVPJJNP-ZSJDYOACSA-N
7789-20-0	Water-d2	[2H]O[2H]	XLYOFNOQVPJJNP-ZSJDYOACSA-N

- Minerals vs. simple salts

1344-48-5	Mercury sulfide (HgS)	S=[Hg]	QXKXDIKCIPXUPL-UHFFFAOYSA-N
19122-79-3	Cinnabarite	S=[Hg]	QXKXDIKCIPXUPL-UHFFFAOYSA-N
1309-56-4	Molybdenite	S=[Mo]=S	CWQXQMHSOZUFJS-UHFFFAOYSA-N
1317-33-5	Molybdenum disulfide	S=[Mo]=S	CWQXQMHSOZUFJS-UHFFFAOYSA-N

- Reaction products

- **CASRN**: 90268-15-8
- **Name**: [1,1'-Biphenyl]-4,4'-diamine, **reaction products with** 4-(6-methyl-2-benzothiazolyl)benzenamine, 4-(6-methyl[2,6'-bibenzothiazol]-2'-yl)benzenamine and sodium sulfide (Na2(Sx))
- **SMILES**: Cc1ccc2c(c1)sc(n2)c3ccc(cc3)N.Cc1ccc2c(c1)sc(n2)c3ccc4c(c3)sc(n4)c5ccc(cc5)N.c1cc(ccc1c2ccc(cc2)N)N.[Na]S[Na]



Data Observations and Challenges

- Polymers

Monomer representation only
- **CASRN**: 69678-94-0
- **Name**: [1,1′-Biphenyl]-4,4′-dicarbonitrile, ***homopolymer***
- **SMILES**: c1cc(ccc1C#N)c2ccc(cc2)C#N (for the monomer)

Polymer representation
- **CASRN**: 108644-22-0
- **Name**: Poly[oxy(hexylmethylsilylene)]
- **SMILES**: *O[Si](*)(C)CCCCCC

- Biologicals

CASRN	Name
81295-09-2	Restriction endodeoxyribonuclease BamHI
80449-04-3	Restriction endodeoxyribonuclease BglI
81295-12-7	Restriction endodeoxyribonuclease BglII
83589-01-9	Restriction endodeoxyribonuclease ClaI

- Alloys
  - It is not possible to interchange details through SMILES for alloys

39344-91-7	Steel, (AISI 1055)	[C].[Si].[P].[S].[Mn].[Fe]	PBCZGXHYZKIUEO-UHFFFAOYSA-N
37268-90-9	Steel 45	[C].[Si].[P].[S].[Mn].[Fe]	PBCZGXHYZKIUEO-UHFFFAOYSA-N
39367-89-0	Steel, (DIN 1.5122)	[C].[Si].[P].[S].[Mn].[Fe]	PBCZGXHYZKIUEO-UHFFFAOYSA-N
288158-84-9	Steel, (DIN 1.5217)	[C].[Si].[P].[S].[V].[Mn].[Fe].[Nb]	UAYKTNKAWYPKPK-UHFFFAOYSA-N

Mapping to our ChemReg QC Levels

- **Highest** quality, manually curated Grade 1 chemicals had >2500 conflicts
- The primary collisions are based on stoichiometry
- 2200 chemicals collided based on No Structures

Wikidata

https://www.wikidata.org/
https://www.wikidata.org/wiki/Wikidata_talk:WikiProject_Chemistry

Matching with Wikidata

- Protocol #1
  1. Check SMILES / InChI consistency (with CDK 2.3)
  2. Search Wikidata by InChIKey
  3. Report CAS mismatches
- Protocol #2
  - Find CAS RN matches in Wikidata
  - Generate QuickStatements
  - Add as references

Results #1

- 1365 InChIs generated from the SMILES has InChIKey mismatches
- 148443 CAS RNs have an InChIKey found in Wikidata
- 8410 InChIKey matches do not have CAS RNs in the Wikidata record
- 7862 InChIKey matches report different CAS RN in Wikidata

Slide credit: Egon Willighagen

Where Wikidata & CAS Common Chemistry do not agree

CAS RN in Wikidata (Q83071553 / FTGVBNYAAXQWQM-UHFFFAOYSA-N) does not match: expected 18924-98-6 but found 50683-27-7

CAS RN in Wikidata (Q82006169 / VCEOZBZYQYAEMK-UHFFFAOYSA-N) does not match: expected 83547-95-9 but found 116821-35-3 --> 7890

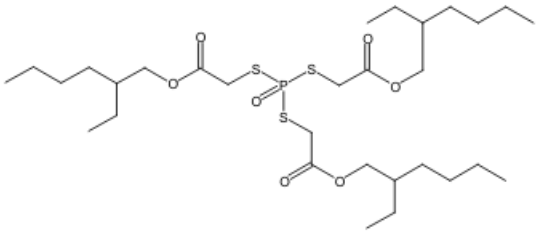
CAS Registry Number
by GZWDer (flood)

116821-35-3

0 references

8-Oxa-3,5-dithia-4-phosphatetradecanoic acid, 10-ethyl-4-[[2-[(2-ethylhexyl)oxy]-2-oxoethyl]thio]-7-oxo-, 2-ethylhexyl ester, 4-oxide

CAS Registry Number®
83547-95-9



CAS Name
8-Oxa-3,5-dithia-4-phosphatetradecanoic acid, 10-ethyl-4-[[2-[(2-ethylhexyl)oxy]-2-oxoethyl]thio]-7-oxo-, 2-ethylhexyl ester, 4-oxide

Molecular Formula
$C_{30}H_{57}O_7PS_3$

Molecular Mass

Other Names and Identifiers

InChI
InChI=1S/C30H57O7PS3/c1-7-13-16-25(10-4)19-35-28(31)22-39-38(34,40-23-29(32)36-20-26(11-5)17-14-8-2)41-24-30(33)37-21-27(12-6)18-15-9-3/h25-27H,7-24H2,1-6H3

InChIKey
VCEOZBZYQYAEMK-UHFFFAOYSA-N

SMILES
CCCCC(CC)COC(=O)CSP(=O)(SCC(=O)OCC(CC)CCCC)SCC(=O)OCC(CC)CCCC

Canonical SMILES
CCCCC(CC)COC(=O)CSP(=O)(SCC(=O)OCC(CC)CCCC)SCC(=O)OCC(CC)CCCC

Other Names for this Substance
- 8-Oxa-3,5-dithia-4-phosphatetradecanoic acid, 10-ethyl-4-[[2-[(2-ethylhexyl)oxy]-2-oxoethyl]thio]-7-oxo-, 2-ethylhexyl ester, 4-oxide

Deleted or Replaced CAS Registry Numbers
116821-35-3

QuickStatements | English | New batch | Last batches | Chat | Git | Help | Egon Willighagen | Your last batches

Batch on Wikidata by Egon Willighagen [Batches]

Status: 0% (0) of 2 done

1	init	aclonifen [Q341945]	ADD	Statement	CAS Registry Number [P231] : "74070-46-5"	
2	init	aclonifen [Q341945]	ADD	Sources to	CAS Registry Number [P231] : "74070-46-5"	stated in [P248]:CAS Common Chemistry [Q18907859] retrieved [P813]:2021-03-31 reference URL [P854]:"https://commonchemistry.cas.org/detail?cas_rn=74070-46-5" InChIKey [P235]:"DDBMQDADIHOWIC-UHFFFAOYSA-N"

First | Page 1 | Last

All errors Init 10

Run | Run in background

CAS Registry Number

by KrBot and SoCalChemBot and Egon Willighagen

74070-46-5 edit

2 references

stated in	Unique Ingredient Identifier
UNII	1762RDA835
language of work or name	English
title	aclonifen (English)
retrieved	14 October 2016

stated in	CAS Common Chemistry
retrieved	31 March 2021
reference URL	https://commonchemistry.cas.org/detail?cas_rn=74070-46-5
InChIKey	DDBMQDADIHOWIC-UHFFFAOYSA-N

Results #2

Adding references to Wikidata

Growth Wikidata validation references

Wikidata Query Service | Examples | Help | More tools | English

```
SELECT ?date (COUNT(?chemical) AS ?count) WHERE {
  ?chemical p:P231 ?casStatement .
  ?casStatement ps:P231 ?cas ;
                prov:wasDerivedFrom ?reference .
  ?reference pr:P248 wd:Q18907859 ;
             pr:P813 ?date .
} GROUP BY ?date ORDER BY ASC(?date)
```

6 results in 341 ms | Code | Download | Link

date	count
26 March 2021	1
31 March 2021	10
1 April 2021	25
2 April 2021	44
3 April 2021	70
4 April 2021	100

https://w.wiki/39bu
https://w.wiki/39h2
https://w.wiki/39jj

Wikipedia

https://en.wikipedia.org/
https://en.wikipedia.org/wiki/Wikipedia_talk:WikiProject_Chemistry

Results #3: Validating Wikipedia

Wikidata Query Service | Examples | Help | More tools

```
SELECT DISTINCT ?date ?chemical ?cas ?link WHERE {
  ?chemical p:P231 ?casStatement .
  ?casStatement ps:P231 ?cas ;
                prov:wasDerivedFrom ?reference .
  ?reference pr:P248 wd:Q18907859 ;
             pr:P813 ?date .
  ?link a            schema:Article ; schema:inLanguage "en" ;
        schema:about ?chemical .
}
```

date	chemical	cas	link
5 April 2021	wd:Q27116093	2477-73-8	<https://en.wikipedia.org/wiki/Hydromadinone_acetate>
5 April 2021	wd:Q27280485	1000025-07-9	<https://en.wikipedia.org/wiki/Vadadustat>
3 April 2021	wd:Q57741744	17795-27-6	<https://commons.wikimedia.org/wiki/Category:Alliin>
31 March 2021	wd:Q180341	59-51-8	<https://commons.wikimedia.org/wiki/Category:Methionine>
31 March 2021	wd:Q180341	59-51-8	<https://en.wikipedia.org/wiki/Methionine>
31 March 2021	wd:Q413598	721-50-6	<https://commons.wikimedia.org/wiki/Category:Prilocaine>
31 March 2021	wd:Q413598	721-50-6	<https://en.wikipedia.org/wiki/Prilocaine>
3 April 2021	wd:Q5047474	50935-04-1	<https://en.wikipedia.org/wiki/Carubicin>
3 April 2021	wd:Q4639569	286834-81-9	<https://en.wikipedia.org/wiki/5-APB>
1 April 2021	wd:Q3473757	142001-63-6	<https://en.wikipedia.org/wiki/Saredutant>
3 April 2021	wd:Q2823221	2905-86-4	<https://en.wikipedia.org/wiki/Beta-Ureidoisobutyric_acid>
3 April 2021	wd:Q3629783	495-02-3	<https://en.wikipedia.org/wiki/Auraptene>
31 March 2021	wd:Q416667	520-85-4	<https://commons.wikimedia.org/wiki/Category:Medroxyprogesterone>
31 March 2021	wd:Q416667	520-85-4	<https://en.wikipedia.org/wiki/Medroxyprogesterone>
1 April 2021	wd:Q27096705	843-55-0	<https://en.wikipedia.org/wiki/Bisphenol_Z>

https://w.wiki/3A8C

72 links validated, 15 suspicious pages:
- category pages
- wrong stereoisomer

Wikipedia: ChemBox/DrugBox updated

Template talk:Chembox CASNo/format

From Wikipedia, the free encyclopedia
< Template talk:Chembox CASNo

Template This template is within the scope of **WikiProject Chemistry**, a collaborative effort to improve the coverage of chemistry on Wikipedia. If you would like to participate, please visit the project page, where you can join the discussion and see a list of open tasks.

This template does not require a rating on the project's quality scale.

Template This template is within the scope of ***WikiProject Chemicals***, a daughter project of **WikiProject Chemistry**, which aims to improve Wikipedia's coverage of **chemicals**. To participate, help improve this template or visit the project page for details on the project.

This template does not require a rating on the project's quality scale.

Template-protected edit request on 2 April 2021 [edit]

This edit request has been answered. Set the `|answered=` or `|ans=` parameter to **no** to reactivate your request.

I like to update the URLs to match the new CAS Common Chemistry database. Egon Willighagen (talk) 13:07, 2 April 2021 (UTC)

@Egon Willighagen:: {{Chembox CASNo/format/sandbox}} now has a copy of current code. You can edit & check it as you like. When OK, pls. reactivate the Edit Request (by setting `|answered=no`). Tests can be entered in /testcases2#CAS_number. -DePiep (talk) 15:11, 2 April 2021 (UTC)

~~I have used~~ ~~`https://commonchemistry.cas.org/results?q=`~~ ~~(for example [1]). (diff).~~ Is that the right approach? -DePiep (talk) 15:53, 2 April 2021 (UTC)

More background: d:Wikidata_talk:WikiProject_Chemistry#Validation_of_CAS_numbers;_collaboration_with_Wikipedia? Wikidata. -DePiep (talk) 15:56, 2 April 2021 (UTC)

Using `https://commonchemistry.cas.org/detail?cas_rn=` (for example [2]). (diff). Is this the ANI? -DePiep (talk) 16:00, 2 April 2021 (UTC)

Yes, that looks right. --Egon Willighagen (talk) 19:04, 2 April 2021 (UTC)

Tests added to /testcases2#CAS_number. While some links might look slow or non-effective, the new URL is an improvement (which had more useless links, I'd say). Request reactivated. -DePiep (talk) 20:31, 2 April 2021 (UTC)

Request reactivated

Please replace all code from .../format/sandbox into live code: diff.

Change: using URL for new public CAS publication. -DePiep (talk) 20:31, 2 April 2021 (UTC)

To DePiep and Egon Willighagen: ✓ **done**, and thank you both very much! *P.I. Ellsworth* ed. put'r there 10:59, 3 April 2021 (UTC)

Conclusions

- CC BY-NC 4.0 License and API (*) allows community validation for 500 thousand CAS Registry Numbers

- There are a lot of CAS Registry Numbers consistent between resources
- There is more difficult chemistry (polymers, solutions, mixtures, …)
  - Covered by chemistry represented in database?
  - Can the intermediate representation (SMILES, InChIKey) do the matching?
- EPA CompTox Chemicals Dashboard (53% CAS RNs known)
- Wikidata / Wikipedia
  - Communities involved: CAS RN mismatches reported and discussed
  - English Wikipedia's ChemBox and DrugBox links to CAS Common Chemistry updated
  - Wikidata: validated CAS Registry Numbers are being annotated
    - Can be used to validate other open databases

(*) we used a spreadsheet instead of the API to speed up things

Advancing the accuracy of open chemical information with CAS Common Chemistry

Egon L. Willighagen (Maastricht Uni/NL), @egonwillighagen, https://orcid.org/0000-0001-7542-0286
Antony J. Williams (US-EPA), @ChemConnector, https://orcid.org/0000-0002-2668-4821
Christopher Grulke (US-EPA)
Ann Richard (US-EPA)
Andrea Jacobs (CAS, a division of the American Chemical Society)

Wednesday

CAS common chemistry and the value of community collaboration for chemical informatics

11:50pm - 12:10am CET – Central European Time - April 15, 2021

Andrea Jacobs | Evan Bolton | Stuart Chalk | Simon Coles | Jeremy Frey | Katherine Hickey | Bonnie Lawlor | Connor McClellan | Leah McEwen | Nathan Patrick | Adam Sanford | Martin Walker | Antony Williams | Dustin Williams | Egon Willighagen

View Archive | No Evaluation | Add to Calendar | Remove from Itinerary