# *Simplified from* [*https://github.com/PathwayAndDataAnalysis/causalpath/blob/master/wiki/InputFormat.md*](https://github.com/PathwayAndDataAnalysis/causalpath/blob/master/wiki/InputFormat.md)

# Inputs and their format for CausalPath

Users need to prepare a minimum of 2 text files to run a CausalPath analysis.

1. Proteomics data
2. Analysis parameters

An example set of input files is available [here](https://github.com/PathwayAndDataAnalysis/causalpath/raw/master/wiki/sample-data-and-parameters.zip).

## Proteomics data

*CausalPath reads the proteomics dataset from a* ***tab-delimited text file*** *where the first row contains the* ***column headers****. Below are description of those columns.*

ID: A **unique text identifier** for each row – ideally: gene symbols and modification sites – used for visualization – should not contain "|" and "@"

Symbols: **HGNC symbol** of the related gene; if > 1 : single space separation

Sites: If **phosphoprotein** **measurement**, should indicate **protein sites** that are affected (1 letter capital aminoacid code, following an integer for the location on the UniProt caconical sequence); if more than 1: pipe (|) separation; if > 1 gene symbole in the row: site group for each symbol, single space separated

Effect: If **phosphoprotein measurement**, this column can contain the **effect of the related phosphorylation on the protein activity** (a for activating, i for inhibiting); if effect too complex for a simple: leave blank (preferred) or use c (but c will remove this row from possible causes; whereas blank CausalPath looks it up from its database)

Value: numeric value; can be > 1 columns, each with unique name; may be normalized protein reads, comparison values like fold changes (nature of values to be specified in the parameters file).

## Parameters file

***Name*** *of the file should be parameters.txt exactly. Each parameter in a* ***separate line****, in the* ***format*** *parameter-name = parameter-value.*

proteomics-values-file: **Name of the proteomics values file**. It should have **at least one ID column and one or more columns for experiment values**. Platform file and values file can be the same file.

proteomics-repeat-values-file: **Name of the proteomics values file for the repeated experiment**, if exists; exact same columns with the original proteomic values file, in the same order. Row IDs have to match with the corresponding row in the original data.

proteomics-platform-file: **Name of the proteomics platform file**. Each row should belong to either a gene's total protein measurement, or a site specific measurement. This file should contain **ID, gene symbols, modification sites, and known site effects**. Platform file and values file can be the same file.

id-column: **Name of the ID column** in platform and values files.

symbols-column: **Name of the symbols column** in platform file.

sites-column: **Name of the sites** column.

effect-column: **Name of the effect** column.

value-transformation: How to **use the values in the proteomics file**; no effect on a single value (1 value column)

* arithmetic-mean
* geometric-mean
* max
* difference-of-means
* fold-change-of-mean
* significant-change-of-mean
* significant-change-of-mean-paired
* signed-p-values
* correlation

value-column: **Name of a value column** (to use when only 1 group of experiments to consider)

control-value-column: **Name of a control value column**. (to use when control and test value columns)

test-value-column: **Name of a test value column** (to use when there are control and test value columns)

do-log-transform: Whether **log transform** ('true' or 'false'; default: false)

rna-expression-file: Name of the **RNA expression file** (tab-delimited text file 1st row = sample names, every other row = a gene, 1st column = gene symbol)

threshold-for-data-significance: **Threshold value for selecting significant data** (to use only **when FDR controlling procedure already performed outside of CausalPath**; can be set for each different data type separately; form: 'thr-val data-type', e,g, '1 phosphoprotein' or '2 protein’)

fdr-threshold-for-data-significance: **FDR for data significance**; can be set for each different data type separately; form: 'fdr-val data-type', e.g. '0.1 phosphoprotein' or '0.05 protein’

pool-proteomics-for-fdr-adjustment: Whether **consider proteomic and phosphoproteomic data as a single dataset during FDR adjustment** ('true' (typically with RPPA data) or 'false' (typically with MS data); default: false)

correlation-value-threshold: Option to **control correlation** with its value (cannot be used with FDR control, but can be used with p-value control)

correlation-upper-threshold: For **filtering correlations** with an upper value (in some types of proteomic data, highest correlations come from errors)

pval-threshold-for-correlation: **P-value threshold for correlation** in a correlation-based causality (to use only **when FDR control performed outside of CausalPath**)

fdr-threshold-for-correlation: **FDR threshold for the correlations** in a correlation-based analysis

stdev-threshold-for-data: **Standard deviation threshold** that can be set for each different data type separately; form: 'stdev-thr data-type', e.g. '0.5 phosphoprotein'

default-missing-value: Option to specify a **default value for the missing values** in the proteomics file

minimum-sample-size: Set **minimum sample size of the control and test sets**; in case there are missing values in proteomic file, the comparisons can have different sample sizes for controls and tests

calculate-network-significance: Whether to **calculate significances of the properties of the graph** (‘true’: a p-value for network size, and also downstream activity enrichment p-values for each gene on the graph are calculated); ignored by webserver, network significance calculated only when using CausalPath locally

permutations-for-significance: **Number of randomizations** performed when testing if the result network is large, or any protein's downstream is enriched; should be reasonably but not too high (e.g, 1000)

fdr-threshold-for-network-significance: **FDR for network significance** calculations for the downstream activity enrichment of genes

use-network-significance-for-causal-reasoning: After calculation of network significances in a non-correlation-based analysis, introduces the **detected active and inactive proteins as data to be used in the analysis** (applies only to the proteins that already have a changed data on them, and have no previous activity data associated)

prioritize-activity-data: When there is an ActivityData associated to a protein (can be user hypothesis or inferred by network significance), **do not use other omic data for evidence of activity change in causal reasoning**

minimum-potential-targets-to-consider-for-downstream-significance: **Exclude genes with few qualifying targets** on the network while calculating downstream significance for each source gene (might not be significant, and BH procedure to control FDR is used); default: 5

do-site-matching: Whether to **force site matching** in causality analysis; default: true

site-match-proximity-threshold: **Allow** **the phosphorylation relation to apply to close-by sites** when change observed in a site of the target protein which is not targeted by the relation, but very close to a known target site

site-effect-proximity-threshold: **Allow** **changing sites with unknown effect to have the same effect with the neighbor site** with known effect. CausalPath has a database of phosphorylation site effects. Default: 0 (=exact usage of site effects); should be used responsibly

built-in-network-resource-selection: **Which network resource** to use; for multiple resources: spaced or comma separated

* PC
* REACH
* PhosphoNetworks
* IPTMNet
* RHOGEF
* PCTCGAConsensus
* TRRUST
* TfactS
* NetworKIN

relation-filter-type: To **limit the results with a specific type of relation**

* no-filter
* phospho-only
* expression-only
* without-expression
* phospho-primary-expression-secondary

gene-focus: To **crop the result network to the neighborhood of certain gene** (gene symbols provided in a row separated by a semicolon, e.g. 'MTOR;RPS6KB1;RPS6')

mutation-effect-file: **Mutation effects** can be provided using this parameter (otherwise all mutations assumed to be inactivating)

color-saturation-value: Specifies the **value where node colors reach most intense color** (should be positive value, used symmetrically); if value-transformation = significant-change-of-mean, value is -log(p) with a sign associated to it

show-all-genes-with-proteomic-data: Put the genes with significant changes that could not make into the network of the result put as **disconnected nodes when the analysis is not correlation based**; default: true.

show-insignificant-data: **Make the insignificant protein data visible** on the result graph (might be better to turn off when too much)

hide-data-not-part-of-causal-relations: **Only data that are in the identified causal relations** are drawn on the result graph

data-type-for-expressional-targets: Control possible **other data types explainable by expressional relations** (e.g. explain RNA changes with expressional relations (more direct explanation than total protein measurement)); typical values: 'rna' and 'protein'; can be used multiple times to use rna and protein data together; default: explanations generated only for proteomic changes

generate-data-centric-graph: Generate **data-centric graph** as result as well (nodes are not the genes (gene-centric) but the data); default: false

gene-activity: Assign a **specific activity or inactivity to a gene** in the analysis.; value should start with gene name and 1 letter code for activity or inactivity,e.g. 'BRAF a', or 'PTEN i'.

tf-activity-file: Name of the file to **input results from an inference for transcriptional factor activities** (e.g. PRECEPTS, MARINa or VIPER); file with 1st column=TF symbol, 2nd column='activated' or 'inhibited’

use-strongest-proteomic-data-per-gene: Use **only the strongest proteomic feature at the upstream** of relations (might be used in case a proteomic experiment outputs too many phosphorylation sites with lots of changes → many proteins have evidences for both activation and inhibition → hard to read networks); default: false

use-missing-proteomic-data-for-test: Use a **G-test to check unequal distribution of missing values;** if opted, and sufficient data, G-test result combined with t-test result with Fisher's method (warning: assumes missing values uniformly distributed to samples; if violated, false positives will appear; set ‘false’ if not sure)

randomized-matrix-directory-for-missing-proteomic-data: To provide use **pre-generated matrices** to compute significances when using randomization as alternative to G-test for interpreting missing data distribution (CausalPath cannot generate those matrices; operation that typically requires a lot of memory)

missing-value-test-data-sufficiency-threshold: **P-value threshold sufficiency test for G-test**; when G-test only used for proteomic row with sufficient data; test for sufficiency: generate an extreme case where missing data is shifted to the smaller group and see if this can provide a p-value small enough; if this extreme shifting cannot make the the p-value small enough (specified with this threshold), then G-test not used for that row

custom-resource-directory: To **customize the resource directory** which stores what CausalPath downloads in the first run; default: .panda

tcga-directory: **add genomic data from TCGA** (only useful when proteomic data have same sample IDs; TCGA data can be loaded into a local directory from Broad Firehose, and the directory provided here; e.g. org.panda.resource.tcga.BroadDownloader in <https://github.com/PathwayAndDataAnalysis/resource> can do that)

hgnc-file: For reproducibility: **HGNC resource file** (for reproducibility)

custom-causal-priors-file: **Custom file for causal priors** (for reproducibility)

custom-site-effects-file: **Custom file for site effects** (for reproducibility)

use-expression-for-activity-evidence: **Testing if RNA expression is a good proxy for protein activity**