http://www.orcid.org/0000-0002-2668-4821

Overview of open resources to support automated structure verification and elucidation

Antony Williams[1] and Emma Schymanski[2]

1. *National Center for Computational Toxicology, US EPA, Research Triangle Park, Durham, NC, USA.*
2. *Luxembourg Centre for Systems Biomedicine (LCSB), University of Luxembourg, Campus Belval, Luxembourg.*

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

March 2018

ACS Spring Meeting, New Orleans

Today's Session

- Our focus for the session:
  - Access to data to support automated structure verification and elucidation – NMR and MS
  - Data quality, curation and validation – and a call to action
  - OPENness is here – Open Access, Data, Source
  - Data standards – we already have them and there are more coming
  - Vendors providing and using available data
  - There are tools USING these data for Structure Elucidation

  - Cannot be an exhaustive review…

An Ideal Scenario…

- All published structures and spectra will be available from all published articles for repurposing and reuse in standard formats (preferably not necessarily Open!)
- Scientists are building open approaches
  - MS fragmentation
  - NMR shift prediction
  - Structure generators
  - Computer-Assisted Structure Elucidation (CASE)
- Are we there yet???

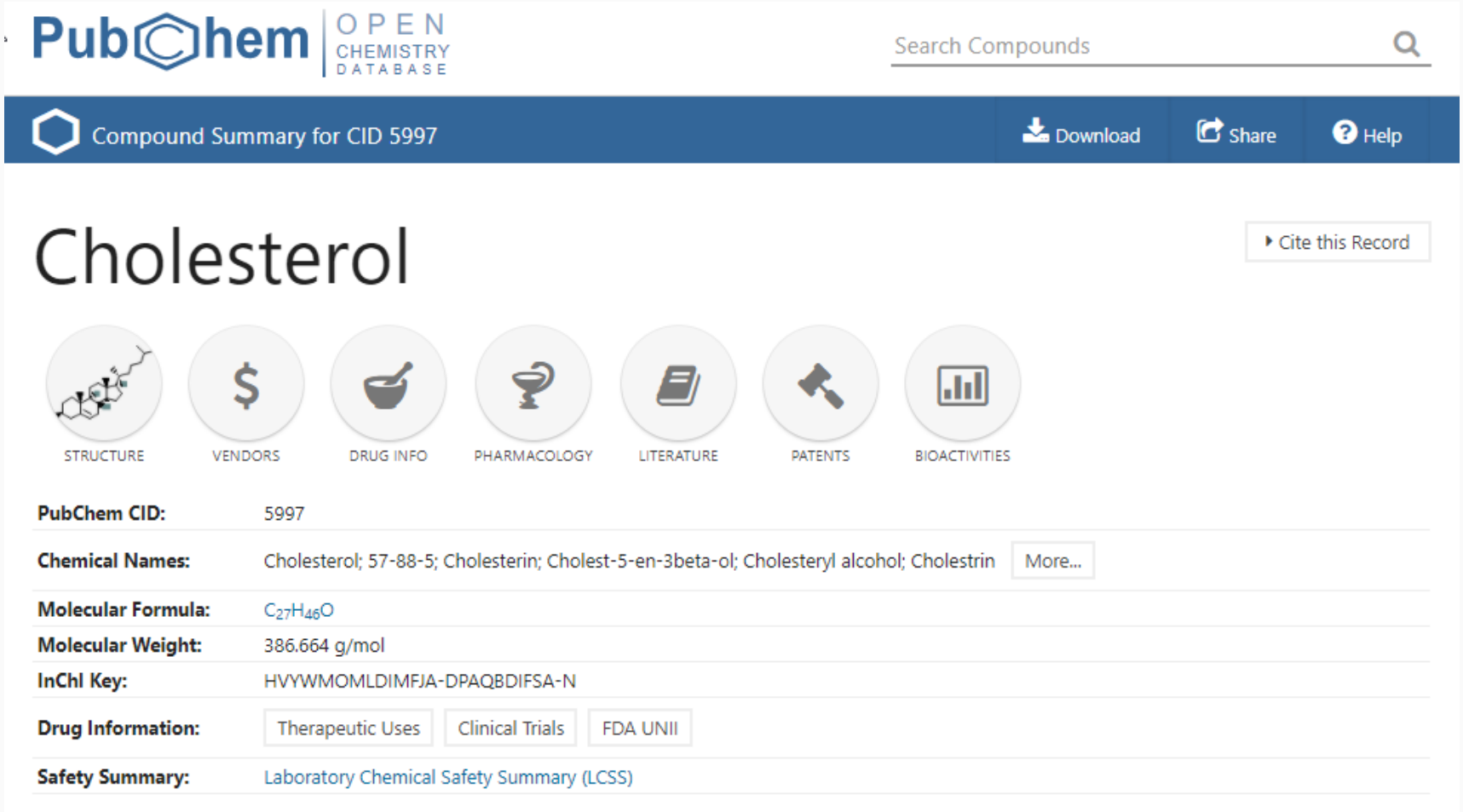
Publishers sharing data

- We have achieved ideal scenario right?
- No – PDF figures in Supplementary Info is still the default position
- There is a need for public databases of spectral data. There ARE some out there.
- Just like Wikipedia we are primarily consumers rather than contributors…

Sites sharing data

- There are many sites that “share” spectral data. Generally in non-open formats
- There are rich resources
- Cannot easily be used to serve automated structure verification and elucidation.

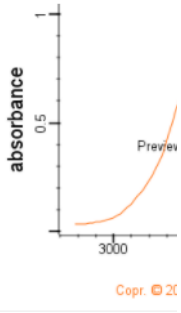
PubChem – Spectral Links



PubChem - Spectral Links

EPA United States Environmental Protection Agency

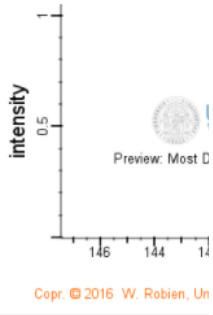
4.4.1 Infrared Spectra

Infrared Spectra: 1 of 1 (ATR-IR Spectra)	
Instrument Name	Bio-Rad FTS
Technique	ATR-Neat (DuraSamplIR II)
Source of Spectrum	Forensic Spectral Resear
Source of Sample	Sigma-Aldrich Company
Catalog Number	C8667
Lot Number	SLBC7554V
Copyright	Copyright © 2014-2017
Thumbnail	

4.4.2 1D NMR Spectra

1D NMR Spectra: 1 of 16 (1H NMR Spectra)	
1H NMR Spectra	1D NMR Spectrum 2491 - 1H NMR Spectrum (HMDB0062453)
13C NMR Spectra	1D NMR Spectrum 3188 - 13C NMR Spectrum (HMDB0062453)

from Human Metabolome Database (HMDB)

1D NMR Spectra: 2 of 16 (13C NMR Spectra)	
Copyright	Copyright © 2016 W. Robien Rights Reserved.
Thumbnail	

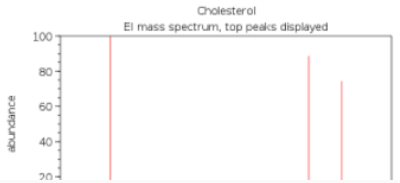
4.4.3 Mass Spectrometry

4.4.3.1 GC-MS

1. GC-MS Spectrum 2260
2. GC-MS Spectrum 4679
3. GC-MS Spectrum 4680
4. GC-MS Spectrum 4681
5. GC-MS Spectrum 4682
6. GC-MS Spectrum 4683

from Human Metabolome Database (HMDB)

1 of 3

NIST Number	332884
Library	Main library
Total Peaks	274
m/z Top Peak	43
m/z 2nd Highest	55
m/z 3rd Highest	57
Thumbnail	

Spectral Links to Partial Data

EPA United States Environmental Protection Agency

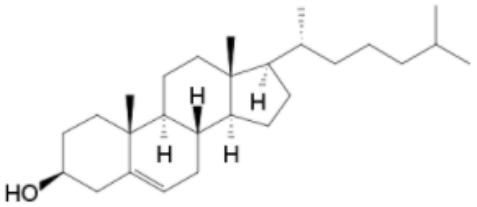
SpectraBase™ Search Buy Now Log In BIO-RAD

Cite Share Follow Feedback

CHOLESTEROL;VITAMIN-D-DERIVATIVE

Compound with one FTIR, fifteen NMR, and one Raman spectra

Download

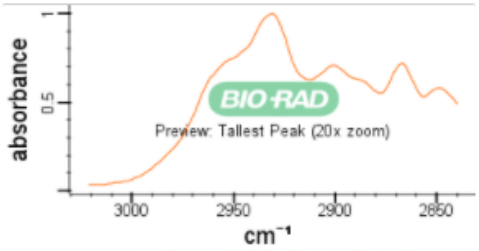


SpectraBase Compound ID	DW7SMDXE5b3
InChI	InChI=1S/C27H46O/c1-18(2)7-6-8-19(3)23-11-12-24-22-10-9-20-17-21(28)13-15-26(20,4)25(22)14-16-27(23,24)5/h9,18-19,21-25,28H,6-8,10-17H2,1-5H3/t19-,21+,22+,23-,24+,25+,26+,27-/m1/s1
InChIKey	HVYWMOMLDIMFJA-DPAQBDIFSA-N Google Search
Mol Weight	386.7 g/mol
Molecular Formula	C27H46O
Exact Mass	386.354866 g/mol

Spectra | Synonyms | Similar Compounds | Literature | Related Ambiguous Compounds

- ^{13}C NMR 15
- ATR-IR 1
- Raman 1

View Full Record



Copyright	Copyright © 2014-2017 Bio-Rad Laboratories, Inc. All Rights Reserved.
Source of Sample	Sigma-Aldrich Company Llc
Source of Spectrum	Forensic Spectral Research
Catalog Number	C8667

SDBS – Free Not Open

Access to this database is free of charge. However we request visitors to our database not to download more than 50 spectra and/or compound information in one day. All accesses are recorded. It is prohibited that you use any information of SDBS for profit-making or commercial use without obtaining proper permission from us. If more spectra are required for some specific purpose or commercial use, you should consult us and describe the intended usage or purpose of our SDBS.

SDBS – Free Not Open

EPA United States Environmental Protection Agency

Spectral Database for Organic Compounds SDBS

Japanese | Introduction | Disclaimer | HELP | Contact | What's New | RIO-DB | FAQ | LINK | AIST

SDBS Information

SDBS No.: 887

Compound Name:
cholesterol

Molecular Formula: $C_{27}H_{46}O$

Molecular Weight: 386.6

CAS Registry No.:
57-88-5

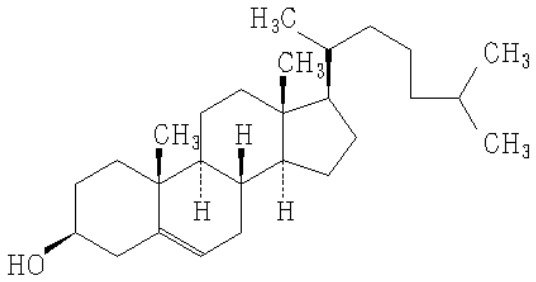
Derivatives:
display in a separate page
SDBS Structures Web (on trial since 2013-05-14)

Spectral Code:
Mass :
^{13}C NMR : in $CDCl_3$
^{1}H NMR : 400 MHz in $CDCl_3$
IR : nujol mull
IR : KBr disc

Chemical Information:

SDBS No:	887	**CAS Registry No.:**	57-88-5
Molecular Formula:	$C_{27}H_{46}O$	**Molecular Weight:**	386.6

SDBS-NO= 887
CHOLESTEROL



Compound Name:
cholesterol
5-cholesten-3beta-ol

SDBS – Free Not Open

EPA United States Environmental Protection Agency

Spectral Database for Organic Compounds SDBS

Japanese | Introduction | Disclaimer | HELP | Contact | What's New | RIO-DB | FAQ | LINK | AIST

SDBS Information

SDBS No.: 887

Compound Name:
cholesterol

Molecular Formula: $C_{27}H_{46}O$

Molecular Weight: 386.6

CAS Registry No.:
57-88-5

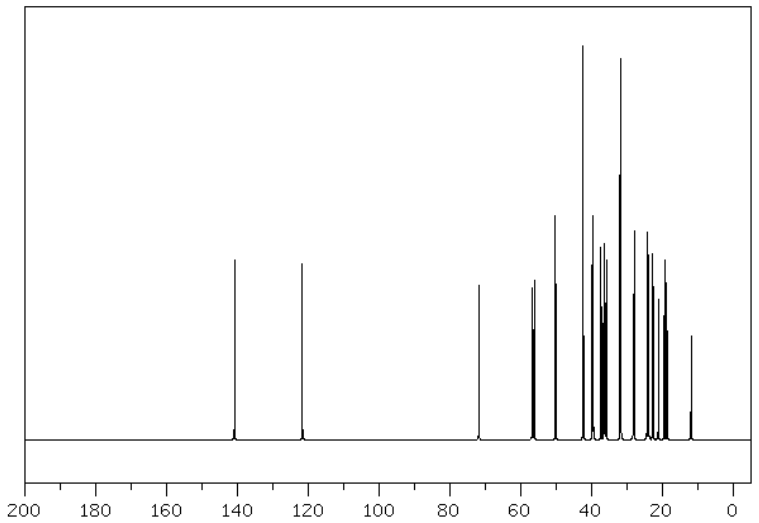
Derivatives:
display in a separate page
SDBS Structures Web (on trial since 2013-05-14)

Spectral Code:
Mass :
^{13}C NMR : in $CDCl_3$
^{1}H NMR : 400 MHz in $CDCl_3$
IR : nujol mull
IR : KBr disc

Chemical Information:

SDBS-^{13}C NMR SDBS No. 887CDS-05-861
$C_{27}H_{46}O$
cholesterol

25.16 MHz
0.141 g : 1.5 ml $CDCl_3$



ChemSpider

EPA United States Environmental Protection Agency

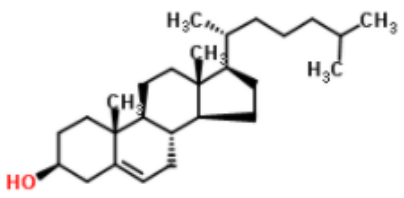
ChemSpider
Search and share chemistry

Simple Structure Advanced History

Found 1 result

Search term: **cholesterol** (Found by approved synonym)

COMMENT ON THIS RECORD



3D

Cholesterol

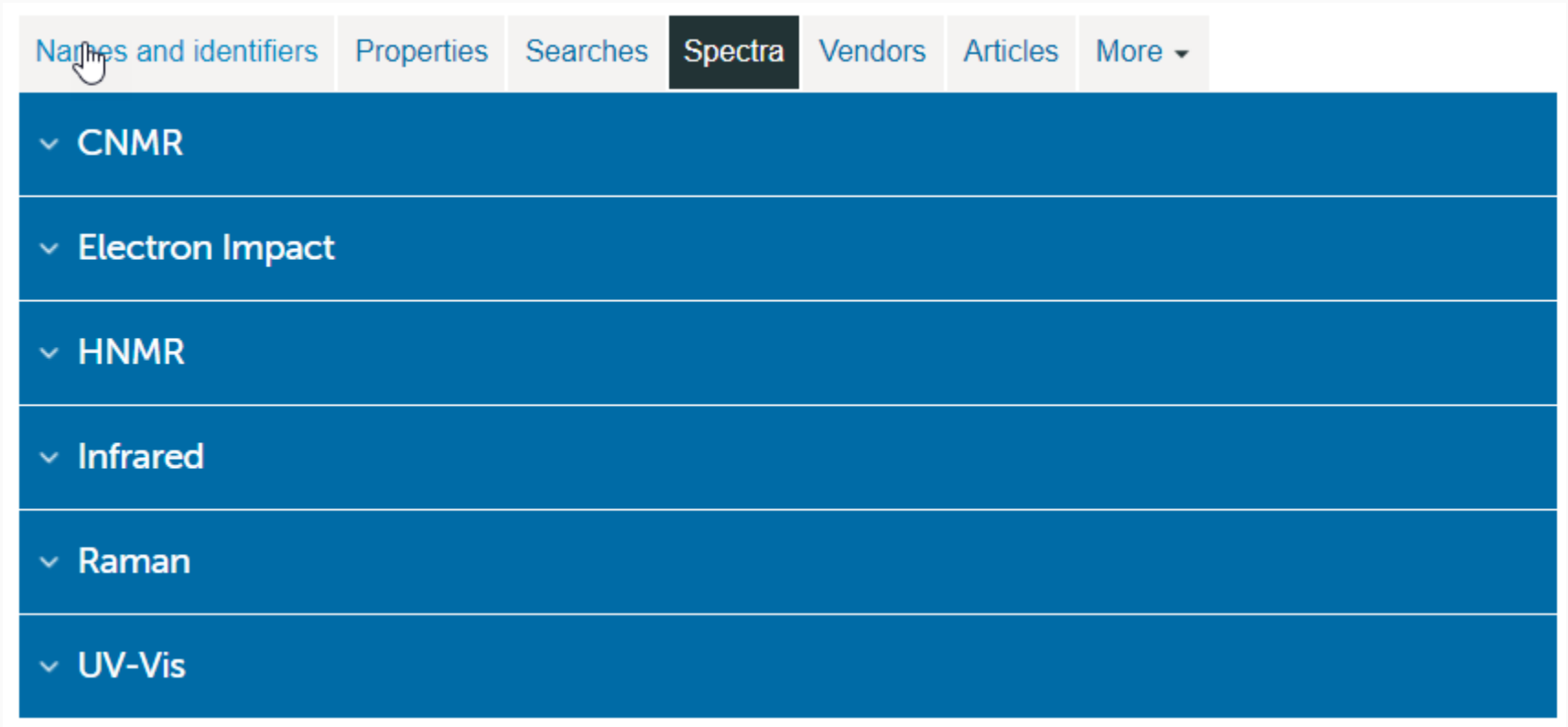
Molecular Formula	$C_{27}H_{46}O$
Average mass	386.654 Da
Monoisotopic mass	386.354858 Da
ChemSpider ID	5775

- 8 of 8 defined stereocentres

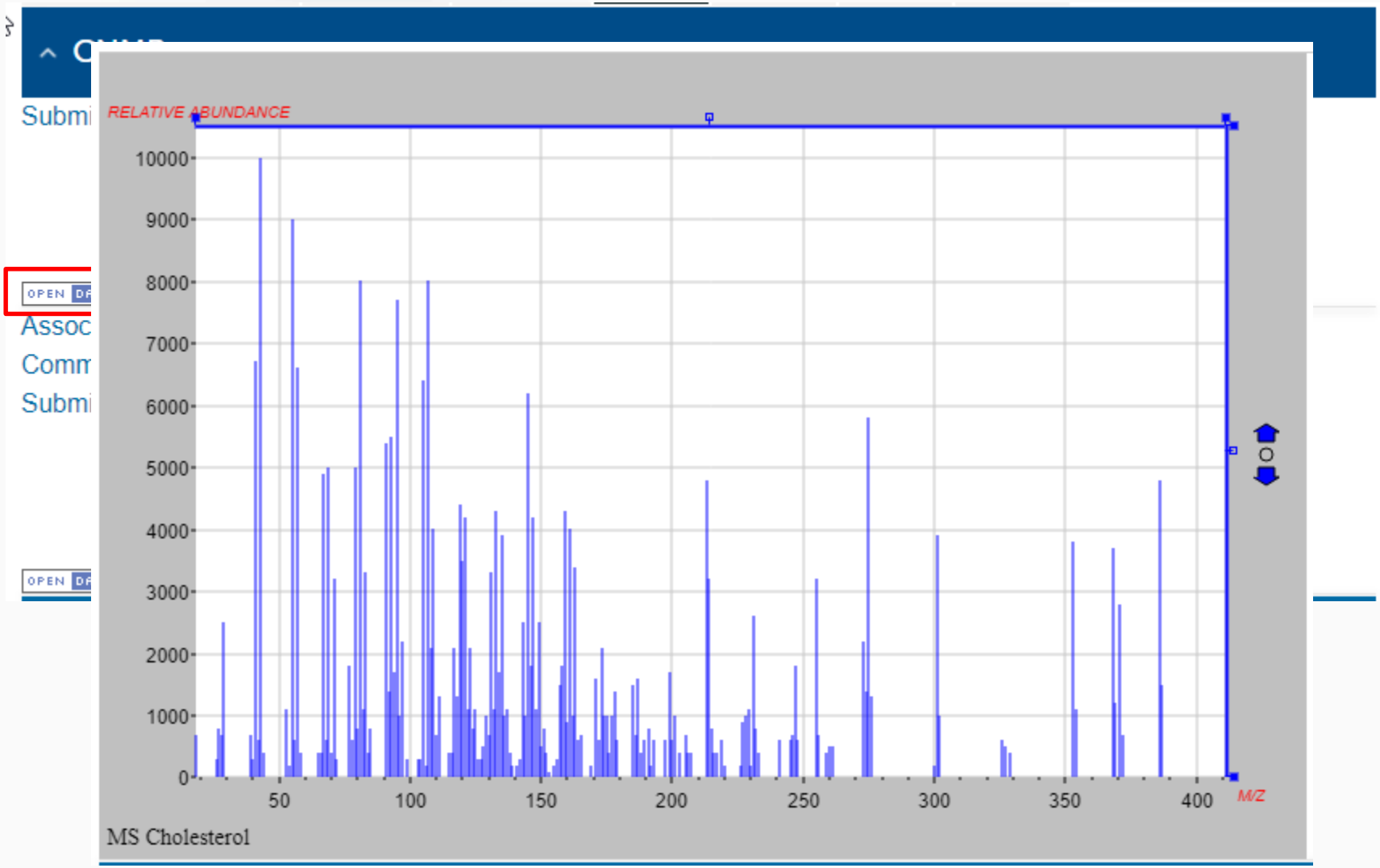
Featured data source

The Merck Index *Online* has more data on this compound

ChemSpider


Names and identifiers
Properties
Searches
Spectra
Vendors
Articles
More
CNMR
Electron Impact
HNMR
Infrared
Raman
UV-Vis

ChemSpider


RELATIVE ABUNDANCE
10000
9000
8000
7000
6000
5000
4000
3000
2000
1000
0
50
100
150
200
250
300
350
400
M/Z
MS Cholesterol

NIST WebBook
https://webbook.nist.gov/chemistry/

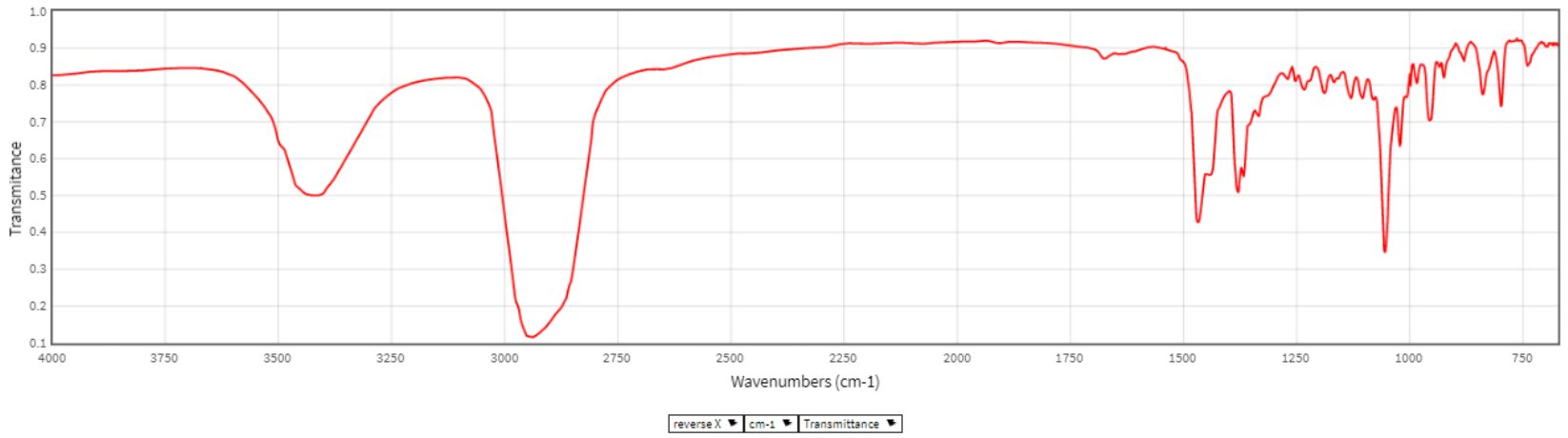
EPA United States Environmental Protection Agency

Condensed Phase Spectrum

Plot | Help / Software credits

Cholest-5-en-3beta-ol

Infrared Spectrum

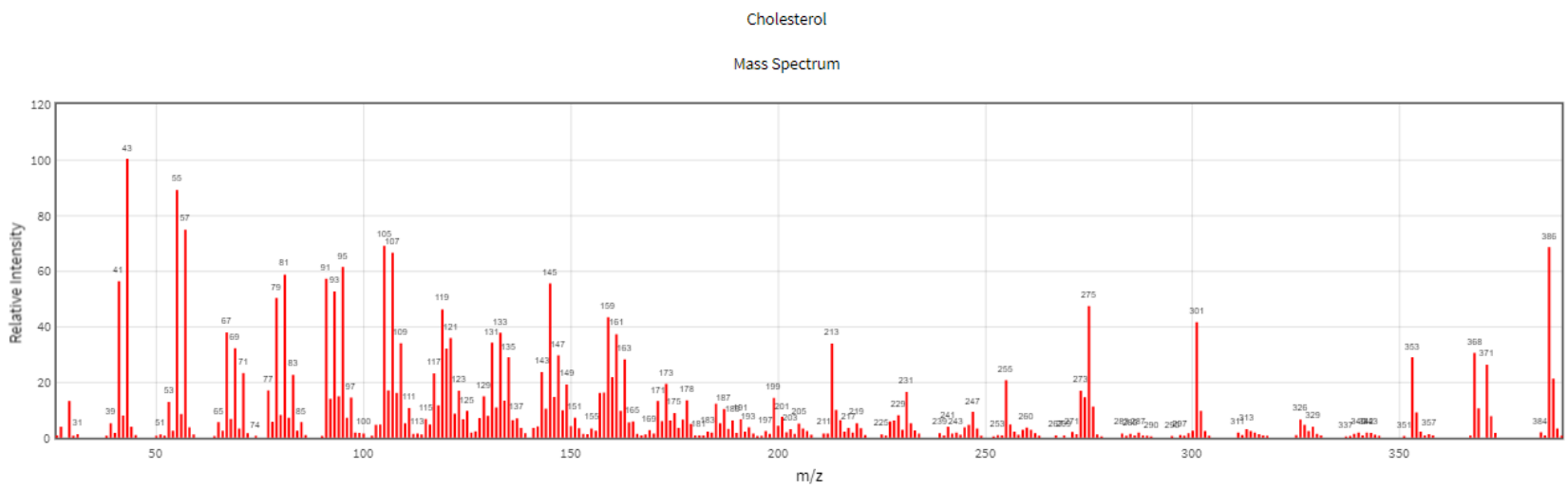


NIST WebBook
https://webbook.nist.gov/chemistry/

EPA United States Environmental Protection Agency

Spectrum

Plot | Help / Software credits



Focused Databases

- Focused databases
  - Compiled focused databases of Open Data are preferable
  - Spectral data for structure elucidation – Open Mass Spec Data especially useful (Emma's talk!)
  - Data can be brought in-house and integrated
  - Algorithms can be derived – e.g. NMR shift prediction

NMRShiftDB
https://nmrshiftdb.nmr.uni-koeln.de/

EPA United States Environmental Protection Agency

NMRShiftDB

Current usage is:

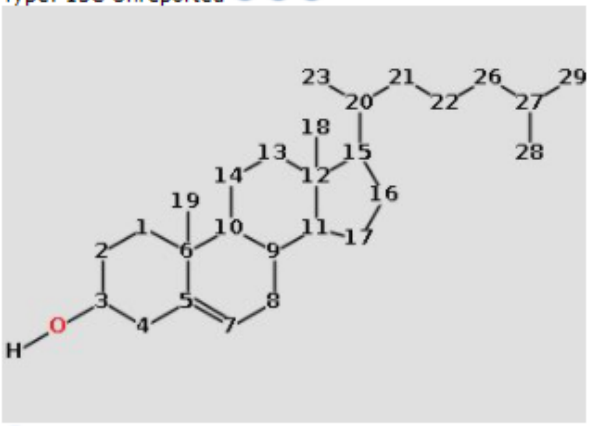
Registered Users: 1216

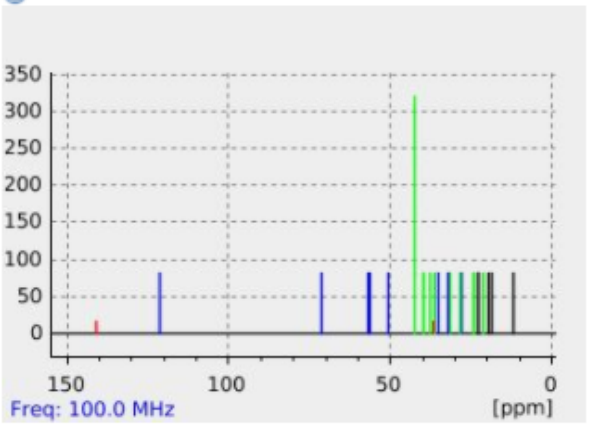
Structures: 43445

Spectra: Measured 52114, calculated 549

Spectral Data | Additional Data | Download

Type: 13C Unreported





Atom No. ⬇	Mult.(coupling const.)	Meas. Shift
1	T	37.5
2	T	31.6
3	D	71.3
4	T	42.4
5	S	141.2
6	S	36.5
7	D	121.3
8	T	32.0
9	D	32.3
10	D	50.5
11	D	56.9
12	T	42.4
13	T	40.0
14	T	21.2
15	D	56.5
16	T	28.3
17	T	24.3
18	Q	12.0
19	Q	19.4
20	D	35.4
21	T	36.4
22	T	24.1
23	Q	18.8
26	T	39.6
27	D	28.0
28	Q	22.8
29	Q	22.5

NMRShiftDB
https://nmrshiftdb.nmr.uni-koeln.de/

EPA United States Environmental Protection Agency

Spectral Data | Additional Data | Download

Molecule	10016303
Chemical name(s)	
Chemical formula	$C_{27}H_{46}O$
Molecular weight	386.653
Number of double bond equivalents (DBEs)	5.0
Number of all rings, size of smallest set of smallest rings	10, 4
Canonical name(s)	• OC1CC2=CCC3C(CCC4(C)C(CCC34)C(... (truncated) (SMILES) • cholest-5-en-3-ol (IUPAC from ACD/Name) • cholest-5-en-3-ol (Index from ACD/Name) • InChI=1S/C27H46O/c1-18(2)7-6-8... (truncated) (INChI) • HVYWMOMLDIMFJA-UHFFFAOYSA-N (InChI Key) • InChI=1/C27H46O/c1-18(2)7-6-8-... (truncated) (InChI with fixed H layer)
CAS-Number	
Additional information	Deposition in PubChem; Links from Unichem: ChEMBL;eMolecules;eMolecules;IBM strategic IP insight platform and the National Institutes of Health;SureChEMBL;PubChem Compounds;Mcule;ACToR;MolPort
Molecule keywords	
13C Spectrum	**10038299 Rating: 10**
Type	13C
Measurement conditions	
Temperature [K]	Unreported
Solvent	Unreported
Field Strength [MHz]	Unreported
Assignment Method	Unreported
Literature	S. Berger; S. Braun; H.-O. Kalinowski: 13-C-NMR-Spektroskopie, New York: Thieme Verlag 1984.
Additional comments	610.MOL; Multiplicities generated automatically from H count
Additional information	
Spectrum categories	

NMRShiftDB
https://nmrshiftdb.nmr.uni-koeln.de/

It is the policy of nmrshiftdb2 to have its content freely available. This means that spectra in nmrshiftdb2 can be accessed and used without restrictions. Access to single items is possible via the software and we also make available the bunch data (without the personal data). Note that, according to our opinion, NMR spectra do not represent "literary and artistic works" as mentioned in Art. 2 (1) of the Berne Convention, the author of a spectrum is, so to say, nature, and that therefore there is no copyright in NMR data. nmrshiftdb2 as a database is a protected work in the sense of article 5 of the WIPO Copyright Convention, authored by its contributors and reviewers. It is available under the **nmrshiftdb2 Data License**.

The nmrshiftdb2 data are available as:

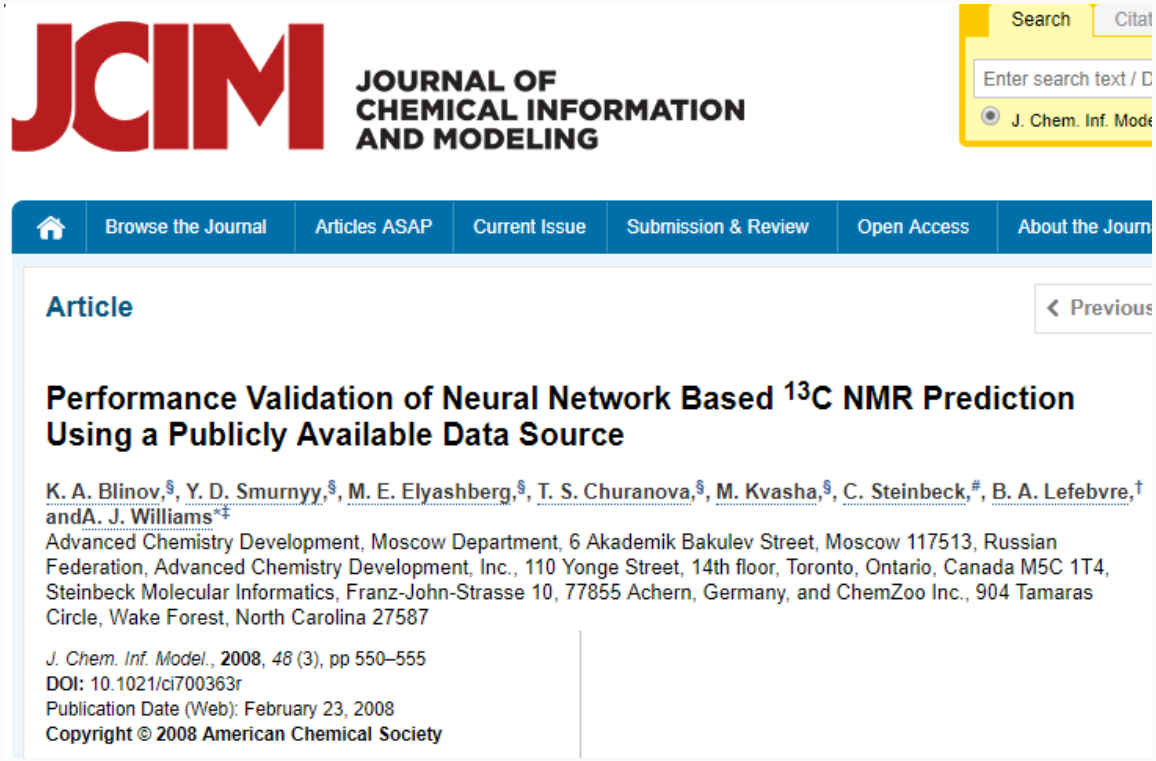
Table 1. Data downloads

-	With 2D coordinates
sdf	**nmrshiftdb2.sd**
sdf with spectra	**nmrshiftdb2withsignals.sd**
cml	**nmrshiftdb2.xml**

Questions about nmrshiftdb2 should be directed to the **administrator**.

Open Resources

- Open Databases offer more value
  - Bring the data in-house, integrate, link
  - Ingest and train algorithms



JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

Search | Citat

Enter search text / D

J. Chem. Inf. Mode

Browse the Journal | Articles ASAP | Current Issue | Submission & Review | Open Access | About the Journ

Article

Previous

Performance Validation of Neural Network Based ^{13}C NMR Prediction Using a Publicly Available Data Source

K. A. Blinov,[§], Y. D. Smurnyy,[§], M. E. Elyashberg,[§], T. S. Churanova,[§], M. Kvasha,[§], C. Steinbeck,[#], B. A. Lefebvre,[†] andA. J. Williams*[‡]

Advanced Chemistry Development, Moscow Department, 6 Akademik Bakulev Street, Moscow 117513, Russian Federation, Advanced Chemistry Development, Inc., 110 Yonge Street, 14th floor, Toronto, Ontario, Canada M5C 1T4, Steinbeck Molecular Informatics, Franz-John-Strasse 10, 77855 Achern, Germany, and ChemZoo Inc., 904 Tamaras Circle, Wake Forest, North Carolina 27587

J. Chem. Inf. Model., **2008**, *48* (3), pp 550–555
DOI: 10.1021/ci700363r
Publication Date (Web): February 23, 2008
Copyright © 2008 American Chemical Society

CSEARCH/NMRPREDICT
http://nmrpredict.orc.univie.ac.at/

EPA United States Environmental Protection Agency

Welcome to the CSEARCH / NMRPREDICT-Server

I have ...	I want ...	The solution to your problem
A structure Proposal	A predicted ^{13}C-NMR Spectrum	Use the CSEARCH-Robot-Referee don't enter any lines
A Structure Proposal and a ^{13}C-NMR Peaklist	A Structure Verification	Use the CSEARCH-Robot-Referee Enter as many lines as necessary Optionally assign as many lines as possible
A Structure Proposal and a ^{13}C-NMR Peaklist	A Structure Verification	Use the Wiley Data Checker based on CSEARCH technology Enter your structure and the peaklist and all will be done for you
A ^{13}C-NMR Peaklist	A Structure Proposal	Use the Spectral Similarity Search You can use this feature directly from Bruker's TOPSPIN
A ^{13}C-NMR Peaklist and hundreds/thousands of Structure Proposals	A Hitlist sorted by Similarity	Use the CSEARCH-Decision Engine

MassBank
https://massbank.eu/MassBank/

EPA United States Environmental Protection Agency

MassBank norman High Resolution Mass Spectral Database

Quick Search Results

Home | Quick Search | Peak Search | Record Index | Statistics | Imprint/Data privacy MassBank ID: Go

Search Parameters :

Compound Name: **cholesterol**

Instrument Type: **CE-ESI-TOF , ESI-FTICR , ESI-ITFT , ESI-ITTOF , ESI-QTOF , HPLC-ESI-TOF , LC-ESI-IT , LC-ESI-ITFT , LC-ESI-ITTOF , LC-ESI-Q , LC-ESI-QFT , LC-ESI-QIT , LC-ESI-QQ , LC-ESI-QTOF , LC-ESI-TOF , UPLC-ESI-QTOF**

MS Type: **All**

Ion Mode: **Positive**

Edit / Resubmit Query

Results : 5 Hit. (1 - 5 Displayed) Open All Tree

First Prev **1** Next Last (Total **1** Page) ▼ Results End

Name	Formula / Structure	ExactMass	ID
Cholesterol 5 spectra	C27H46O 	386.35486	
LC-ESI-QTOF; MS2; CE:10 eV; [M+H]+			CO000101
LC-ESI-QTOF; MS2; CE:20 eV; [M+H]+			CO000102
LC-ESI-QTOF; MS2; CE:30 eV; [M+H]+			CO000103
LC-ESI-QTOF; MS2; CE:40 eV; [M+H]+			CO000104
LC-ESI-QTOF; MS2; CE:50 eV; [M+H]+			CO000105

MassBank
https://massbank.eu/MassBank/

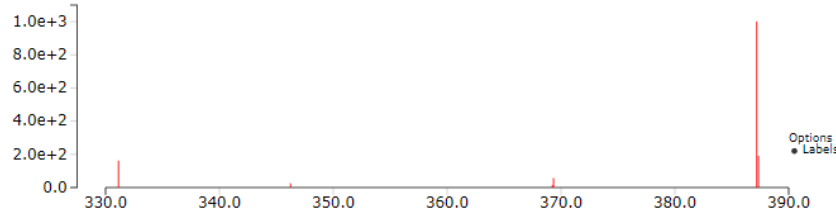
EPA United States Environmental Protection Agency

MassBank Record: CO000101

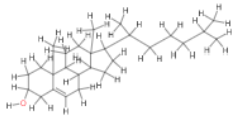
Home | Quick Search | Peak Search | Record Index | Statistics | Imprint/Data privacy MassBank ID: Go

Cholesterol; LC-ESI-QTOF; MS2; CE:10 eV; [M+H]+

Mass Spectrum



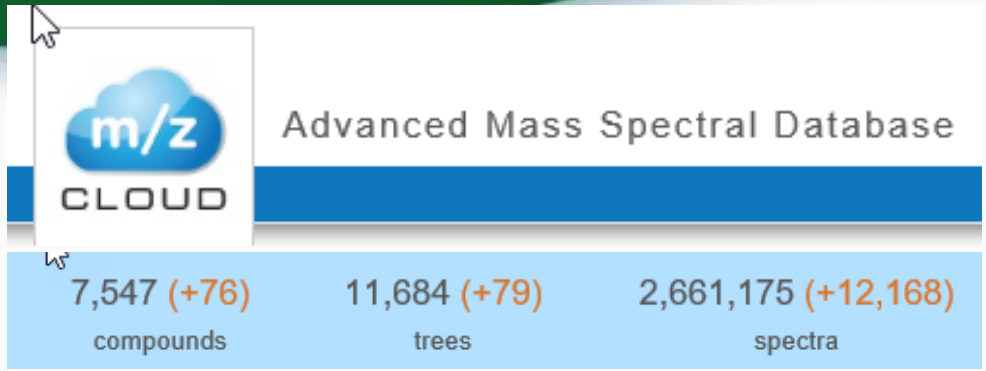
Chemical Structure




```
ACCESSION: CO000101
RECORD_TITLE: Cholesterol; LC-ESI-QTOF; MS2; CE:10 eV; [M+H]+
DATE: 2016.01.19 (Created 2008.07.15, modified 2012.11.20)
AUTHORS: Dennis W. Hill, Tzipporah M. Kertesz, Robert Friedman, David F. Grant
LICENSE: CC BY-SA
PUBLICATION: Mass Spectral Metabonomics beyond Elemental Formula: Chemical Database Querying by Matching Experimental with Computational Fragmentation Spectra, http://pubs.acs.org/doi/abs

CH$NAME: Cholesterol
CH$COMPOUND_CLASS: N/A
CH$FORMULA: C27H46O
CH$EXACT_MASS: 386.35487
CH$SMILES: [H]C(C([H])([H])C([H])([H])C([H])([H])C(C(C4([H])[H])([H])C(C([H])([H])[H])(C3([H])[H])C(C4([H])[H])([H])C([H])(C([H])(C3([H])[H])1)C([H])([H])C([H])=C(C([H])([H])2)C1(C([H])(
CH$IUPAC: InChI=1S/C27H46O/c1-18(2)7-6-8-19(3)23-11-12-24-22-10-9-20-17-21(28)13-15-26(20,4)25(22)14-16-27(23,24)5/h9,18-19,21-25,28H,6-8,10-17H2,1-5H3
CH$LINK: PUBCHEM CID:304
```

m/z CLOUD
https://www.mzcloud.org/



Cholest-4-en-3-one

enlarge

Systematic / IUPAC Name: (8*S*,9*S*,10*R*,13*R*,14*S*,17*R*)-10,13-dimethyl-17-[(2*R*)-6-methylheptan-2-yl]-1,2,6,7,8,9,11,12,14,15,16,17-dodecahydrocyclopenta[a]phenanthren-3-one

ID: Reference231

Other Names: 4-Cholesten-3-one;
Cholestenone;
3-Oxo-4-cholestene;
3-Keto-4-cholestene;
(+)-4-Cholesten-3-one ; more

Formula: $C_{27}H_{44}O$

Class: Endogenous Metabolites

Spectral Data

Cholest-4-en-3-one mass spectral data can be found in a separate interface. The data are manually curated and of the highest quality.

View Mass Spectra

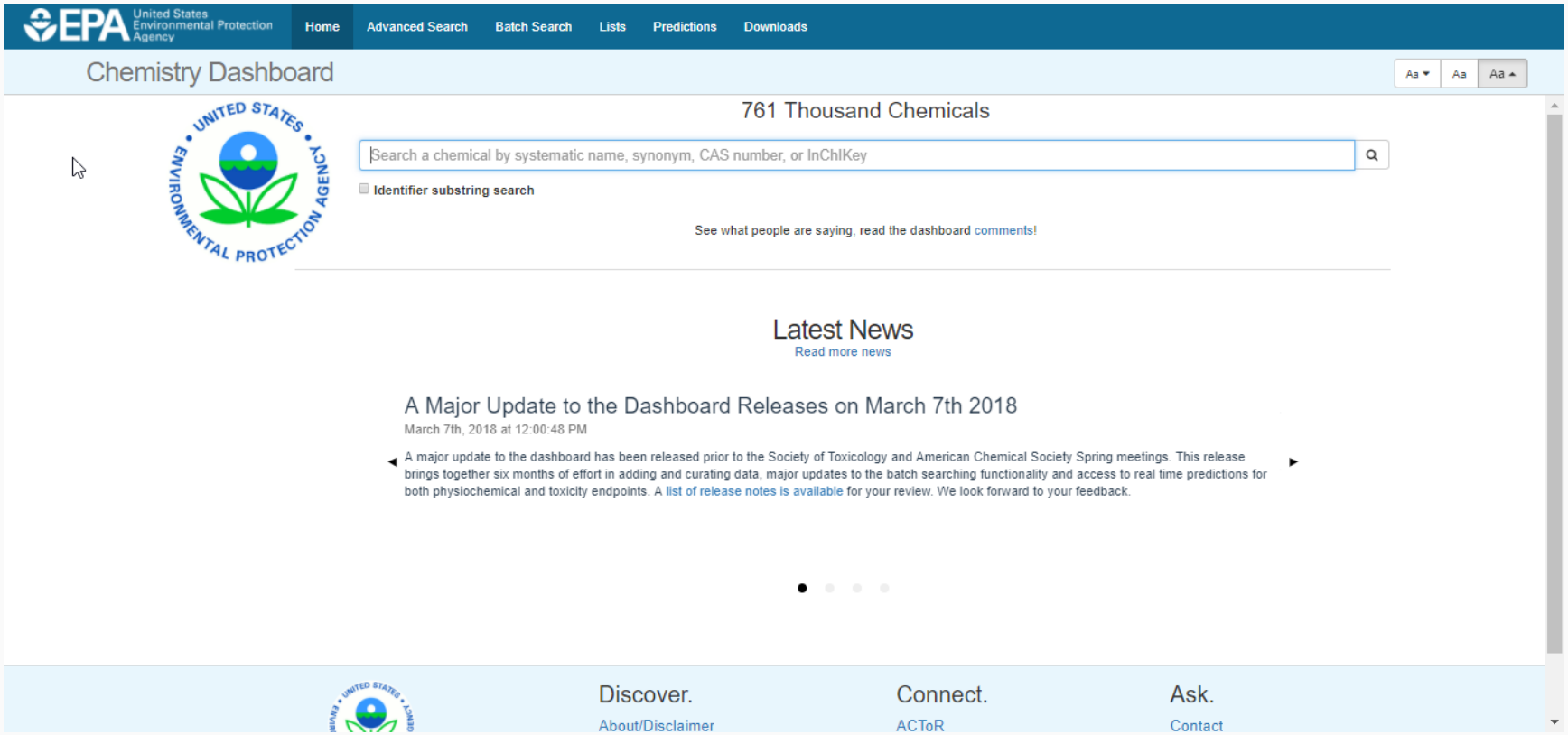
Available MS Data

Used Instruments	Orbitrap Elite
No. of Spectral Trees	1
No. of Spectra	107
Tandem Spectra	MS^1, MS^2, MS^3
Ionization Methods	ESI
Analyzers	FT

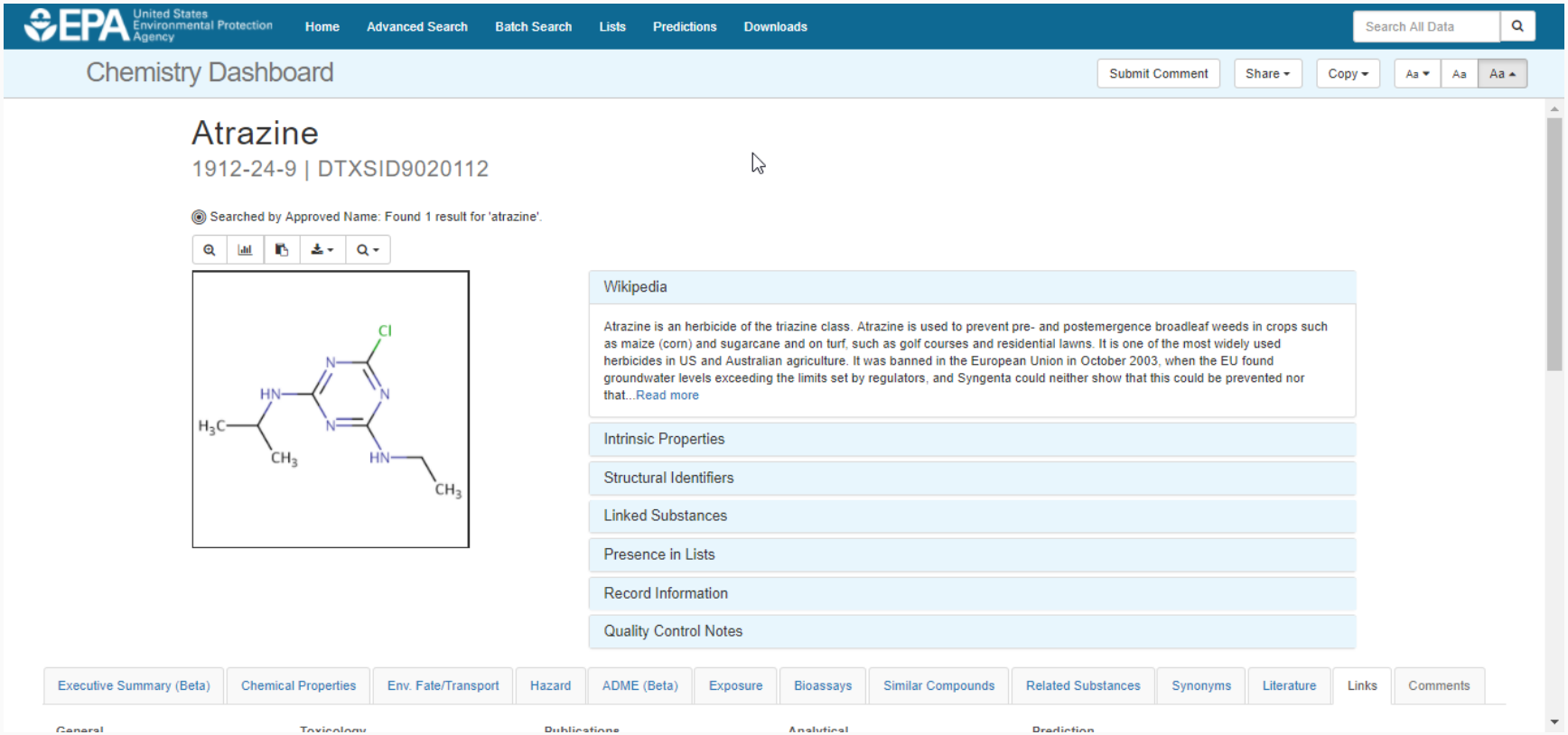
Integrating Data and Services

- Integration:
  - Use simple URL linking for navigation
  - Provide simple services for real time prediction

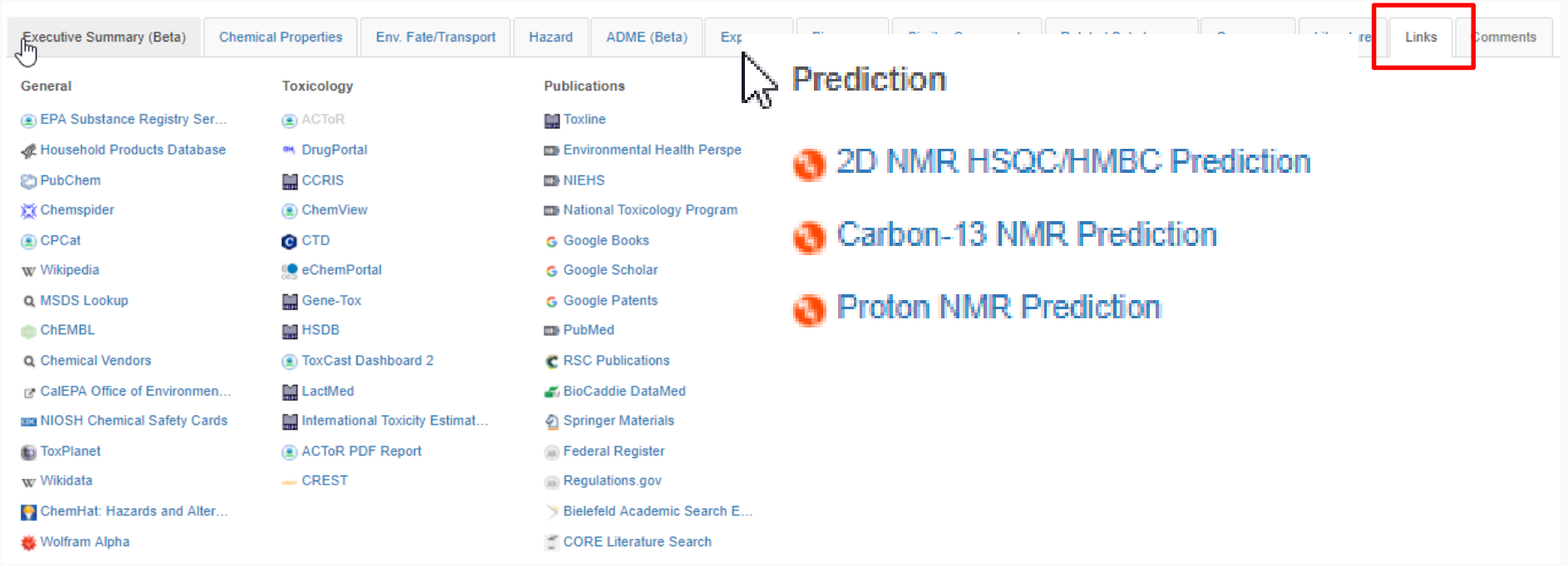
Example Integrations
CompTox Chemistry Dashboard



Link-Based Access



Link Access


Executive Summary (Beta)
Chemical Properties
Env. Fate/Transport
Hazard
ADME (Beta)
Exp
Links
Comments
General
EPA Substance Registry Ser...
Household Products Database
PubChem
Chemspider
CPCat
Wikipedia
MSDS Lookup
ChEMBL
Chemical Vendors
CalEPA Office of Environmen...
NIOSH Chemical Safety Cards
ToxPlanet
Wikidata
ChemHat: Hazards and Alter...
Wolfram Alpha
Toxicology
ACToR
DrugPortal
CCRIS
ChemView
CTD
eChemPortal
Gene-Tox
HSDB
ToxCast Dashboard 2
LactMed
International Toxicity Estimat...
ACToR PDF Report
CREST
Publications
Toxline
Environmental Health Perspe
NIEHS
National Toxicology Program
Google Books
Google Scholar
Google Patents
PubMed
RSC Publications
BioCaddie DataMed
Springer Materials
Federal Register
Regulations.gov
Bielefeld Academic Search E...
CORE Literature Search
Prediction
2D NMR HSQC/HMBC Prediction
Carbon-13 NMR Prediction
Proton NMR Prediction

nmrdb.org
Tools for NMR spectroscopists

Universidad del Valle

EPA United States Environmental Protection Agency

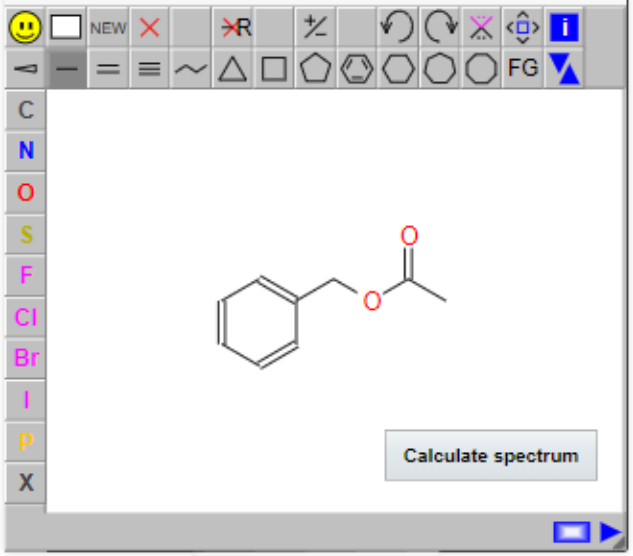
ABOUT | PREDICT 1H NMR | PREDICT 13C NMR | PREDICT 2D

Molfile or SMILES

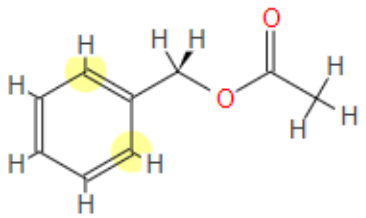
Paste or drop a molfile or SMILES

Help

Draw a chemical structure to predict



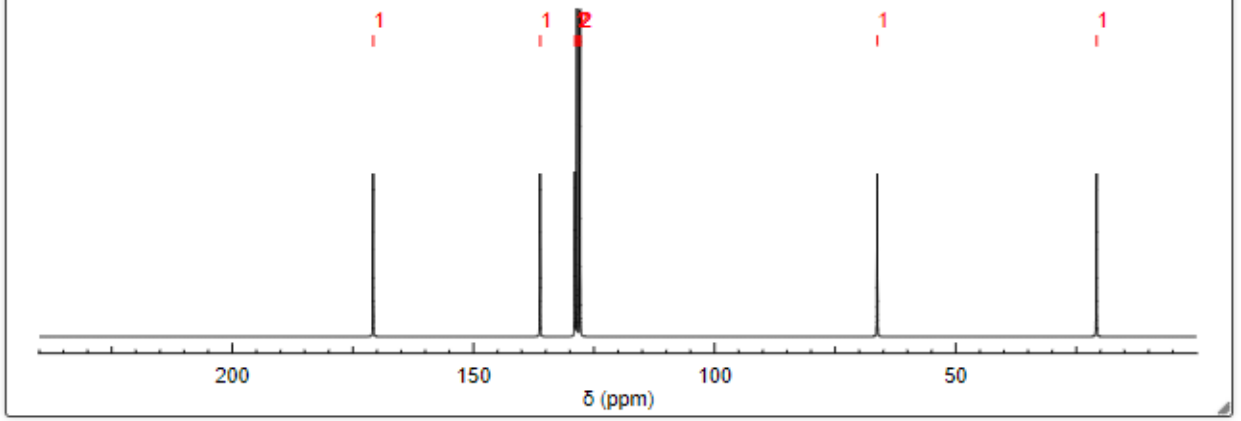
Chemical structure with hydrogen exploded



δ (ppm)
128.9
20.8
127.9
127.9
66.2
128.5

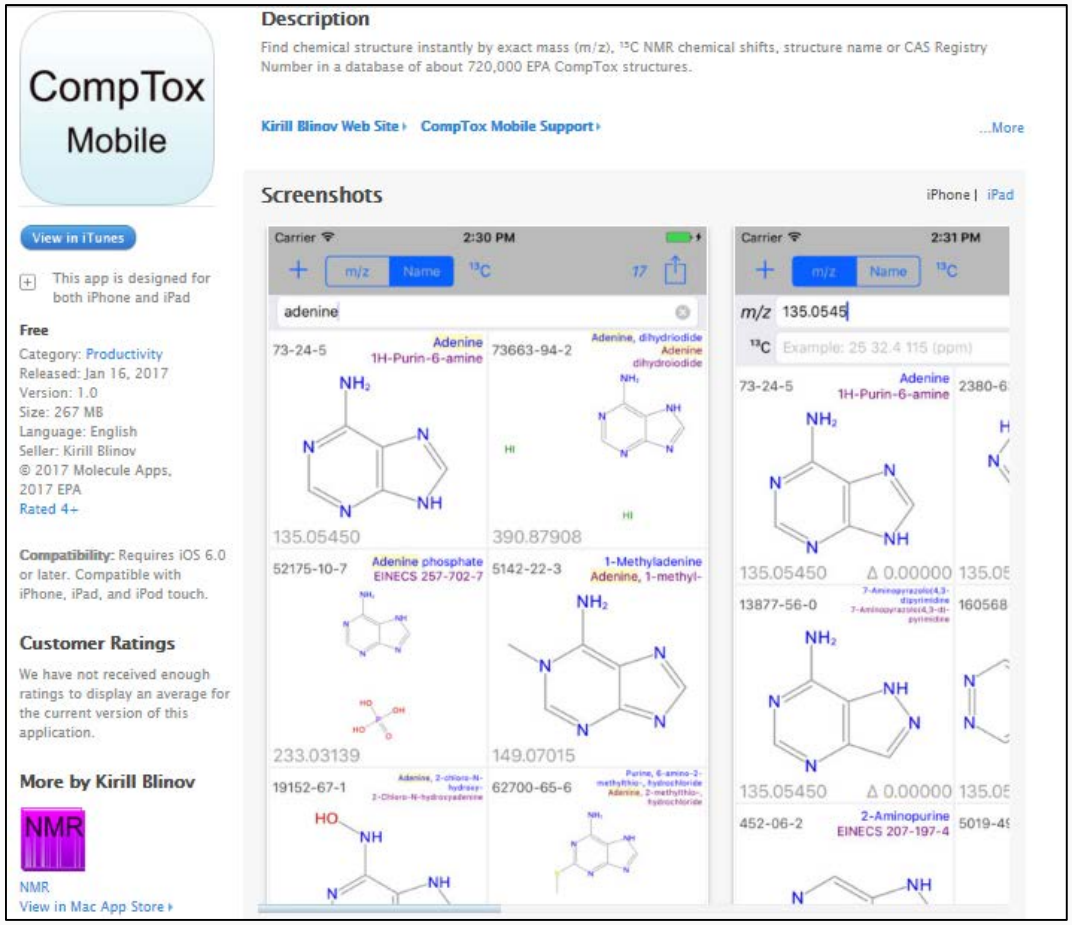
Drop or paste a JCamp file

13C NMR spectra



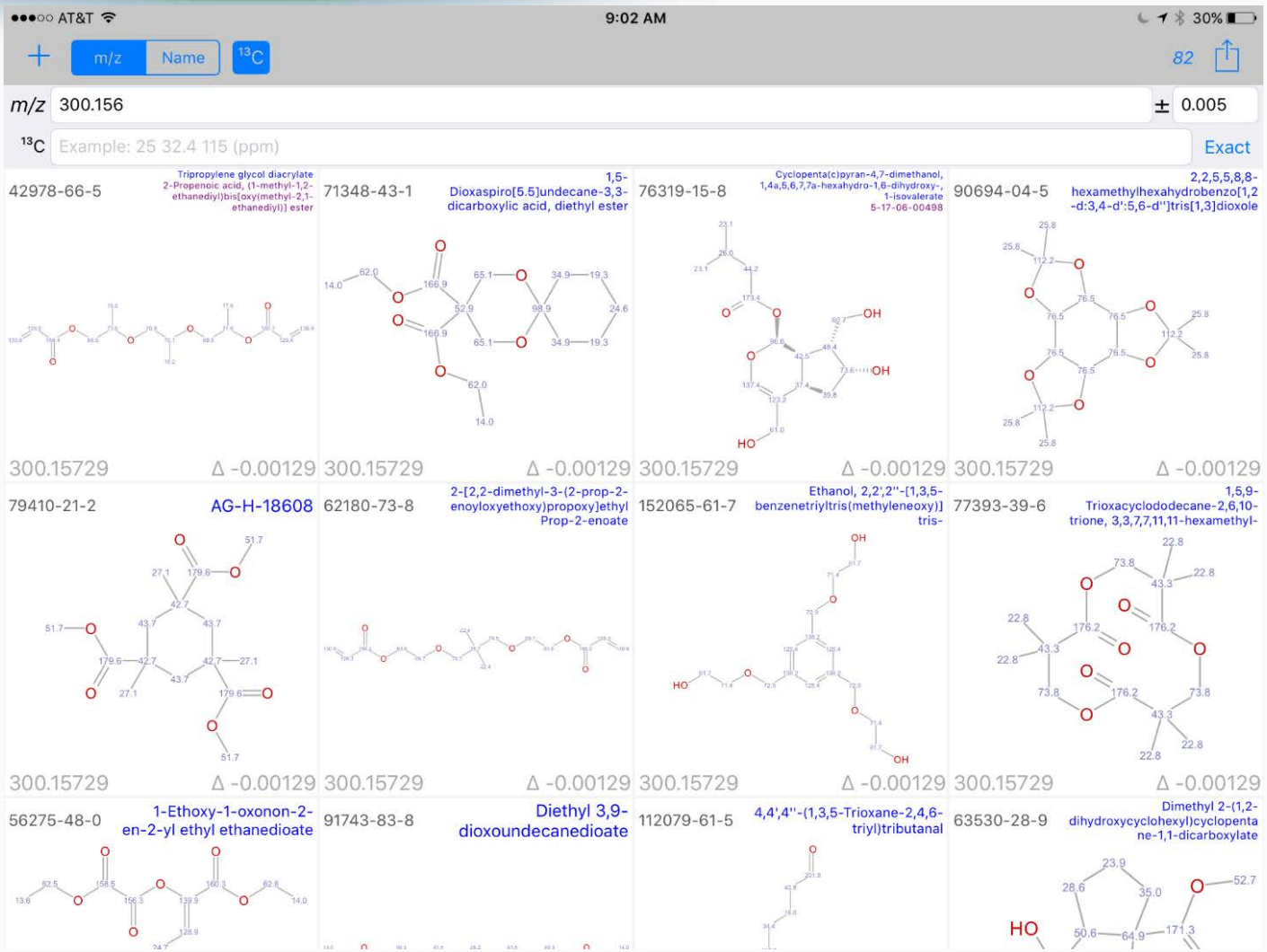
Draw a chemical structure and click on "Calculate spectrum".
You may also DRAG / DROP a molfile !
You will get an interactive NMR spectrum.

Open Data For Bulk Predictions

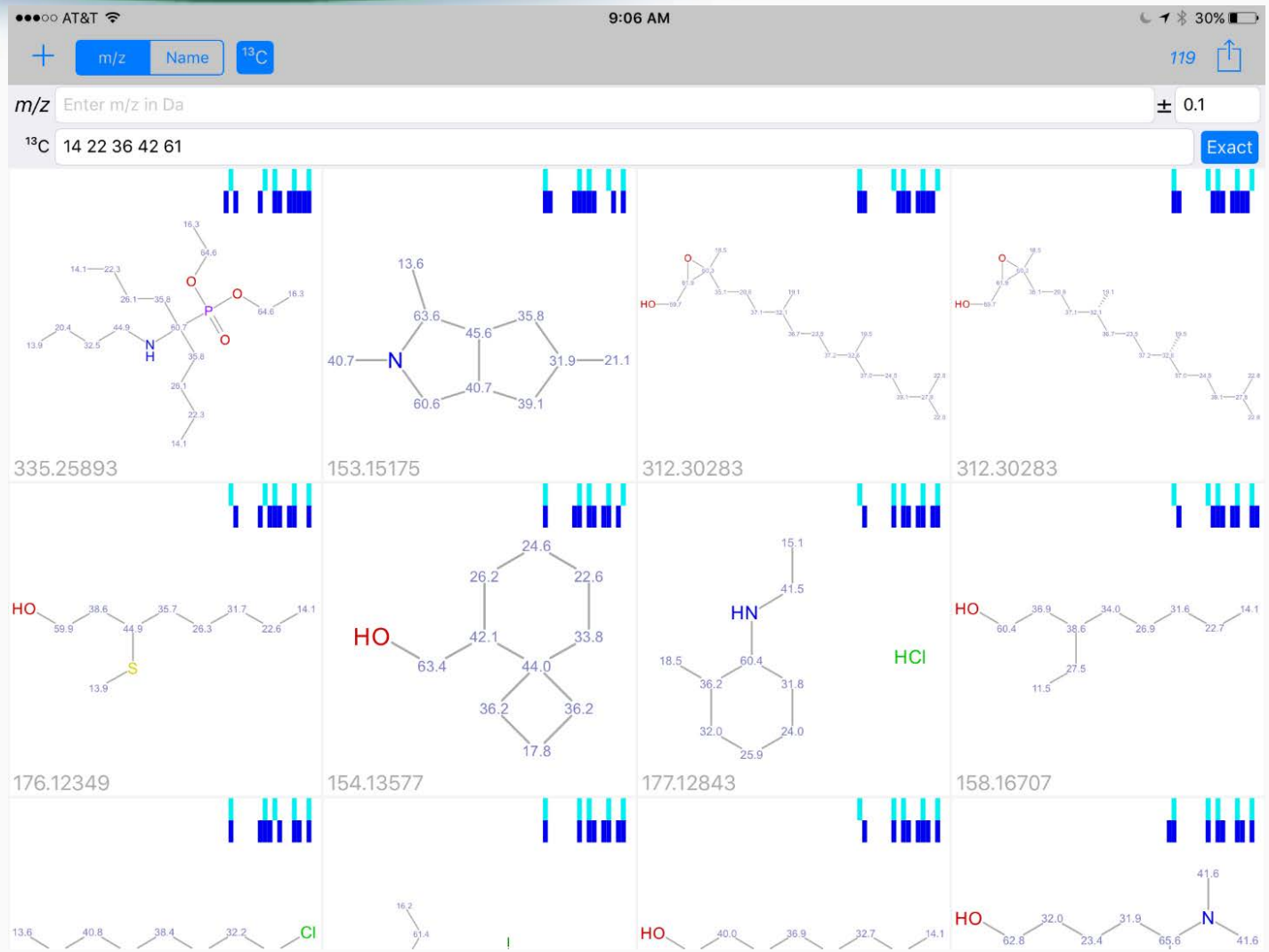


- Open Data for apps
  - Structures
  - CAS Registry Numbers
  - Names
  - Formulae
  - Mass
- iOS app including predicted C13 NMR

Mass Searching



Mass and CNMR Searching



Important Standards in our efforts

- Structures – Molfile, SDF file, InChIs (standard and non-standard)
- NMR – JCAMP and all its variants
- MS – mzML, MSP (and all its variants), MassBank

There are more coming

NMReDATA, a standard to report the NMR assignment and parameters of organic compounds

Marion Pupier[a], Jean-Marc Nuzillard[b], Julien Wist[c], Nils E. Schlörer[d], Stefan Kuhn[d], Mate Erdelyi[e], Christoph Steinbeck[f], Antony J. Williams[g], Craig Butts[h], Tim D. W. Claridge[i], Bozhana Mikhova[j], Wolfgang Robien[k], Hesam Dashti[l], Hamid R. Eghbalnia[l], Christophe Farès[m], Kessler Pavel[n], Fabrice Moriaud[o], Mikhail Elyashberg[p], Dimitris Argyropoulos[q], Manuel Pérez[r], Patrick Giraudeau[s,t], Roberto R. Gil[u], Paul Trevorrow[v], Damien Jeannerat[a,]*

Conclusion

- The abundance of online data continues to grow
- There are “integrated data”, there are databases, there are online tools, there are mobile apps
- Data Quality is critical and OPENness is enabling
  - Open Data
  - Open Standards
  - Open Source

- The rest of the day will expand on these efforts…

Contact

Antony Williams
US EPA Office of Research and Development
National Center for Computational Toxicology (NCCT)
Williams.Antony@epa.gov
ORCID: https://orcid.org/0000-0002-2668-4821