Walk-through ThETA

Table of content

Summary

The Transcriptome-driven Efficacy estimates for gene-based TArget discovery (ThETA) R package provides an easy way to implement novel transcriptome-based efficacy scores of target(gene)-disease associations, which are defined in Failli et al. 2019. Here, we describe

- how to compile tissue-specific and modulation scores for disease-gene associations;
- how to integrate these scores to Open Target-based scores;
- how to compile annotations and tissue-specific networks associated to top selected targets;

1. Introduction

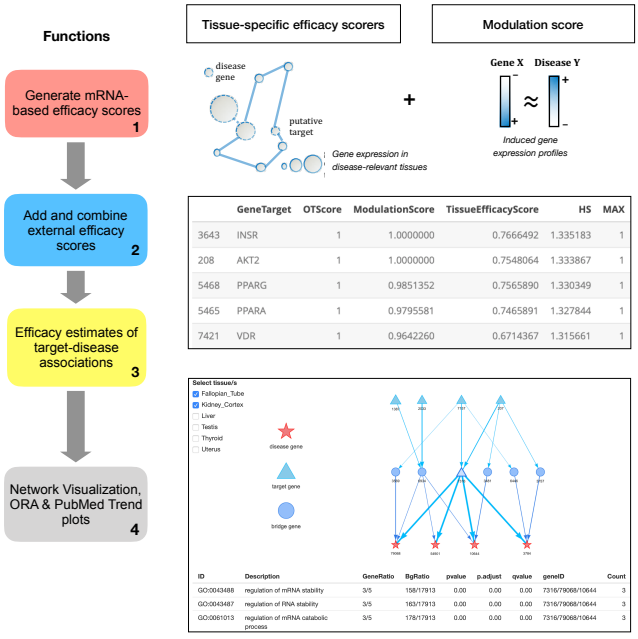


Figure 1 - Graphical illustration of the steps implemented in the R package ThETA. 1) Generate disease-gene association scores by using two novel mRNA-based efficacy scores. 2) Add and combine efficacy scores obtained from different computational methods. 3) Compile the efficacy estimates for all annotated disease-gene pairs. 4) Visualization tools to visualize identified drug targets(genes) in tissue-specific networks and enrichment results of biological pathways.

Figure 1 shows the workflow implemented in the R package ThETA. Current functions provided by ThETA are listed below:

- Compile tissue-specific expression networks by using GTEx and StringDB (Human PPI).
- Compile disease-relevant tissues by implementing the algorithm proposed by Kitsak et al. 2016 (https://www.nature.com/articles/srep35241).
- Extract disease-relevant genes from DisGeNET and mark these genes on the disease-relevant tissue-specific gene expression.
- Compile the tissue-specific efficacy scores on disease-relevant tissues.
- Compile the modulation score, which estimates the likelihood of a gene perturbation (e.g., knockout and knockdown) to result in specific reversion of disease gene-expression profiles (lists of down- and up-regulated genes are downloaded from Enrichr: https://amp.pharm.mssm.edu/Enrichr/).
    - Integrate multiple efficacy scores with the max function and the harmonic sum (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5210543/).
    - Build igraph objects including tissue-specific networks and paths connecting selected drug-targets (or genes) and disease-relevant genes (it also includes info on the gene modulation scores).

2. Generate tissue-specific efficacy scores for a given 'disease'

In order to compile tissue-specific efficacy estimates of drug-target disease associations we need to:

- Collect tissue-specific gene expression profiles from GTEX.
- Define a protein-protein interaction network.
- Identify disease-associated genes (from DisGeNET) and disease relevant tissues.
- Compile node centrality scores

These steps are computationally expensive! Therefore, ThETA provides pre-compiled .rda files that can be used to rapidly generate tissue-specific efficacy scores.

.rda file	**description**
`gtexv7_zscore.rda`	z-scores compiled from log transformed TPM expression profiles of GTEx.
`ppi_strdb_700.rda`	human protein-protein interaction network extracted from StringDB (combined scores >= 700)
`dis_vrnts.rda`	disease-associated genes (from DisGeNET; score >= 0.6)
`disease_tissue_zscores.rda`	significances (z-scores) of disease-tissue associations
`dis_vrnts.rda`	tissue-specific node centrality scores (integration of degree, clust. coeff. and betweenness

Therefore, the users first upload the R package ThETA and the .rda files needed for the calculation of the tissue-specific efficacy scores.

```
library(ThETA)
data(gtexv7_zscore)
data(ppi_strdb_700)
data(dis_vrnts)
data(disease_tissue_zscores)
data(centrality_score)
```

Then, given a disease (i.e. Diabetes Mellitus Type II - T2DM), we can compile the tissue-specific efficacy scores. The disease must be specified by indicating the corresponding EFO-ID. The scope of the Experimental Factor Ontology (EFO) is to combine biological ontologies, such as anatomy, disease and chemical compounds, and to support the annotation, analysis and visualization of data handled by many groups at the EBI and as the core ontology for OpenTargets.org. The EFO-ID has been selected as identifier for disease terms in order to facilitate the integration with the Open Target scores for disease-gene associations. Therefore, the users can identify the correct EFO-id for their disease terms by using the following web-form: https://www.ebi.ac.uk/ols/ontologies/efo.

Therefore, we can now use the EFO-id related to T2DM and select the corresponding variant genes from the pre-compiled .rda file *dis_vrnts*. The disease gene variants were retrieved from DisGeNET.

```
T2DM_genes = dis_vrnts[[which(names(dis_vrnts) == "EFO:0001360")]]
```

Then, we used the pre-compiled files "*disease_tissue_zscores*" to find significant tissues for T2DM.

```
T2DM_rel_tissue_scores =
disease_tissue_zscores$z[which(rownames(disease_tissue_zscores$z) == "EFO:0001360"),]
```

A tissue-specific efficacy (TSE) score is then estimated for all genes that are expressed in the tissues that are relevant for T2D. It should be noted that the following script is computer-intensive. Indeed, we specified only two genes in input. However, it is highly recommended to use the whole set of T2D-relevant genes.

```
T2DM_Tscores <- tissue.specific.scores(T2DM_genes$entrez[1:2],
                                       ppi_network = ppi_strdb_700,
                                       directed_network = FALSE,
                                       tissue_expr_data = gtexv7_zscore,
                                       dis_relevant_tissues = T2DM_rel_tissue_scores,
                                       W = centrality_score$borda.disc,
                                       cutoff = 4, verbose = TRUE)
```

The ThETA R package also provides a parallel version of this functions when compiling the tissue-specific genes for a list of disease terms, namely "*list.tissue.specific.scores*".

The verbose-related messages for "*tissue.specific.scores*" indicate progress notifications:

```
"6 tissue/s significant for given disease."
"Compiling the tissue-specific efficacy scores for disease-genes in Fallopian_Tube."
"Compiling the tissue-specific efficacy scores for disease-genes in Kidney_Cortex."
"Compiling the tissue-specific efficacy scores for disease-genes in Liver."
"Compiling the tissue-specific efficacy scores for disease-genes in Testis."
"Compiling the tissue-specific efficacy scores for disease-genes in Thyroid."
```

The output is a *data.frame* object containing the TSE score for all genes-tissue pairs.

	Fallopian_Tube	Kidney_Cortex	Liver	Testis	Thyroid	Uterus	avg_tissue_score
1080	0.6030391	0.6152350	0.6311006	0.6045757	0.5482383	0.5823116	0.5974167
6356	0.4022889	0.5029331	0.6117747	0.5240006	0.4777560	0.4218802	0.4901056
5734	0.5131348	0.5667641	0.5722064	0.5692878	0.7050607	0.4737322	0.5666977
3308	0.6161507	0.5682854	0.5524524	0.6434554	0.5724015	0.6165954	0.5948901
136	0.5348355	0.5012205	0.5654343	0.5561065	0.6517366	0.4500510	0.5432307

Finally, we can order the disease-gene associations based on the "*avg_tissue_score*" and select the top 50 gene targets.

```
T2DM_top50 <- T2DM_Tscores[order(T2DM_Tscores$avg_tissue_score, decreasing =
TRUE)[1:50],]
```

3. Generate modulation-based efficacy scores for a given 'disease'

Since the modulation score is not time-consuming, ThETA provides a function to compile the modulation scores for all available gene sets: disease and gene perturbations. An important input of this process is the set of up- and down-regulated gene sets identified in disease and gene perturbations. Currently, ThETA includes lists of up- or down-regulated gene sets retrieved from EnrichR.
(https://amp.pharm.mssm.edu/Enrichr/).

However, the users could compile the modulation score based on a different set of up- and down-regulated gene sets. The input need to compile the modulation score should be a list of four different gene sets:

- Disease Perturbations – genes up.
- Disease Perturbations – genes down.
- Single Gene Perturbations – genes up.
- Single Gene Perturbations – genes down.

We first upload the dysregulated gene lists by using the .rda file "*geo_gene_sets*".

```
data(geo_gene_sets)
```

The ThETA package provides a function to calculate the modulation score for all the genes (since it is not a computationally intensive task).

```
modulation_scores <- modulation.score(geneSets = geo_gene_sets, verbose = TRUE)
```

EnrichR provides either Disease Ontology (DO) ids or Concept Unique Identifiers (CUIs) to label disease perturbations. Use of different types of disease id may cause gene-disease pair repetitions in the final output of the *modulation.score* (different DO ids might be associated with the same disease). To overcome this issue, a .csv file containing manually curated mapping between either DO ids or CUIs and EFO ids is available in the data folder.

The following code shows how to:

* Cross-link the output of *modulation.score* with the .csv file in data.
* Remove duplicated gene-disease pairs from the output.

First, DO ids or CUIs are replaced with EFO ids.

```
enrichr_to_efo <- read.csv(system.file("conversion_enrichr_efo.csv",
                                       package = "ThETA"), row.names = 1,
                                       stringsAsFactors = F)
modulation_scores$disease.id <-
enrichr_to_efo[modulation_scores$disease.id,'disease.id']
```

Then, gene symbols need to be converted to Entrez Gene IDs in order to facilitate the integration between TSE and modulation scores.

```
library(org.Hs.eg.db)
modulation_scores$target.entrez <- AnnotationDbi::mapIds(org.Hs.eg.db,
modulation_scores$target.id,'ENTREZID','SYMBOL')
modulation_scores <- modulation_scores[modulation_scores$disease.id != '' &
                                       !is.na(modulation_scores$target.entrez),]
```

```
library(data.table)
modul_score <- data.table::as.data.table(modulation_scores)
modul_score <- as.data.frame(modul_score[, .SD[which.max(modscore)],
                                         by=list(disease.id, target.entrez)])
```

Let's now select the modulation scores for T2D.

```
T2DM_Mscores = data.frame(modul_score[modul_score$disease.id=='EFO:0001360',
                                      c("target.entrez", "modscore")], row.names = 1)
```

	modscore
5376	0.8885534
28996	0.6405335
7403	0.6394576
9126	0.6087785
2908	0.5864897
4209	0.7700771

Here we show how to integrate the tissue-specific efficacy scores with the modulation scores.

```
common_t2d_genes <- intersect(rownames(T2DM_Mscores), rownames(T2DM_Tscores))
T2DM_Iscores <- data.frame("Mscore" = T2DM_Mscores[common_t2d_genes,],
                           "TSEscore" = T2DM_Tscores[common_t2d_genes,],
                           row.names = common_t2d_genes)
```

	Mscore	TSEscore.Fallopian_Tube	TSEscore.Kidney_Cortex	TSEscore.avg_tissue_score
5376	0.8885534	0.4080139	0.3245370	0.3405319
28996	0.6405335	0.6021129	0.6343235	0.5993317
7403	0.6394576	0.4560764	0.3476318	0.4438615
9126	0.6087785	0.6297193	0.5543101	0.5976275
2908	0.5864897	0.6044743	0.5592048	0.5856605

4. Benchmark efficacy scoring methods with real-world targets in DrugBank and TTD

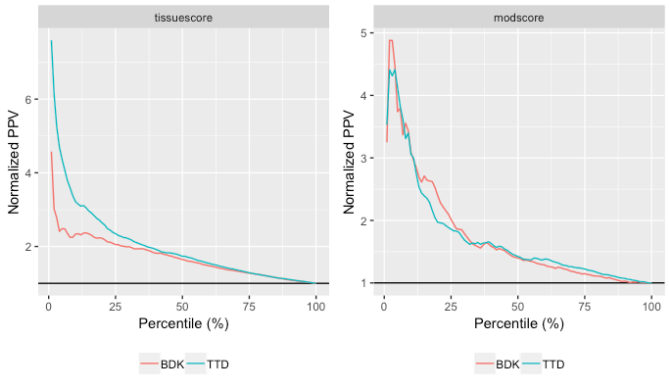
The R package ThETA also provides plotting functions to visualize positive predictive value (PPV) curves when comparing the efficacy estimates of target-disease associations with a golden standard (a set of true/known target-disease associations). The current version of ThETA includes three golden standards: the Therapeutic Target Database (Chen et al. 2002), the Comparative Toxicogenomic Database (Davis et al. 2017) and DrugBank (Wishart et al. 2018).

```
vm <- ppvpercents(list(TTD=GS_TTD, BDK=GS_DBK), modul_score,
                  entrez.col = "target.entrez",
                  disease.col = "disease.id",
                  score.col = c("modscore"))


vt <- ppvpercents(list(TTD=GS_TTD, BDK=GS_DBK), avg_tissue_score,
                  entrez.col = "target.entrez",
                  disease.col = "disease.id",
                  score.col = c("tissuescore"))

gridExtra::grid.arrange(vt + theme(legend.position="bottom"),
                        vm + theme(legend.position="bottom"), ncol=2)
```

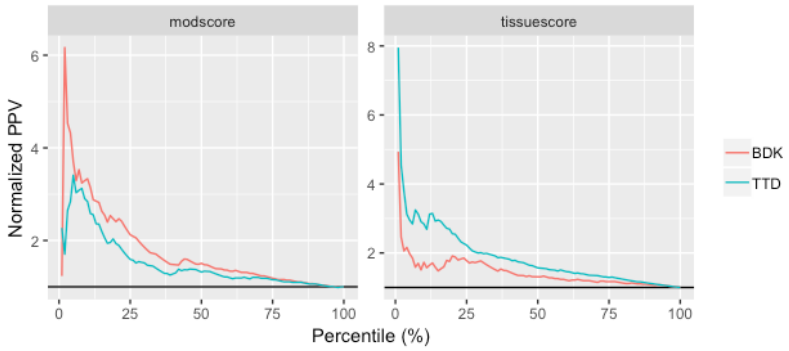
Please, note that in order to compile the following plots the vectors *module_score* and *avg_tissue_score* must be provided. These two datasets of efficacy estimate can be compiled by using the provided vignette files.



In the following example, we show the PPV curves for the target-disease associations having a modulation and a tissue-specific efficacy score.

```
modtis <- merge(modul_score[,c(1,2,4)],
                avg_tissue_score[,c(1,3,4)],
                by=c("disease.id","target.entrez"))

va <- ppvpercents(list(TTD=GS_TTD, BDK=GS_DBK), modtis,
                  entrez.col = 'target.entrez',
                  disease.col = 'disease.id',
                  score.col = c('modscore','tissuescore'))
```



5. Combine the novel mRNA-based efficacy scores with the Open Targets scores

The Open Targets (OT) platform scores target-disease associations across different data sources and types (https://docs.targetvalidation.org/faq/association-score). In a recent study we have shown that combining these scores with our novel efficacy scores improves the accuracy in predicting in predicting known target-disease associations. In addition, novel gene-disease associations are discovered. (Failli et al. 2019) The OT Platform REST API allows access to data available on the OT Platform.

The following examples shows how to retrieve disease-gene association scores from the OT platform.

```
server <- 'https://platform-api.opentargets.io/v3/platform'
endpoint_prmtrs <- '/public/association/filter'
optional_prmtrs <-
'?size=10000&disease=EFO_0001360&fields=disease.id&fields=target.gene_info.symbol&fiel
ds=association_score.overall&fields=disease.efo_info.label'
uri <- paste(server,endpoint_prmtrs,optional_prmtrs,sep='')
```

Then, we use a `GET` request to pull raw data into our environment. Pulled data, in the JavaScript Object Notification (JSON) format, are subsequently converted into a usable format.

```
if("httr" %in% rownames(installed.packages()) == FALSE) {install.packages("httr")}
if("jsonlite" %in% rownames(installed.packages()) == FALSE)
{install.packages("jsonlite")}
library(httr)
library(jsonlite)

get_association_json <- httr::content(httr::GET(uri),'text')
get_association_usable <- jsonlite::fromJSON(get_association_json, flatten = TRUE)

OT_score <- get_association_usable$data[,c(2:3,1,4)]
OT_score$disease.id <- gsub('_',':',OT_score$disease.id)
colnames(OT_score)[c(1,4)] <- c('target.id', 'disease.name')

# remove duplicated gene symbols
OT_score = OT_score[-which(duplicated(OT_score$target.id)),]
```

We then convert the Gene symbols to Entrez Gene IDs in order to align the OT scores with those provided by ThETA.

```
library(org.Hs.eg.db)
OT_score$target.entrez <-
AnnotationDbi::mapIds(org.Hs.eg.db,OT_score$target.id,'ENTREZID','SYMBOL')
OT_score <- OT_score[!is.na(OT_score$target.entrez),]
```

target.id	disease.id	association_score.overall	disease.name	target.entrez
PPARG	EFO:0001360	1	type II diabetes mellitus	5468
KCNJ11	EFO:0001360	1	type II diabetes mellitus	3767
INSR	EFO:0001360	1	type II diabetes mellitus	3643

The scores obtained from the OT platform are first concatenated to the TSE and modulation scores.

```
all_scores <- base::merge(OT_score, T2DM_Iscores, by.x = "target.entrez", by.y =
"row.names", all = TRUE)
```

Then, we use the function *integrate.scores* to provide merged scores: harmonic sum or maximum score.

```
T2DM_allsc <- integrate.scores(all_scores, c("association_score.overall",
                                             "Mscore",
                                             "TSEscore.avg_tissue_score"))

T2DM_allsc <- T2DM_allsc[order(T2DM_allsc$HS, decreasing = TRUE),]
rownames(T2DM_allsc) <- T2DM_allsc[,1]

# let's simplify the final table of the disease-gene association scores
tab_score <- T2DM_allsc[,c("target.id","association_score.overall", "Mscore",
                           "TSEscore.avg_tissue_score", "HS","MAX")]
colnames(tab_score)[1:4] <-
c("GeneTarget","OTScore","ModulationScore","TissueEfficacyScore")
```

Here, we show the top 6 gene target based on an integration of different disease-gene associations scores.

	GeneTarget	OTScore	ModulationScore	TissueEfficacyScore	HS	MAX
5468	PPARG	1	0.9851352	0.6542443	1.318978	1
3643	INSR	1	1.0000000	0.6189753	1.318775	1
208	AKT2	1	1.0000000	0.6138788	1.318209	1
79068	FTO	1	0.6372124	0.9868802	1.317521	1
5465	PPARA	1	0.9795581	0.6051955	1.312134	1
7421	VDR	1	0.9642260	0.5808759	1.305598	1

6. Visualize selected gene targets in tissue-specific networks and related biological pathways

An important step of the tissue-specific-efficacy scoring method is the calculation of the shortest paths connecting a putative gene-target to known disease-genes. These pathways, which are compiled for each disease-relevant tissue, could provide further information about why an identified top gene-target is relevant for a given disease. In this regard, the ThETA package provides utility functions (i) to build tissue-specific network involving a pre-selected set of gene-targets and to identify overrepresented biological annotations/pathways.

The following example shows how to use the function *build_tissue_specific_networks* which returns

- tissue-specific networks (*igraph objects*);
- shortest-paths linking a set of gene targets (e.g. top 5 from the tissue-specific efficacy score) to known disease-genes;
- a list of genes closely related to the set of the specified gene targets.

```
tsrwr = build.tissue.specific.networks(tissue_scores = T2DM_Tscores,
                                       disease_genes = T2DM_genes$entrez,
                                       ppi_network = ppi_strdb_700,
                                       tissue_expr_data = gtexv7_zscore,
                                       top_targets = rownames(T2DM_top50)[1:5])
```

Then, ThETA provides functions to compile

- biological annotations which are significantly associated with a set of genes (by using over-representation analysis);
- plots for interpreting the ORA analysis;
- pubmed trend plots based on a set of gene targets.

```
library(org.Hs.eg.db)
T2D_ora_data_shp = generate.ora.data(tsrwr$shp[[1]], databases = "KEGG")
T2D_ora_data_rwr = generate.ora.data(tsrwr$rwr, databases = "KEGG")
```

```
T2D_ora_plot_rwr = generate.ora.plots(T2D_ora_data_rwr,
                                      set_plots = c("dotplot","cnetplot"),
                                      showCategory = 5, font_size = 10)
```

```
figure <- ggpubr::ggarrange(plotlist = T2D_ora_plot_rwr[1:2], nrow = 2, ncol = 1,
                            common.legend = TRUE, legend = "bottom",
                            labels=names(T2D_ora_plot_rwr)[1:2])
figure
```

```
library(org.Hs.eg.db)
pmc_genes = as.character(AnnotationDbi::mapIds(org.Hs.eg.db,
rownames(T2DM_allsc)[c(1:5)], 'SYMBOL', 'ENTREZID'))
print(pmc_genes)
T2D_pmd_plot_top = novelty.plots(pmc_genes, font_size = 14, pubmed = c(2010,2018))
T2D_pmd_plot_top
```

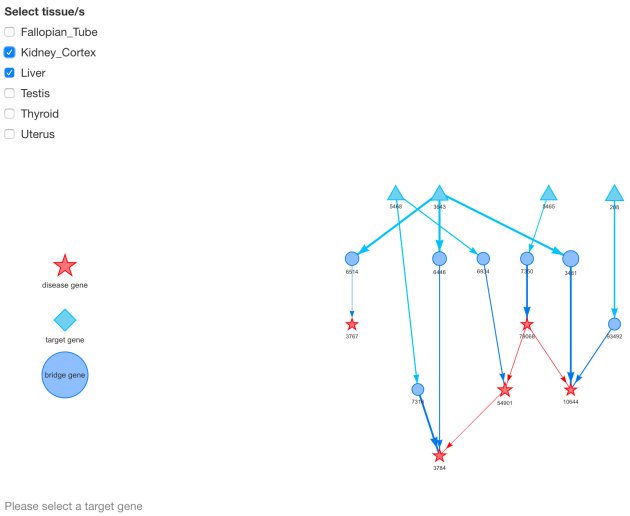
Visualize selected enrichment results of biological pathways linked to selected gene targets

How to visualize tissue-specific networks and biological annotations of selected drug(gene) targets

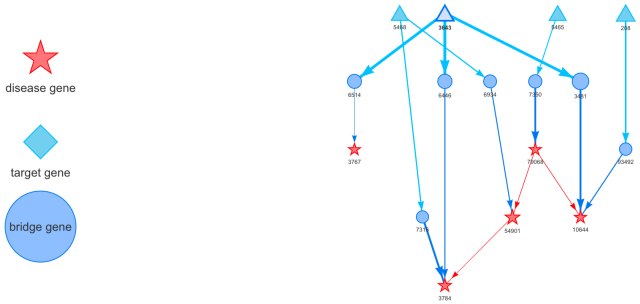
An R-shiny-based application was built for the visualization of tissue-specific gene networks highlighting connections between disease-genes and drug-targets/genes.

```
library(shiny)
library(visNetwork)
library(org.Hs.eg.db)

visualize.graph(tissue_scores = T2DM_Tscores,
                disease_genes = T2DM_genes$entrez[1:5],
                ppi_network = ppi_strdb_700,
                tissue_expr_data = gtexv7_zscore,
                top_targets = rownames(T2DM_allsc)[1:5],
                db='BP')
```



The over representation analysis of biological annotations is compiled over the genes included in the shortest patch connecting the selected target to the disease-genes.



ID	Description	GeneRatio	BgRatio	pvalue	p.adjust	qvalue	geneID	Count
GO:0022600	digestive system process	3/7	94/17653	0.00	0.00	0.00	6514/6446/3784	3
GO:0060453	regulation of gastric acid secretion	2/7	11/17653	0.00	0.00	0.00	6446/3784	2
GO:0045725	positive regulation of glycogen biosynthetic process	2/7	15/17653	0.00	0.00	0.00	3643/3481	2
GO:0007586	digestion	3/7	134/17653	0.00	0.00	0.00	6514/6446/3784	3
GO:0070293	renal absorption	2/7	16/17653	0.00	0.00	0.00	6446/3784	2

Bibliography

Chen, X., Ji, Z.L. and Chen, Y.Z. 2002. TTD: therapeutic target database. *Nucleic Acids Research* 30(1), pp. 412–415.

Davis, A.P., Grondin, C.J., Johnson, R.J., et al. 2017. The Comparative Toxicogenomics Database: update 2017. *Nucleic Acids Research* 45(D1), pp. D972–D978.

Failli, M., Paananen, J. and Fortino, V. 2019. Prioritizing target-disease associations with novel safety and efficacy scoring methods. *Scientific Reports* 9(1), p. 9852.

Wishart, D.S., Feunang, Y.D., Guo, A.C., et al. 2018. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Research* 46(D1), pp. D1074–D1082.