ComptoxR: An R Package to Retrieve Compound Information from US EPA Products and APIs for Rapid Chemical Hazard, Risk, Toxicological Evaluation and Screening

Sean Thimons, Antony Williams

## Abstract

The US Environmental Protection Agency’s (EPA) CompTox Chemicals Dashboard (CCD) is a web-based application that has integrated diverse types of relevant domain data through a cheminformatics layer, built upon a database of curated substances linked to chemical structures. These data include physicochemical, environmental fate and transport, exposure, usage, in vivo toxicity, and in vitro bioassay data, surfaced through an integration hub with link-outs to additional EPA data and public domain online resources. Since April 2016, there have been over 10 releases which have increased the data available and functionality. The current release includes a rich API which facilitates new applications to be developed. This short communication describes ComptoxRan R package that can query the underlying databases that support CCD and related tools, extract out the data into R objects, and perform several analyses . ComptoxR includes wrappers for a number of capabilities from US EPA Cheminformatics Modules such as the Hazard Comparison and ToxPrint Chemotypes Exploring Chemical Feature Enrichments as well as the the GenRA web-based appplication.

The package is available through CRAN? and Github.com

.

## *Depending on journal – do you need a software availability section here?with licence information, system requirements and a one liner installation instruction e.g. install.packages or something equivalent*

## Introduction

### Implementation and Design Details

ComptoxR is written in R and relies upon several packages to function effectively. Several other packages were used for operation speed or because they offered time-saving implementations. Functions are grouped by .R file bearing the same name. An associated documentation file is generated when building the package that an end-user can readily access, along with several vignettes detailing suggested use-cases.

Most functions that work to retrieve data on a substance? are also implemented in a ‘batch’ function to reduce query times. Lists of compounds that number over 200 will automatically be batched to accommodate API throughput limits and seamlessly assemble the data into a single dataframe object. Typically, any function will use a RESTful GET request for a single compound, and a POST request for everything else.

When possible, the data returned is returned as-is from the API. We feel this an appropriate design choice as this reduces design debt that comes with software development and allows us to accommodate future additions or changes to our data sources. Some functions (e.g.: ct\_details() offer different subsets of data from the endpoint. Where this is an option, the default option is specified to the user, with optional arguments detailed in the appropriate help file.

### Data Sources

ComptoxR comes with several existing data sources packaged within for analysis and comparison purposes. Some are required, as they function as “dictionary” files for searching purposes, while others (i.e.: dtx\_list) are examples that can be used to demonstrate abilities. There are four main API endpoints that ComptoxR targets: three public-facing APIs from US EPA (CCD, GenRA, and CTS) and one for PubChem to retrieve Globally Harmonized System of Classification and Labeling of Chemicals (GHS) data through the webchem package.

## Installation and initial setup

The CCD API requires an API token for authentication. After requesting one, create an environmental variable as detailed below. A user can test if their key is present by running the ct\_api\_key() function. A restart of the R session may be needed. Upon loading the package, a user must set the server pathway using the comptox\_server() function. This is designed to accomodate upstream shifts in server URLS, or allow access to private servers when available. When running the function, the base URL of the request will be output to the console.

library(ComptoxR)  
  
Sys.setenv('ccte\_api\_key' = [TOKEN HERE])  
  
ct\_api\_key()  
  
comptox\_server()

## Use cases

### Querying names, supported identifiers, and details of compounds

Often, a user needs to find a particular compound, but may only have a portion of a name or is using an identifier such as a CAS-RN or InChIKey string. The ct\_name() function provides several ways to search for compounds for their preferred name and DTXSID. Three search parameters (starts-with, equals, substring) offer different ways of searching for compounds.

#String search  
ct\_name('atraz', param = 'start-with') %>% select(searchName, rank, preferredName) %>% head(n = 10)

Parameter(s) declared: start-with  
Requesting valid names by provided search parameters....

# A tibble: 10 x 3  
 searchName rank preferredName   
 <chr> <int> <chr>   
 1 Approved Name 9 Atrazine   
 2 Approved Name 9 Atrazine-2-Ethoxy   
 3 Approved Name 9 Atrazine-acetochlor mixt.   
 4 Approved Name 9 Atrazine-d5   
 5 Approved Name 9 Atrazine mercapturate   
 6 Approved Name 9 Atrazine-mesotrione-S-metolachlor mixt.  
 7 Approved Name 9 Atrazine mixture with butylate   
 8 Approved Name 9 Atrazine mixture with metolachlor (1:1)  
 9 Approved Name 9 Atrazine mixture with pendimethalin   
10 Approved Name 9 Atrazine mixture with terbutryn

#InChlKey search  
ct\_name('MXWJVTOOROXGIU', param = 'start-with') %>% select(searchValue, rank, preferredName) %>% head(n = 10)

Parameter(s) declared: start-with  
Requesting valid names by provided search parameters....

# A tibble: 6 x 3  
 searchValue rank preferredName   
 <chr> <int> <chr>   
1 MXWJVTOOROXGIU-DETAZLGJSA-N 13 6-Chloro-N~2~-ethyl-N~4~-(propan-2-yl)-1,3,~  
2 MXWJVTOOROXGIU-NTDPGYSYSA-N 13 Atrazine-ring-UL-14C   
3 MXWJVTOOROXGIU-SGEUAGPISA-N 13 Atrazine-d5   
4 MXWJVTOOROXGIU-TTXLGWKISA-N 13 6-chloro-N2-ethyl-N4-(propan-2-yl)(13C3)-1,~  
5 MXWJVTOOROXGIU-UHFFFAOYSA-N 13 Atrazine   
6 MXWJVTOOROXGIU-ZSJDYOACSA-N 13 6-Chloro-N~2~-ethyl-N~4~-(propan-2-yl)-1,3,~

#CAS\_RN search  
ct\_name('1912-24', param = 'contain') %>% select(searchValue, preferredName, casrn)

Parameter(s) declared: contain  
Requesting valid names by provided search parameters....

# A tibble: 5 x 3  
 searchValue preferredName casrn  
 <chr> <chr> <chr>  
1 121912-24-1 5-(Perfluoroheptyl)pentanamine hydrochloride (1:1) 1219~  
2 21912-24-3 5-(2-Chloroethanesulfonyl)bicyclo[2.2.1]heptane-2-carbonyl~ 2191~  
3 321912-24-7 ß-Alanine, N-[5-[[4-chloro-6-[[8-hydroxy-3,6-disulfo-7-[2-~ 3219~  
4 1261912-24-6 6-(5-Formylthiophen-2-yl)pyridine-3-carboxylic acid 1261~  
5 1912-24-9 Atrazine 1912~

Additionally, a function to return all verified synonyms of a given compound is available. A user may find this useful, as old CAS-RNs, product names, or variations on proper names can be returned.

ct\_synonym(query = 'DTXSID7020182') %>% head(n = 10)

Searching for compound synonym....

dtxsid value search rank  
1 DTXSID7020182 4,4’-Propane-2,2-diyldiphenol valid 1  
2 DTXSID7020182 Bisfenol A valid 1  
3 DTXSID7020182 BPA valid 1  
4 DTXSID7020182 EC No.: 201-245-8 valid 1  
5 DTXSID7020182 80-05-7 valid 1  
6 DTXSID7020182 4,4'-(Propane-2,2-diyl)diphenol valid 1  
7 DTXSID7020182 Phenol, 4,4'-(1-methylethylidene)bis- valid 1  
8 DTXSID7020182 Bisphenol A valid 1  
9 DTXSID7020182 DTXSID7020182 good 2  
10 DTXSID7020182 Bisphenol-A del\_casrn 3

After a compound’s DTXSID has been located, a user may want to inquire on various details for the compound. The ct\_details() function provides an executive summary of a compound, with over sixty variables being available. Here, a user can easily ascertain important details about the compound in question, ranging from the quality of curation about a compound, the number of active assays in the ToxCast assay program, to knowing the monoisotopic mass, molecular formula or SMILES string.

# A selection of the data available for the query  
  
df %>% select(dtxsid,preferredName, smiles, molFormula, monoisotopicMass,toxcastSelect,octanolWaterPartition,qcLevelDesc) %>% glimpse()

Rows: 1  
Columns: 8  
$ dtxsid <chr> "DTXSID8031865"  
$ preferredName <chr> "Perfluorooctanoic acid"  
$ smiles <chr> "OC(=O)C(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)(F)C(F)(F~  
$ molFormula <chr> "C8HF15O2"  
$ monoisotopicMass <dbl> 413.9737  
$ toxcastSelect <chr> "133/1396"  
$ octanolWaterPartition <dbl> 3.10642  
$ qcLevelDesc <chr> "Level 1: Expert curated, highest confidence in ~

### Querying for physio-chemical properties and environmental fate/ transport data

A common, initial step for hazard identification is screening by physio-chemical properties or potential longevity within a given environment. The function ct\_prop() returns both experimental and predictive physical-chemical properties, with accompanying source documentation. Users can then filter or subset the data as appropriate.

| name | unit | experimental | predicted |
| --- | --- | --- | --- |
| Boiling Point | °C | 190.200 | 193.251 |
| Density | g/cm^3 | 1.800 | 1.724 |
| Flash Point | °C |  | 67.998 |
| Henry's Law | atm-m3/mole |  | 0.000 |
| Index of Refraction |  |  | 1.289 |
| LogKoa: Octanol-Air |  |  | 4.156 |
| LogKow: Octanol-Water |  | 3.104 | 5.683 |
| Melting Point | °C | 55.571 | 24.268 |
| Molar Refractivity | cm^3 |  | 42.893 |
| Molar Volume | cm^3 |  | 237.268 |
| Polarizability | Å^3 |  | 17.004 |
| Surface Tension | dyn/cm |  | 16.819 |
| Thermal Conductivity | mW/(m\*K) |  | 65.258 |
| Vapor Pressure | mmHg | 0.988 | 0.243 |
| Water Solubility | mol/L | 0.014 | 0.753 |
| pKa Acidic Apparent |  | 3.150 |  |

For environmental fate/ transport data, the ct\_env\_fate() function also returns experimental and predictive data where available.

| endpointName | unit | predicted | experimental |
| --- | --- | --- | --- |
| Atmos. Hydroxylation Rate | cm3/molecule\*sec | 0.000 |  |
| Bioaccumulation Factor |  | 7,674.000 |  |
| Bioconcentration Factor |  | 993.599 |  |
| Biodeg. Half-Life | days | 4.943 |  |
| Fish Biotrans. Half-Life (Km) | days | 1.345 |  |
| Soil Adsorp. Coeff. (Koc) | L/kg | 1,161.658 |  |

### Hazard Comparison

Following the logic primarily outlined by @vegosen\_2020, the hc\_table() function is provided to compare a set of compounds across a multitude of endpoints. Each endpoint is binned into a range of categorical tiers (‘VH’, ‘H’, ‘M’, ‘L’, ‘NA’). These tiers are based off of numerical data (e.g. LD50 values) and categorical data (e.g.: from GHS data). Within the hc\_table() function, a number of other functions are called to retrieve data. Data is then aggregated and converted to a categorical value based on the criteria as detailed by @vegosen\_2020.

In this implementation, we use a variation on the ‘trumping’ methodology. Briefly, we search for numerical, authoritative data sources as a primary search parameter. This entails experimental, historical data that has been curated by CCD. For compounds that are data-poor, we also search for GHS data for the relevant ‘H codes’ and across curated, aggregated lists (ECHA SVCH, IARC, OEHHA Prop 65) for hazard and risk classifications. Lists are pulled from CCD or directly from the parent agency’s site, and stored as a RDS object. Since these do not (typically) contain numerical threshold data, we convert the ranged values of the from binning criteria into a median value and assign it the compound with the corresponding categorical classification. Finally, as a last resort we query against QSAR-based models (e.g.: TEST) and assign a categorical value from the prediction. We feel this is an appropriate workflow, as historical data and GHS rankings are based off of experimental data, while QSAR-based models are statistical extrapolations and subject to applicability domain restrictions.

Not included in the original implementation or the current hosted web-application, we include several endpoints that risk-assessors may find useful under the larger family of ‘Safety’. Endpoints of ‘Ignitability’, ‘RxnWater’ (reaction with water), and ‘SelfRxn’ (self-reacting) were added and derived based off of physio-chemical properties (flash point, boiling point) and GHS ‘H codes’ classifications. We also opted to include a ‘Mobility’ endpoint under the ‘Environmental Fate’ family that incorporates ECHA’s definitions of ‘persistence, mobility and toxicity (PMT/vPvM)’ compounds based off of the organic carbon-water coefficient, as this is a well-described property in the literature and able to be predicted by QSAR model.

The function hc\_table() function has a number of useful arguments that may be of interest. While the query argument is required for live searching, the archive or save arguments are not. If on an initial query the ‘save’ argument is set to TRUE, the underlying data that is generated is saved in the local directory as an .Rds file. In parallel, the arguments archive\_string and save\_string allow the user to supply a custom string for a file name.

| Human Health Effects | | | | Ecotoxicity | | | | | Safety | | | | | | Env. Fate | | | | |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Preferred  Name | Oral | Dermal | Inhalation | | Carcinogenicity | Mutagenicity | Endocrine\_Disruption | Reproductive | | Developmental | Acute\_Aquatic\_Toxicity | Chronic\_Aquatic\_Toxicity | Ignitability | RxnWater | | SelfRxn | Persistence | Bioaccumulation | Mobility | |
| Perfluorooctanoic acid | H |  | H | | H | M |  | H | | H | M |  |  |  | |  | L | VH | M | |
| Benzene | L |  | L | | VH | H | L | H | | H | H | M | M |  | |  | L | VH | VH | |
| Dazomet | M | L | L | |  | M | L | H | | H | VH | VH |  |  | |  | L | L | VH | |
| (Z)-1,2-Dichloroethylene | M |  | M | |  | H |  |  | |  | M | M | M |  | |  | L | L | VH | |
| Tamoxifen | M |  |  | | VH |  |  | H | | H | VH | VH |  |  | |  | L | H | L | |
| Bisphenol A | M |  |  | |  | M | H | H | | H | H | M |  |  | |  | L |  | M | |

Data completeness functions are also able to be ran to better assess the relative quality of both the compounds and endpoints being assessed.

The ‘compound coverage’ score is an assessment of how many endpoints for each compound were able to be queried for, displayed as a fraction of 1.0 (analogous to 100%). Here, all compounds that were included in the initial query has at-least 50% of the endpoints represented (12 out of 16).

hc\_cc <- hc\_compound\_coverage(table = df, ID = 'preferredName', suffix = '\_amount')  
  
hc\_cc

# A tibble: 6 x 2  
 preferredName data\_coverage  
 <chr> <dbl>  
1 Perfluorooctanoic acid 0.647  
2 Benzene 0.824  
3 Dazomet 0.765  
4 (Z)-1,2-Dichloroethylene 0.588  
5 Tamoxifen 0.588  
6 Bisphenol A 0.588

Conversely, the ‘endpoint coverage’ score returns the percentage of compounds being available per each endpoint. This table is critical to the ToxPi calculation, as it is responsible for the weighting parameter. By default, the weighting per each endpoint is defaulted to a ‘1.0’ (i.e.: equal). We recommend that a user run this function with an explicitly declared filter parameter for the score set to ‘0’ to see all relative responses before selecting an appropriate cutoff value. Failing to explicitly specify a filter cutoff will default the value to ‘0.5’.

# A tibble: 16 x 3  
 endpoint score weight  
 <chr> <dbl> <dbl>  
 1 oral\_amount 1 1  
 2 dermal\_amount 0.167 1  
 3 inhalation\_amount 0.667 1  
 4 cancer\_amount 0.5 1  
 5 geno\_amount 0.833 1  
 6 endo\_amount 0.5 1  
 7 reproductive\_amount 0.833 1  
 8 developmental\_amount 0.833 1  
 9 acute\_aq\_amount 1 1  
10 chronic\_aq\_amount 0.833 1  
11 ignite\_amount 0.333 1  
12 rxnwater\_amount 0 1  
13 selfrxn\_amount 0 1  
14 persistance\_amount 1 1  
15 bac\_amount 1 1  
16 mobile\_amount 1 1

Defaulting to 0.5% score for cutoff!

# A tibble: 12 x 3  
 endpoint score weight  
 <chr> <dbl> <dbl>  
 1 oral\_amount 1 1  
 2 inhalation\_amount 0.667 1  
 3 cancer\_amount 0.5 1  
 4 geno\_amount 0.833 1  
 5 endo\_amount 0.5 1  
 6 reproductive\_amount 0.833 1  
 7 developmental\_amount 0.833 1  
 8 acute\_aq\_amount 1 1  
 9 chronic\_aq\_amount 0.833 1  
10 persistance\_amount 1 1  
11 bac\_amount 1 1  
12 mobile\_amount 1 1

After assessing the quality of data, a user may manually adjust the weight response per each endpoint via manual subsetting.

The helper ct\_archive() function can be ran to load the data easily into local working memory for closer examination. If the such a file exists AND the ‘archive’ flag is set to true, the function uses the data to recreate the table, saving query request time. The underlying data structure unpacks to the global environment should a user want to examine the data further.

#### ToxPi Ranking

As detailed by @reif\_2010, we have implemented the Toxicological Priority Index (ToxPi) scoring algorithm within this package. ToxPi relied upon weight-of-evidence to subjectively rank compounds across various domains yielding a unitless score.

For each endpoint from the hazard comparison table, the responses are scaled through a min-max normalization to [0-1] (Eq 1). The score for each endpoint is then summed per compound, yielding a unitless score, where the higher score indicates higher, relative hazard and risk (Eq 2).

Eq. 1:

Eq. 2:

It should be noted that low performing (or compounds with a score of zero) compounds do **not** indicate a *lack* of associate hazard or risk, but rather a lack of information regarding these compounds.

Eq. 1 is detailed in the function tp\_single\_score(), where a single vector (e.g.: a single column from a data frame) is evaluated. An additional parameter (disabled by default) is also included, which ‘backfills’ any compound that has a zero for that endpoint to the specified number. This could be the lowest, non-zero response, the average, or the max of each endpoint. We feel this is an important option that should be used with care, as it could adversely affect the conclusions a user may draw.

Eq. 2 is detailed in the function tp\_combined\_score(). Once a hazard comparison table has been generated, a user can request a relative prioritization based on the bias table. Should no bias table be specified, the function will proactively call the hc\_endpoint\_coverage() function with filter cutoff parameter specified to 0.1. The ellipse argument allows for the back\_fill argument to be specified from the tp\_single\_score() function.

As an example, a subset from an included data set from the package (the object dtx\_list) was used. An analysis of this suggests that along the endpoints we have selected for (n = 12), Dazomet shows the highest relative risk, with (Z)-1,2-Dichloroethylene showing the lowest relative risk, when comparing along similar exposure amounts. This scoring is partially due to the amount of data being returned by each compound, with Dazomet having ~76% of the endpoints documented, and (Z)-1,2-Dichloroethylene having ~56% (see **@tbl-hc-compound-coverage**)

#Backfill with 50% of the lowest non-zero response  
df\_tp <- tp\_combined\_score(table = df, ID = 'preferredName', bias = bias, back\_fill = 'half\_min')  
  
df\_tp %>% select(preferredName, score) %>% arrange(desc(score)) %>% flextable()

| preferredName | score |
| --- | --- |
| Dazomet | 6.9455 |
| Benzene | 6.6619 |
| Perfluorooctanoic acid | 6.1537 |
| Bisphenol A | 5.5471 |
| Tamoxifen | 5.4079 |
| (Z)-1,2-Dichloroethylene | 2.8528 |

If we were to re-run the analysis to *only* examine endpoints that all shared inclusive coverage of all of the compounds, the results are detailed below. Of interest, PFOA and BPA are raised in ranking while Benzene is dropped in ranking within the list, indicating some overlapping parameter that may be of interest in the right context. While we recognize that this subset of compounds is not a realistic comparison, it still outlines the useful and subjectiveness of using the ToxPi approach for prioritization.

bias2 <- hc\_endpoint\_coverage(table = df, ID = 'preferredName', suffix = '\_amount')

Defaulting to 0.5% score for cutoff!

bias2$weight[bias2$score != 1] <- 0  
  
df\_tp2 <- tp\_combined\_score(table = df, ID = 'preferredName', bias = bias2, back\_fill = 'half\_min')  
df\_tp2 %>% select(preferredName, score) %>% arrange(desc(score)) %>% flextable()

| preferredName | score |
| --- | --- |
| Dazomet | 2.5514 |
| Perfluorooctanoic acid | 2.1537 |
| Bisphenol A | 1.9119 |
| Benzene | 1.8972 |
| Tamoxifen | 1.4079 |
| (Z)-1,2-Dichloroethylene | 1.2463 |

### Generalized Read-Across (GenRA) and ToxPrint Enrichment

#### GenRA

Read-Across is an approach thoroughly documented in the literature (@patlewicz\_2023) for extrapolating from a data-rich ‘community’ to data-poor ‘individual’ for predictive purposes. We offer the functionality available to the public found at <https://comptox.epa.gov/genra/> through this package for both single compound and batch-search functionality. We demonstrate here a typical workflow for using GenRA functions that would be analogous to the website experience.

##### Basic functionality

Like nearly all of the functions in this package, the starting point is a DTXSID for referencing a compound. Here, the DTXSID of Trimethylene glycol diacrylate is used. The other arguments are the fingerprint to find analogues by (Morgan [default], Torsion, ToxPrints, ToxCast, or ToxRef), filters to select the analogues by (compounds with or without data) and the number of analogs to return. In the example, we’ve used the ToxPrint, and requested 10 analogs. The first row is our searched compound, with the most similar compounds following, as described by the Jaccard similarity metric.

nn <- genra\_nn(query = 'DTXSID2026107', fp = 'chm\_mrgn', sel\_by = 'tox\_txrf', n = 10) %>% select(jaccard, name, dsstox\_sid, casrn, mol\_weight)  
  
nn

# A tibble: 11 x 5  
 jaccard name dsstox\_sid casrn mol\_weight  
 <dbl> <chr> <chr> <chr> <dbl>  
 1 1 Tetraethylene glycol diacrylate DTXSID20261~ 1783~ 302.   
 2 0.483 Ethyl acrylate DTXSID40205~ 140-~ 100.   
 3 0.306 Ethylene glycol monoethyl ether acetate DTXSID90219~ 111-~ 132.   
 4 0.267 Triethylene glycol dimethyl ether DTXSID80262~ 112-~ 178.   
 5 0.267 Triethylene glycol DTXSID40213~ 112-~ 150.   
 6 0.25 2-(2-Ethoxyethoxy)ethanol DTXSID20219~ 111-~ 134.   
 7 0.25 Acrylamide DTXSID50200~ 79-0~ 71.1  
 8 0.233 Diethylene glycol dimethyl ether DTXSID10246~ 111-~ 134.   
 9 0.231 2-(2-Butoxyethoxy)ethanol DTXSID80215~ 112-~ 162.   
10 0.229 N,N'-Methylenebisacrylamide DTXSID80255~ 110-~ 154.   
11 0.222 Diethylene glycol monomethyl ether DTXSID30250~ 111-~ 120.

We then request the read-across using the next function, with similar arguments. A subset of the output is displayed below. GenRA returns data that belongs to either the ToxCast database or the ToxRef database, where it is grouped by study outcomes. The groups are:

* chronic toxicity (CHR)
* developmental toxicity (DEV)
* developmental neurotoxicity (DNT)
* multigenerational toxicity (MGR)
* neurotoxicity (NEU)
* reproductive toxicity (REP)
* acute toxicity (ACU)
* subacute toxicity (SAC)
* subchronic toxicity (SUB)
* other (OTH) toxicity testing studies

The column ‘observation’ details the queried compound’s experimental results for that study design (if available), while the ‘pred resp’ is a predicted response (operating in a binary outcome).

Sending request for: DTXSID2026107  
Detected binary response choice!  
  
Request complete!

| name | effect\_desc | observation | pred resp |
| --- | --- | --- | --- |
| CHR:adrenal gland | Chronic adrenal gland | no\_data | Neg |
| CHR:body weight | Chronic body weight | no\_data | Pos |
| CHR:bone | Chronic bone: Example effects include fibro-osseous lesions, hyperplasia, hyperostosis, hypertrophy, myelofibrosis, inflammation, osteodystrophy, osteofibrosis, osteosarcoma, and osteosclerosis. | no\_data | Neg |
| CHR:brain | Chronic brain | no\_data | Neg |

Should a user desire it, a generic read-across function read\_across() is provided to generate a prediction based on other data (e.g.: LD50 values) leveraging the GenRA’s ability to fingerprint and find similar compounds, but currently relies on a user to provide the data.

#### ToxPrint Enrichment

Accurate depiction of chemicals is an important when operating in the cheminformatic space. As such, through both the CCD and GenRA, the ability to generate molecular fingerprints are provided. These fingerprints provide ways of describing complex 3D molecular features into a machine-readable format. One such method is the open-source XML-based ‘Chemical Subgraphs and Reactions Markup Language (CSRML)’ . As detailed by @yang\_2015, these predefined molecular descriptors describe not only atom, bonds, electron systems, and molecules, but can also address issues such as topology or complex connectivity. Part of this novel method of describing compounds allows it to be extended to the extrapolation of *in-vitro* bioassays (@wang\_2019a). Here, we’ve reproduced the ToxPrint Enrichment workflow as found on the Cheminformatics web application (<https://www.epa.gov/chemical-research/cheminformatics>). A user can upload compounds, where the Toxprint fingerprints are retrieved by the GenRA service.

### Non-Targeted Analysis Compound Searching

To better assist in the hazard and risk management space, several functions are suggested for use. For searching by mass range or by a generic formula, the functions ct\_search\_mass() and ct\_search\_formula() are able to be leveraged. Users are able to then subset the resulting data to their liking.

### Case study implementation of package

To demonstrate the potential workflow of this package, we offer two hypothetical situations. In the first, risk assessors would be interested in the prioritization of sites that contain storage tanks of compounds. The hypothetical situation would revolve around selecting for highest threats to human and ecosystem, while taking into account site-specific volumes. For this, the risk assessors will incorporate various endpoints, including physio-chemical and environmental fate/ transport parameters to assess the potential if a tank were to be ruptured. The second scenario would examine an industrial waste stream being discharged into a river system. Risk assessors would examine two paradigms, one focused on acute threats to both human and ecosystem, the other on a long-term site surveillance for compounds known to be persistent, bioaccumulative, and have potential for chronic exposures.

#### Scenario 1: Storage tank prioritization

In scenario 1, a regulatory permit writer is looking to prioritize sites for enhanced risk management planning. A data set of the current listed compounds is assembled via the CCD’s curated list. For compound production volumes, we’ve taken the average 2019 value from the ‘Exposure - Production Volume’ section from CCD for each compound (where available) as user-supplied data metric. This data is generated from the Chemical Data Reporting (CDR) Rule, issued under the Toxic Substances Control Act (TSCA), and represents the total volumes within the United States.

Next, the permit writer can utilize the hc\_table() function and pass the list into it. The production volumes for each compound is joined against the table, and a bias table for weighing is created to work with the prioritization function. A 2x weight is applied against persistence, BAC, mobility factor, volume endpoints while other endpoints are rejected due to poor coverage. Finally, the ToxPi priortitization scheme is employed, with the top 10 compounds being returned. Not all compounds that were in this top 10 ranking had reported production volumes, suggesting that other endpoints should be considered and that total volumes within a given site may not be as important.

| compound | preferredName | volume\_amount | score |
| --- | --- | --- | --- |
| DTXSID5020023 | Acrolein | 375,000,000 | 7.5571 |
| DTXSID7021106 | Pentachlorophenol | 14,843,844 | 7.4988 |
| DTXSID6023787 | Nickel sulfate | 30,000,000 | 6.8450 |
| DTXSID2040317 | Aluminum sesquisulfate | 3,000,000,000 | 6.6511 |
| DTXSID2035069 | Lead nitrate | 300,000 | 6.4591 |
| DTXSID70872953 | Ammonium dichromate | 1,000,000 | 6.4577 |
| DTXSID8020044 | Allyl alcohol | 550,000,000 | 6.3423 |
| DTXSID6034479 | Copper sulfate | 3,000,000,000 | 5.9316 |
| DTXSID5020029 | Acrylonitrile | 3,000,000,000 | 5.8757 |
| DTXSID0020446 | Diuron | 39,888 | 5.8593 |

#### Scenario 2: River exposure to industrial waste stream

In scenario 2, a regulatory authority is taking samples from a river system that had been exposed to a inadvertent discharge of an industrial waste water. Similar to scenario 1, the source data contains individual sites with several compounds detected at each site in varying concentrations. Of interest is what is the immediate, acute threat, and what is the threat to long-term health for both the ecosystem and human health. The full code used to demonstrate this is available on the Github and the supplementary material.

For the first part of the analysis, the assessor chooses to prioritize oral, dermal, inhalation, and the acute aquatic toxicity as these pose an immediate threat to clean-up efforts as well as broad spectrum environmental threat to the river system. Several compounds are detected at multiple locations, with Site ‘C’ having the highest weighted mean score. In the top 25 samples (by score), three compounds (Pentachlorophenol, Chlorpyrifos, Caffeine) were detected at all four sites.

For the second part of the analysis, the assessor chose the inverse of the acute hazard assessment, choosing to focus on threats from cancer, genotoxic, endocrine, reproductive, developmental, and chronic aquatic toxicity, while also considering the long-term threat by incorporating persistence, bioaccumulation, and mobility characteristics. Similar to the acute prioritization, Site ‘C’ showed the highest weighted mean score. With the top 25 samples (by score), five compounds were detected at all four locations (Benzyl butyl phthalate, Pentachlorophenol, Bisphenol A, 1,4-Dichlorobenzene, Hydroquinone), offering several canidates for long-term monitoring efforts considering the endpoints assessed.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| | location | count | avg\_score | | --- | --- | --- | | C | 5 | 3.814612 | | B | 5 | 2.962880 | | A | 5 | 2.946431 | | D | 10 | 2.894122 | | | preferredName | count | avg\_score | | --- | --- | --- | | Pentachlorophenol | 4 | 4.0248 | | Chlorpyrifos | 4 | 3.7323 | | Caffeine | 4 | 2.4336 | | Hydroquinone | 2 | 2.6405 | | 1,4-Dichlorobenzene | 2 | 2.5242 | | Benzyl alcohol | 1 | 3.0867 | | 2,4-Dichlorophenoxyacetic acid | 1 | 3.0788 | | Ethylbenzene | 1 | 2.6468 | | Benzoic acid | 1 | 2.6241 | | Triphenyl phosphate | 1 | 2.3182 | | Ethyl acetate | 1 | 2.2377 | | Methyl ethyl ketone | 1 | 2.2270 | | Ethylene glycol | 1 | 2.2057 | | Benzyl butyl phthalate | 1 | 2.2006 | | | preferredName | count | avg\_score | | --- | --- | --- | | Benzyl butyl phthalate | 4 | 5.350200 | | Pentachlorophenol | 4 | 5.281500 | | Bisphenol A | 4 | 4.782900 | | 1,4-Dichlorobenzene | 4 | 4.560350 | | Hydroquinone | 4 | 4.478400 | | Chlorpyrifos | 3 | 4.412633 | | Caffeine | 1 | 4.536600 | | 2,4-Dichlorophenoxyacetic acid | 1 | 4.120200 | | | location | count | avg\_score | | --- | --- | --- | | C | 7 | 4.986809 | | D | 6 | 4.882052 | | B | 6 | 4.786349 | | A | 6 | 4.745915 | |

## Comparisons to other relevant R packages

### webchem

webchem is a “R package to retrieve chemical information from the web…” and provides access to “a suite of web APIs to retrieve chemical information” from a variety of disparate databases. Some of the the databases (e.g.: ChemSpider) require a separate API token to access. Curation quality of these databases is not reported.

Comparatively, ComptoxR provides a single point of access to a variety of already curated and integrated databases under the ‘umbrella’ of CCD. Users can query one compound and retrieve results from upwards of 30 **DOUBLE CHECK** databases and aggregate results at their leisure.

Currently ComptoxR relies on webchem to pull GHS data, but aims to depreciate this dependency as internal efforts at US EPA continue to work towards hosting and curating GHS data.

standartox

### ECOTOXr

ECOTOXr allows for a local download and storage of the US EPA’s ‘ECOTOX’ environmental database in a SQLite format for query. As of writing, ECOTOX does not have an API for querying this database.

ECOTOX is covered under the CCD ‘umbrella’ of databases and is thus covered under this package. ComptoxR does not natively store data automatically, but returned queries can be saved in a variety of R storage formats and recalled later.

### toxpiR

toxpiR is a native implementation of the Toxicological Priority Index (ToxPi) prioritization algorithm, and purports to offer more flexibility than the ToxPi GUI.

Inherently, the math behind the ToxPi algorithm is not hard to reproduce (a min-max normalization scheme), but the toxpiR package utilizes the R6 reference class and may not be as easily accessible to all users. In comparison, the tp function made available through ComptoxR takes two data frame objects (a weighing table and a table holding the parent data) and performs the normalization scheme. We also offer the ability to back-fill data gaps with the minimum, mean, maximum (as supported by a meta-analysis @to\_2018), as well as 50% of the lowest non-zero value of a given endpoint, and provide meta-statistics on the quality of data being analyzed.

## Open science

Given the open-source nature of this project, all source code is currently hosted on an US EPA Github, with future plans to submit to CRAN when development is stable. Code is licensed under the [**INSERT**] license.

## Further Development

Looking forward, we aim to continue to add new data sources. While available, but not currently built out as-of the writing of this paper is the bio-activity data sets from CCD. We also look to adding data endpoints such as product usage, product volume, as well as GHS data. Finally, we envision addition of other ‘Hazard Comparison’ endpoints to better suit risk-assessor needs.

## Conclusions

ComptoxR is a package that was designed to assist with retrieving high quality curated chemical data for various purposes.

##### Disclaimer

This article was supported in part by an appointment to the Research Participation Program at the Office of Research and Development, Center for Environmental Solutions and Emergency Response, U.S. Environmental Protection Agency, administered by the Oak Ridge Institute for Science and Education through an interagency agreement between the U.S. Department of Energy and EPA.The views expressed in this article are those of the author(s) and do not necessarily represent the views or the policies of the U.S. Environmental Protection Agency. EPA is distributing this information solely as a public service.

This R package is a compilation of information sourced from many databases and literature sources, including U.S. Federal and state sources and international bodies, which can save the user time by providing information in one location. The data are not fully reviewed by USEPA – the user must apply judgment in use of the information. You should consult the original scientific paper or data source if possible. Reference herein to any specific commercial products, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government. The views and opinions of the developers of the site expressed herein do not necessarily state or reflect those of the United States Government, and shall not be used for advertising or product endorsement purposes With respect to documents available from this server, neither the United States Government nor any of their employees, makes any warranty, express or implied, including the warranties of merchantability and fitness for a particular purpose, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights.

##### Acknowledgements

We are grateful to for their comments on the manuscript.

###### References