US EPA TOXCAST DATA RELEASE JULY 2019 Summary Files

This file describes the contents of the July 2019 ToxCast data release. The zip file contains the following summary-level files:

```
##  [1] "ac50_Matrix_190708.csv"
##  [2] "AllResults_flags_190708.csv"
##  [3] "Assay_Quality_Detailed_Stats_190708.csv"
##  [4] "Assay_Quality_Summary_Stats_190708.csv"
##  [5] "Assay_Summary_190708.csv"
##  [6] "Chemical_Summary_190708.csv"
##  [7] "cyto_dist_190708.csv"
##  [8] "fitc_Matrix_190708.csv"
##  [9] "hitc_Matrix_190708.csv"
## [10] "logc_max_Matrix_190708.csv"
## [11] "logc_min_Matrix_190708.csv"
## [12] "m4id_Matrix_190708.csv"
## [13] "max_mean_Matrix_190708.csv"
## [14] "max_med_Matrix_190708.csv"
## [15] "modl_ac10_Matrix_190708.csv"
## [16] "modl_acb_Matrix_190708.csv"
## [17] "modl_acc_Matrix_190708.csv"
## [18] "modl_ga_Matrix_190708.csv"
## [19] "modl_gw_Matrix_190708.csv"
## [20] "modl_la_Matrix_190708.csv"
## [21] "modl_lw_Matrix_190708.csv"
## [22] "modl_Matrix_190708.csv"
## [23] "modl_rmse_Matrix_190708.csv"
## [24] "modl_tp_Matrix_190708.csv"
## [25] "neglogac50_Matrix_190708.csv"
## [26] "oldstyle_ac50_Matrix_190708.csv"
## [27] "oldstyle_neg_log_ac50_Matrix_190708.csv"
## [28] "Sample_Summary_190708.csv"
## [29] "spid_Matrix_190708.csv"
## [30] "tested_Matrix_190708.csv"
## [31] "tested_mc_Matrix_190708.csv"
## [32] "tested_sc_Matrix_190708.csv"
## [33] "zscore_Matrix_190708.csv"
```

In addition to the above listed files, the ToxCast program also released a MySQL dump file containing all data and the 2.0 version of the R package (tcpl) that interacts with the MySQL database used to process all of the data for this release. For information/data not included in the listed summary files, users will need to download and interact with the MySQL database. We also encourage the database users to utilize the 'tcpl' R package containing numerous queries and functionality for easily loading and visualizing the data. At the bottom of this file is an R script to produce all of the listed files, utilizing the MySQL database and 'tcpl' R package.

All information in the summary-level files is reported at the chemical level. When more than one sample existed for a given chemical-assay pair, logic incorporating the distribution of activity calls, the shape of the curves, the cautionary flags, and the potency across samples was used to select a single sample. For more information, see the 'tcplSubsetChid' function in the 'tcpl' R package.

Each of the matrix files, indicated by "_Matrix" in the name, contain rows of distinct chemicals and columns of assay endpoints, where each cell contains data for a single chemical-endpoint pair. The first column in

the matrix files gives the chemical code; column names correspond to assay endpoint name. The zip files contains matrices for 23 of the variables captured at level 4 and level 5 of the analysis:

1. "fitc" – the fit category
2. "hitc" – the activity or hit call, 1 indicates active
3. "m4id" – the level 4 id (from database) for the selected sample
4. "logc_max" – log base 10 of the maximum concentration tested
5. "logc_min" – log base 10 of the minimum concentration tested
6. "max_mean" – the maximum of the means at each concentration
7. "max_med" – the maximum of the medians at each concentration
8. "modl_ac10" – the activity concentration at 10% of the modeled top value (AC10)
9. "modl_acb" – the activity concentration at baseline
10. "modl_acc" – the activity concentration at cutoff
11. "modl_ga" – the gain AC50
12. "modl_gw" – the gain hill coefficient
13. "modl_la" – the loss AC50
14. "modl_lw" – the loss hill coefficient
15. "modl" – the winning model
16. "modl_rmse" – the root mean square error (RMSE)
17. "modl_tp" – the modeled top of the curve
18. "spid" – the sample id for the selected sample
19. "tested" – whether the chemical was tested, 1 indicates tested
20. "tested_mc" – whether the chemical was tested, 1 indicates tested in the multiple-concentration format
21. "tested_sc" – whether the chemical was tested, 1 indicates tested in the single-concentration format
22. "zscore" – the zscore of AC50 values based on the chemical-specific cytotoxicity distribution file
23. "ac50" - a modified AC50 table (in non-log units) where assay/chemical pairs that were not tested, or tested and had a hitcall of 0 or -1 have the value 1e6.
24. "neglogac50" - log(AC50/1e6) where assay/chemical pairs that were not tested, or tested and had a hitcall of 0 or -1 have the value 0.

NOTE: For all matrix files directly from 'tcpl' R package, concentrations are given in log base 10 micromolar units.

Two new matrix files have been generated. These files are in the same format as the files above. However, these files use a combination of three matrix files, "modl_ga", "hitc", and "tested", to create an "old-style AC50 matrix similar to the files released in previous generations of ToxCast data release.

25. "oldstyle_ac50" – modl_ga (AC50) in micromolar [not log10 value] for all active and tested chemical-assay combinations. A million value represented all inactive and tested chemical-assay combinations and NA values represent not tested.
26. "oldstyle_neg_log_ac50" – the negative log10 transformation of the above AC50 matrix anchored to zero; -log10(AC50/1000000)

All parameters beginning with "modl" are derived from the winning model. The complete set of parameters for all models is available in the MySQL database. NA values in the matrix files have different meanings, depending on the file. NA either means we did not test the chemical, or we could not compute the parameter. For example, when the constant model wins, we cannot compute a gain AC50. Similarly, if the Hill model wins, the loss AC50 is not applicable. NA in the hit-call matrix ("hitc") means the chemical did not get tested in the multiple concentration format. However, the chemical may have been tested in an initial screen at a single concentration and was not selected for further testing. The "tested" matrix indicates whether the chemical has been tested in an assay, and reflects both the single-concentration and multiple-concentration screening formats.

The hit-call matrix contains NA, 0, 1, and -1 values. "NA" indicates the chemical was not tested in the multiple-concentration screening format, "0" indicates the chemical was determined inactive, "1" indicates

the chemical was determined active, and "-1" indicates that the activity could not be determined. Only chemicals with less than 4 concentrations of data received the "-1" designation.

The parameters for the winning model are given regardless of hit-calling; therefore, many inactive chemicals have a gain AC50 value in the "modl_ga" file.

The other two data files contain the cautionary flags for all the selected samples in the matrix files and the cytotoxicity distribution by chemical. The flag file contains many database id values, basic chemical information, the assay endpoint name (aenm), and the flag information. The flag information includes the flag database id (l6_mthd_id), the flag output (flag), and the flag value/unit (fval/fval_unit) when applicable. Not all flags have an associated value.

The cytotoxicity distribution file contains basic chemical information, the median (med) and MAD (mad) of gain AC50 across the cytotoxicity assays (in log base 10 micromolar units), and the number of cytotoxicity assays with an active hit-call (nhit). The global MAD (global_mad) is defined as the median of all the MAD values, excluding NA values. The cytotoxicity assays are indicated by the "burst_assay" field in the assay summary file. When a chemical hits less than two cytotoxicity assays, the cytotoxicity point (cyto_pt) is defined as 3, otherwise the cytotoxicity point is the cytotoxicity median (med) for the chemical. "cyto_pt_um" and "lower_bnd_um" are the cytoxicity point and the cytotoxicity point minus three times the global MAD in micromolar units, respectively.

The chemical summary file contains the mapping from "code" to CASRN (casn), DSSTox_GSID (chid), and chemical name (chnm) for all unique chemicals. The assay summary files contain all annotations and quality statistics for the currently released assay set. More detailed descriptions of these files are available in the assay information README file, "README_INVITRODB_V3_2_ASSAYINFO".

Detailed information about all of the parameters is available in the tcpl R package vignette, "Pipeline_Overview.pdf" (ToxCast Data Pipeline Overview).

For questions or concerns, please contact Monica Linnenbrink at: linnenbrink.monica@epa.gov.

R Script to produce July 2019 ToxCast Tox21 Data Release

```
# Connect to database using tcplConf before running
rm(list = ls())
library(tcpl)
library(data.table)
library(parallel)
post <- format(Sys.Date(), "_%y%m%d.csv")
post <- format(Sys.Date(), "_%y%m%d.csv")
mainDir <- paste0(toupper(format(Sys.time(), "%b%Y")),"_TOXCAST_EXTERNAL_RELEASE")
subDir <- paste0(toupper(str_extract(tcplConfList()$TCPL_DB,"invitrodb.*")),"_SUMMARY")
dir.create(file.path(getwd(),mainDir))
dir.create(file.path(getwd(),mainDir,subDir))

## Write the matrix files and cytotoxicity distribution file
res <- tcplVarMat(
  add.vars = c("modl_ga", "hitc", "modl_tp", "modl_la", "modl","max_mean",
    "modl_acc", "modl_acb", "modl_ac10", "max_med", "logc_max",
    "logc_min", "spid", "m4id", "modl_gw", "modl_lw", "fitc",
    "modl_rmse"
  ),
  odir = file.path(
    getwd(),
    mainDir,
```

```
    subDir
  )
)
names(res) <- vars
# generate toxcast data matrices
mat_ga <- res[["modl_ga"]]
mat_hitc <- res[["hitc"]]
mat_test <- res[["tested"]]
cns <- colnames(mat_ga)
rns <- rownames(mat_ga)
mat_test <- mat_test[rns, cns]
mat_ac <- 10^mat_ga
mat_ac[mat_hitc == 0] <- 1e6
mat_ac[is.na(mat_ga) & mat_test == 1] <- 1e6
mat_lac <- -log10(mat_ac / 1e6)
fdate <- format(Sys.Date(), "%y%m%d.csv")
#get the data for cyto_dist table
cyto_dist <- tcplPrepOtpt(tcplCytoPt())
write.csv(cyto_dist, file.path(
  getwd(),
  mainDir,
  subDir,
  paste0("cyto_dist_", fdate)
),
row.names = FALSE
)

write.csv(mat_test, file.path(
  getwd(),
  mainDir,
  subDir,
  paste0("tested_Matrix_", fdate)
),
row.names = TRUE
) # overwrites default tested matrix to make sure nrow is the same across all
write.csv(mat_ac, file.path(
  getwd(),
  mainDir,
  subDir,
  paste0("oldstyle_ac50_Matrix_", fdate)
),
row.names = TRUE
)
write.csv(mat_lac, file.path(
  getwd(),
  mainDir,
  subDir,
  paste0("oldstyle_neg_log_ac50_Matrix_", fdate)
),
row.names = TRUE
)
## Write the chemical summary files
clib <- tcplLoadClib()
```

```
clib <- clib[, list(clib = paste(clib, collapse = "|")), by = chid]
chem <- tcplLoadChem(include.spid = FALSE)
chem <- merge(x = chem, y = clib, by = "chid", all.x = TRUE)
write.csv(chem, file.path(
  getwd(),
  mainDir,
  subDir,
  paste0("Chemical_Summary", post)
),
row.names = FALSE
)

sample <- tcplQuery("SELECT spid, stkc, stkc_unit, tested_conc_unit,
                    chemical.*
                    FROM sample, chemical WHERE sample.chid =
                    chemical.chid")
write.csv(sample, file.path(
  getwd(),
  mainDir,
  subDir,
  paste0("Sample_Summary", post)
),
row.names = FALSE
)
## Write the flag file
m4ids <- res[["m4id"]][!is.na(res[["m4id"]])]
write.csv(tcplPrepOtpt(tcplLoadData(
  lvl = 6L,
  fld = "m4id",
  val = m4ids
)),
file.path(
  getwd(),
  mainDir,
  subDir,
  paste0("AllResults_flags", post)
),
row.names = FALSE
)
```