US EPA TOXCAST DATA RELEASE OCTOBER 2021 Model (level 4), Activity Calls (level 5), & Flag (level 6) Export Spreadsheet Files

This file describes the contents of the October 2021 ToxCast data release. The zip file contains the following detailed export files per assay source, not including this README file:

```
##  [1] "EXPORT_LVL5&6_ASID11_CLD_210920.csv"
##  [2] "EXPORT_LVL5&6_ASID12_CCTE_PADILLA_210920.csv"
##  [3] "EXPORT_LVL5&6_ASID13_TANGUAY_210920.csv"
##  [4] "EXPORT_LVL5&6_ASID14_STM_210920.csv"
##  [5] "EXPORT_LVL5&6_ASID17_CCTE_210920.csv"
##  [6] "EXPORT_LVL5&6_ASID1_ACEA_210920.csv"
##  [7] "EXPORT_LVL5&6_ASID20_CCTE_SHAFER_210920.csv"
##  [8] "EXPORT_LVL5&6_ASID21_CPHEA_STOKER_210920.csv"
##  [9] "EXPORT_LVL5&6_ASID24_CCTE_GLTED_210920.csv"
## [10] "EXPORT_LVL5&6_ASID25_UPITT_210920.csv"
## [11] "EXPORT_LVL5&6_ASID28_ERF_210920.csv"
## [12] "EXPORT_LVL5&6_ASID2_APR_210920.csv"
## [13] "EXPORT_LVL5&6_ASID31_CCTE_MUNDY_210920.csv"
## [14] "EXPORT_LVL5&6_ASID3_ATG_210920.csv"
## [15] "EXPORT_LVL5&6_ASID4_BSK_210920.csv"
## [16] "EXPORT_LVL5&6_ASID5_NVS_210920.csv"
## [17] "EXPORT_LVL5&6_ASID6_OT_210920.csv"
## [18] "EXPORT_LVL5&6_ASID7_TOX21_210920.csv"
## [19] "EXPORT_LVL5&6_ASID8_CEETOX_210920.csv"
## [20] "EXPORT_LVL5&6_ASID9_LTEA_210920.csv"
```

In addition to the above listed files, the ToxCast program also released a MySQL dump file containing all data and a beta 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. For more information on how to on data retrieval with tcpl please see the data retrieval vignette. https://cran.r-project.org/web/packages/tcpl/vignettes/Data_retrieval.html

All information in the summary file is reported at the sample level and a single file has been produced per assay source name (asnm) or assay source id (asid). Each row in this file contains a unique combination of sample (spid) and assay endpoint (aeid) with all of the model information applied to the underlying concentration response data.

References:

Filer et al. 2017 (DOI: https://doi.org/10.1093/bioinformatics/btw680) Additional tcpl vignettes:

Introduction - https://cran.r-project.org/web/packages/tcpl/vignettes/Introduction_Appendices.html

Data Processing - https://cran.r-project.org/web/packages/tcpl/vignettes/Data_processing.html

All columns starting with 'cnst' refer to parameters from the constant model. All columns starting with 'hill' refer to parameters from the hill model. All columns starting with 'gnls' refer to parameters from the gain-loss model (the product of 2 hill models, one with a negative hill slope, and that share a top/upper-asymptote). For more information on the specifics of the modeling process please refer to the data analysis R package documentation (Filer et al. 2017 DOI: https://doi.org/10.1093/bioinformatics/btw680). All columns starting with 'modl' refer to the winning models parameters. Below is the list of the 91 columns exported for this dataset from Level 4 (modeling), 5 (model selection & hit calling), 6 (flagging) data processing and 7 (statistical bootstrapping).

1. m5id = unique id level 5 processing
2. spid = sample id (id blindly provided to vendors)
3. chid = chemical id (DSSTox GSID) 1:1 with casrn
4. casn = CAS Registry number
5. chnm = chemical name
6. code = CAS Registry number (excel protected)
7. aeid = assay endpoint id (unique id)
8. aenm = assay endpoint name
9. m4id = unique id for level 4 processing
10. bmad = baseline median absolute deviation (noise around baseline)
11. resp_max = maximal single replicate response
12. resp_min = minimal single replicate response
13. max_mean = maximal mean response at a given concentration
14. max_mean_conc = corresponding concentration of max_mean
15. max_med = maximal median response at a given concentration
16. max_med_conc = corresponding concentration of max_med
17. logc_max = maximum tested log concentration (log uM)
18. logc_min = minimum tested log concentration (log uM)
19. cnst = constant model successfully run (1 or 0)
20. hill = hill model successfully run (1 or 0)
21. hcov = hill model covariance
22. gnls = gain-loss model successfully run (1 or 0)
23. gcov = gain-loss model covariance
24. cnst_er = constant model error term
25. cnst_aic = constant model AIC (used to select winning model)
26. cnst_rmse = constant model RMSE
27. cnst_prob = constant model probability (based on AIC)
28. hill_tp = hill model top of curve
29. hill_tp_sd = hill model top standard deviation
30. hill_ga = hill model logAC50 (gain logAC50)
31. hill_ga_sd = hill model AC50 standard deviation
32. hill_gw = hill model slope
33. hill_gw_sd = hill model slope standard deviation
34. hill_er = hill model error term
35. hill_er_sd = hill model error standard deviation
36. hill_aic = hill model AIC (used to select winning model)
37. hill_rmse = hill model RMSE
38. hill_prob = hill model probability (based on AIC)
39. gnls_tp = gain-loss top of curve
40. gnls_tp_sd = gain-loss top of curve standard deviation
41. gnls_ga = gain-loss model gain logAC50
42. gnls_ga_sd = gain-loss model gain logAC50 standard deviation
43. gnls_gw = gain-loss model gain slope (positive)
44. gnls_gw_sd = gain-loss model gain slope standard deviation
45. gnls_la = gain-loss model loss logAC50
46. gnls_la_sd = gain-loss model loss logAC50 standard deviation
47. gnls_lw = gain-loss model loss slope (negative)
48. gnls_lw_sd = gain-loss model loss slope standard deviation
49. gnls_er = gain-loss model error term
50. gnls_er_sd = gain-loss model error standard deviation
51. gnls_aic = gain-loss model AIC (used to select model winner)
52. gnls_rmse = gain-loss model RMSE
53. gnls_prob = gain-loss model probability (based on AIC)
54. nconc = number of tested concentrations

55. npts = number of data points
56. nrep = number of replicates
57. nmed_gtbl = number of median values greater than baseline
58. hitc = hit call (based on 'coff' and winning model) positive =1, negative =0, not enough data to fit = -1; max_med must exceed coff
59. modl = winning model
60. fitc = fit category (defined by many parameters)
61. coff = response cutoff (used to define hit-call)
62. actp = activity probability (1-cnst_prob)
63. modl_er = winning model error term
64. modl_tp = winning model top of curve (where applicable)
65. modl_ga = winning model gain logAC50 (where applicable)
66. modl_gw = winning model gain slope (where applicable)
67. modl_la = winning model loss logAC50 (where applicable)
68. modl_lw = winning model loss slope (where applicable)
69. modl_rmse = winning model RMSE
70. modl_prob = winning model probability
71. modl_acc = winning model log concentration at 'coff'
72. modl_acb = winning model log concentration at 'bmad'
73. resp_unit = response units
74. flag_id = concatenated list of flag ids
75. flag = concatenated list of flag names
76. chit = chemical-level hit call
77. stkc = stock concentration of sample
78. stkc_unit = stock concentration unit, typically mM
79. test_conc_unit = tested concentration unit, typically uM
80. spid_legacy = legacy sample id
81. gsid_rep = representative sample; based on tcplSubsetChid()
82. hit_pct = Total percent of hit calls made after 1000 bootstraps
83. total_hitc = Total number of hit calls made after 1000 bootstraps
84. modl_ga_min = Low bound of the 95% confidence interval for the AC50
85. modl_ga_max = Upper bound of the 95% confidence interval for the AC50
86. modl_ga_med = Median AC50 after 1000 bootstraps
87. modl_gw_med = Median gain Hill coefficient for 1000 bootstraps
88. modl_ga_delta = AC50 confidence interval width in log units
89. cnst_pct = Percent of 1000 bootstraps that the constant model was selected as the winning model
90. hill_pct = Percent of 1000 bootstraps that the Hill model was selected as the winning model
91. gnls_pct = Percent of 1000 bootstraps that the gain-loss was selected as the winning model

The parameters for the winning model are given regardless of hit-calling; therefore, many inactive chemicals have a gain AC50 chemical in the "modl_ga" column, for example. The "hitc" column provides the activity call (1=active, 0=inactive, -1=unable to model)

For questions or concerns, please contact Jason Brown at: brown.jason@epa.gov

R Script to produce October 2021 ToxCast Tox21 Data Release

```
library(tcpl)
library(data.table)
library(parallel)
library(stringr)
# must appropriately connect to tcpl database and set working directory correctly
```

```
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.*")),"_LEVEL5")
dir.create(file.path(getwd(),mainDir))
dir.create(file.path(getwd(),mainDir,subDir))


tcplExportAsidFits <- function(asid) {
  aeids <- tcplLoadAeid("asid", asid)$aeid
  dat5 <- tcplPrepOtpt(tcplLoadData(5L, fld = "aeid", aeids))
  agg5 <- tcplSubsetChid(dat5)[, list(m5id, gsid_rep = 1)]
  dat6 <- tcplLoadData(6L, fld = "aeid", aeids)
  dat7 <- tcplLoadData(7, fld = "aeid", aeids)
  agg6 <- dat6[, list(
    flag_ids = paste(mc6_mthd_id, collapse = "|"),
    flags = paste(flag, collapse = "|"),
    flag_length = .N
  )
  ,
  by = m5id
  ]
  dat5 <- merge(dat5,dat7, by = c("m4id","aeid"), all.x = T)
  setkey(dat5, "m5id")
  setkey(agg5, "m5id")
  setkey(agg6, "m5id")
  dat5 <- agg5[dat5]
  dat5 <- agg6[dat5]
  dat5[is.na(gsid_rep), gsid_rep := 0]

  sample <- tcplQuery("SELECT * FROM sample")
  setkey(sample, "spid")
  setkey(dat5, "spid")
  dat5 <- sample[dat5]

  dat5 <- dat5[, c(
    "m4id", "m5id",
    "spid", "stkc", "stkc_unit", "tested_conc_unit",
    "chid", "casn", "chnm", "code",
    "aeid", "aenm", "resp_unit",
    "nconc", "npts", "nrep", "nmed_gtbl",
    "hitc", "modl", "fitc", "coff", "bmad",
    "gsid_rep", "chit",
    "flag_ids", "flags","flag_length",
    "resp_max", "resp_min", "max_mean", "max_mean_conc",
    "max_med", "max_med_conc",
    "logc_max", "logc_min", "actp",
    "modl_er", "modl_tp", "modl_ga", "modl_gw", "modl_la",
    "modl_lw", "modl_rmse", "modl_prob", "modl_acc", "modl_acb",
    "cnst", "hill", "hcov", "gnls", "gcov",
    "cnst_er", "cnst_aic", "cnst_rmse", "cnst_prob",
    "hill_tp", "hill_tp_sd", "hill_ga", "hill_ga_sd", "hill_gw",
    "hill_gw_sd", "hill_er", "hill_er_sd", "hill_aic", "hill_rmse",
    "hill_prob",
```

```
    "gnls_tp", "gnls_tp_sd", "gnls_ga", "gnls_ga_sd", "gnls_gw",
    "gnls_gw_sd", "gnls_la", "gnls_la_sd", "gnls_lw", "gnls_lw_sd",
    "gnls_er", "gnls_er_sd", "gnls_aic", "gnls_rmse", "gnls_prob","hit_pct","total_hitc","modl_ga_min",
    "modl_ga_med",   "modl_gw_med",   "modl_ga_delta", "cnst_pct",  "hill_pct", "gnls_pct"
  ),
  with = FALSE
  ]

  setkeyv(dat5, cols = c("aeid", "chid", "gsid_rep", "spid"))

  fname <- file.path(
    getwd(),
    mainDir,
    subDir,
    paste("EXPORT_LVL5&6",
      paste0("ASID", asid),
      tcplLoadAsid("asid", asid)$asnm,
      format(Sys.Date(), "%y%m%d.csv"),
      sep = "_"
    )
  )
  write.csv(dat5, fname, row.names = FALSE)
}
# asids <- tcplLoadAsid()$asid
asids <- unique(tcplLoadAeid(fld = "aeid", val = unique(tcplLoadData(5)$aeid), add.fld = "asid")$asid)
mclapply(asids, tcplExportAsidFits, mc.cores = length(asids))
```