

Lessons Learned in Building the CompTox Chemicals Dashboard

Engineering a more sustainable web-based chemical database

Christopher Grulke¹

Antony Williams¹

Amar Singh¹

Jeremy Dunne¹

Ann Richard¹

Jeff Edwards¹



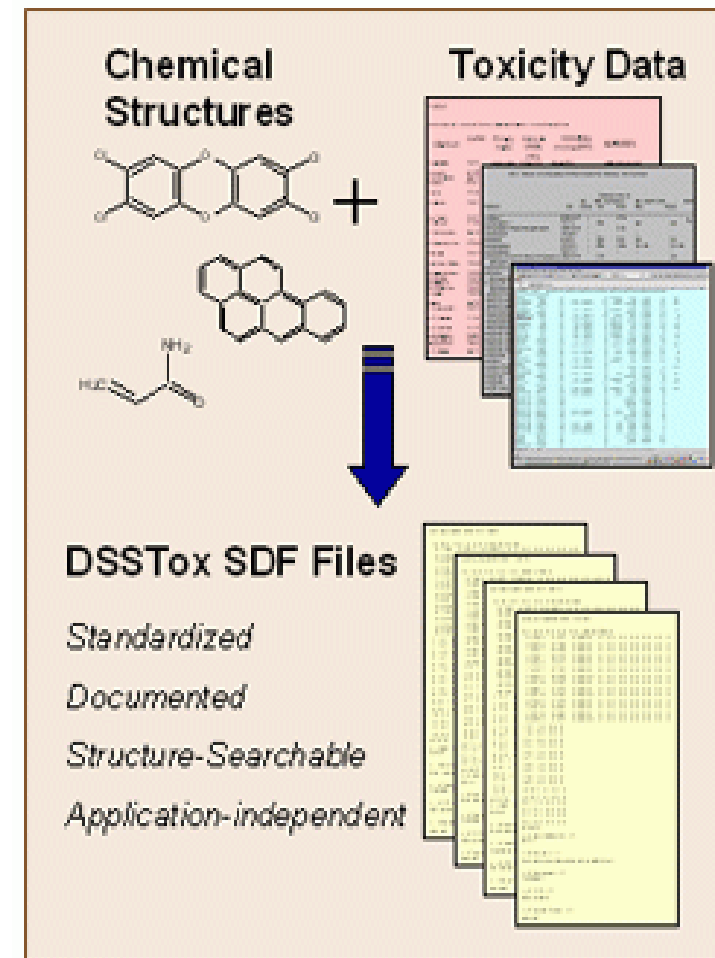
American Chemical Society Meeting, Fall 2019
28 August 2019, San Diego, FL

1. National Center of Computational Toxicology, U.S. EPA
(soon to be the Center for Computational Toxicology and Exposure)

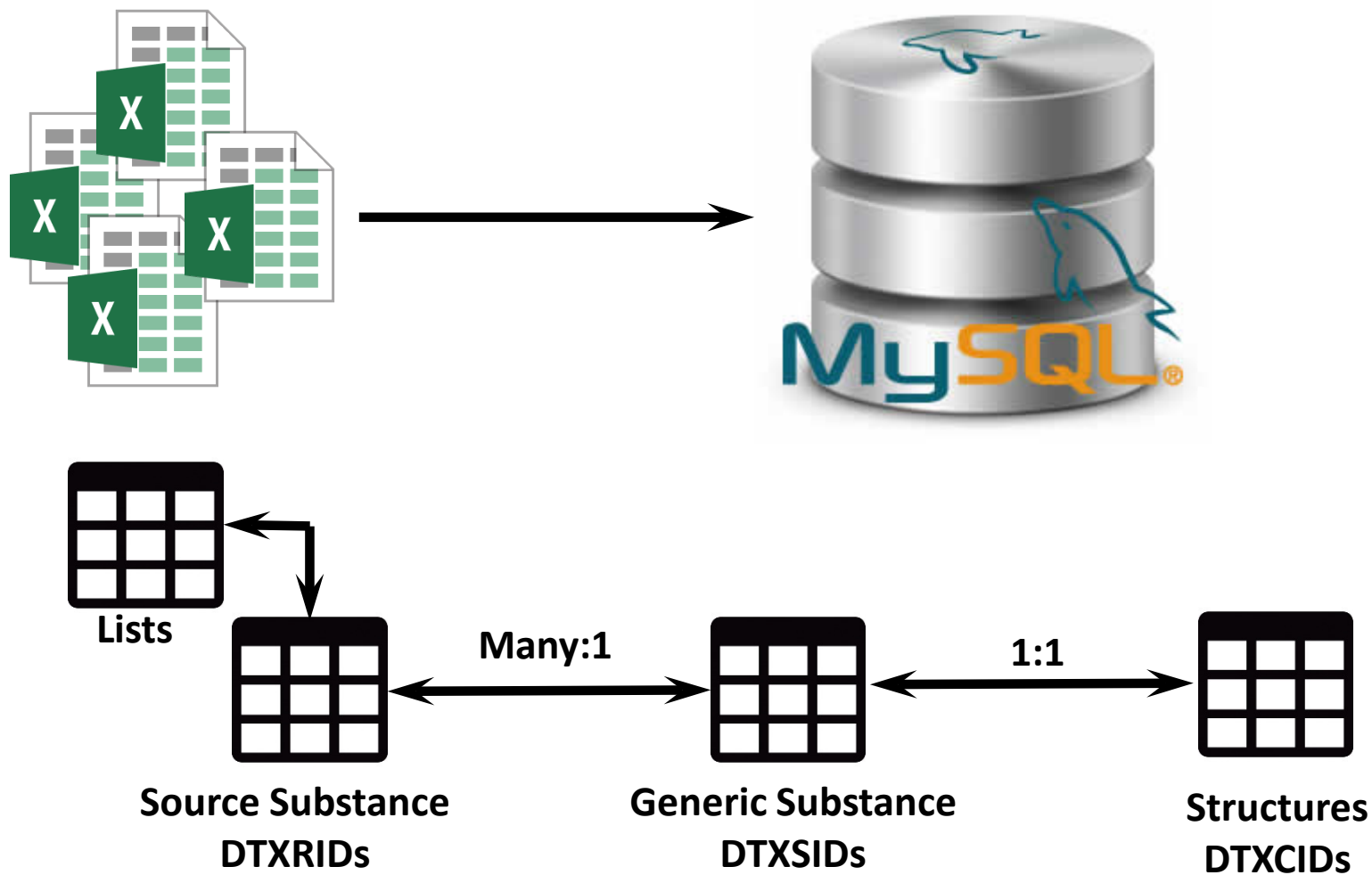
DSSTox Ancient History

Goal: Linking chemical structures to data enabling SAR

- First release of data files in 2004
- Focused on high impact sets of data
 - Carcinogenic Potency Database
 - Drinking water disinfection by-products
 - EPA's Integrated Risk Information System
 - FDA's Maximum Daily Dose dataset
 - EPA's Fat Head Minnow Toxicity dataset
 - etc...
- Managed all chemical registration for ToxCast and Tox21 chemicals
- By 2014, roughly 20K manually curated substance records



Generalized DSSTox Storage Architecture



Chemical Registration

ACToR-DSSTox Chemical Registration

**View/Edit a
Single Record**

Structure
Search

Browse/Curate
Records

Export DSSTox

Chemotypes

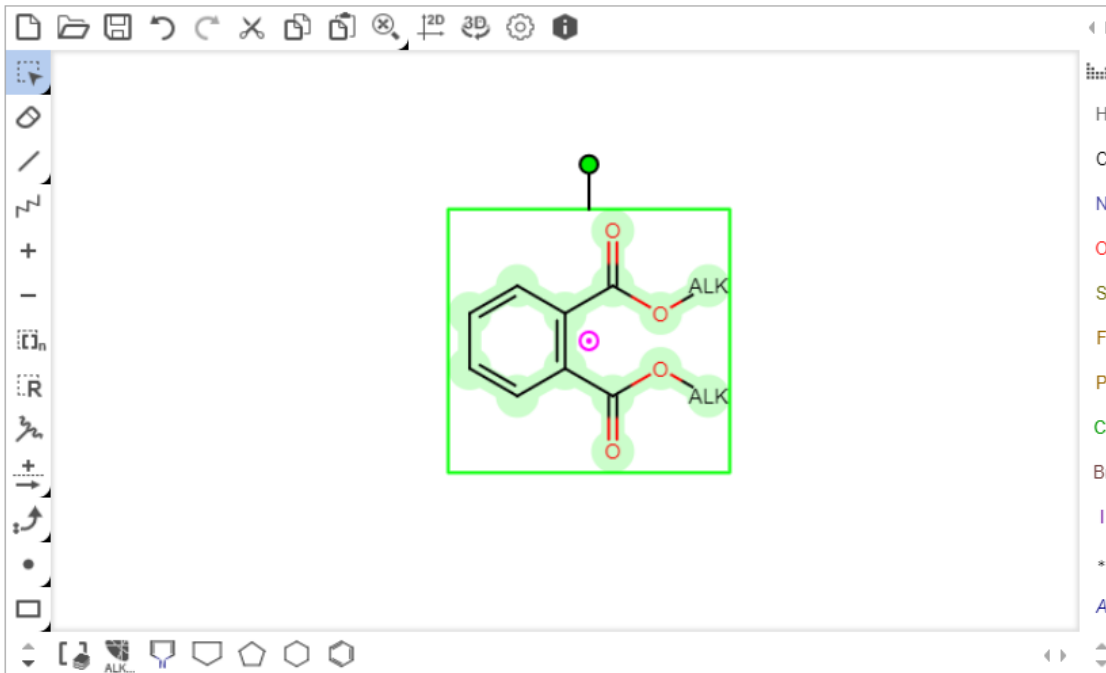
Manage
Chemical Lists

Manage
Property Data

Add Deleted
Casrns

CAS-RN matched
null
You are viewing the
record associated with
DTXSID5028665
CASRN: 68515-48-0

Q 68515-48-0



Calculate from Structure

Substance_ID: DTXSID5028665

CAS: 68515-48-0

Name: DINP branched

Substance Type: Mixture/Formulation

QC Level: DSSTox_High

Data Source: STN(DSSTox)

mixture of dinonyl phthalates

QC Notes:

Compound_ID:

Chemical Shown:

Markush Query

Private Notes:

Source of CAS-Compound:

Public

Double Stereo:

None

Chiral Stereo:

Unspecified

Chemical Form:

Organic

Chemical List Registration

Hits						
	ssCAS-RN	ssName	Hit Desc	Hit Substance_ID	Hit Casrn	Hit Name
☐		Norgestrel	Structure matched SMILES	DTXSID10859541	NOCAS_859541	13-Ethyl-17-ethynyl-17-hydroxy-1,2,6,7,8,9,10,11,12,13,14,15,16,17-tetradecahydro-3H-cyclopenta[a]phenanthren-3-one (non-preferred name)
☐		Norgestrel	Mapped Identifier matched NAME1	DTXSID3036496	797-63-7	Levonorgestrel
☐		Norgestrel	Mapped Identifier matched NAME1	DTXSID3047477	6533-00-2	dl-Norgestrel

SMILES

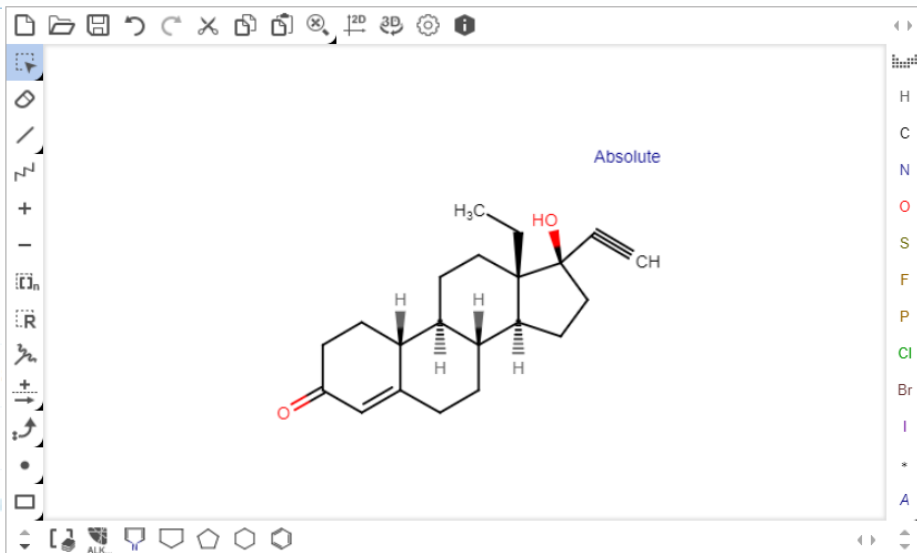
Valid Synonym matched **NAME2**
Preferred Name matched **NAME1**
Valid Synonym matched other record: **NAME1**
Unique Synonym matched other record: **NAME2**
Unique Synonym matched other record: **NAME2**

Mapped Identifier matched **NAME1**

Mapped Identifier matched **NAME2**

Name2Structure

Export All



Calculate from Structure

Substance_ID: DTXSID3036496

CAS: 797-63-7

Name: Levonorgestrel

Substance Type: Single Compound

QC Level: DSSTox_High

Data Source: STN(DSSTox)

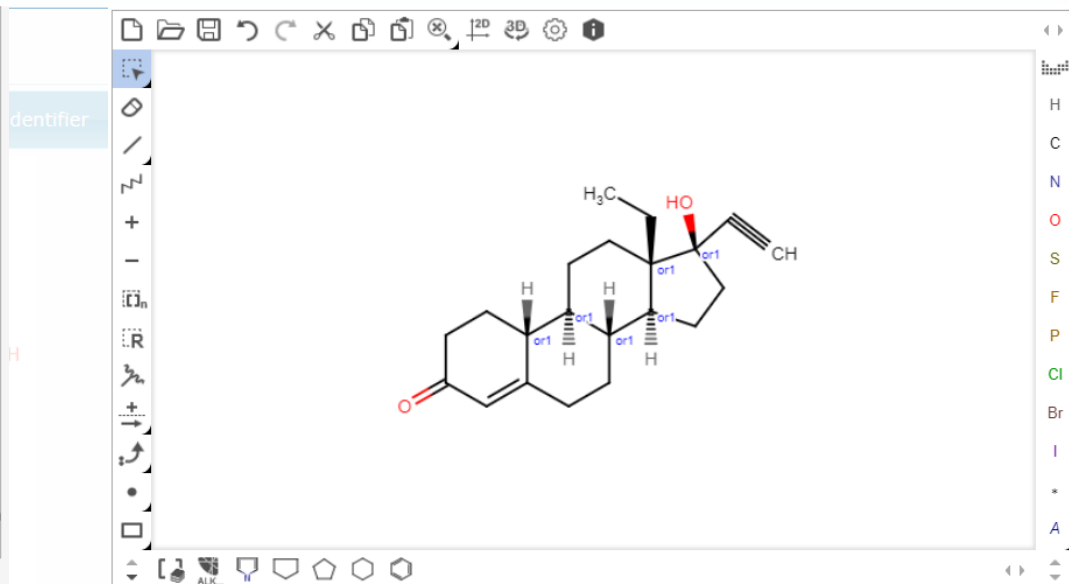
Compound_ID:

Chemical Shown:

Private Notes:

Source of CAS-Compound:

Double Stereo:



Calculate from Structure

Substance_ID: DTXSID3047477

CAS: 6533-00-2

Name: dl-Norgestrel

Substance Type: Mixture of Stereoisomers

QC Level: DSSTox_High

Compound_ID:

Chemical Shown:

Private Notes:

DTXSID

-UHFFFAOYNA-N

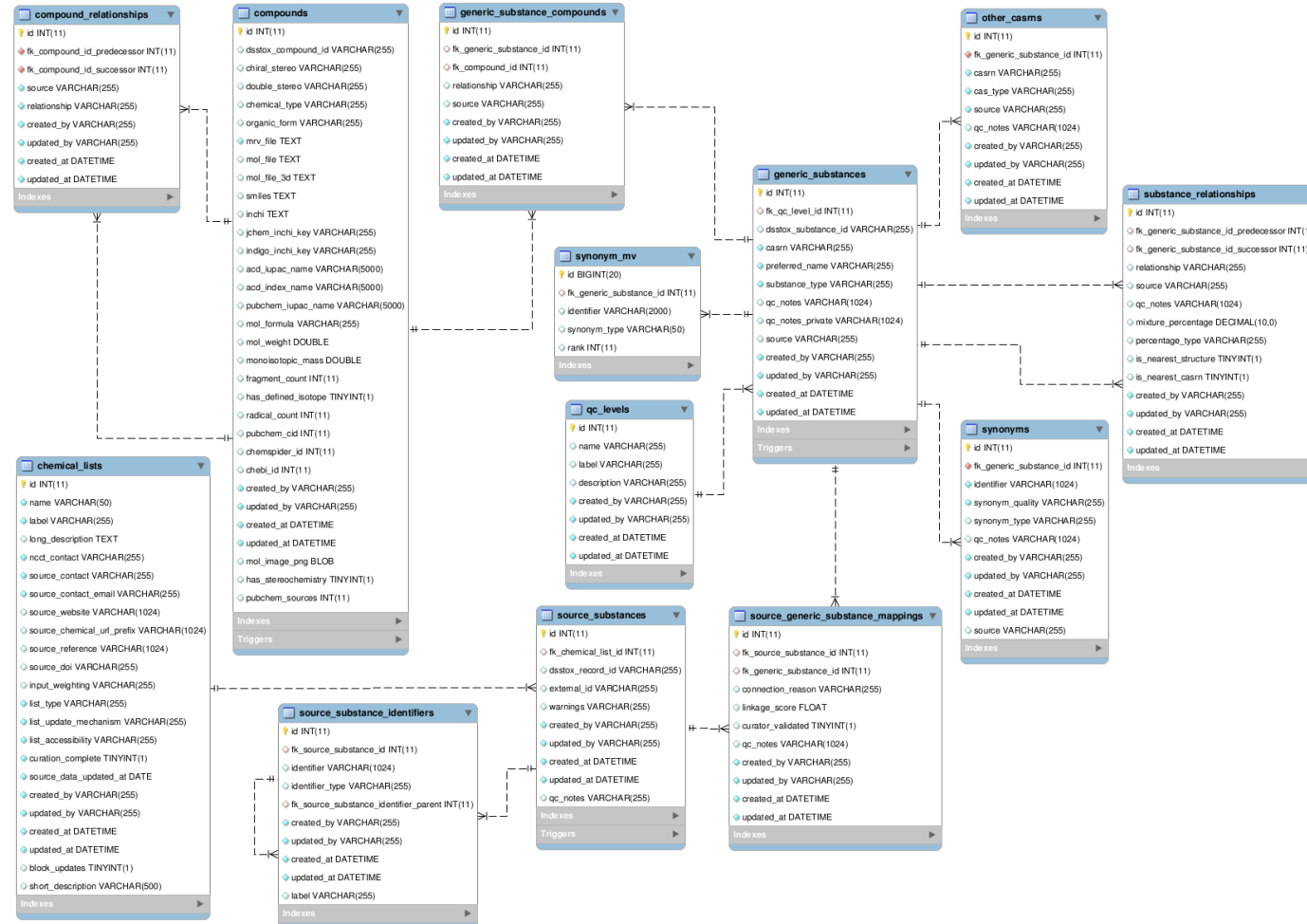
5,17-
ed name)

5,17-
ed name)

Hits

DSSTox Chemistry Schema

19 Tables



ChemTrack Bottles


ChemTrack
Inventory ▾
Partnerships ▾
Ordering ▾
Data

Hi Chris !

Uploaded MOSAIC Files **25**
Mapped & Available Bottles **29850**
Unmapped Bottles **449**
External Bottles **98**
Add New MOSAIC
Create External Bottle

Mapped & Available Bottles

Show entries

Search:

Barcode Type	Barcode	COA Summary	Compound Name	CAS	QTY Available	Units	Vendor	SAM	CPD	Can Plate?	Comment	
INVALID_SUPTX0013222_INVALID	ALID_SUPTX0013222_INVALID	2660	Aluminium phthalocyanine chloride	14154-42-8	1000	mg	Sigma Chemical Company	SAM004888816	CPD003650672	Yes		Edit
INVALID_SUPTX0013294_INVALID	ALID_SUPTX0013294_INVALID	2735	Nickle(III) carbonate basic hydrate	12607-70-4	250000	mg	Sigma Chemical Company	SAM004888887	CPD003650704	Yes		Edit
EPA_Vial_Source_Storage	BF00079587	18768	Pentabromophenol	608719	178	mg	Sigma Chemical Company	SAM006061824	CPD001224527	Yes		Edit
EPA_Vial_Source_Storage	BF00079581	18774	Phthalic anhydride	85-44-9	193	mg	Sigma Chemical Company	SAM006061820	CPD001252223	Yes		Edit
EPA_Vial_Source_Storage	BF00079583	18772	TRIPOLI	7631-86-9	196	mg	Sigma Chemical Company	SAM006061823	CPD001252283	Yes		Edit
EPA_Vial_Source_Storage	BF00079582	18773	4,4'-Methylenebis(2-chloroaniline)	101144	188	mg	Sigma Chemical Company	SAM006061819	CPD001307314	Yes		Edit
EPA_Vial_Source_Storage	BF00079585	18770	4-sec-Butyl-2,6-di-tert-butylphenol	17540-75-9	188	mg	Sigma Chemical Company	SAM006061825	CPD004560495	Yes		Edit
EPA_Vial_Source_Storage	BF00079584	18771	Creosote	8001-58-9	178	mg	Sigma Chemical Company	SAM006061821	CPD004757028	Yes		Edit
EPA_Vial_Source_Storage	BF00079586	18769	tert-Amyl methyl ether	994-05-8	194	mg	Sigma Chemical Company	SAM006061822	CPD004757029	Yes		Edit
EPA_Vial_Source_Storage	BF00079580	18775	3,3'-	91-94-	188	mg	Sigma Chemical Company	SAM006061826	CPD004757030	Yes		Edit

ChemTrack Search

Find Chemicals

ChemTrack

Inventory Partnerships Ordering Data

Single Chemical Search

Info! Find by multiple bottles

aspirin

bpa

50-00-0

tylenol

tce

Additional Filters

All

Solution

Neat

Concentration

Minimum

Minimum Amount

Amount

Multiple Chemical Search

Show 10 entries

Search:

	Searched By	Found By	DTXSID	Name	CASRN	Neat(mg)	0-24mM Stock(ul)	24-100mM Stock(ul)	Number of Bottles
+	tylenol	Synonym from Valid Source	DTXSID2020006	Acetaminophen	103-90-2	HIGH	HIGH	HIGH	15
+	bpa	Expert Validated Synonym	DTXSID7020182	Bisphenol A	80-05-7	HIGH	HIGH	HIGH	39
-	tce	Expert Validated Synonym	DTXSID2021319	Tetrachloroethylene	127-18-4	NONE	NONE	NONE	3

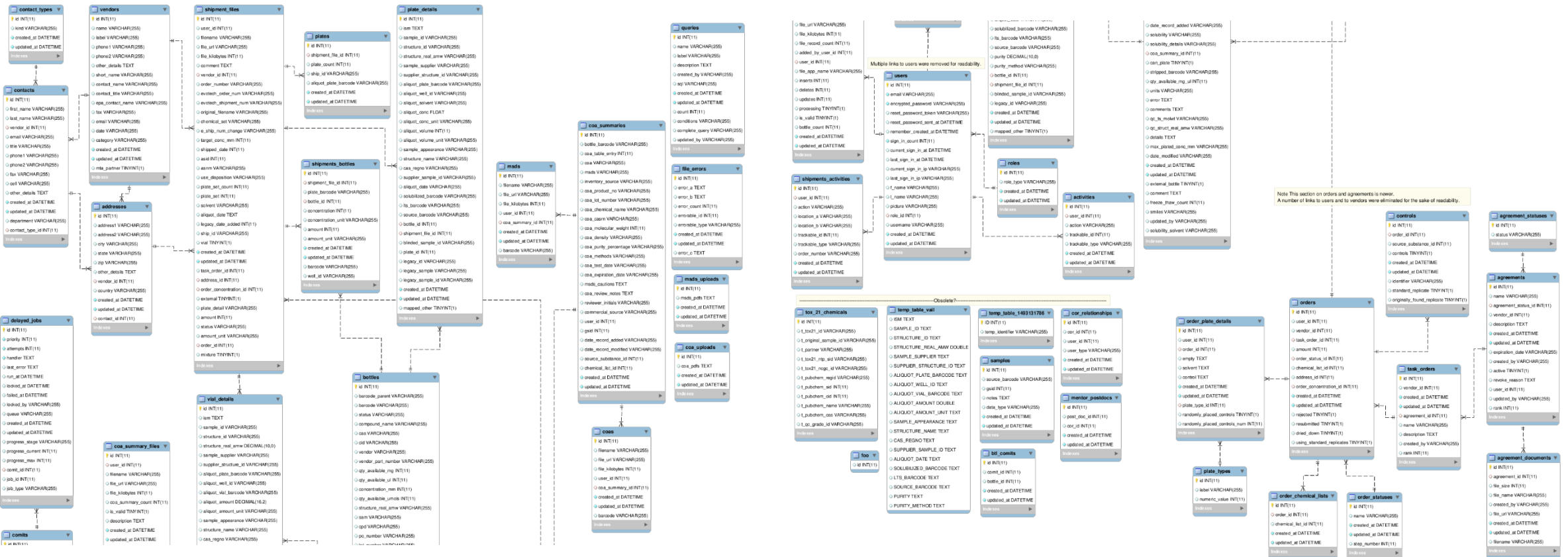
Barcode	Supplier	QTY	Units	Concentration (mM)	Solubility Solvent
TX009583	LightBiologicals	-	mg	-	-
TX009584	LightBiologicals	-	mg	-	-
Tox21_201196_legacy	-	-	-	-	-

	50-00-0	CAS-RN	DTXSID7020637	Formaldehyde	50-00-0	NONE	NONE	NONE	0
-	aspirin	Approved Name	DTXSID5020108	Aspirin	50-78-2	HIGH	HIGH	HIGH	14

Barcode	Supplier	QTY	Units	Concentration (mM)	Solubility Solvent
00891165	Enamine	-	mg	-	-
TX003515	Sigma Chemical Company	17831	ul	20	DMSO
TX003516	Sigma Chemical Company	19	mg	-	-
TX016586	Sigma Chemical Company	2742	ul	99	DMSO

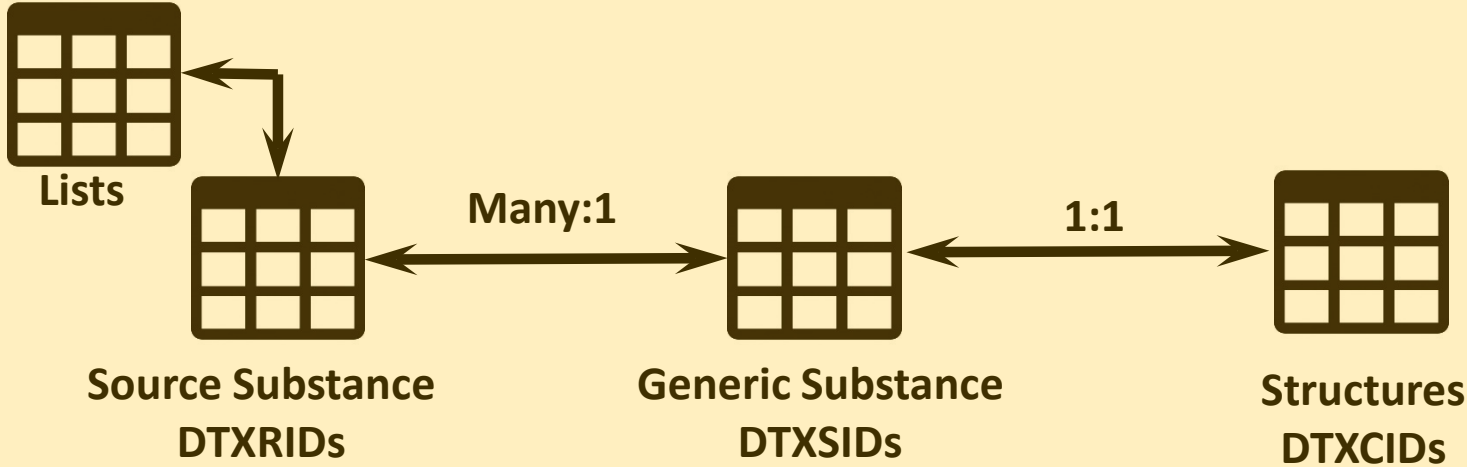
11/11/2019

57 Tables

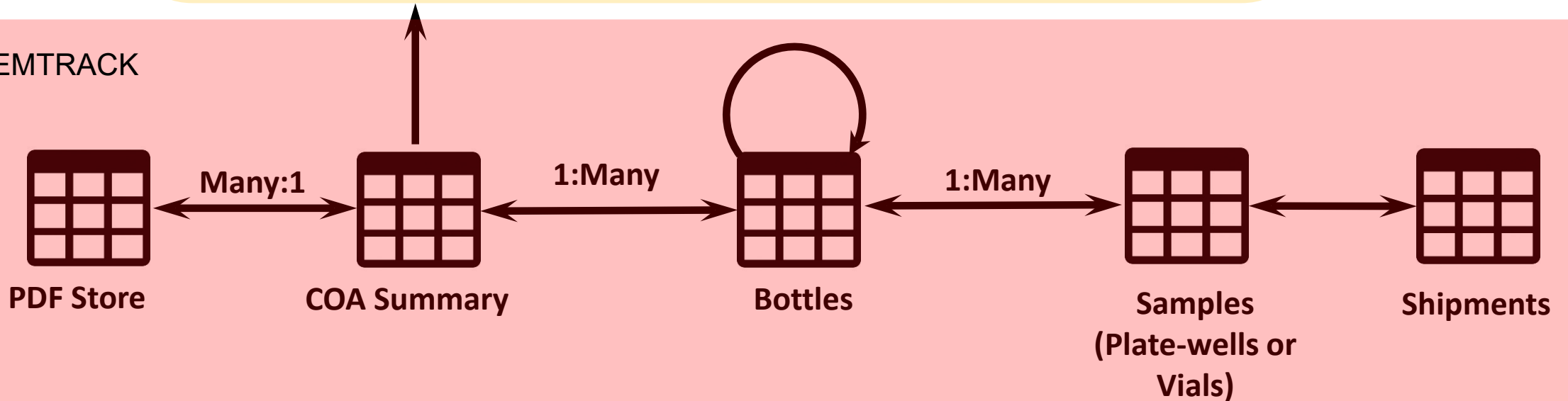


Generalized Data Model

DSSTOX CHEMISTRY



CHEMTRACK



Sharing the DSSTox Database

ChemTrack

Multiple Chemical Search

Search:

Show 19 entries

Searched By	Found By	DTXSID	Name	CASRN	Neat(mg)	0-24mM Stock(ul)	24-100mM Stock(ul)	Number of Bottles
tylenol	Synonym from Valid Source	DTXSID2020006	Acetaminophen	103-90-2	HIGH	HIGH	HIGH	15
lpa	Expert Validated Synonym	DTXSID7020182	Bisphenol A	80-05-7	HIGH	HIGH	HIGH	39
tce	Expert Validated Synonym	DTXSID2021319	Tetrachloroethylene	127-18-4	NONE	NONE	NONE	3
Barcode	Supplier	QTY	Units	Concentration (mM)	Solubility	Solvent		
TX009583	Enamine	-	mg	-	-	-	-	-
TX009584	Sigma Chemical Company	17831	ul	20	DMSO	-	-	-
TX021_201194_legacy	Sigma Chemical Company	19	mg	-	-	-	-	-
50-00-0	CAS-RN	DTXSID7020637	Formaldehyde	50-00-0	NONE	NONE	NONE	0
aspirin	Approved Name	DTXSID5020108	Aspirin	50-78-2	HIGH	HIGH	HIGH	14
Barcode	Supplier	QTY	Units	Concentration (mM)	Solubility	Solvent		
0891165	Enamine	-	mg	-	-	-	-	-
TX003515	Sigma Chemical Company	17831	ul	20	DMSO	-	-	-
TX003516	Sigma Chemical Company	19	mg	-	-	-	-	-
TX003586	Sigma Chemical Company	17831	ul	20	DMSO	-	-	-

ACToR-DSSTox Chemical Registration

View/Edit a Single Record | Structure Search | Business/Calendar | Export DSSTox Records | Chemtypes | Manage Chemical Lists | Manage Property Data | Add Deleted | Add Deleted |

CAS-RN matched 40-mg(ul) 0- You are viewing the record associated with DTXSID5020665 CASRN: 68515-48-0

Q: 68515-48-0

ChemReg

Substance ID: DTXSID5020665

CAS: 68515-48-0

Name: DIMP branched

Substance Type: Mixture/Formulation

QC Level: DSSTox_High

Data Source: STN(DSSTox)

QC Notes: mixture of diisopropyl phthalate

Compound ID: Chemical Shown: Markush Query

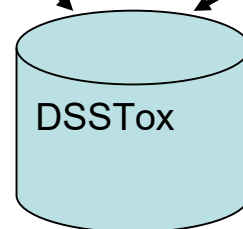
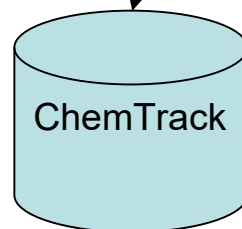
Private Notes:

Source of CAS-Compound: Public

Double Stereo: None

Chiral Stereo: Unspecified

Chemical Form: Organic



“ChemSpiderMan” comes to NCCT (2015)

- DSSTox: ~700K substance records (25K manually curated)



This data is great!

We need to make it
available to the
public!

Sharing the DSSTox Database

ChemTrack

Multiple Chemical Search

Searched By	Found By	DTXSID	Name	CASRN	Neat(mg)	0-24mM Stock(ul)	24-100mM Stock(ul)	Number of Bottles
tylenol	Synonym from Valid Source	DTXSID0202006	Acetaminophen	100-90-2	HIGH	HIGH	HIGH	15
lps	Synonym from Valid Source	DTXSID0201882	Streptol A	80-90-0	HIGH	HIGH	HIGH	39
tea	Synonym from Valid Source	DTXSID0201882	Streptol A	80-90-0	HIGH	HIGH	HIGH	3

Barcode	Supplier	QTY	Units	Concentration (mg)	Solubility Solvent
TX009583	LightBioscience	-	mg	-	-
TX009584	LightBioscience	-	mg	-	-
Tx021_201194_legacy	-	-	-	-	-

Barcode	Supplier	QTY	Units	Concentration (mg)	Solubility Solvent
00861105	Crumine	-	mg	-	-
TX003515	Sigma Chemical Company	17831	ul	20	DMSO
TX003516	Sigma Chemical Company	19	mg	-	-
TX014595	Sigma Chemical Company	3343	ul	80	DMSO

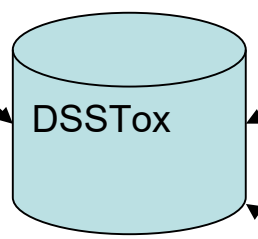
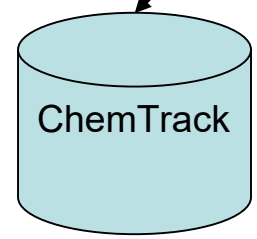
ChemReg

ACToR-DSSTox Chemical Registration

CAS-RN matched 437441-00
You are viewing the record associated with DTXSID0528665
CASRN: 68515-48-0

Substance_ID: DTXSID0528665
CAS: 68515-48-0
Name: DIMP branched
Substance Type: Mixture/Formulation
QC Level: DSSTox_High
Data Source: STN(DSSTox)
QC Notes: mixture of dimethyl phthalate

Compound_ID:
Chemical Shown:
Markush Query:
Private Notes:
Source of CAS-Compound: Public
Double Stereo: None
Chiral Stereo: Unspecified
Chemical Form: Organic



Dashboard

DL-Alanine
302-72-7 | DTXSID6031255
Searched by Synonym from Valid Source.

DETAILS

- EXECUTIVE SUMMARY
- PROPERTIES
- ENV. FATE/TRANSPORT
- HAZARD
- ADME
- EXPOSURE
- BIOACTIVITY
- SIMILAR COMPOUNDS
- GENRA (BETA)
- RELATED SUBSTANCES
- SYNONYMS
- LITERATURE
- LINKS

Wikipedia

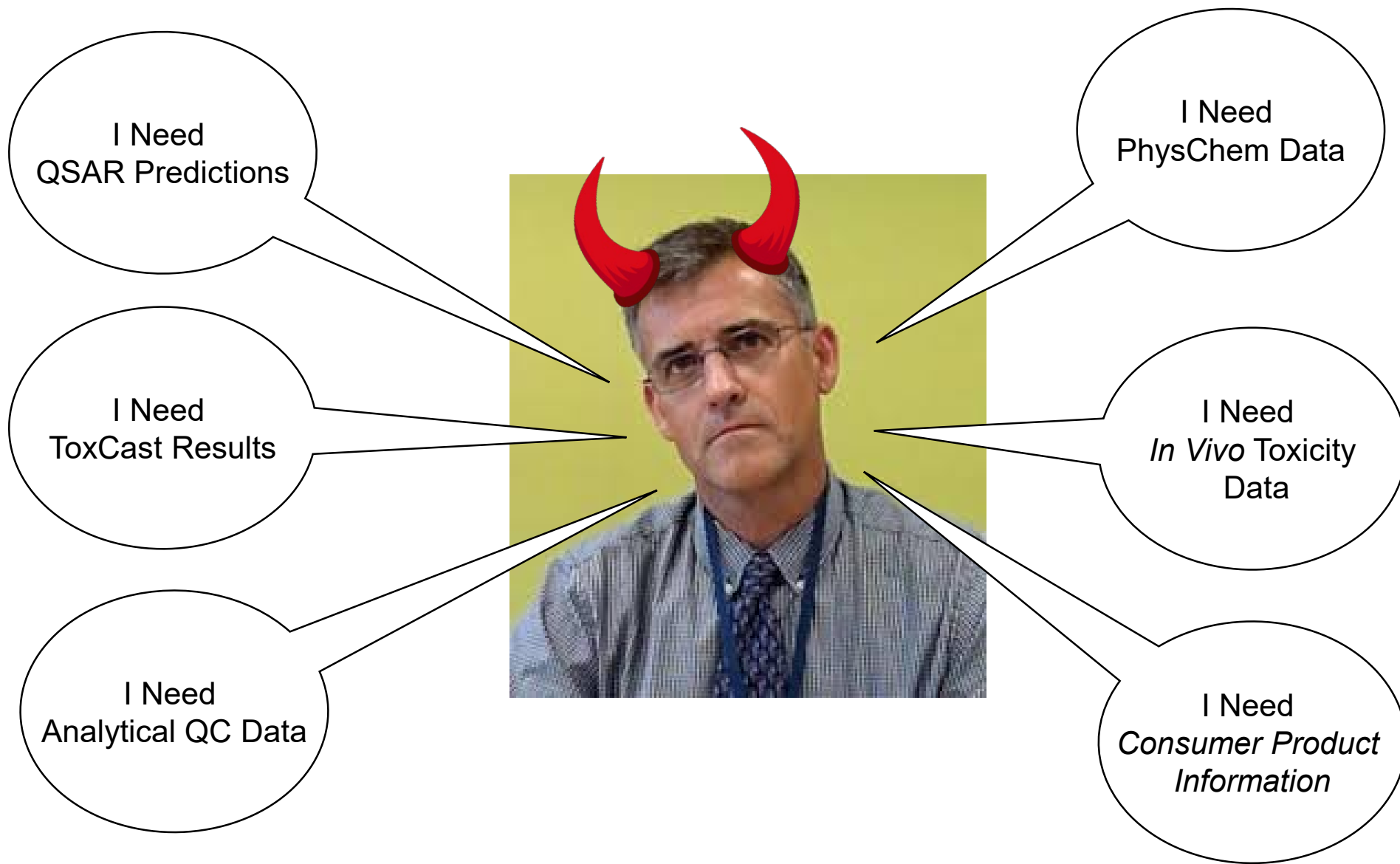
DL-Alanine (symbol **Ala** or **Al**) is an α-amino acid that is used in the biosynthesis of proteins. It contains group, both attached to the central carbon atom which also carries a methyl group side chain. Cond 2-aminopropanoic acid, and it is classified as a nonpolar, aliphatic α-amino acid. Under biological c form with its amine group protonated (as -NH₃⁺).

Intrinsic Properties

Structural Identifiers

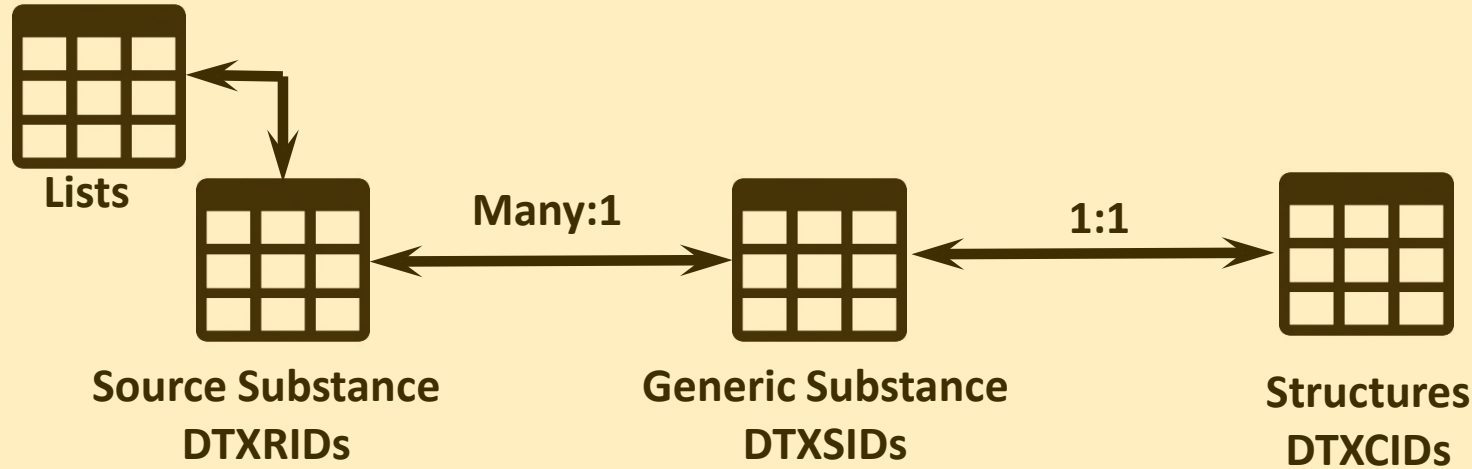
- IUPAC Name: Alanine
- SMILES: CC(N)C(=O)O
- InChI String: InChI=1/C3H7NO2/c1-2(4)3(5)6/h2H,4H2,1H3,(1,5,6)
- InChIKey: QNAYBMKLOCPYGI-UHFFFAOYSA-N

The Dashboard Taskmaster

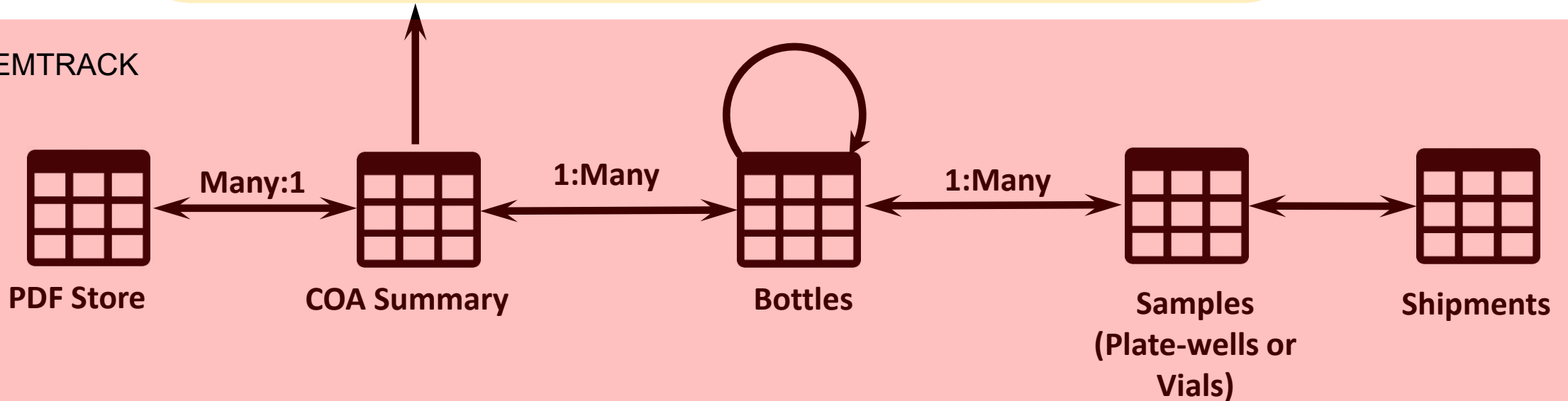


I Need PhysChem Data...

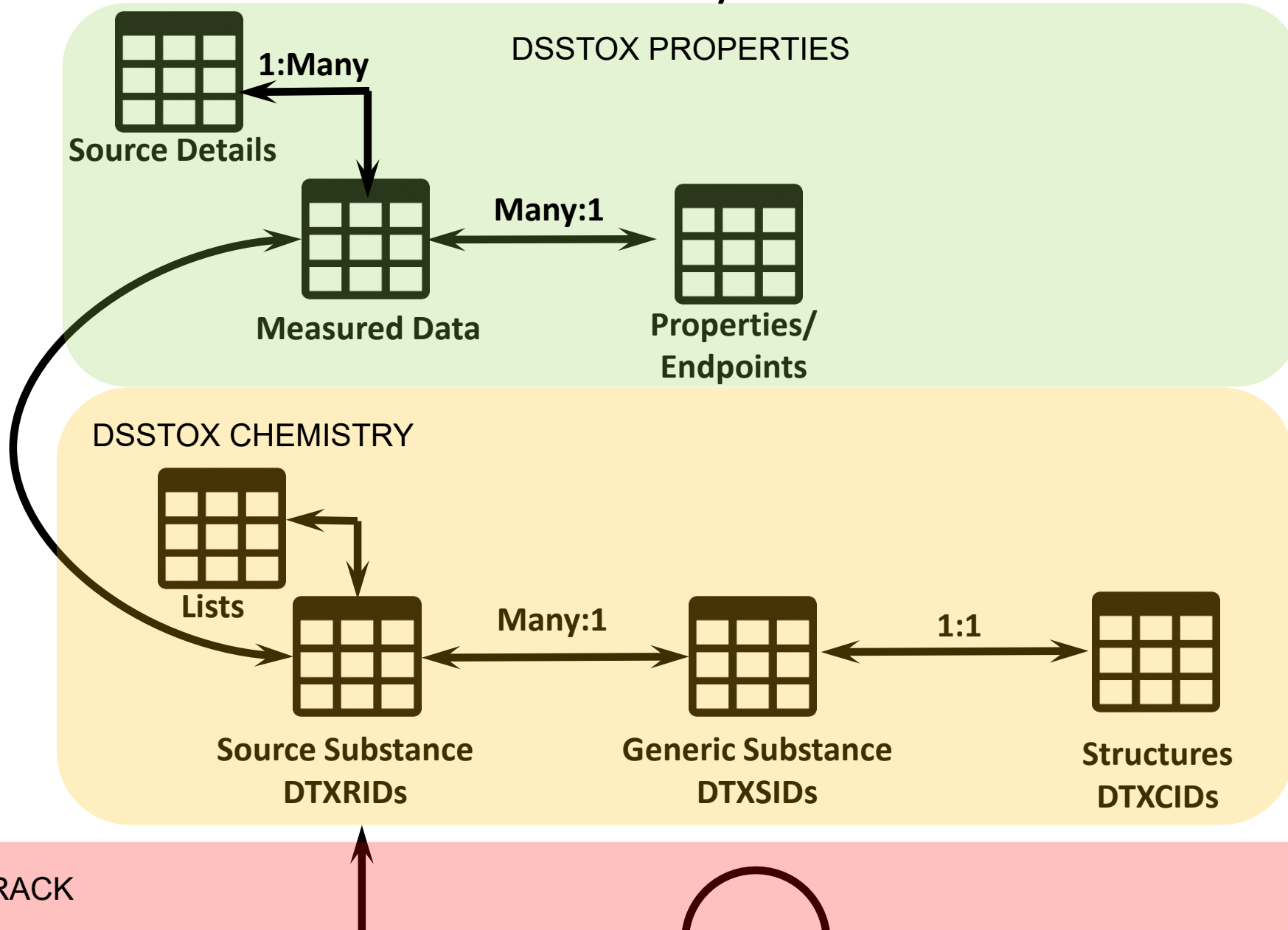
DSSTOX CHEMISTRY



CHEMTRACK



I Need PhysChem Data...



Sharing the DSSTox Database

ChemTrack

Multiple Chemical Search

Searched By	Found By	DTXSID	Name	CASRN	Neat(mg)	0-24mM Stock(ul)	24-100mM Stock(ul)	Number of Bottles
tylenol	Synonym from Valid Source	DTXSID0202006	Acetaminophen	100-90-2	HIGH	HIGH	HIGH	15
aspirin	Synonym from Valid Source	DTXSID0201882	Aspirin	50-78-2	HIGH	HIGH	HIGH	39
formaldehyde	Synonym from Valid Source	DTXSID0201108	Formaldehyde	50-00-0	NONE	NONE	NONE	0

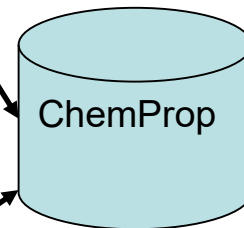
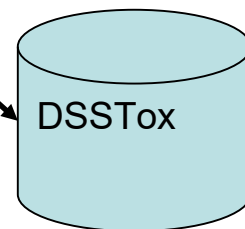
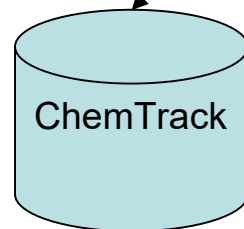
ChemReg

ACToR-DSSTox Chemical Registration

CAS RN matched: 68515-48-0
You are viewing the record associated with: DTXSID05028665
CASRN: 68515-48-0

Substance ID: DTXSID05028665
CAS: 68515-48-0
Name: DIMP branched
Substance Type: Mixture/Formulation
QC Level: DSSTox High
Data Source: STN(DSSTox)
QC Notes: mixture of dimethyl phthalate

Chemical Structure: CC1=CC=C(C=C1)C(=O)OC2=CC=CC=C2



Dashboard

DL-Alanine
302-72-7 | DTXSID6031255
Searched by Synonym from Valid Source.

DETAILS

- EXECUTIVE SUMMARY
- PROPERTIES
- ENV. FATE/TRANSPORT
- HAZARD
- ADME
- EXPOSURE
- BIOACTIVITY
- SIMILAR COMPOUNDS
- GENRA (BETA)
- RELATED SUBSTANCES
- SYNONYMS
- LITERATURE
- LINKS

Wikipedia

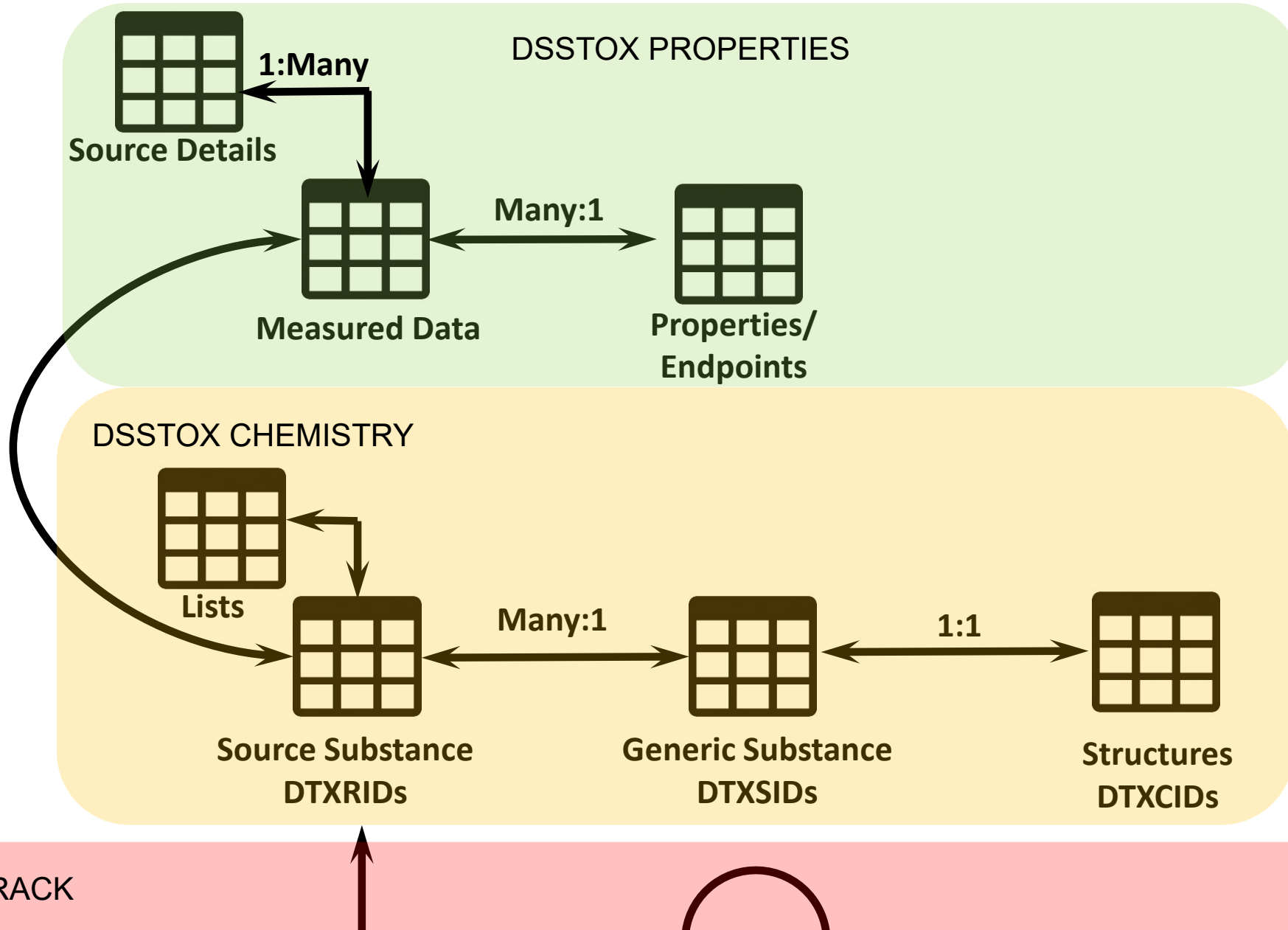
Alanine (symbol **Ala** or **Al**) is an α-amino acid that is used in the biosynthesis of proteins. It contains group, both attached to the central carbon atom which also carries a methyl group side chain. Cond. 2-aminopropanoic acid, and it is classified as a nonpolar, aliphatic α-amino acid. Under biological conditions, it is found in its zwitterion form with its amine group protonated (as -NH₃⁺).

Intrinsic Properties

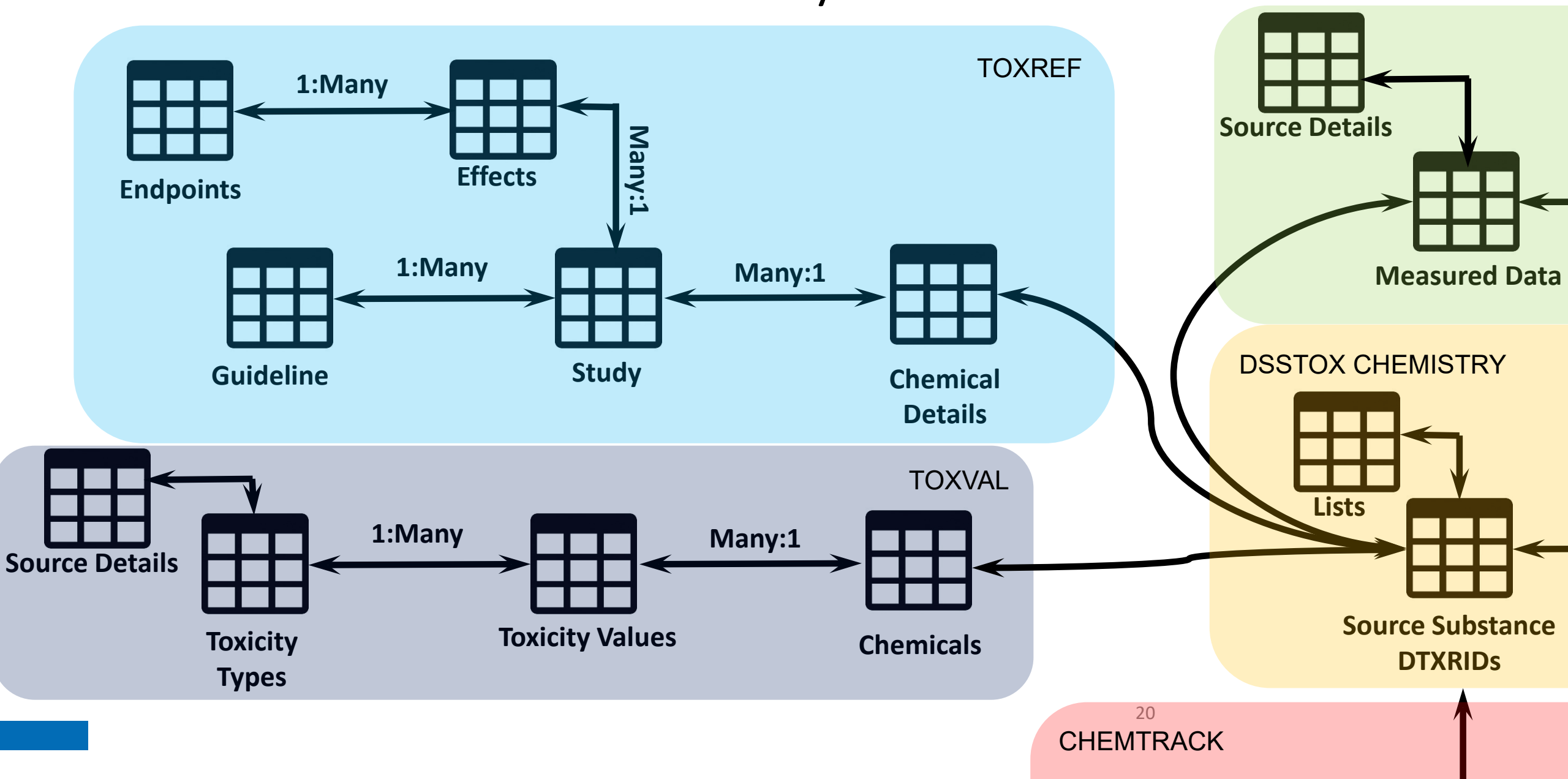
Structural Identifiers

- IUPAC Name: Alanine
- SMILES: CC(N)C(=O)O
- InChI String: InChI=1/C3H7NO2/c1-2(4)3/5(6)/2H,4H(2,1H3,1,5,6)
- InChIKey: QNAYBMKOCPPVGJ-UNFFFAOYSA-N

I Need *In Vivo* Toxicity Data...



I Need *In Vivo* Toxicity Data...



Sharing the DSSTox Database

ChemTrack

Multiple Chemical Search

Searched By	Found By	DTXSID	Name	CASRN	Neat(mg)	0-24mM Stock(ul)	24-100mM Stock(ul)	Number of Bottles
tylenol	Synonym from Valid Source	DTXSID0202006	Acetaminophen	103-90-2	HIGH	HIGH	HIGH	15
lps	Synonym from Valid Source	DTXSID020182	Streptol A	80-50-0	HIGH	HIGH	HIGH	39
tea	Synonym from Valid Source	DTXSID020182	Streptol A	80-50-0	HIGH	HIGH	HIGH	3

Barcode	Supplier	QTY	Units	Concentration (mg)	Solubility Solvent
TX009583	LightBioscience	-	mg	-	-
TX009584	LightBioscience	-	mg	-	-
Tx021_201104_legacy	-	-	-	-	-

Barcode	Supplier	QTY	Units	Concentration (mg)	Solubility Solvent
50-00-0	CAS-RN	DTXSID7020637	Formaldehyde	50-00-0	NONE
aspirin	Approved Name	DTXSID0020108	Aspirin	50-78-2	HIGH

Barcode	Supplier	QTY	Units	Concentration (mg)	Solubility Solvent
00861105	Crumine	-	mg	-	-
TX003515	Sigma Chemical Company	17831	ul	20	DMSO
TX003516	Sigma Chemical Company	19	mg	-	-
TX014595	Sigma Chemical Company	3343	ul	80	DMSO

ChemReg

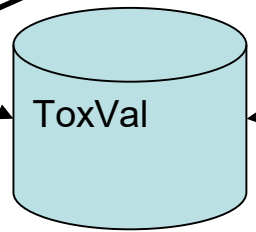
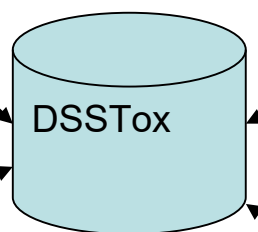
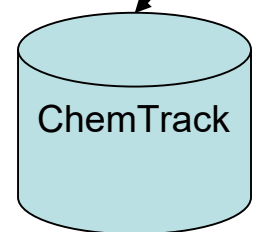
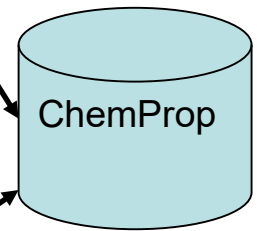
ACToR-DSSTox Chemical Registration

CAS-RN matched 43-79-6-9
You are viewing the record associated with DTXSID0020605
CASRN: 68515-48-0

Chemical Structure: CC1=CC=C(C=C1)C(=O)N

Substance ID: DTXSID0020605
CAS: 68515-48-0
Name: DIMP branched
Substance Type: Mixture/Formulation
QC Level: DSSTox_High
Data Source: STN(DSSTox)
QC Notes: mixture of dimethyl phthalate

Compound ID: Chemical Shown
Private Notes: Markush Query
Source of CAS-Compound: Public
Double Stereo: None
Chiral Stereo: Unspecified
Chemical Form: Organic



Dashboard

EPA United States Environmental Protection Agency

DL-Alanine
302-72-7 | DTXSID6031255
Searched by Synonym from Valid Source.

Chemical Structure: CC(N)C(=O)O

DETAILS

- EXECUTIVE SUMMARY
- PROPERTIES
- ENV. FATE/TRANSPORT
- HAZARD
- ADME
- EXPOSURE
- BIOACTIVITY
- SIMILAR COMPOUNDS
- GENRA (BETA)
- RELATED SUBSTANCES
- SYNONYMS
- LITERATURE
- LINKS

Wikipedia

Alanine (symbol **Ala** or **Al**) is an α -amino acid that is used in the biosynthesis of proteins. It contains group, both attached to the central carbon atom which also carries a methyl group side chain. Cond. 2-aminopropanoic acid, and it is classified as a nonpolar, aliphatic α -amino acid. Under biological conditions it is found in its zwitterion form with its amine group protonated (as $-\text{NH}_3^+$).

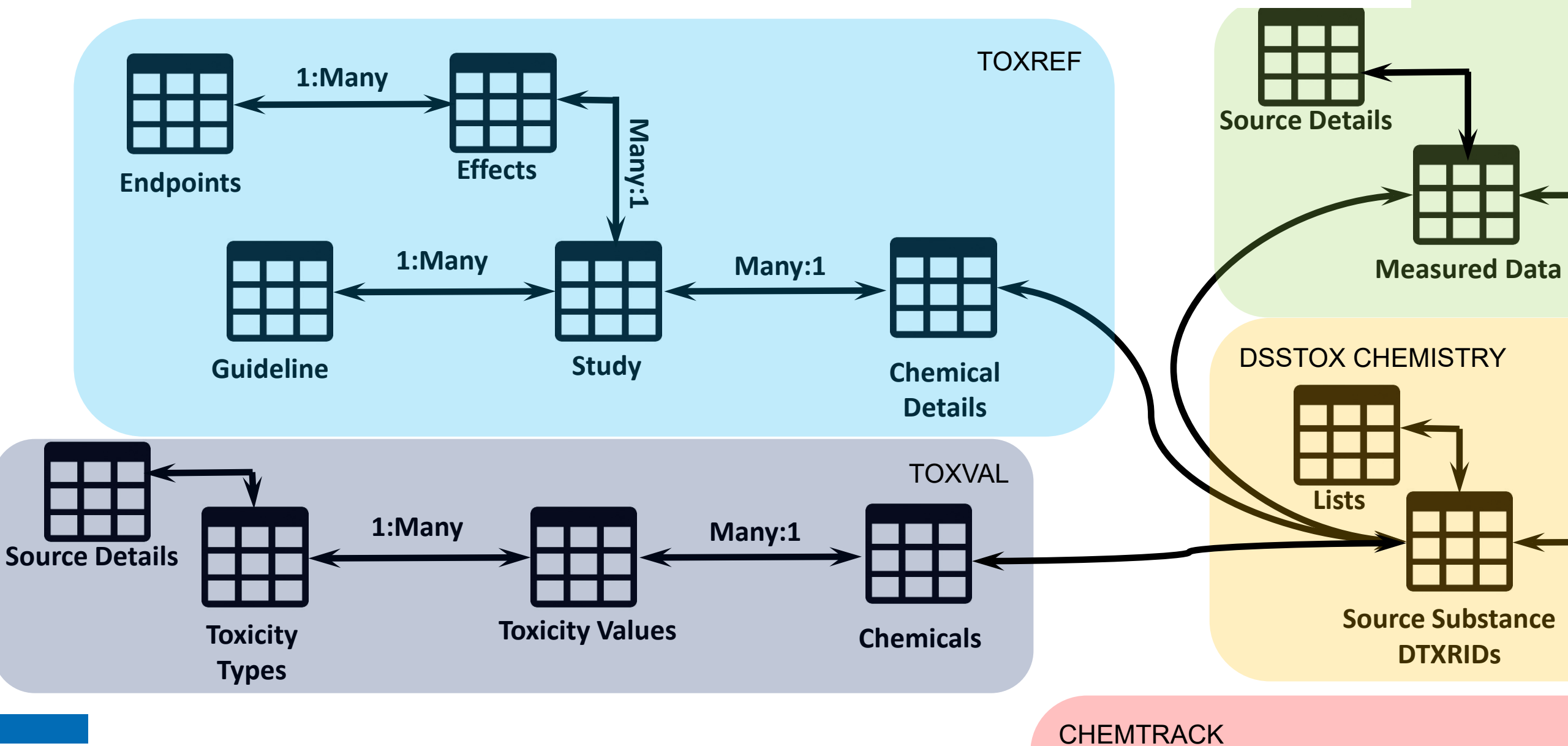
Intrinsic Properties

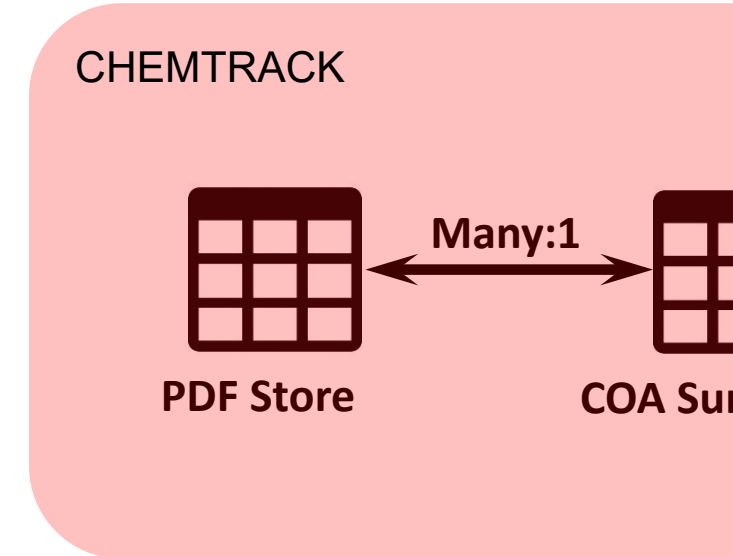
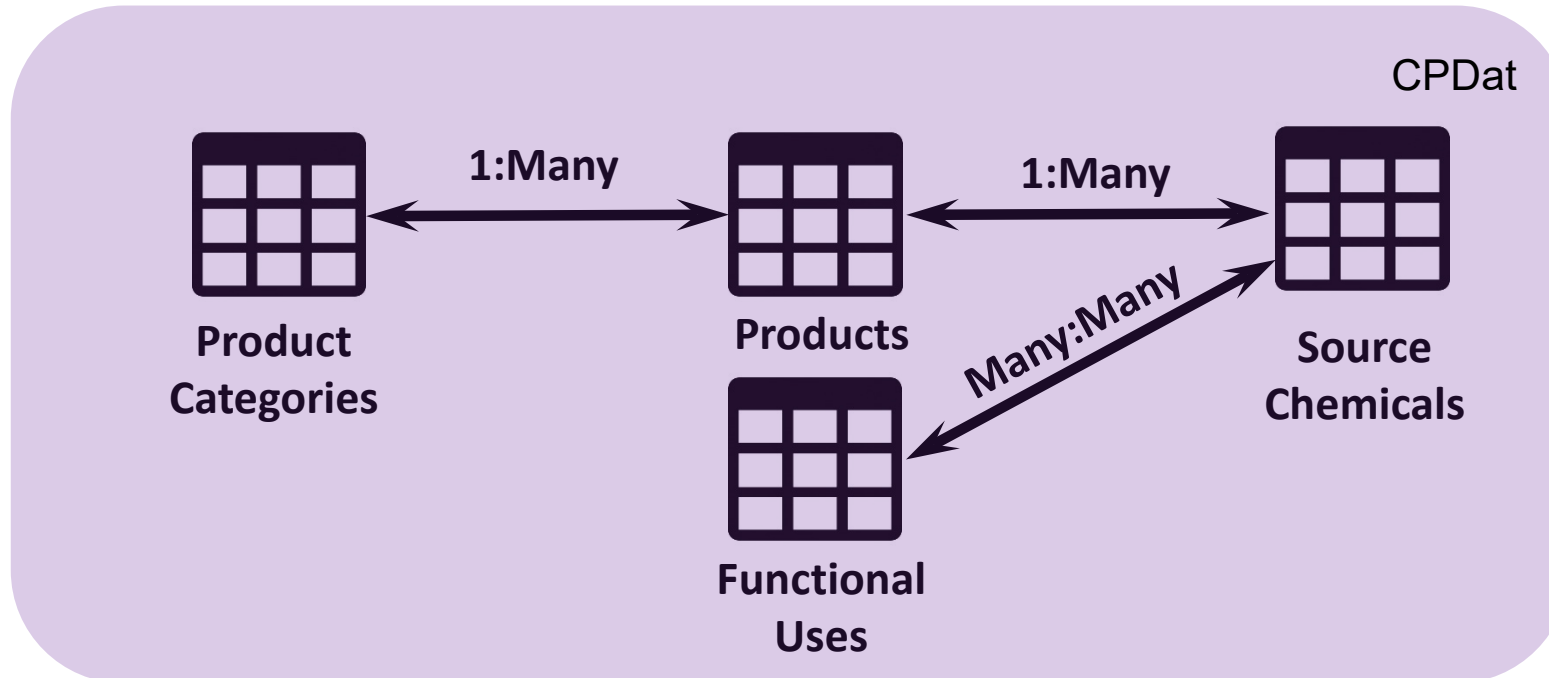
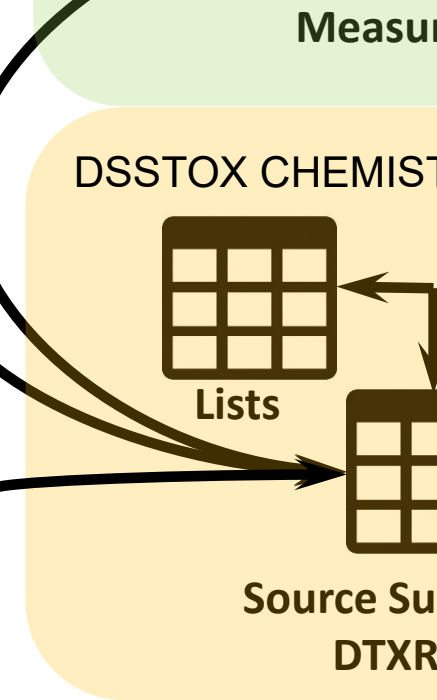
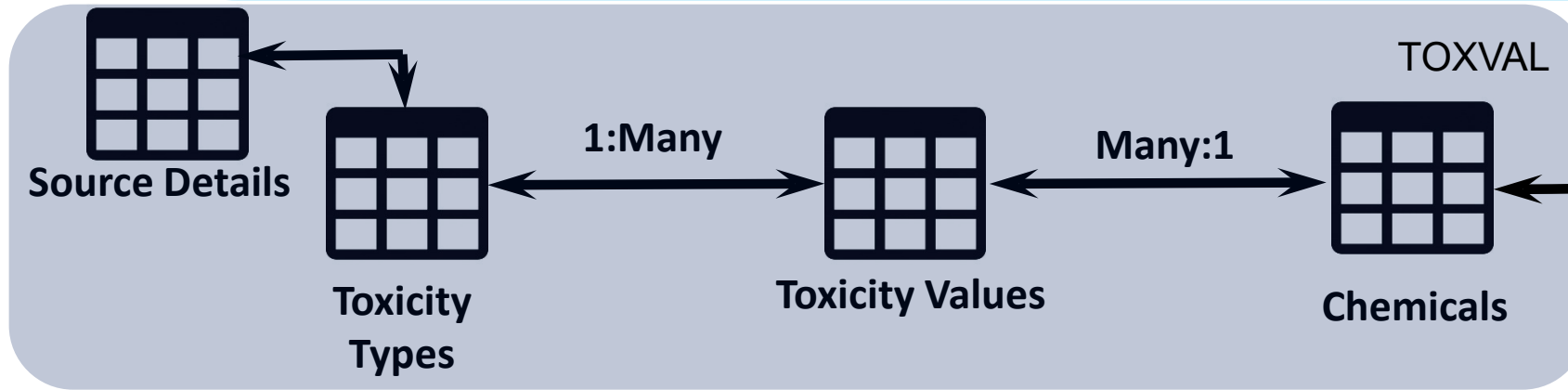
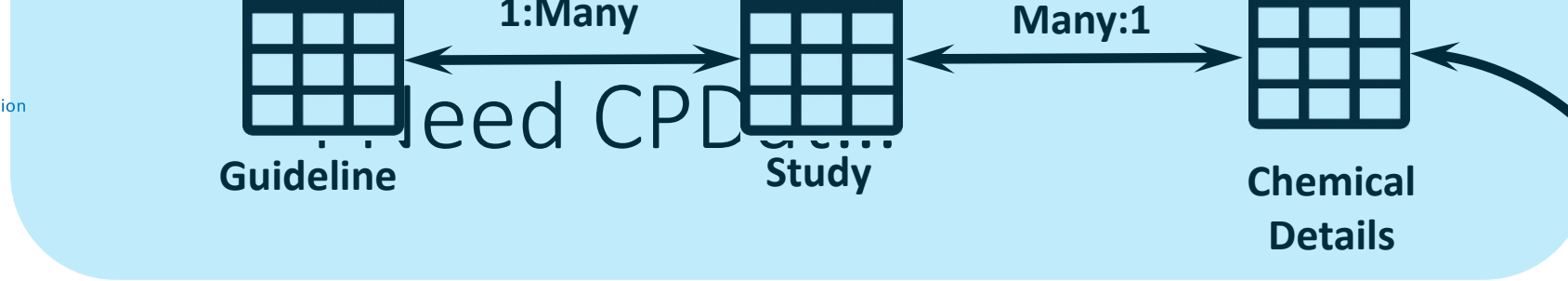
Structural Identifiers

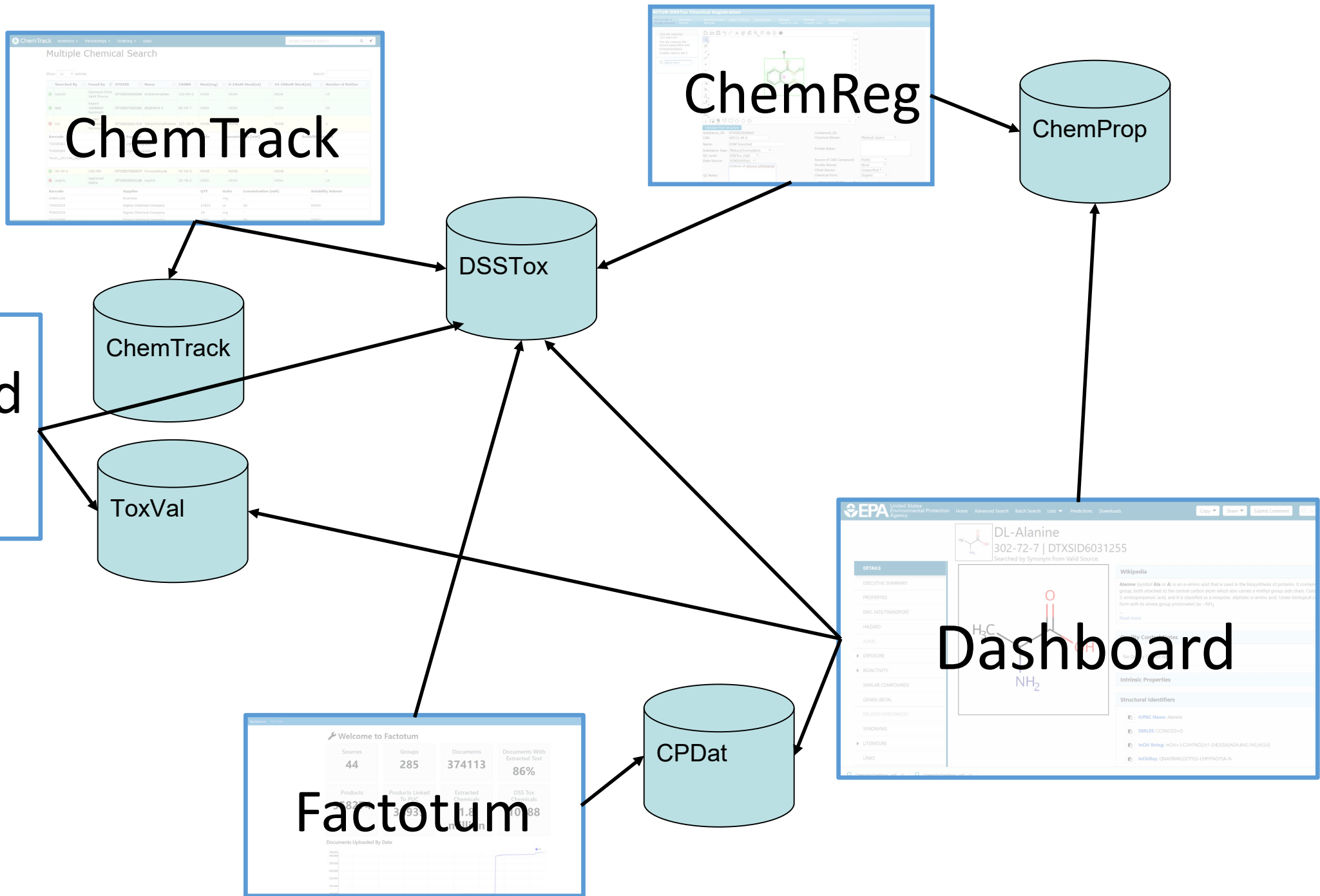
- IUPAC Name: Alanine
- SMILES: CC(N)C(=O)O
- InChI String: InChI=1/C3H7NO2/c1-2(4)3/5(6)/2H,4H(2,1H3,1,5,6)
- InChIKey: QNAYBMKLOCPYGI-UHFFFAOYSA-N

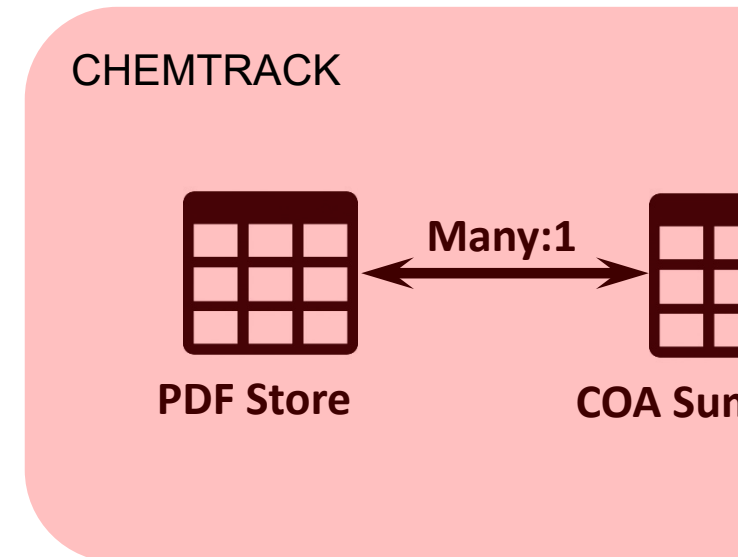
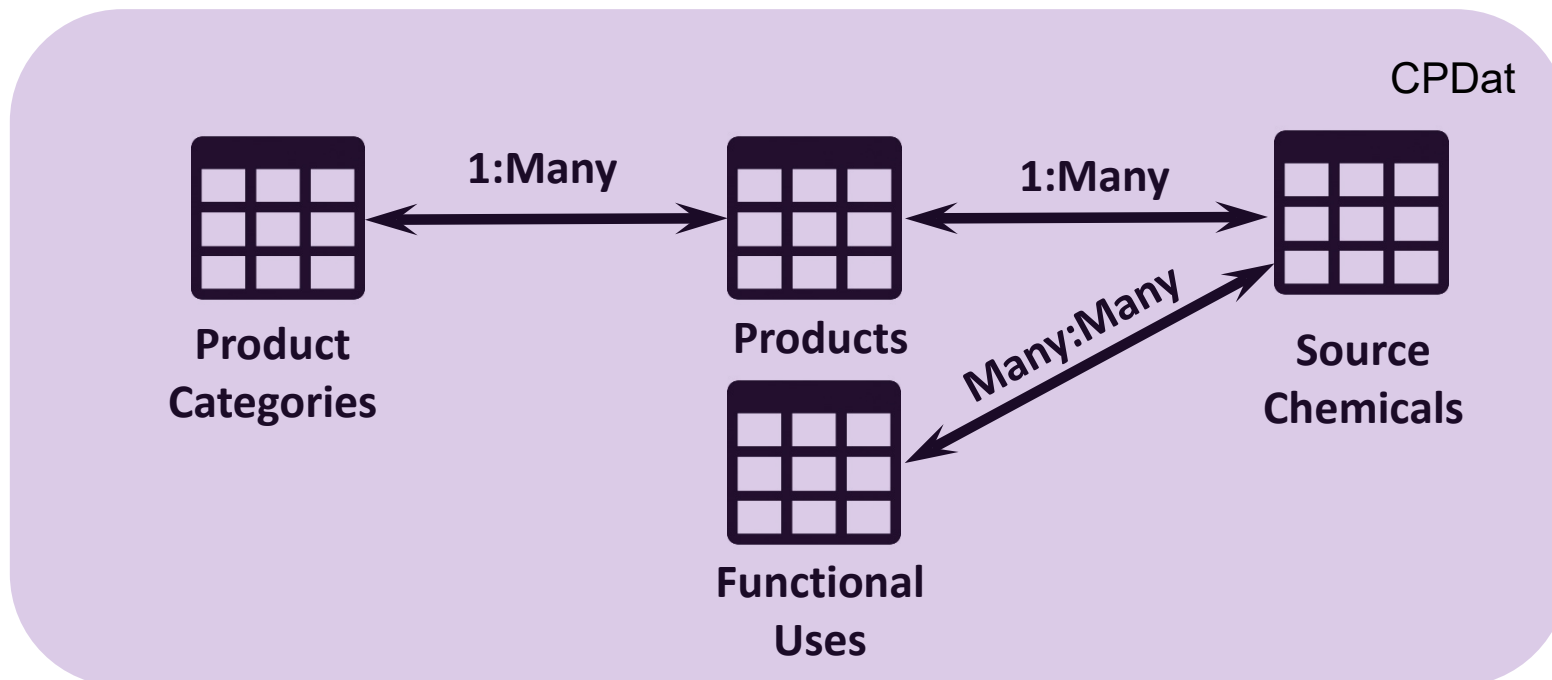
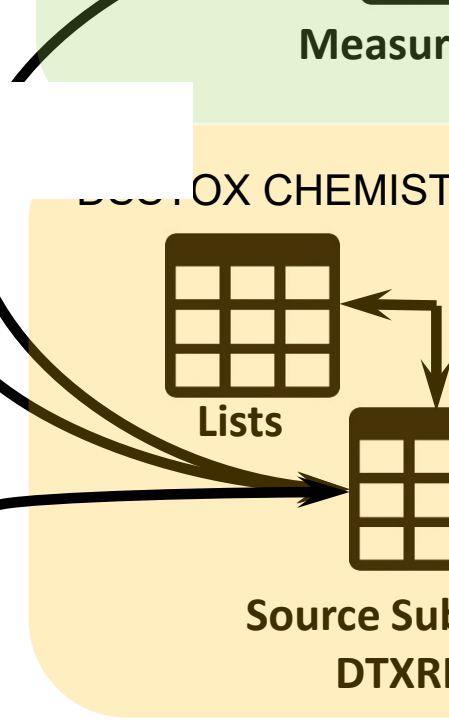
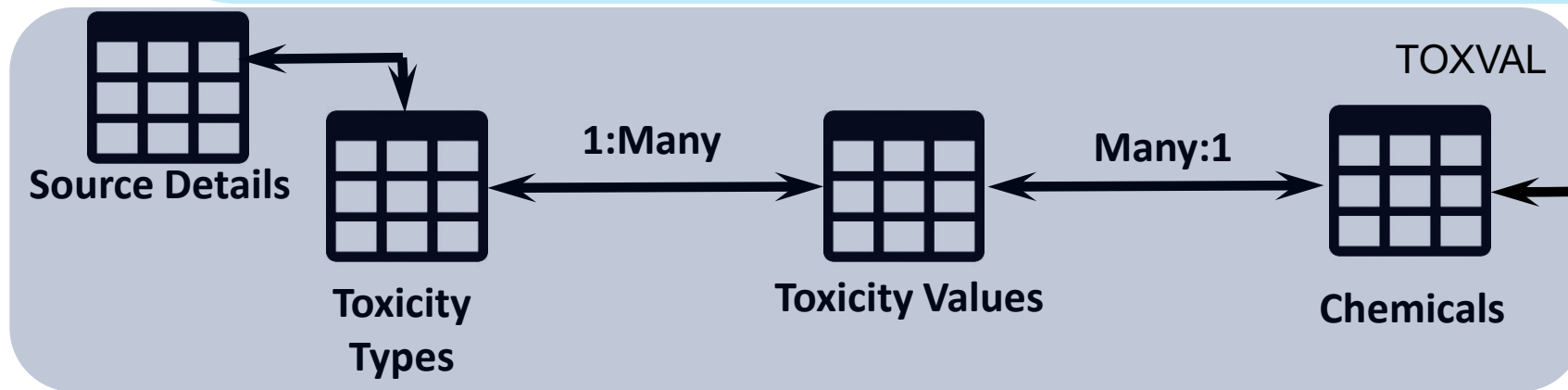
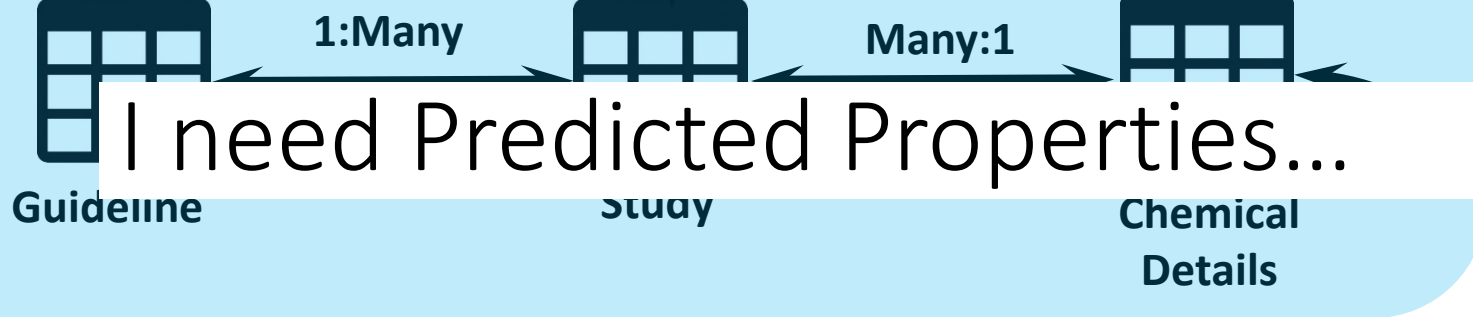
ToxVal Load Scripts

I Need CPDat...

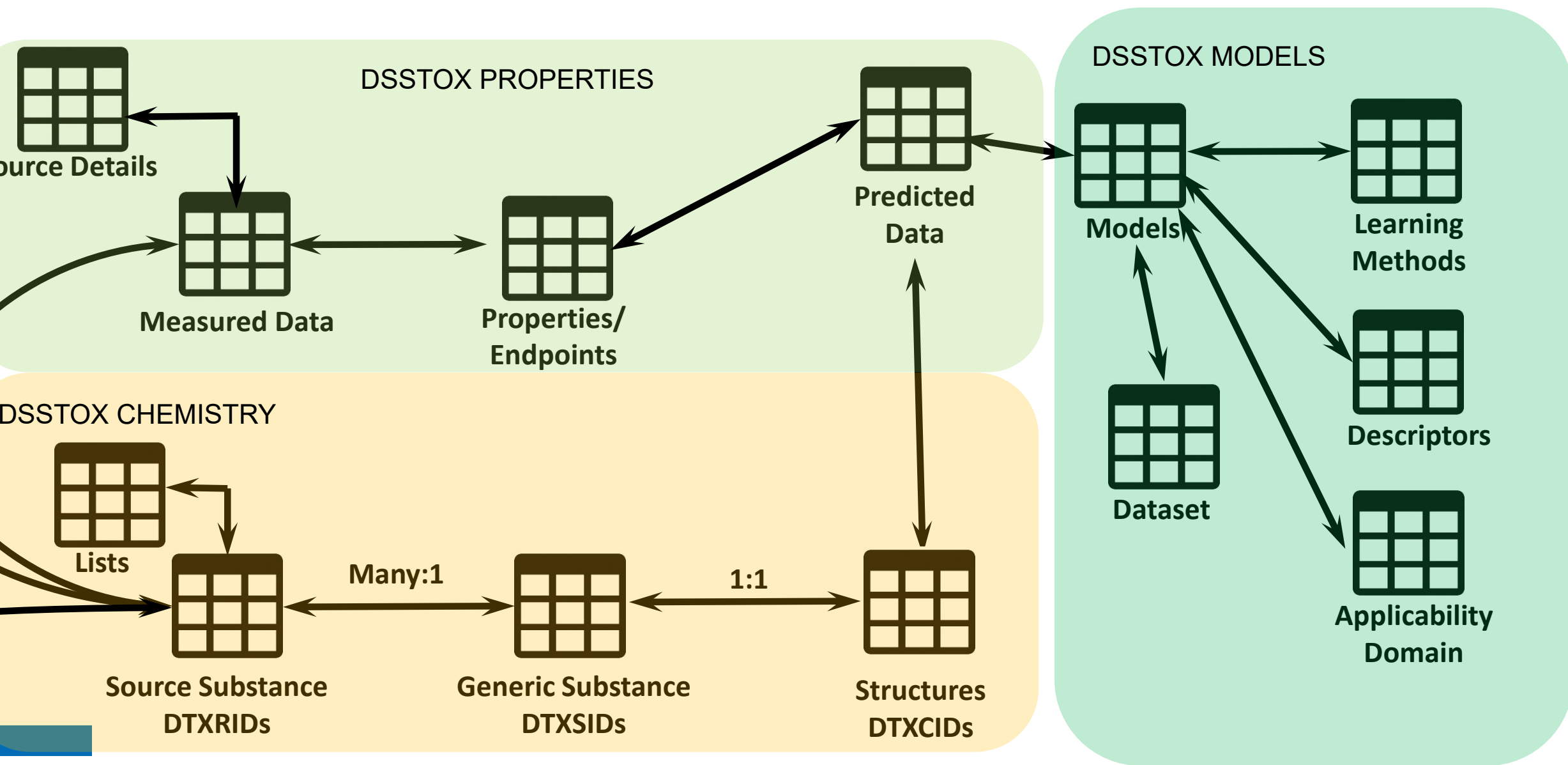






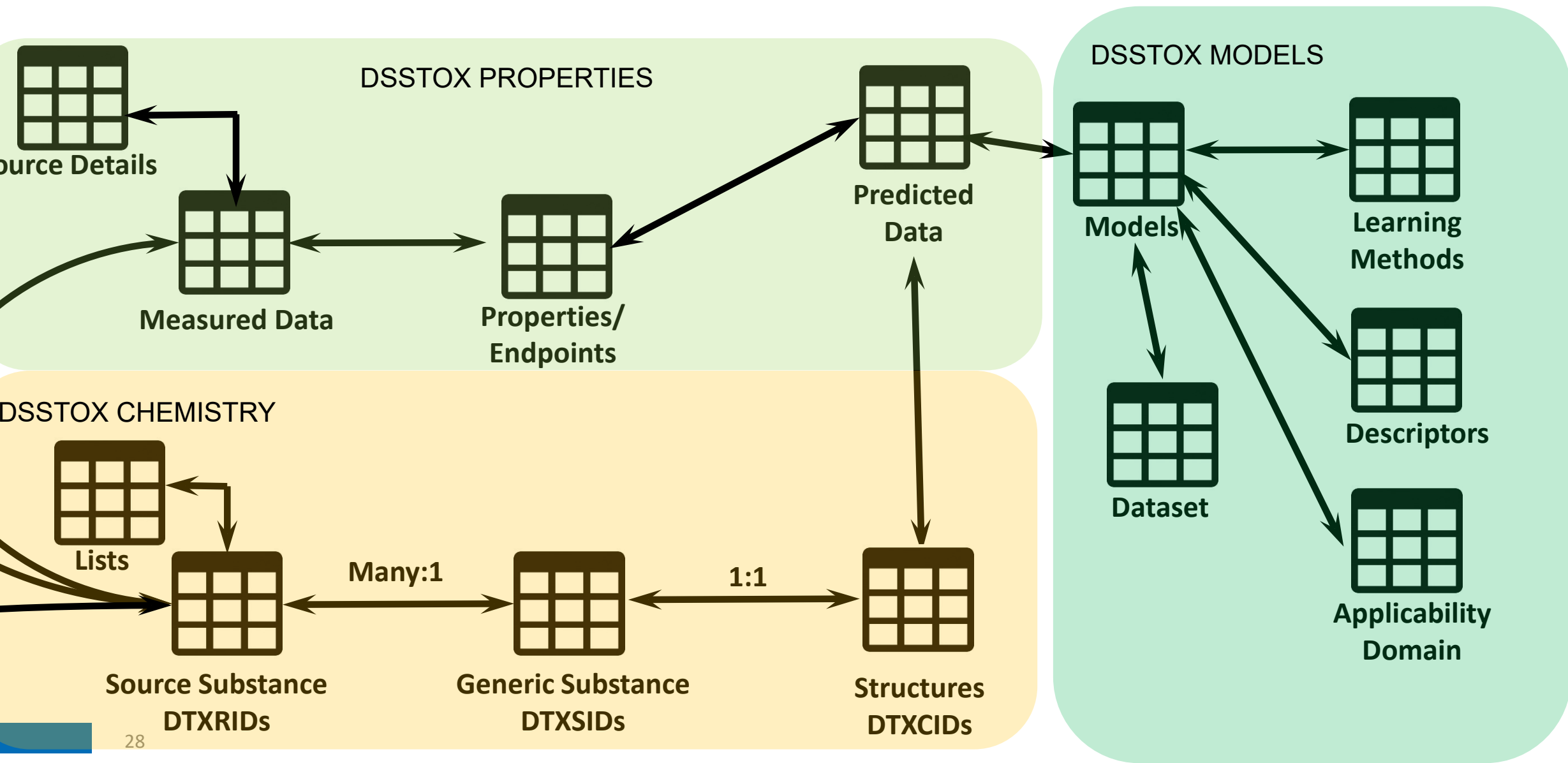


I need Predicted Properties...

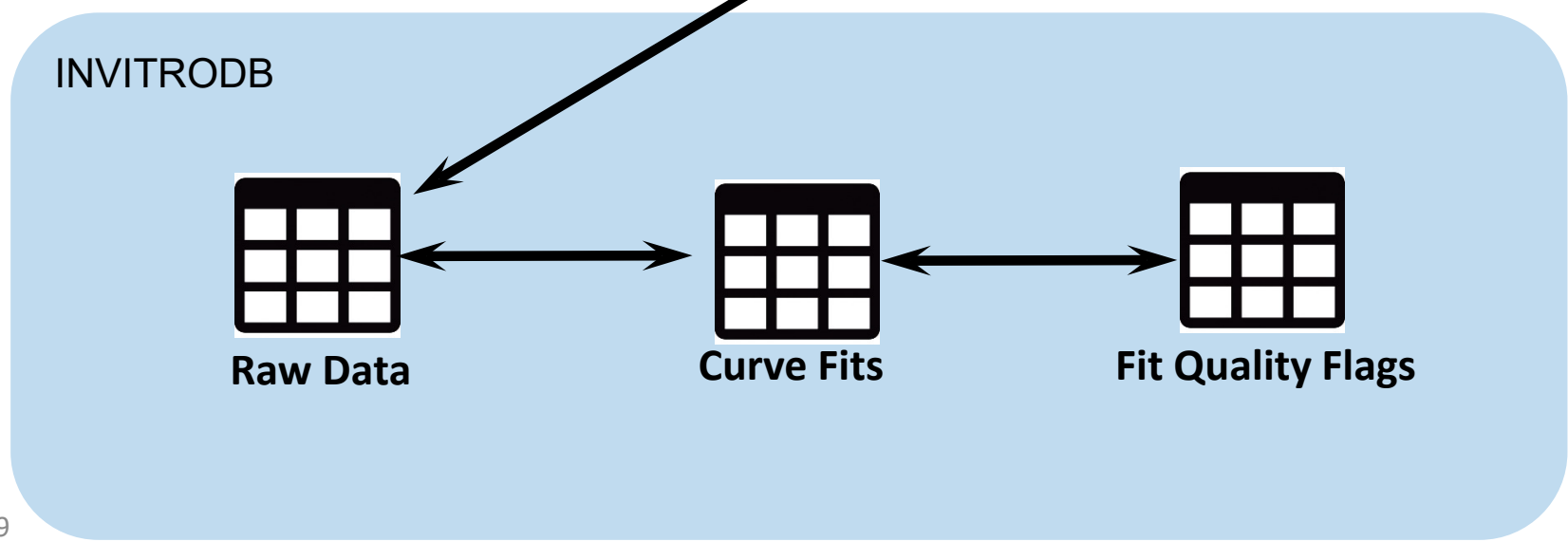
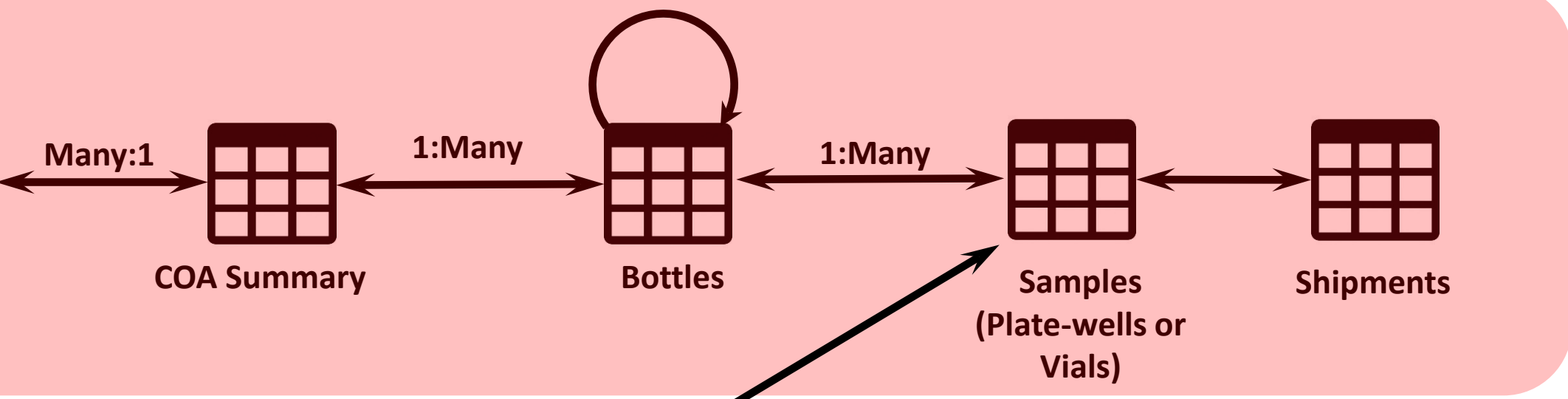


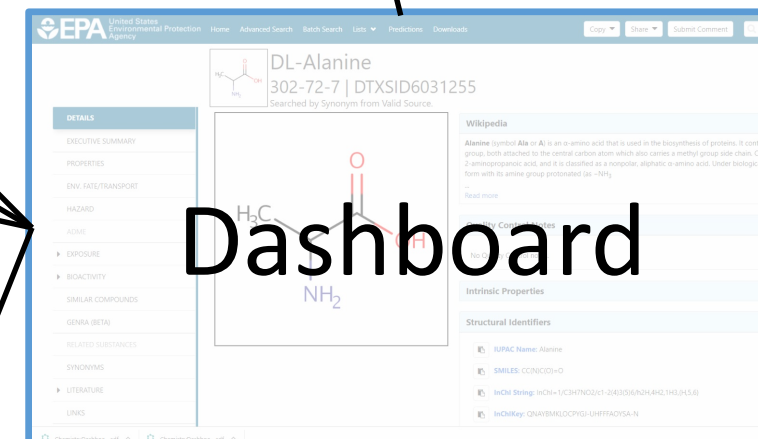
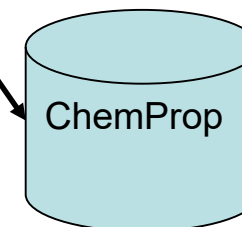
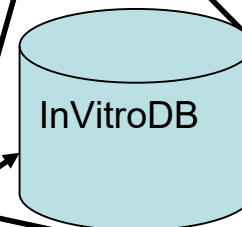
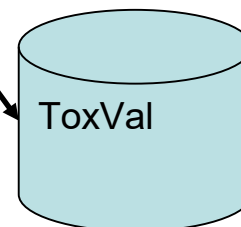
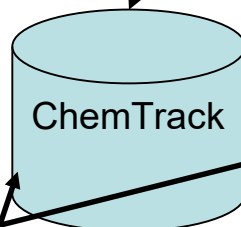
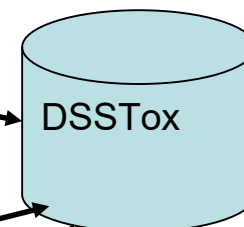
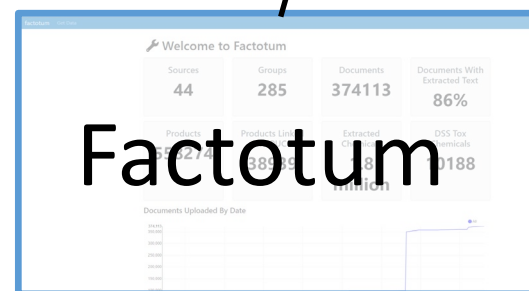
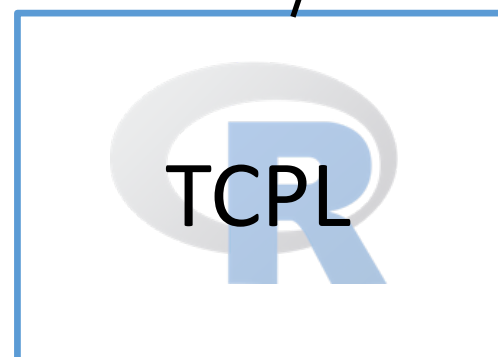
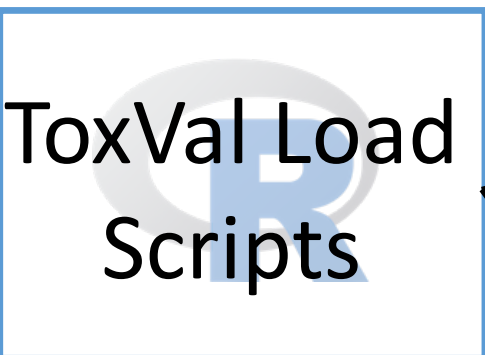


I need ToxCast Results...

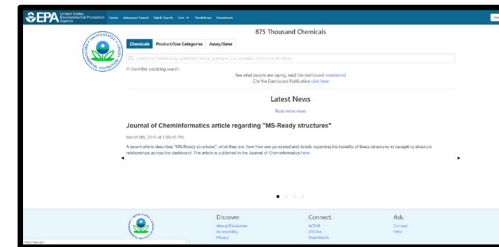


I need ToxCast Results...





Proposed EPA-ORD Data Structure



Production Application Layer

Virtual or Physical Application Databases
– Data Structured to efficiently support
our external production applications

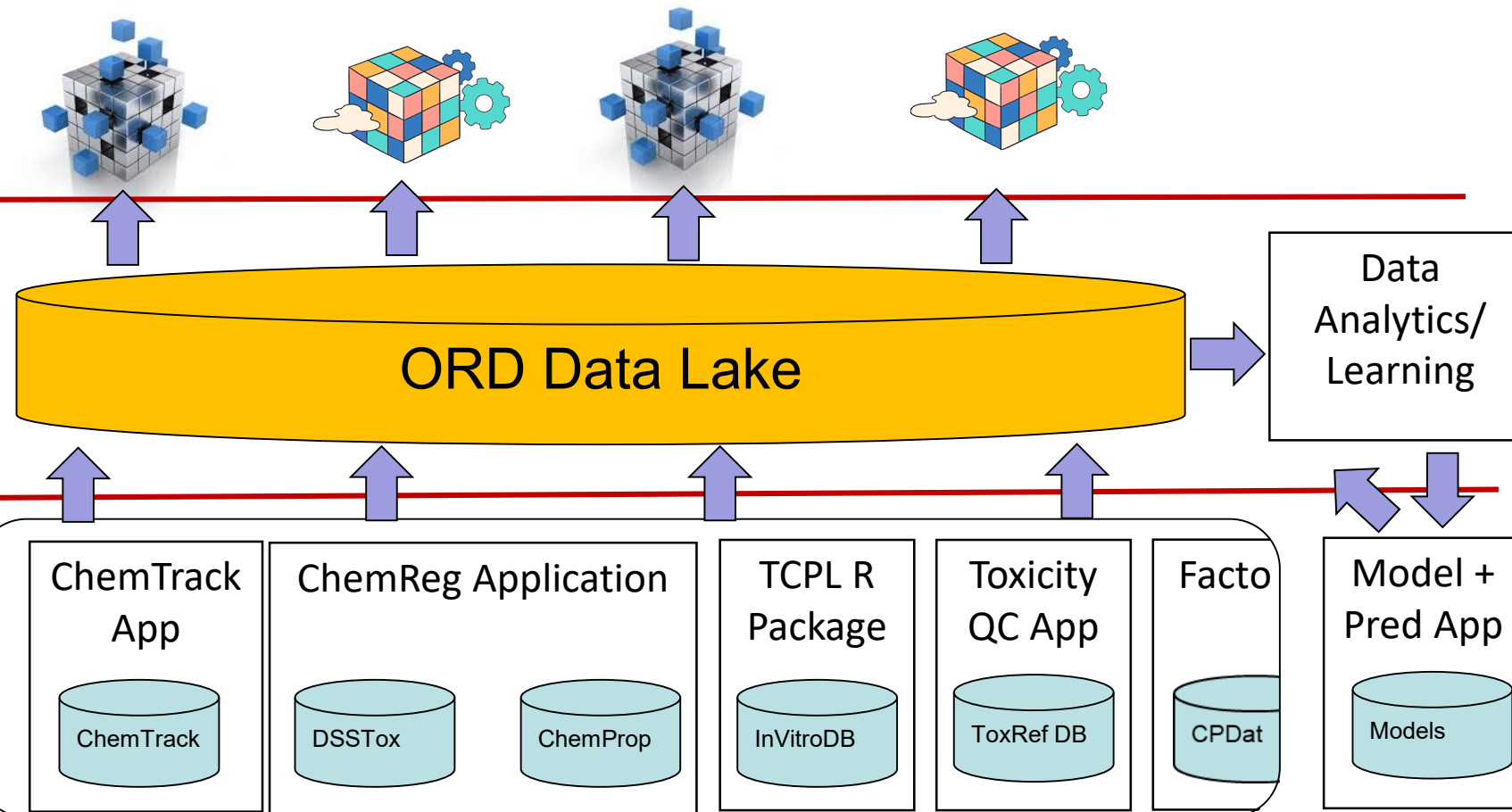
Presentation Layer (Fit-for-Purpose)

Generalized data model
providing integration
across the domains for
scientific exploration

Integration Layer

Transactional Layer

Raw data processing
and data curation,
quality tools



Conclusions

- The Comptox Chemicals Dashboard is great
- The desire to make our data available resulted in a web of tight coupling
- This tight coupling inhibits improvement and flexibility
- Separation of concerns in different data layers adds complexity and overhead but removes restrictions
- APIs provide greater flexibility for maintaining backward compatibility than direct database querying

Acknowledgements



Credit: the Research Triangle Foundation

Software Development

Jeff Edwards

Jeremy Dunne

Modeling Software

Kamel Mansouri

Todd Martin

DSSTox

Inthirany Thillainadarajah

Sakuntala Sivasupramaniam

Ann Richard

EPA's National Center for Computational Toxicology Research Triangle Park, NC



Questions?