Does bigger mean better in the world of chemistry databases?

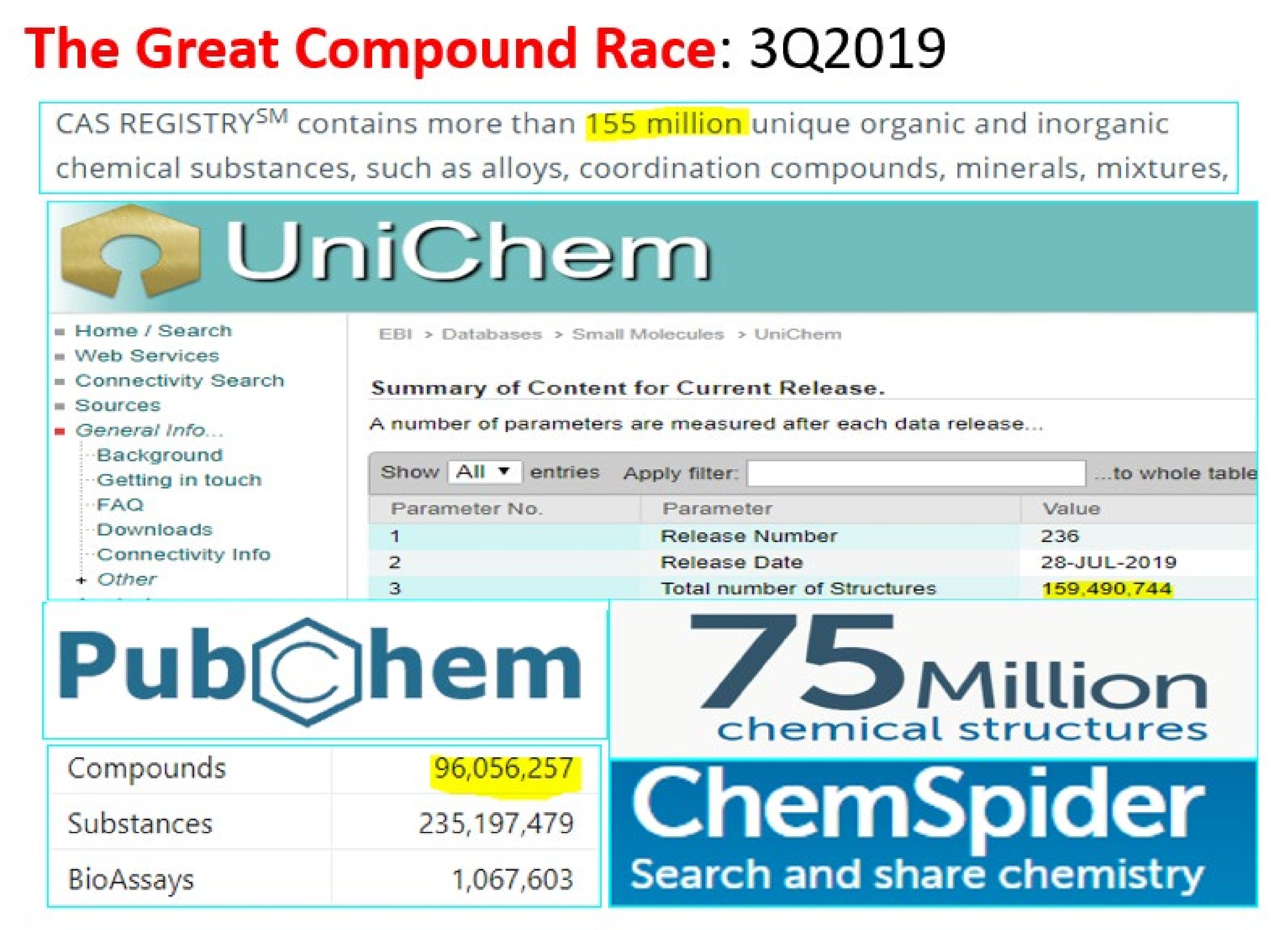
Antony Williams[1] and Christopher Southan[2]

1) National Center for Computational Toxicology, U.S. Environmental Protection Agency, RTP, NC, USA
ORCID ID 0000-0002-2668-4821

2) TW2Informatics Ltd, Göteborg, Sweden 42166
ORCID ID 0000-0001-9580-0446

Views expressed in this presentation are those of the authors and do not necessarily reflect the views or policies of the U.S. EPA

The Good News….



The Great Compound Race: 3Q2019

CAS REGISTRY^{SM} contains more than 155 million unique organic and inorganic chemical substances, such as alloys, coordination compounds, minerals, mixtures,

Summary of Content for Current Release.

A number of parameters are measured after each data release...

Parameter No.	Parameter	Value
1	Release Number	236
2	Release Date	28-JUL-2019
3	Total number of Structures	159,490,744

Compounds	96,056,257
Substances	235,197,479
BioAssays	1,067,603

75 Million chemical structures

The Good News….

- We've never had it so good (UniChem~160 million)

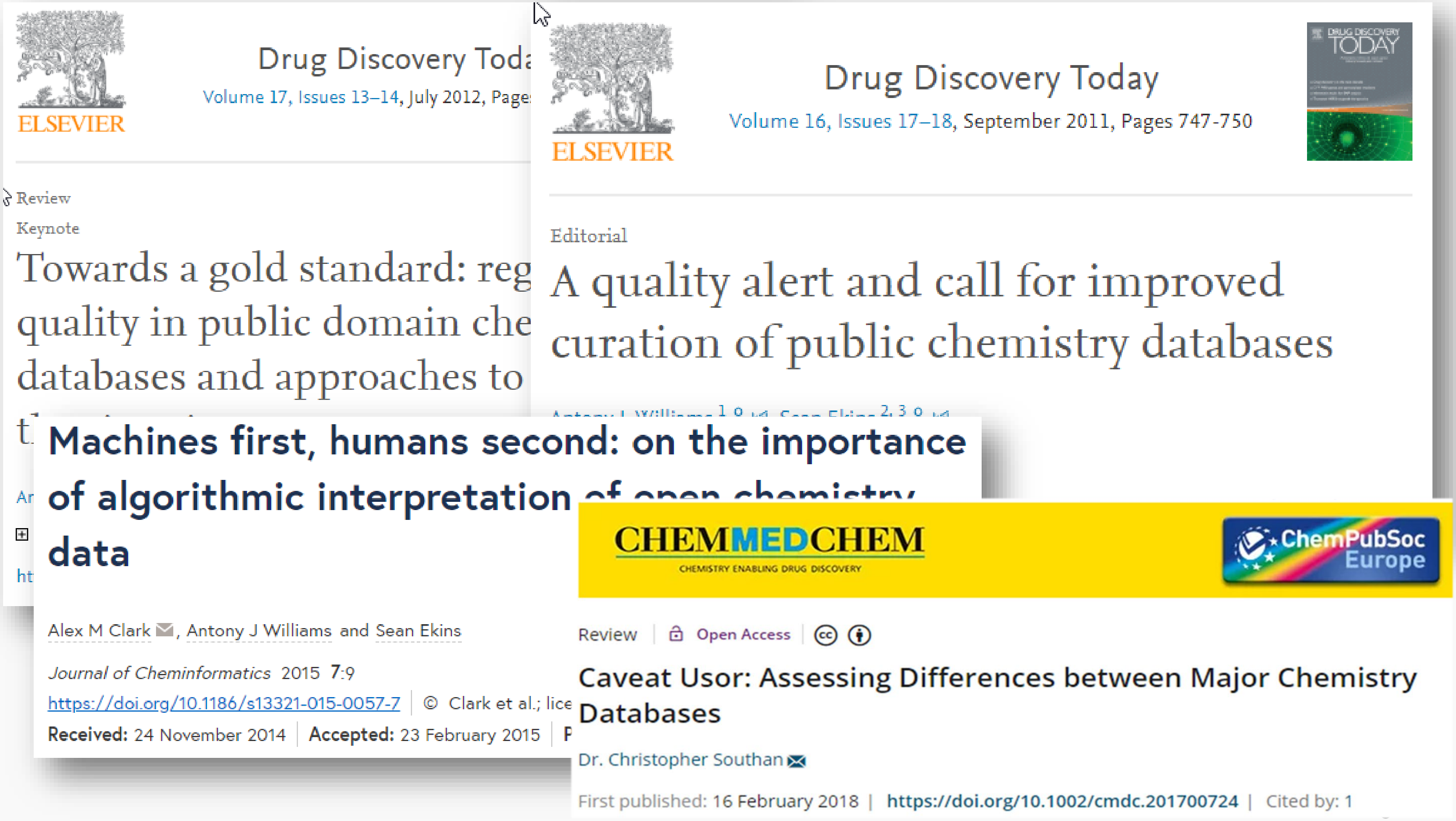
- Sustained growth - since 2Q2017
  - Scifinder +25 million
  - ChemSpider +16 million
  - UniChem +14 million
  - PubChem +6 million

- Massively enabling for chemistry and bioactivity

- All four should be **congratulated**! Public databases in particular (where InChI is the great enabler)

Data quality in public domain databases is challenging…

- Data quality in free web-based databases!



Drug Discovery Toda
Volume 17, Issues 13–14, July 2012, Page

Review
Keynote

Towards a gold standard: reg
quality in public domain che
databases and approaches to
t

Ar
ht

Drug Discovery Today
Volume 16, Issues 17–18, September 2011, Pages 747-750

Editorial

A quality alert and call for improved curation of public chemistry databases

Antony J. Williams 1, 9, Sean Ekins 2, 3, 9

Machines first, humans second: on the importance of algorithmic interpretation of open chemistry data

Alex M Clark, Antony J Williams and Sean Ekins

Journal of Cheminformatics 2015 7:9
https://doi.org/10.1186/s13321-015-0057-7 | © Clark et al.; lice
Received: 24 November 2014 | **Accepted:** 23 February 2015 | P

CHEMMEDCHEM
CHEMISTRY ENABLING DRUG DISCOVERY

ChemPubSoc Europe

Review | Open Access

Caveat Usor: Assessing Differences between Major Chemistry Databases

Dr. Christopher Southan

First published: 16 February 2018 | https://doi.org/10.1002/cmdc.201700724 | Cited by: 1

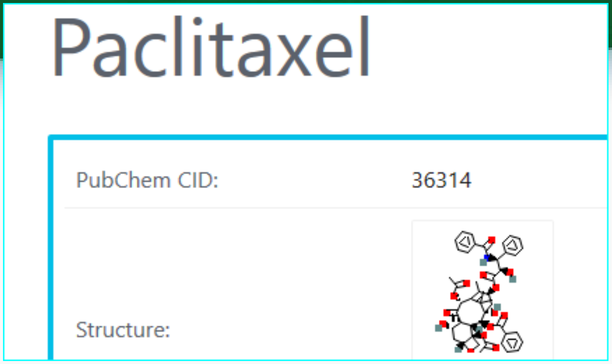
Database Quality and Noise

- Intuitively understood but difficult to quantitate

- Some aspects inherently cheminformatically **challenging** (e.g. Tautomer handling, Kekulisation of complex cycles, atroposiomers, exotic metalloorganic compounds, challenging layout and renderings)

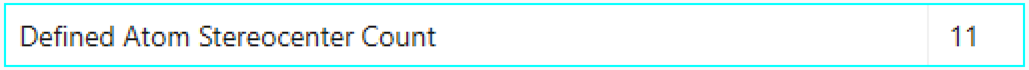
- Other challenges are just **difficult** (e.g. which stereo enumerations were experimentally confirmed or did the bioassays use undefined racemates)

Taxol (Paclitaxel) is noisy….

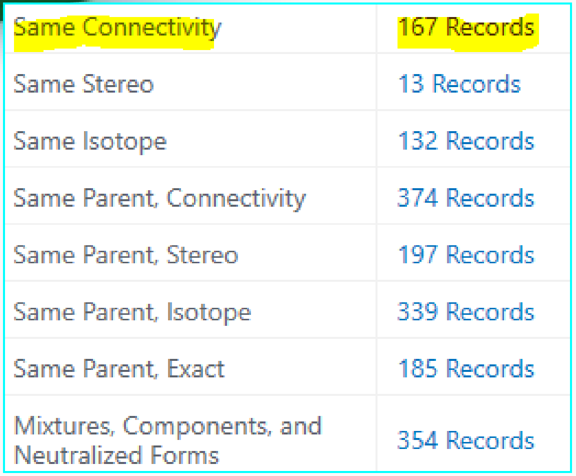


Paclitaxel

PubChem CID:	36314
Structure:	



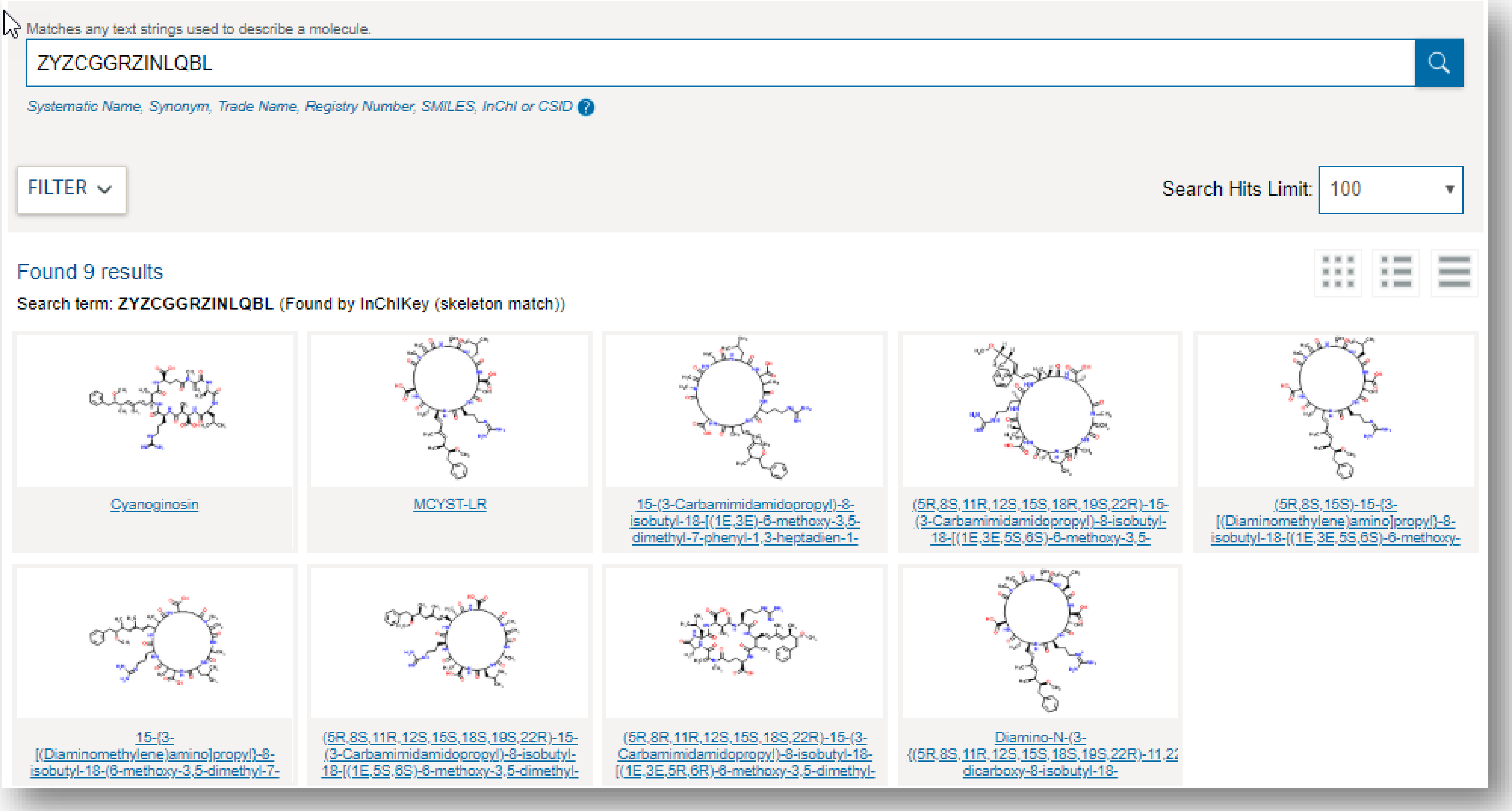
Defined Atom Stereocenter Count	11



Same Connectivity	167 Records
Same Stereo	13 Records
Same Isotope	132 Records
Same Parent, Connectivity	374 Records
Same Parent, Stereo	197 Records
Same Parent, Isotope	339 Records
Same Parent, Exact	185 Records
Mixtures, Components, and Neutralized Forms	354 Records

- CID 36314 most "popular" with **304 singleton submissions and 532 mixtures**
- First submitted by NIAID on 2004-09-15 as SID: 598380 (but is it correct?)
- 154 have different stereo (some MAY be correctly synthesized)
- 34 have different isotopes and 12 of these have same stereo
- 532 mixture SIDs merge to 354 distinct CID mixtures and components
- 66 vendors will sell you CID 36314
- 59 vendors will sell you one of the other 166 different CIDs
- Sigma-Aldrich **submitted the identical structure 10 times** (as different SIDs)
- ZINC links to vendors for 17 of the 167
- 64 of 167 CIDs are single-sources, 17 of which are vendors
- **12 of 167** CIDs include **RN 33069-62-4** as a synonym
- **12 of 167** are flagged as active in **different** BioAssays

Will the correct Microcystin LR Stand Up? ChemSpider Skeleton Search



Matches any text strings used to describe a molecule.

ZYZCGGRZINLQBL

Systematic Name, Synonym, Trade Name, Registry Number, SMILES, InChI or CSID

FILTER

Search Hits Limit: 100

Found 9 results

Search term: **ZYZCGGRZINLQBL** (Found by InChIKey (skeleton match))

Cyanoginosin

MCYST-LR

15-(3-Carbamimidamidopropyl)-8-isobutyl-18-[(1E,3E)-6-methoxy-3,5-dimethyl-7-phenyl-1,3-heptadien-1-

(5R,8S,11R,12S,15S,18R,19S,22R)-15-(3-Carbamimidamidopropyl)-8-isobutyl-18-[(1E,3E,5S,6S)-6-methoxy-3,5-

(5R,8S,15S)-15-{3-[(Diaminomethylene)amino]propyl}-8-isobutyl-18-[(1E,3E,5S,6S)-6-methoxy-

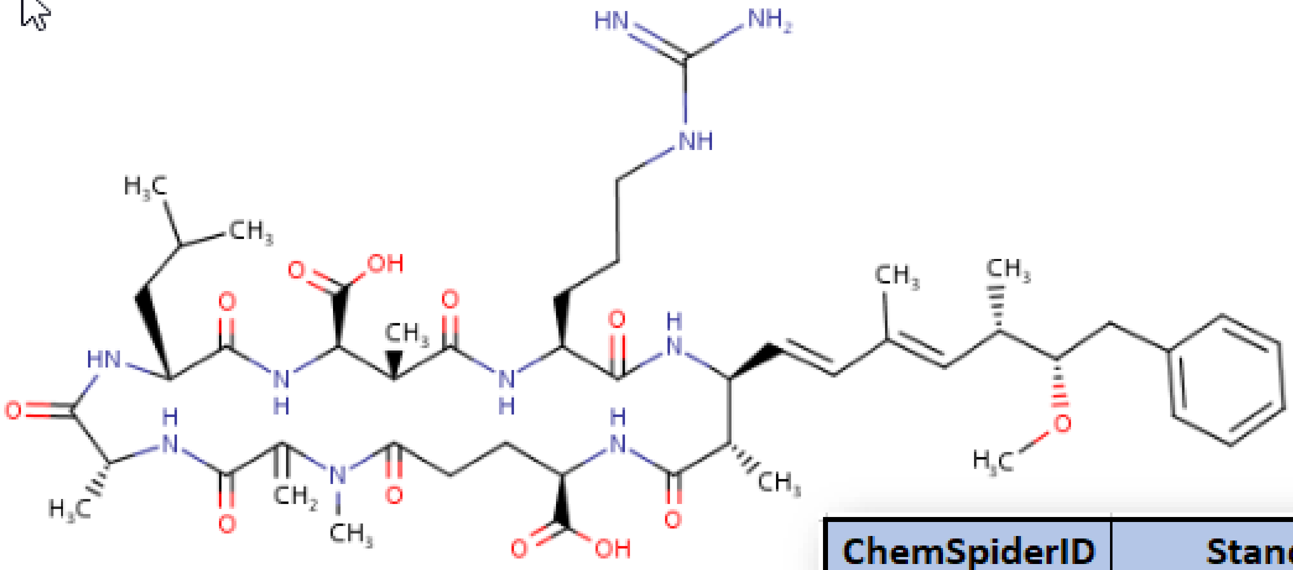
15-{3-[(Diaminomethylene)amino]propyl}-8-isobutyl-18-(6-methoxy-3,5-dimethyl-7-

(5R,8S,11R,12S,15S,18S,19S,22R)-15-(3-Carbamimidamidopropyl)-8-isobutyl-18-[(1E,5S,6S)-6-methoxy-3,5-dimethyl-

(5R,8R,11R,12S,15S,18S,22R)-15-(3-Carbamimidamidopropyl)-8-isobutyl-18-[(1E,3E,5R,6R)-6-methoxy-3,5-dimethyl-

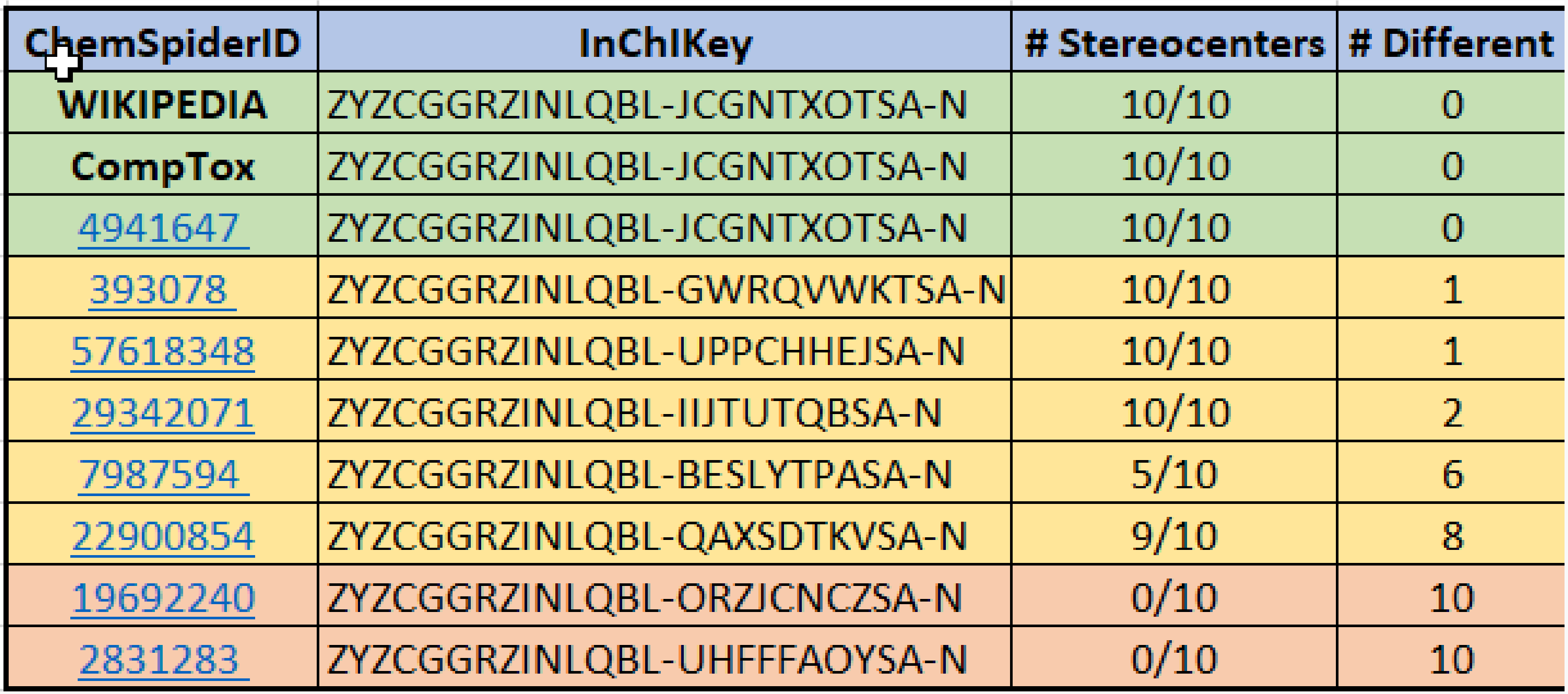
Diamino-N-(3-{(5R,8S,11R,12S,15S,18S,19S,22R)-11,2 dicarboxy-8-isobutyl-18-

Comparing ChemSpider Structures



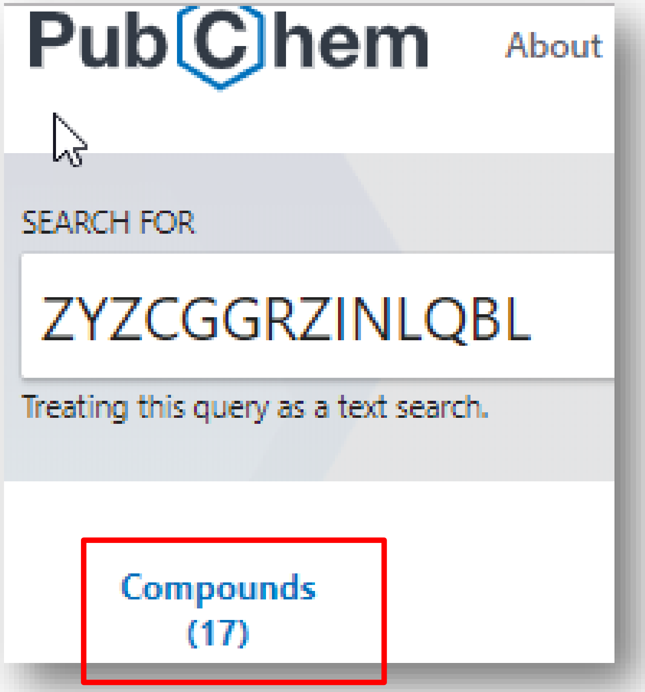
ChemSpiderID	Standard InChIKey Stereolayer
WIKIPEDIA	t28-,29-,30-,31+,34-,35-,36+,37+,38-,40+
CompTox	t28-,29-,30-,31+,34-,35-,36+,37+,38-,40+
4941647	t28-,29-,30-,31+,34-,35-,36+,37+,38-,40+
393078	t28-,29-,30-,31+,34-,35-,36+,**37-**,38-,40+
57618348	t28-,29-,30-,31+,34-,35-,36+,**37-**,38-,40+
29342071	t28-,29-,30-,31+,**34+**,35-,36+,**37-**,38-,40+
7987594	t28-,**29?**,**30?**,31+,**34?**,35-,**36?**,**37-**,38-,**40?**
22900854	t28-,**29?**,**30+**,**31-**,**34+**,**35+**,**36-**,**37-**,38-,**40-**
19692240	NONE
2831283	NONE

Comparing ChemSpider Structures



ChemSpiderID	InChIKey	# Stereocenters	# Different
WIKIPEDIA	ZYZCGGRZINLQBL-JCGNTXOTSA-N	10/10	0
CompTox	ZYZCGGRZINLQBL-JCGNTXOTSA-N	10/10	0
4941647	ZYZCGGRZINLQBL-JCGNTXOTSA-N	10/10	0
393078	ZYZCGGRZINLQBL-GWRQVWKTSA-N	10/10	1
57618348	ZYZCGGRZINLQBL-UPPCHHEJSA-N	10/10	1
29342071	ZYZCGGRZINLQBL-IIJTUTQBSA-N	10/10	2
7987594	ZYZCGGRZINLQBL-BESLYTPASA-N	5/10	6
22900854	ZYZCGGRZINLQBL-QAXSDTKVSA-N	9/10	8
19692240	ZYZCGGRZINLQBL-ORZJCNCZSA-N	0/10	10
2831283	ZYZCGGRZINLQBL-UHFFFAOYSA-N	0/10	10

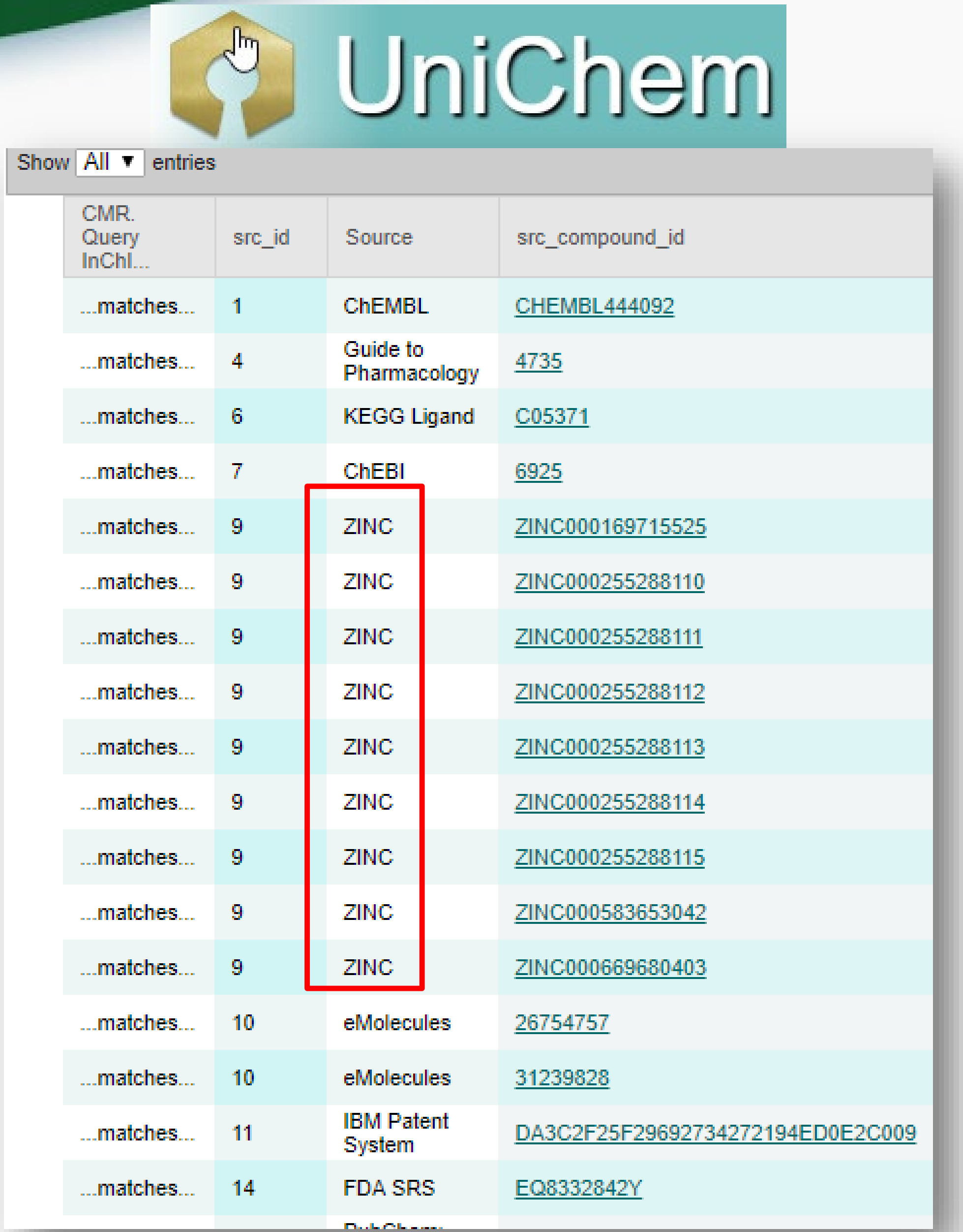
Other Searches



ZYZCGGRZINLQBL

Treating this query as a text search.

Compounds (17)

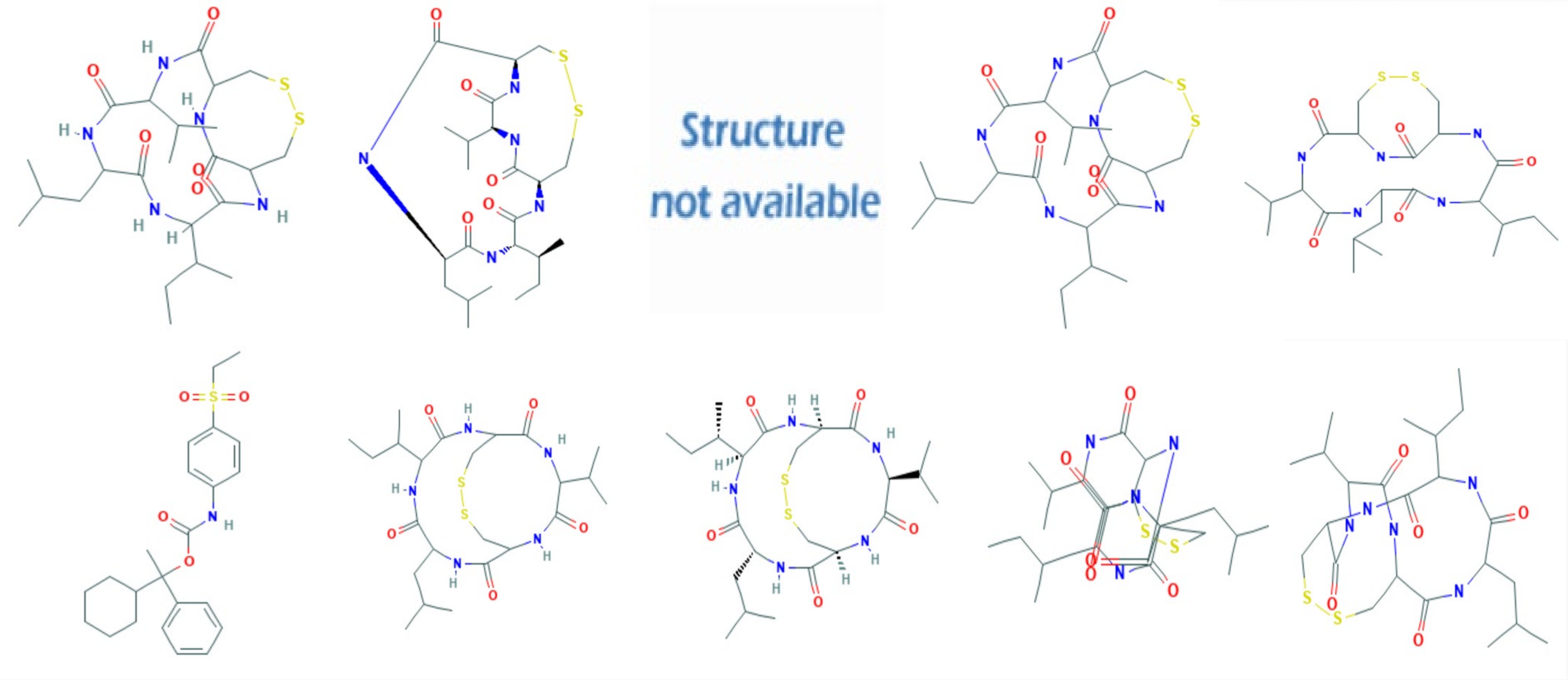


Show All ▾ entries

CMR. Query InChI...	src_id	Source	src_compound_id
...matches...	1	ChEMBL	CHEMBL444092
...matches...	4	Guide to Pharmacology	4735
...matches...	6	KEGG Ligand	C05371
...matches...	7	ChEBI	6925
...matches...	9	ZINC	ZINC000169715525
...matches...	9	ZINC	ZINC000255288110
...matches...	9	ZINC	ZINC000255288111
...matches...	9	ZINC	ZINC000255288112
...matches...	9	ZINC	ZINC000255288113
...matches...	9	ZINC	ZINC000255288114
...matches...	9	ZINC	ZINC000255288115
...matches...	9	ZINC	ZINC000583653042
...matches...	9	ZINC	ZINC000669680403
...matches...	10	eMolecules	26754757
...matches...	10	eMolecules	31239828
...matches...	11	IBM Patent System	DA3C2F25F29692734272194ED0E2C009
...matches...	14	FDA SRS	EQ8332842Y

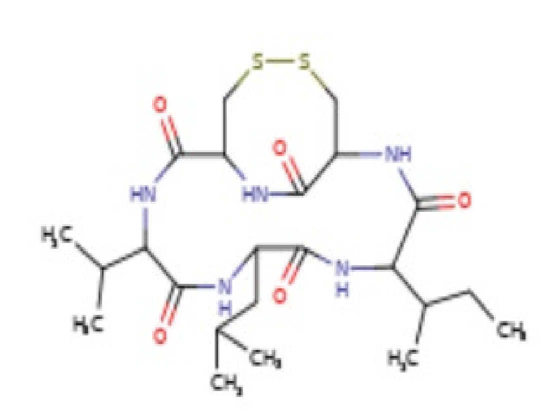
CASRN 3022-92-2 on PubChem

https://pubchem.ncbi.nlm.nih.gov/#query=3022-92-2

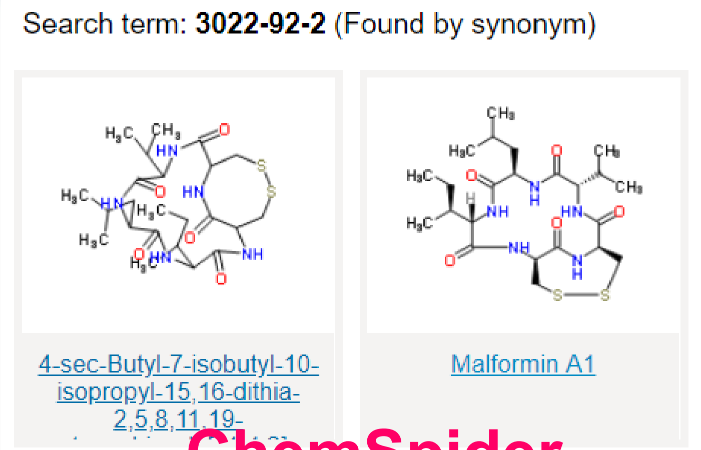


FOUR Different structures, THREE different skeletons

Comparisons…



ChemIDPlus

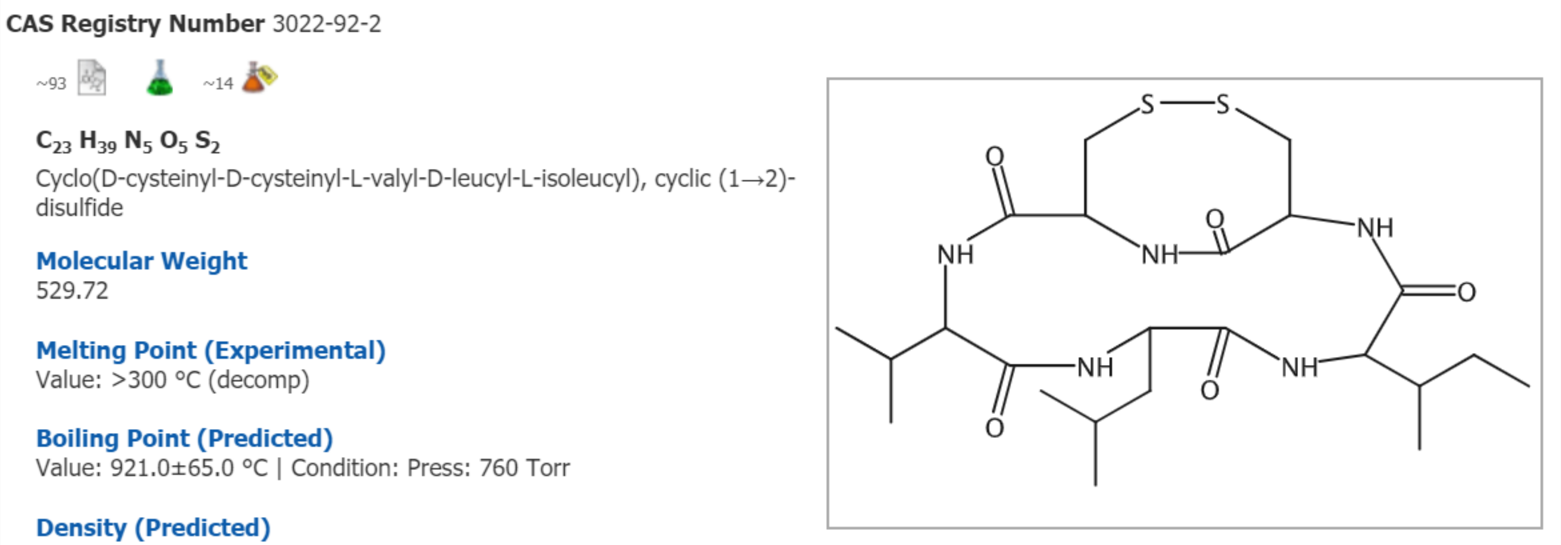


Search term: **3022-92-2** (Found by synonym)

4-sec-Butyl-7-isobutyl-10-isopropyl-15,16-dithia-2,5,8,11,19-

Malformin A1

ChemSpider



CAS Registry Number 3022-92-2

~93 ~14

$C_{23}H_{39}N_5O_5S_2$

Cyclo(D-cysteinyl-D-cysteinyl-L-valyl-D-leucyl-L-isoleucyl), cyclic (1→2)-disulfide

Molecular Weight

529.72

Melting Point (Experimental)

Value: >300 °C (decomp)

Boiling Point (Predicted)

Value: 921.0±65.0 °C | Condition: Press: 760 Torr

Density (Predicted)

SciFinder

Database Quality and Noise

- Common problems: source errors for **CAS-RN mappings**, name-to-structure **conversion errors**, authors ignoring **IUPAC rules** for chemical naming

- We **accept** some intrinsically noisy sources for their **value** compromise (e.g. large vendor aggregations and automated document extraction feeds)

- Some databases index substances without structures: antibodies, large peptides and molasses – not currently mappable but may have linked data

Known issues with public databases (1)

- Different sets of chemistry rules and submission filters
- Operations seem to be focussed on data expansion but less effort into quality
- No inter-resource intersection statistics
- Some useful boutique databases do not submit
- Massive coverage gaps from the literature are not extracted into the public databases
- Coverage gaps from non-document sources (e.g. open drug discovery ELNs)
- Not all are fully open, searchable and downloadable

Known issues with public databases (II)

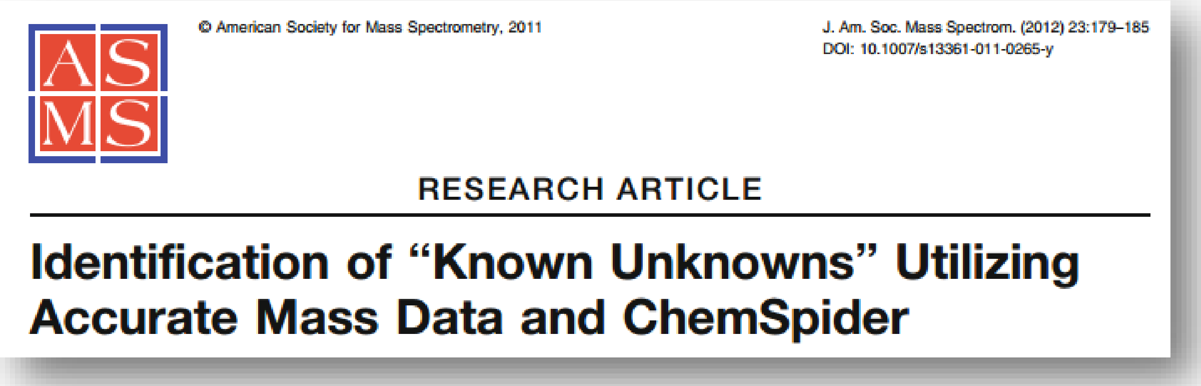
- Unknown extent of contamination by virtuals
- Confounding circularity – identical submissions between systems, with consequent degradation of mappings
- Expert chemical curation, biocuration and crowd-source fixing does not scale
- Public databases are susceptable to exploitation by opportunistic and low-quality submitters
- Large databases aggregate different types of errors
- No real indication of collaboration between the public databases to solve the issues of data quality

Quality has many aspects

- Getting structures to round-trip (Molfile, IUPAC, SMILES, InChI String and Keys all concordant and rendered at least reasonably) – but no surprise
  - Issues of v2000/v3000 exchange and molfiles imperfect
  - InChI is powerful but imperfect and extensions are underway
  - Manually generated IUPAC Names can be very low quality
- Submission filtering rules to ensure plausible structures (e.g. ”Chessboardanes”)
- Tracking molecular ”multiplexing” (i.e. InChIKey inner layer)
- Automated document extraction of chemistry is noisy (SureChEMBL, IBM, Springer, Thieme)

Applications of public databases to non-targeted analysis

- Non-targeted analysis for structure identification and forensics analysis
- Number of hits retrieved based on mass/formula searches explodes based on poorly represented chemicals – especially stereo issues
- The number of hits makes it much harder to rank candidate collections based on meta-data



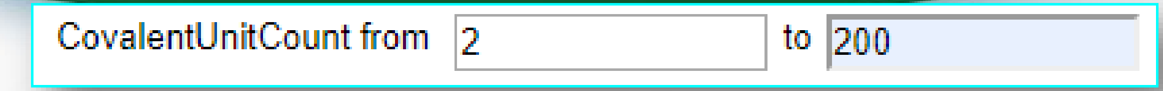
© American Society for Mass Spectrometry, 2011

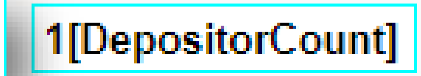
J. Am. Soc. Mass Spectrom. (2012) 23:179–185
DOI: 10.1007/s13361-011-0265-y

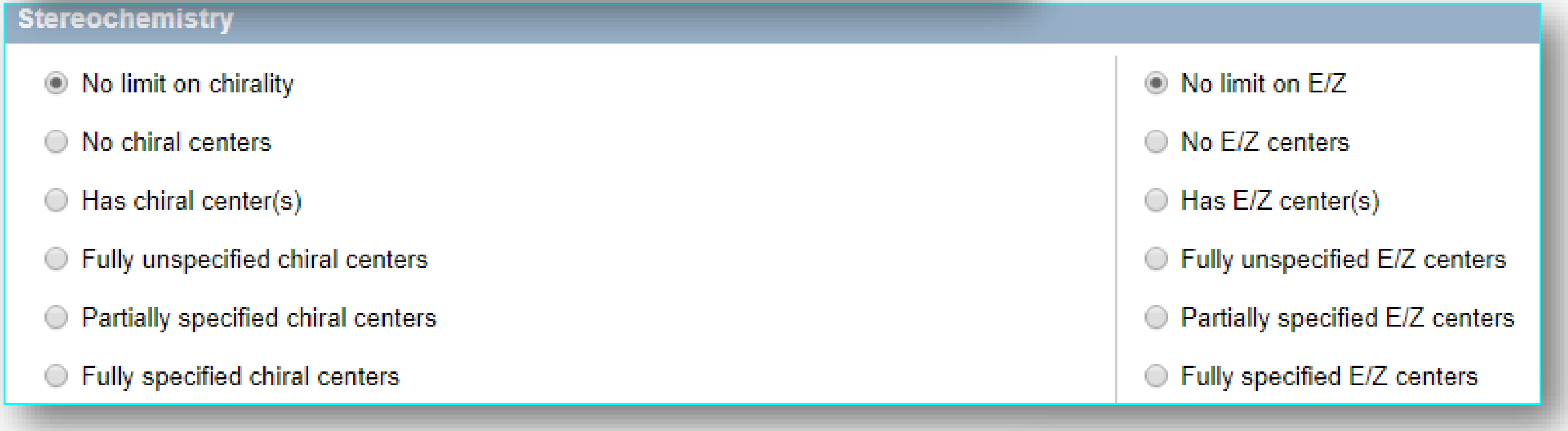
RESEARCH ARTICLE

Identification of "Known Unknowns" Utilizing Accurate Mass Data and ChemSpider

Quantifying noise in PubChem
No other database offers this!


CovalentUnitCount from 2 to 200


1[DepositorCount]

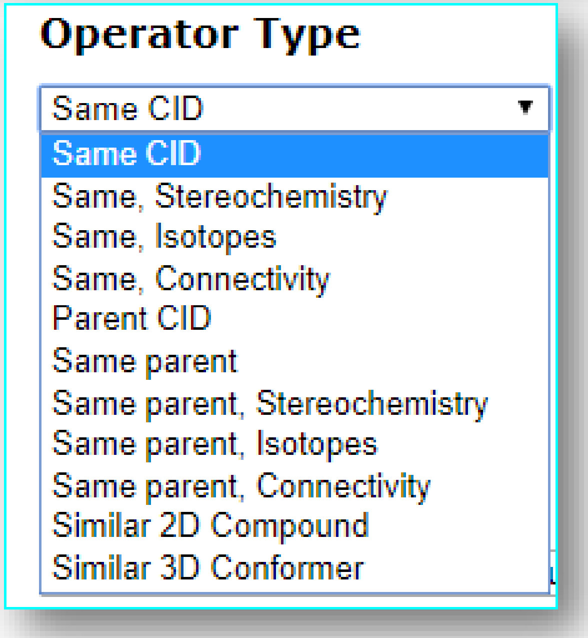

Stereochemistry

● No limit on chirality
○ No chiral centers
○ Has chiral center(s)
○ Fully unspecified chiral centers
○ Partially specified chiral centers
○ Fully specified chiral centers

● No limit on E/Z
○ No E/Z centers
○ Has E/Z center(s)
○ Fully unspecified E/Z centers
○ Partially specified E/Z centers
○ Fully specified E/Z centers

PubChem chemistry rules not perfect but are transparent and can be sliced and diced in useful detail, e.g.

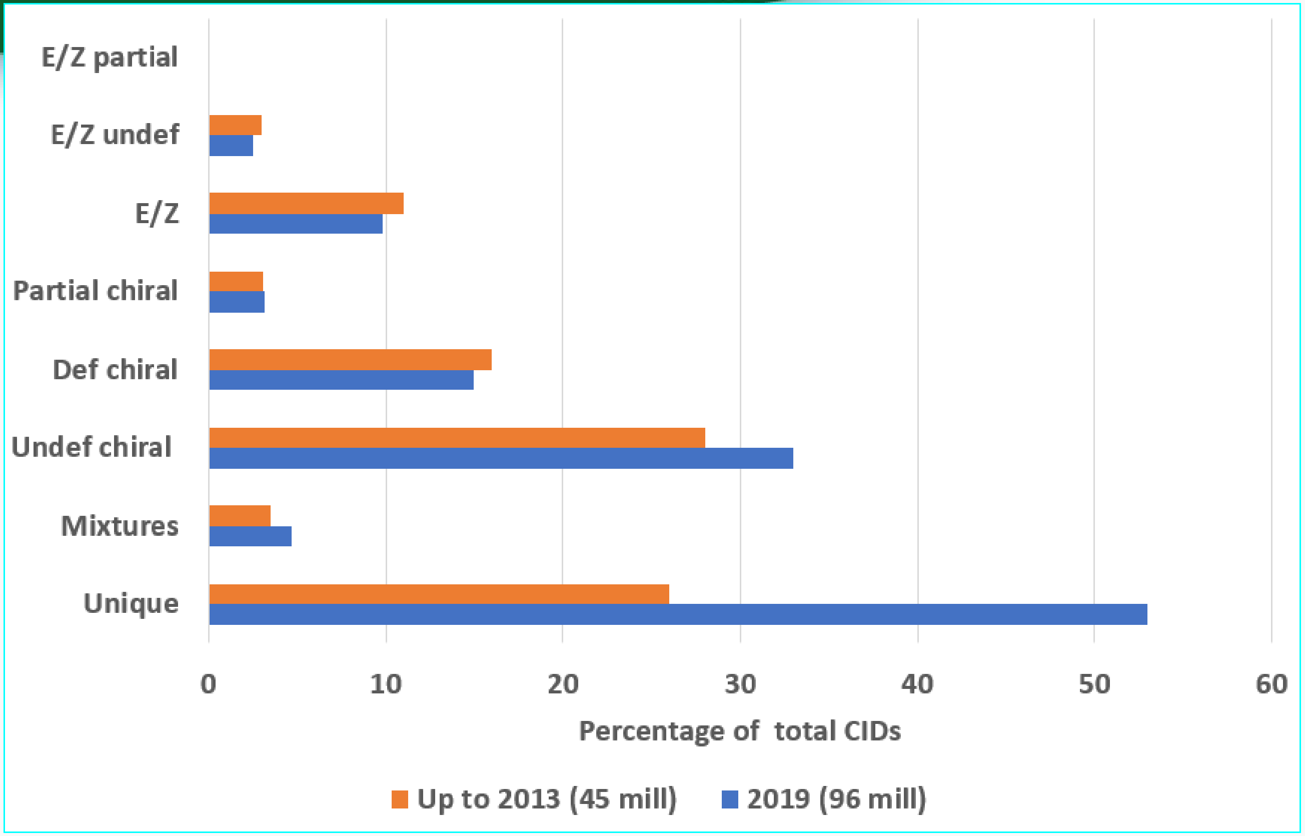
- Mixture counts (covalent units <1)
- Explicit interogation of stereo
- Counts of unique structures (single-source)
- Relationship mapping via individual entries and the PubChem Identifier Exchange Service (up to ~5K)
- These types of stats are informative but should not be overinterpreted


Operator Type

Same CID ▼

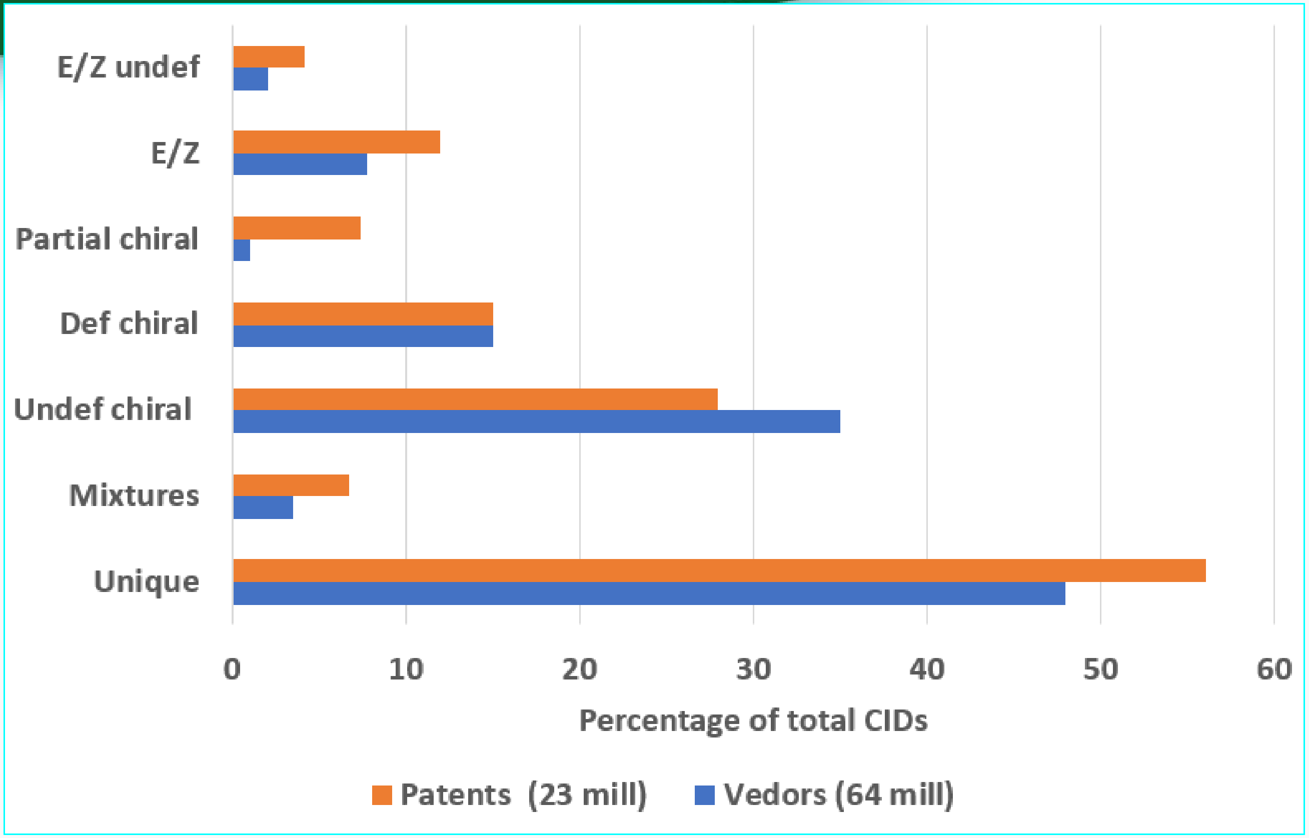
Same CID
Same, Stereochemistry
Same, Isotopes
Same, Connectivity
Parent CID
Same parent
Same parent, Stereochemistry
Same parent, Isotopes
Same parent, Connectivity
Similar 2D Compound
Similar 3D Conformer

Surprising result (I)



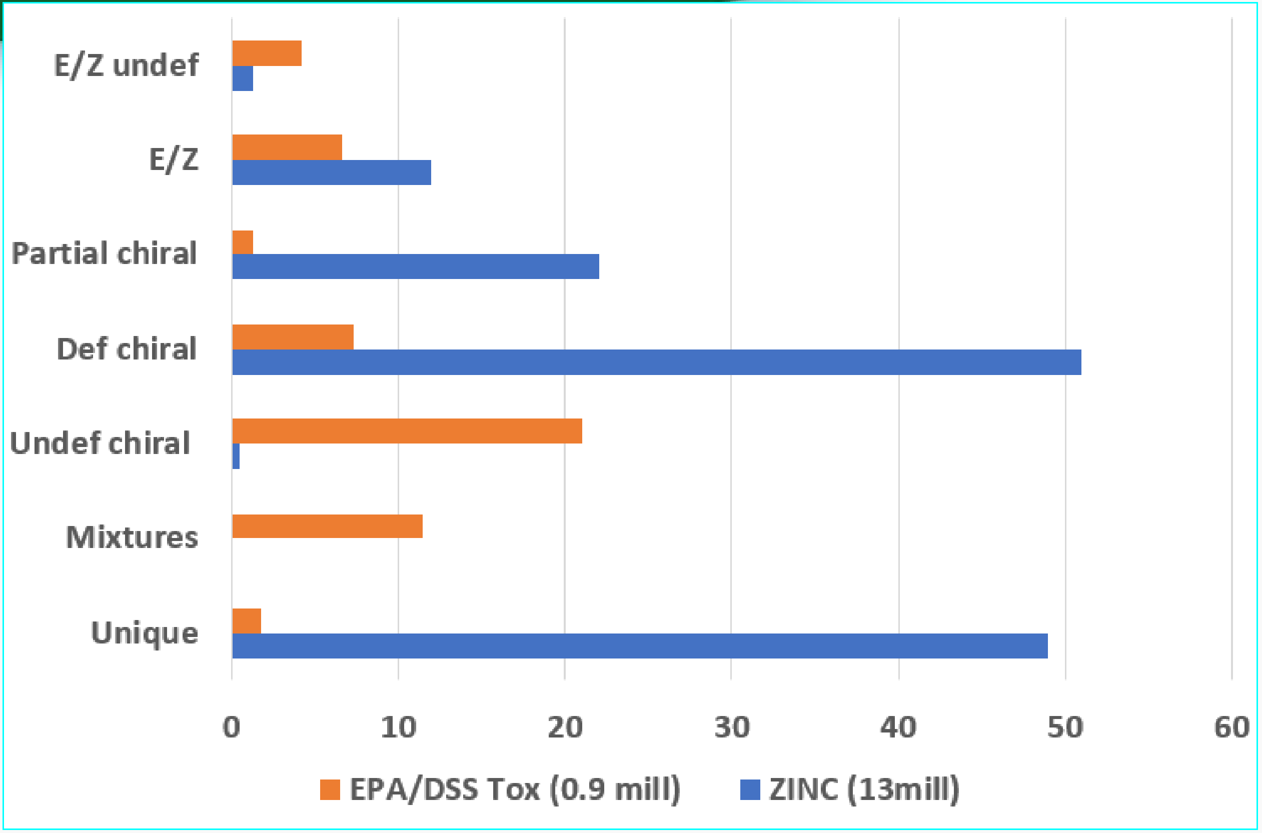
- A big increase in unique single-source content
- Judging by metrics above PubChem has not changed much from doubling in content since 2013
- Except big < uniqueness plus slight < undefined chirality

Surprising result (II)



- Patents high in mixtures
- Vendors low for partial chirality
- Uniqueness in patents is underestimated (i.e. millions of structures extracted by SureChEMBL and IBM but only those two)

Not such a surprising result



- Sources can be quite different e.g. comparison between ZINC and EPA/DSSTox above
- ZINC virtually enumerates stereo which < uniqueness
- The intersect is 275,000 CIDs

Challenges with making improvements

- No quick fixes – we've been discussing it for over a decade...
- ***Acknowledging*** quality and noise issues gives us a chance of not being confounded by them
- But this is problematic for less experienced users
- PubChem allows you to filter just about anything, either pre- or post-analysis

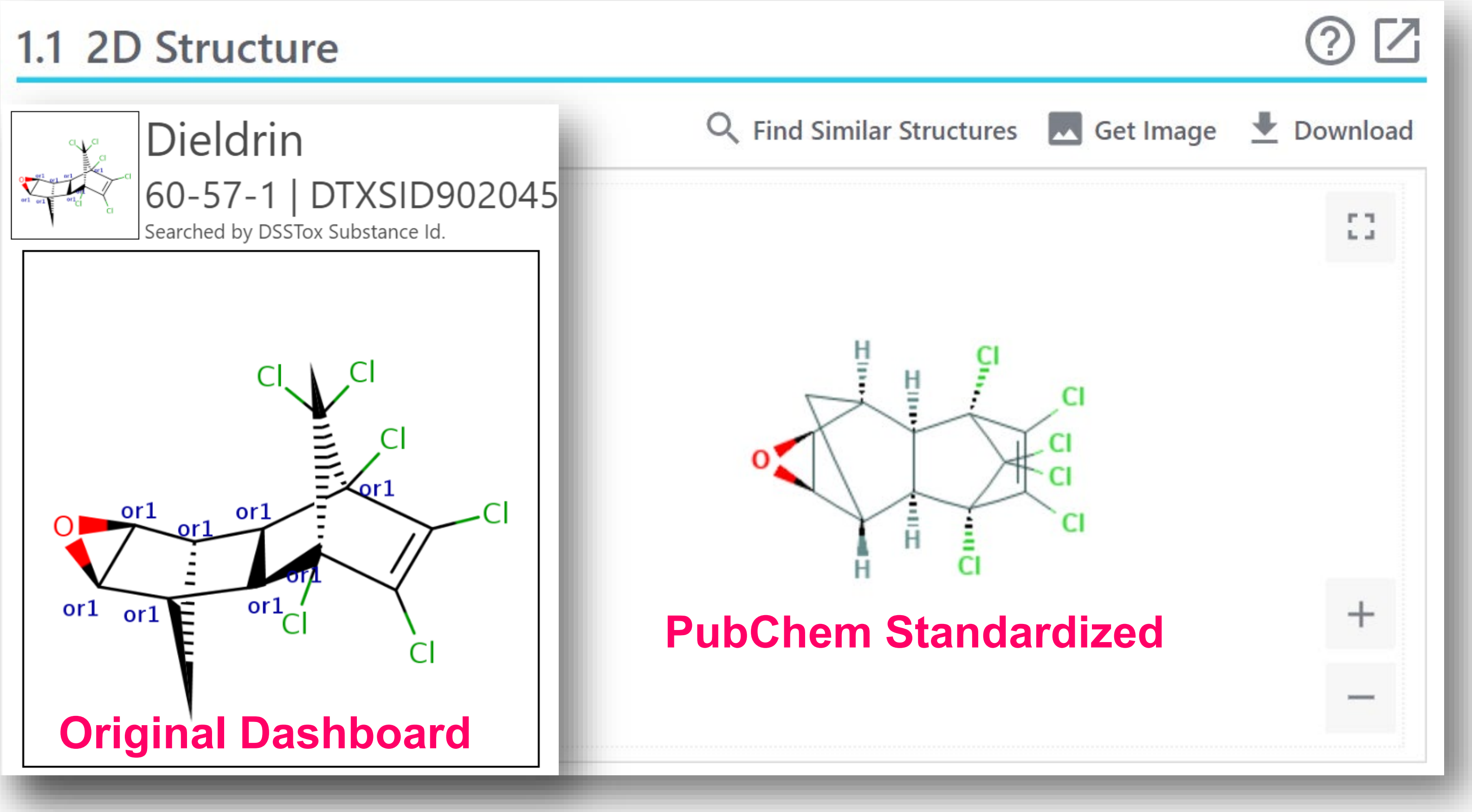
Challenges with making improvements

- Uniqueness is a two-edged sword - value or junk?

- Would be nice if *someone* made a widget that gave a quick quality stats overview for chemicals sets
  - Chemical structures vs. CASRNs va names and other identifiers

- Standalone curated databases can give cleaner results compared with the same content registered elsewhere. e.g. 875k chemicals from CompTox Chemicals Dashboard nested in 96 million in PubChem. Standardization is not lossless...

Standardization and standards V3000 Stereochemistry Support



1.1 2D Structure

Dieldrin

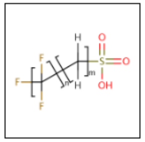
60-57-1 | DTXSID902045

Searched by DSSTox Substance Id.

Original Dashboard

PubChem Standardized

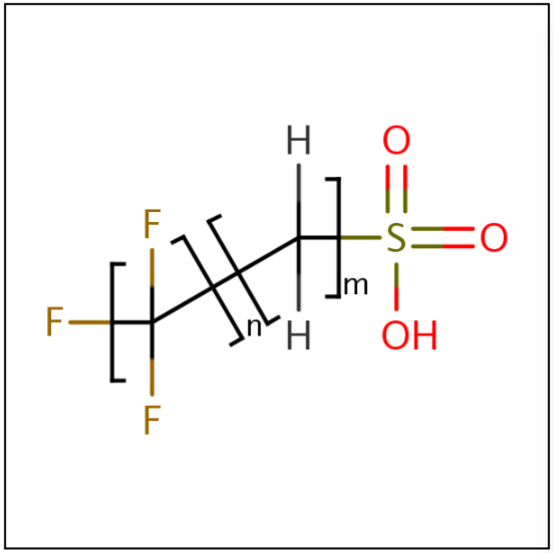
Standardization and standards
Markush Representations

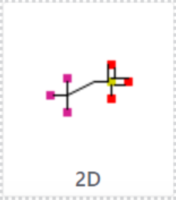


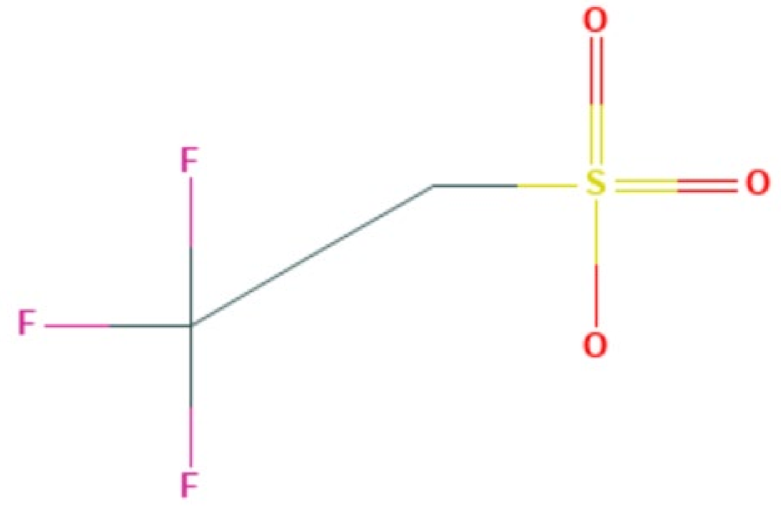
Fluorotelomer (linear) sulfonic acids

NOCAS_892558 | DTXSID50892558

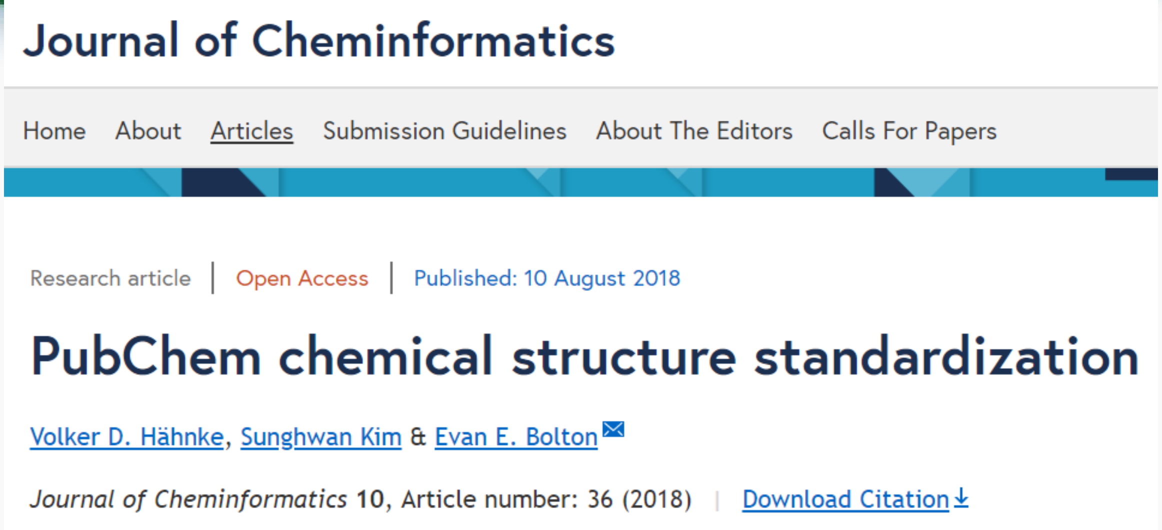
Searched by DSSTox Substance Id.



PubChem SID:	384442688
PubChem CID:	3014047 (2,2,2-Trifluoroethanesulfonic acid) Related Records
Structure:	 2D
Source:	
External ID:	



Standardization Efforts



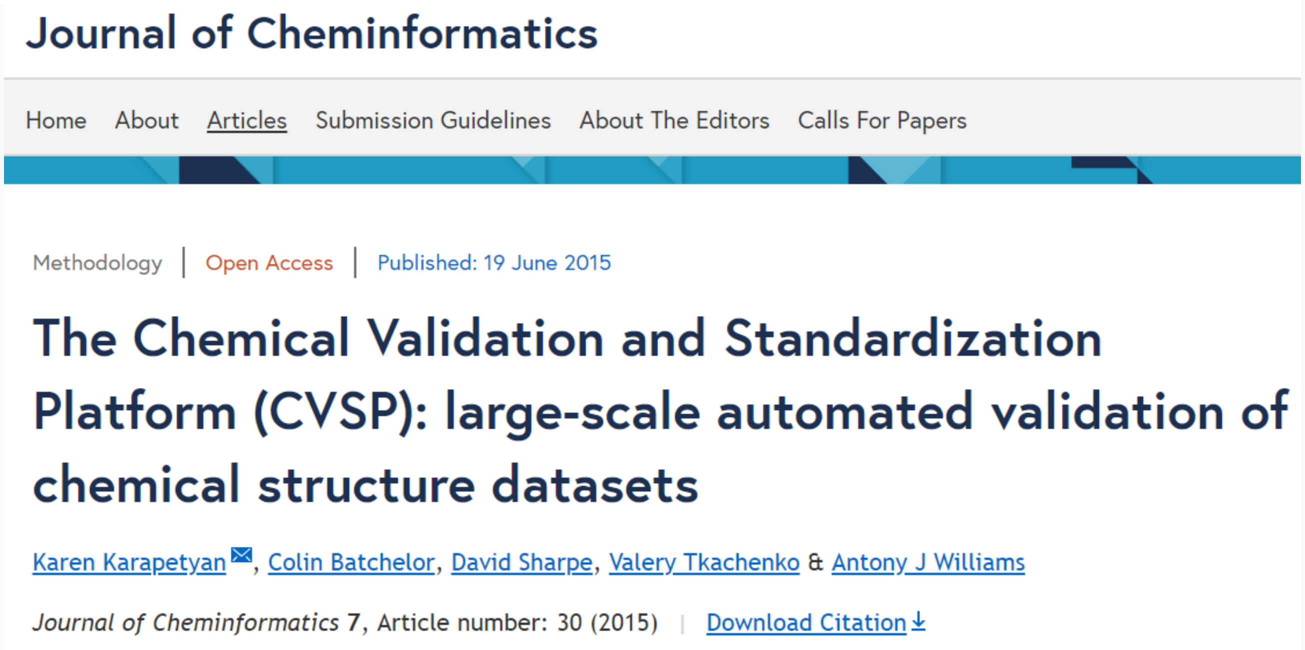
Journal of Cheminformatics

Research article | Open Access | Published: 10 August 2018

PubChem chemical structure standardization

Volker D. Hähnke, Sunghwan Kim & Evan E. Bolton

Journal of Cheminformatics 10, Article number: 36 (2018)



Journal of Cheminformatics

Methodology | Open Access | Published: 19 June 2015

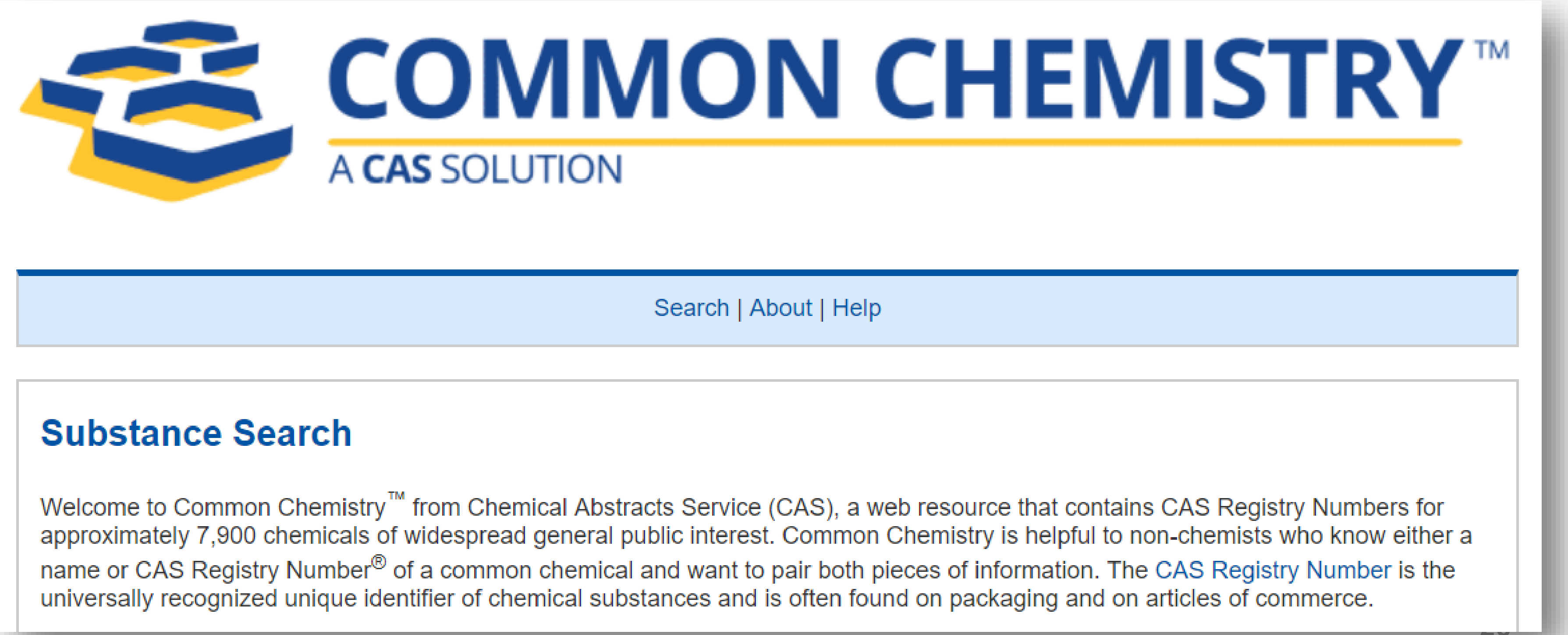
The Chemical Validation and Standardization Platform (CVSP): large-scale automated validation of chemical structure datasets

Karen Karapetyan, Colin Batchelor, David Sharpe, Valery Tkachenko & Antony J Williams

Journal of Cheminformatics 7, Article number: 30 (2015)

The Power but Confusion of CASRNs

- CASRNs have only one true validation path
- CommonChemistry was a **GREAT START** for Wikipedia CAS Validation – but out of date



Substance Search

Welcome to Common Chemistry™ from Chemical Abstracts Service (CAS), a web resource that contains CAS Registry Numbers for approximately 7,900 chemicals of widespread general public interest. Common Chemistry is helpful to non-chemists who know either a name or CAS Registry Number® of a common chemical and want to pair both pieces of information. The CAS Registry Number is the universally recognized unique identifier of chemical substances and is often found on packaging and on articles of commerce.

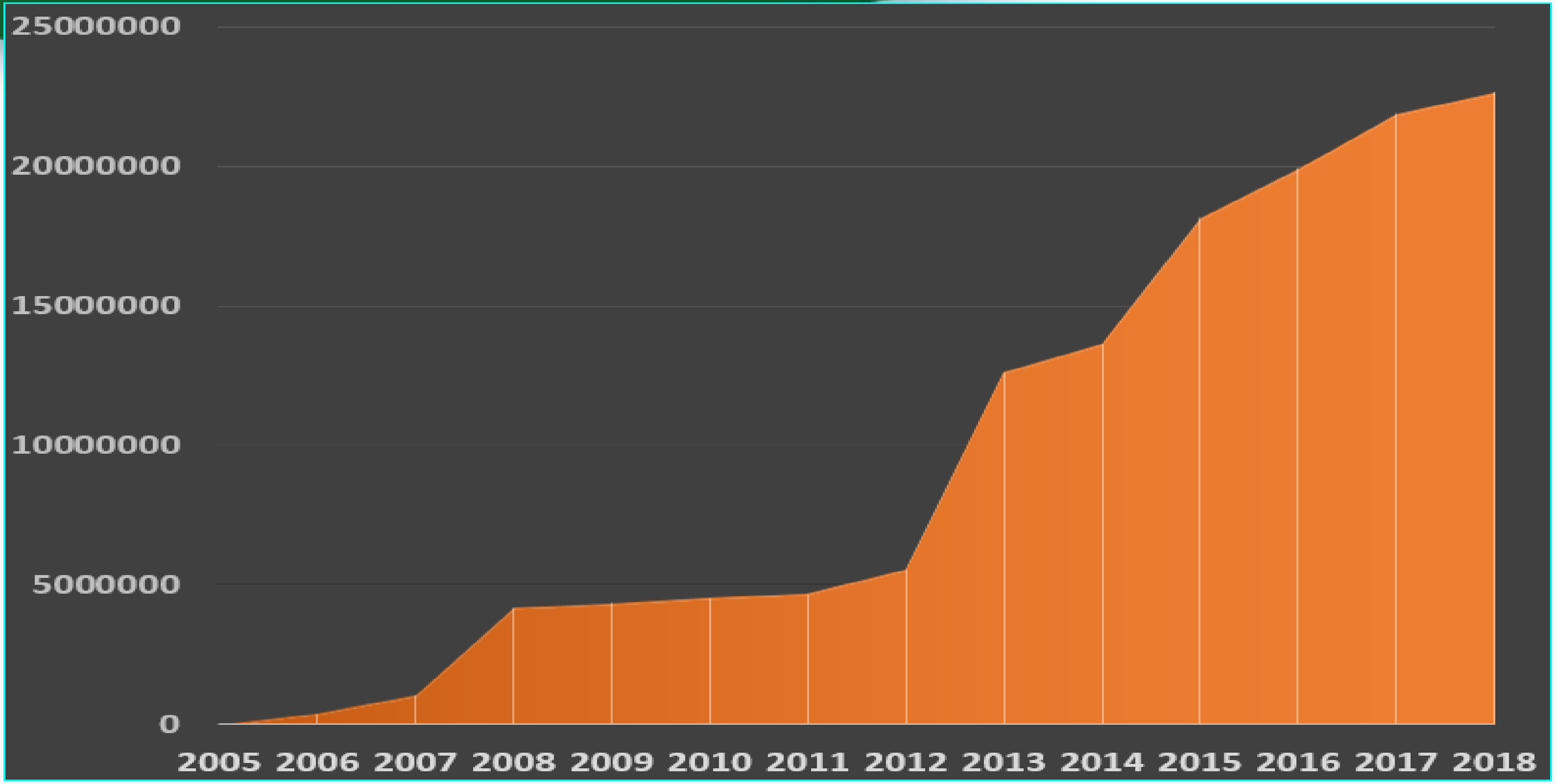
Validation of CASRNs

- Automated bulk validation of CASRNs is possible only with assistance from CAS

Automated Patent Extraction

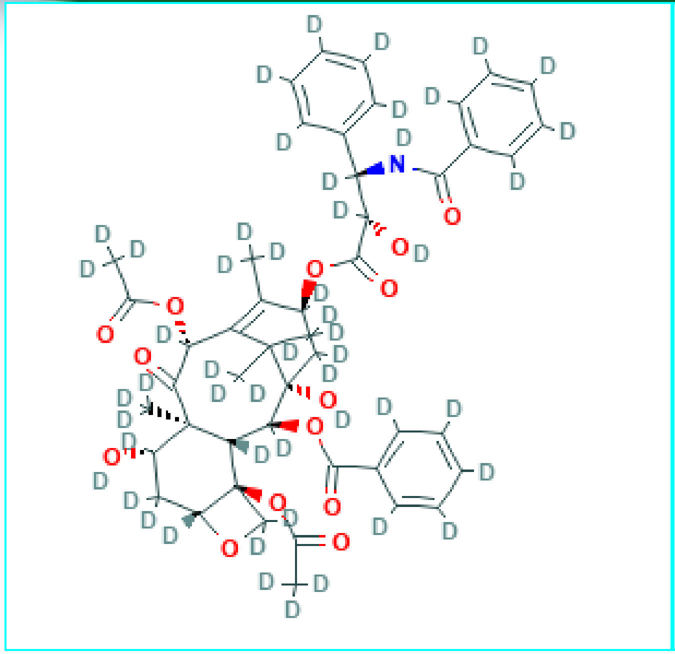
- Classic dilema between very high value and noise
- ChemSpider chose to forego patent data because of quality issues
- PubChem have done a herculean job on their feeds from IBM, SCRIPDB, SureChEMBL and NextMove! (e.g. indexing 3 mill patent documents in the new interface)

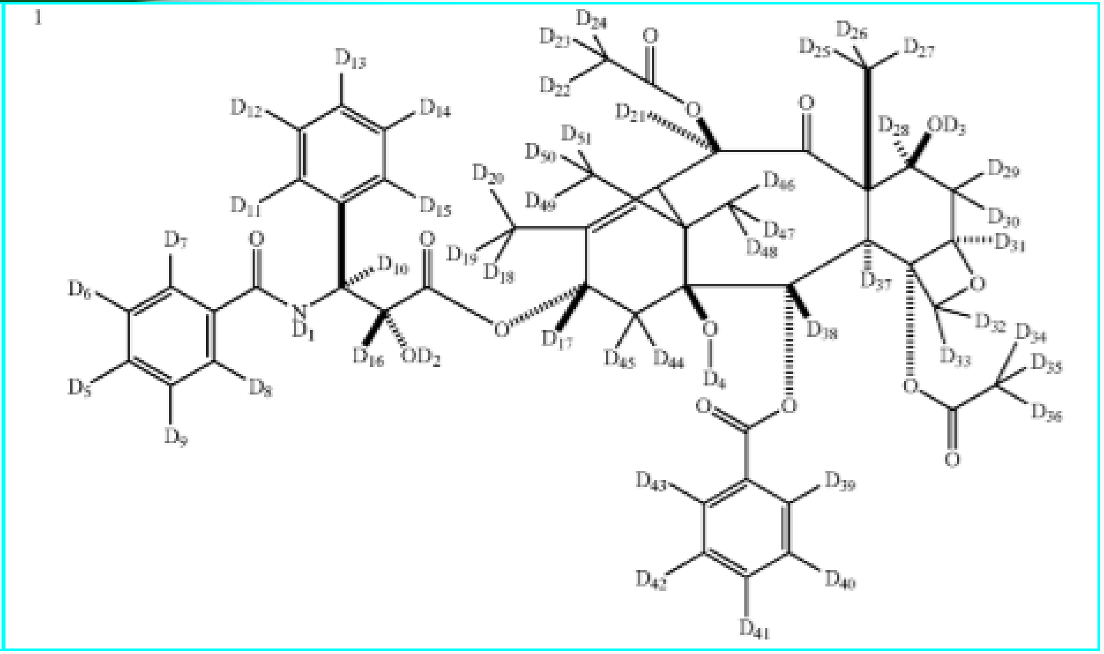
Patent CIDs by year (cumulative)



- SureChEMBL is the only major source regularly updating
- Will there be a post-2017 IBM refresh?
- “**News flash**” Google Patents has started incorporating searchable chemistry extraction – so will this become a complementary feed?

Virtual deuteration: Is there really d-51 Paclitaxel??





- Left:PubChem CID42599845 drawn by Thomson/Derwent
- Right: Exemplification in US20090069410 from Protia
- Filed **100s** of deuterated drug patents 2008/9, Czarnik sole inventor (but no evidence he actually made ‘em)
- Protia, Auspex and Concert filings have led to 1000s of virtually deuterated drugs > PubChem

Observations

- Our massively-valuable open chemical database ecosystem is **noisy, vulnerable** and **under-resourced** – so we need to engage collectively for enhancements

- Expansion of big databases is good but unless they push back against the primary quality of submitters it’s a losing battle

- Crowdsourcing does not scale – so could artificial intelligence/machine learning improve some of strutural standardisation/noise/quality issues?

Observations

- Are 64 million/50% unique, vendor compounds in PubChem too much? (e.g. cap the number of suppliers for common compounds?)
- None of us would have a problem with virtual ”make on demand” compounds if they are clearly tagged
- Springer and Theime index their automatically extracted chemistry against documents – so what about ACS, RSC, Wiley, Elsevier, ChemRxiv, others?
- Data changes - ChemSpider **July 2016:** 57 million from 517 sources; **August 2019** 75 Million from 270 sources

Conclusions

- How do we get the situation to change???
  - More collaboration?
  - More sharing?
  - More standards?

- For now the biggest shift is likely education – the community needs awareness of the issues in large public resources

Acknowledgements

- All of the contributors of data to the public databases
- The hosts (and funders) of the individual databases
- The PubChem and ChemSpider team for answering queries