



# HHS Public Access

Author manuscript

*Curr Protoc.* Author manuscript; available in PMC 2024 January 15.

Published in final edited form as:

*Curr Protoc.* 2022 January ; 2(1): e355. doi:10.1002/cpz1.355.

## Getting Started with the IDG KMC Datasets and Tools

Eryk Kropiwnicki<sup>1,†</sup>, Jessica Binder<sup>2,†</sup>, Jeremy Yang<sup>2</sup>, Jayme Holmes<sup>2</sup>, Alexander Lachmann<sup>1</sup>, Daniel J. B. Clarke<sup>1</sup>, Timothy Sheils<sup>3</sup>, Keith Kelleher<sup>3</sup>, Vincent Metzger<sup>2</sup>, Cristian G. Bologa<sup>2</sup>, Tudor I. Oprea<sup>2,\*</sup>, Avi Ma'ayan<sup>1,\*</sup>

<sup>1</sup>Department of Pharmacological Sciences, Mount Sinai Center for Bioinformatics, Icahn School of Medicine at Mount Sinai, One Gustave L. Levy Place, Box 1603, New York, NY 10029, USA

<sup>2</sup>Translational Informatics Division, Department of Internal Medicine, University of New Mexico, Albuquerque, NM, USA.

<sup>3</sup>National Center for Advancing Translational Science, 9800 Medical Center Drive, Rockville, MD 20850, USA

### Abstract

The Illuminating the Druggable Genome (IDG) consortium is a National Institutes of Health (NIH) Common Fund program designed to enhance our knowledge of understudied proteins. More specifically, proteins unannotated within the three most commonly drug-targeted protein families: G-protein coupled receptors, ion channels, and protein kinases. Since 2014, the IDG Knowledge Management Center (IDG-KMC) has generated several open-access datasets and resources that jointly serve as a highly translational machine learning ready knowledgebase focused on human protein-coding genes and their products. The goal of the IDG-KMC is to develop comprehensive integrated knowledge for the druggable genome to illuminate the uncharacterized or poorly annotated portion of the druggable genome. The tools derived from the IDG-KMC provide either user-friendly visualizations or ways to impute the knowledge about potential targets using machine learning strategies. In the following protocols, we describe how to use each web-based tool for researchers to accelerate illumination in understudied proteins.

**Basic Protocol 1:** Interacting with the Pharos user interface

---

\*Corresponding authors: [toprea@salud.unm.edu](mailto:toprea@salud.unm.edu), [avi.maayan@mssm.edu](mailto:avi.maayan@mssm.edu).

†Equal contribution

Author Contributions

Eryk Kropiwnicki: Software, Data curation, Writing-original draft, Writing-review & editing.

Jessica Binder: Writing-original draft, Writing-review & editing.

Jeremy Yang: Data curation, Software, Visualization.

Daniel J. B. Clarke: Methodology, Software, Validation, Visualization.

Jayme Holmes: Data curation, Resources, Software.

Alexander Lachmann: Formal analysis, Investigation, Methodology, Software, Validation.

Vincent Metzger: Software, Writing-review & editing.

Timothy Sheils: Methodology, Software, Validation, Visualization.

Keith Kelleher: Data curation, Software.

Cristian Bologa: Data curation, Methodology, Software, Writing-review & editing.

Tudor Oprea: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing-original draft, Writing-review & editing

Avi Ma'ayan: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing-original draft, Writing-review & editing

Conflict of Interest Statement

The authors declare no conflict of interest.

**Basic Protocol 2:** Accessing the data in Harmonizome

**Basic Protocol 3:** The ARCHS4 resource

**Basic Protocol 4:** Making predictions about gene function with PrismExp

**Basic Protocol 5:** Using Geneshot to illuminate knowledge about under-studied targets

**Basic Protocol 6:** Exploring understudied targets with TIN-X

**Basic Protocol 7:** Interacting with the DrugCentral user interface

**Basic Protocol 8:** Estimating Anti-SARS-CoV-2 activities with DrugCentral REDIAL-2020

**Basic Protocol 9:** Drug Set Enrichment Analysis using Drugmonizome

**Basic Protocol 10:** The Drugmonizome-ML Appyter

**Basic Protocol 11:** The Harmonizome-ML Appyter

**Basic Protocol 12:** GWAS target illumination with TIGA

**Basic Protocol 13:** Prioritizing kinases for lists of proteins and phosphoproteins with KEA3

**Basic Protocol 14:** Converting PubMed searches to drug sets with the DrugShot Appyter

## Keywords

bioinformatics; druggable genome; drug targets; disease ontology; drug discovery; data visualization; web applications

## INTRODUCTION

There are approximately 25,000 protein-coding genes (Venter et al., 2001) in the human genome. Abnormal protein expression is associated with many human diseases, which makes proteins critical targets for therapeutic agents. Approximately 15% of protein-coding genes are considered part of the “druggable genome”. This means that these proteins can modulate cellular behavior when targeted by experimental small molecule compounds (Hopkins and Groom, 2002; Lipinski et al., 2001; Russ and Lampel, 2005; Johns et al., 2012). Moreover, only a few hundred targets represent the existing clinical pharmacopeia, leaving a massive swath of pharmacology that remains unexploited. Therefore, 85% of druggable proteins remain to be explored as potential therapeutic targets. Much of the druggable genome encodes three critical protein families: non-olfactory G-protein-coupled receptors (GPCRs), ion channels, and protein kinases. Critically, we currently lack crucial knowledge about the function of many proteins from these families and their roles in health and disease. A better understanding of these proteins, structurally or functionally, could shed light on new avenues of investigation for basic science and therapeutic discovery (Oprea et al., 2018).

In this article, we provide several protocols to guide users through the use of IDG tools that accomplish specific computational tasks related to illuminating the druggable genome. In Basic Protocol 1, we describe how users can query the Pharos web interface (Sheils



et al., 2021) to search for data related to gene targets. Basic Protocol 2 explains how to use Harmonizome (Rouillard et al., 2016), a web application that stores gene-attribute associations from various sources that can be readily visualized and leveraged for machine learning. Basic Protocol 3 describes ARCHS4 (Lachmann et al., 2018), a web application that provides easy access to RNA-sequencing data from human and mouse experiments and also includes gene landing pages for all human genes with gene function predictions based on mRNA co-expression. Basic Protocol 4 describes PrismEXP (Lachmann et al., 2021), a machine learning Appyter (Clarke et al., 2021) that improves gene function predictions from gene co-expression correlation data by sharding the global gene-gene co-expression matrix used by ARCHS4. Basic Protocol 5 teaches the user how to use Geneshot (Lachmann et al., 2019), a web application that facilitates querying of biomedical search terms to retrieve prioritized lists of genes related to the search terms. In Basic Protocol 6 we introduce TIN-X (Cannon et al., 2017), the Target Importance and Novelty eXplorer. We demonstrate to users how to query and explore interesting disease-target associations based on novelty and importance metrics derived from natural language processing (NLP) of PubMed abstracts. Basic Protocol 7 describes DrugCentral (Avram et al., 2021), a comprehensive database of approved drugs that includes information relating to drug side effects, mode of action, indications, pharmacologic action, and other information. Basic Protocol 8 explains REDIAL-2020 (Kc et al., 2021), an ensemble machine learning platform that extends the information available in DrugCentral to predict drugs and small molecules that may have anti-SARS-CoV-2 activity. In Basic Protocol 9 we discuss Drugmonizome (Kropiwnicki et al., 2021), a web application that facilitates drug set enrichment analysis and allows users to submit a drug set of interest to retrieve enriched terms that all, or most, of the members of the input set share. Basic Protocol 10 describes Drugmonizome-ML (Kropiwnicki et al., 2021), an Appyter that extends the information available in Drugmonizome to build on-the-fly machine learning models for predicting novel drug and small molecule attributes. In a similar vein, Basic Protocol 11 discusses Harmonizome-ML, an Appyter that enables users to utilize the datasets from Harmonizome to build machine learning models that predict novel gene-attribute associations. Basic Protocol 12 includes a discussion of TIGA (Yang et al., 2021), Target Illumination GWAS Analytics, a tool that summarizes gene-trait associations derived from genome wide association studies (GWAS) with rational and intuitive evidence metrics. In Basic Protocol 13 we describe how users can submit an input list of genes or differentially phosphorylated proteins to KEA3 for kinase enrichment analysis (Kuleshov et al., 2021) to infer kinases associated with the input list. Basic Protocol 14 explains how to use DrugShot, an Appyter that allows for the querying of biomedical search terms to retrieve known and predicted lists of drugs and small molecules related to the query term.

### **Basic Protocol 1: Interfacing with the Pharos user interface**

Pharos is the user interface to the Knowledge Management Center (KMC) for the IDG program, providing facile access to most data types collected by the KMC (Nguyen et al., 2017; Sheils et al., 2020). Given the complexity of the data surrounding any target, efficient and intuitive visualization has been a high priority for users to navigate and summarize search results and rapidly identify patterns. Underlying the interface is a GraphQL API that

provides programmatic access to all KMC data, enabling the incorporation of IDG resources with other applications.

## Necessary Resources

### Hardware

- Desktop or a laptop computer, or a mobile device, with a fast Internet connection

### Software

- An up-to-date web browser such as Google Chrome (<https://www.google.com/chrome/>), Mozilla Firefox (<https://www.mozilla.org/en-US/firefox/>), Apple Safari (<https://www.apple.com/safari/>), or Microsoft Edge (<https://www.microsoft.com/en-us/edge>).

## Protocol steps and annotations

### Search Targets

1. Navigate to Pharos (<https://pharos.nih.gov>).
2. To search for a target, click on the search box on the main page or in the top left corner of subsequent pages. Enter 'STAT3'. Note that multiple search types are available in the dropdown menu. (Figure 1)
3. It is possible to search by pathway or view a list of diseases or ligands associated with a target. Additionally, pressing enter or return will allow a text-based search, which will return a list of results featuring 'stat3' anywhere in the text.
4. Press 'enter', 'return' or click the magnifying glass icon to search for the 'stat3' text string.
5. A list of 81 targets is returned, with 'STAT3' being at the top of the list. The rest of the targets will have the phrase 'stat3' somewhere within the target details. (Figure 2)
6. Click on the STAT3 card to view the target details.

### View target details

7. Follow the steps from above, or alternatively, click on the STAT3 (Target) option from the search box auto-complete. This will navigate directly to the STAT3 target details page.
8. The target details page is divided into several sections that highlight an area of knowledge about the target.
9. Scroll down to the "Protein Summary" section. A brief description of the target, as well as several identifiers is available. In addition, the central radar plot charts the relative knowledge of a target compared to the rest of TCRD on a 0 to 1 scale. This data is sourced from the Harmonizome, which will be discussed further (Figure 3).

10. Scroll down to the next section, “IDG Development Level Summary”. Displayed here is the current development level . Each level has the criteria listed, as well as links to the data for each property (Figure 4).
11. On the left side panel, click on “Disease Associations by Source”. This will navigate within the page to a section displaying disease associations from a variety of sources.
12. Scroll down to the “Disease Novelty (Tin-x)” section, just below Disease Associations. A scatterplot is visible that shows Tin-x data. This data is explained in Protocol 6. Briefly, it is natural language processed PubMed abstracts that chart a target’s importance to a disease, as well as the novelty of that target to the disease. A dense chart indicates a large amount of knowledge about a target and its disease associations, whereas a sparser chart would indicate that target is not frequently studied and has fewer disease associations (Figure 5).
13. Scroll down to the next section “GWAS Traits”. Here a table of GWAS traits is displayed. This list focuses on scoring and ranking protein-coding genes associated with traits from genome-wide association studies. This allows both the discovery of traits most associated with a target, but also lesser emphasized traits (Figure 6).

#### **Finding a list of Understudied targets that share disease associations with STAT3**

14. From the STAT3 target details page, click on “Disease Associations by Source” on the left panel.
15. Click on the “Find Similar Targets” button, directly under the panel header (Figure 7).
16. The targets list page is now shown, with a target similarity filter applied, showing 17,876 targets (Figure 8).
17. To refine this list for targets of interest to the IDG program (mentioned in Protocol 1), click on the “Refined (2020)” checkbox in the IDG Target Lists filter panel on the left side of the page. The list of targets shown is reduced to 290.
18. To find only dark targets in this list, click the “Tdark” value in the Target Development Level filter panel, returning 48 targets (Figure 9).
19. Click on the “click for details...” text on the TMEM63A target card to view a list of associated diseases that this target shares with STAT3 (Figure 10).

#### **Download target list**

20. Click on the downward facing arrow on the right side of the Targets header (Figure 11).
21. A window will pop open displaying a list of fields that can be selected (Figure 12).

22. Click on the Associated Diseases checkbox. Note that many fields are deactivated, to reduce the overall file size.
23. Click on Name and Target Development Level under the Single Value Fields heading.
24. Click the Run Download Query Button. A file download dialog will open. Depending on the complexity of the target list and the fields selected, it may take some time.
25. After the file is downloaded, this list of targets can be used as a starting point for many of the protocols listed below.

### GraphQL queries

26. Click on API on the main Pharos header.
27. A code “sandbox” is now visible, allowing testing of GraphQL queries to fetch complex data from Pharos. A distinct feature of GraphQL is the ability of the consumer to determine the exact fields returned from the query, as opposed to a SQL query, where the data returned is determined by the database developer.
28. Click the “Play” button in the top center to run a sample query. A list of Drugs associated with DRD2 is returned.
29. Click on the “Docs” tab on the right side of the page. A menu will open up that displays the queries available, the inputs required, and the responses and properties returned. Click on the “Docs” tab again to close the menu.
30. Replace the text in the left column with this query:

```
query PaginateData {
  batch(
    filter: {
      facets: [
        { facet: "Target Development Level", values: ["Tdark"] }
        { facet: "IDG Target Lists", values: ["Refined (2020)"] }
      ]
      similarity: "(P40763, Associated Disease)"
    }
  ) {
    results: targetResult {
      count
      targets(skip: 0, top: 100) {
        name
        gene: sym
        accession: uniprot
        idgTDL: tdl
        similarityDetails: similarity {
```

```

        commonOptions
      }
    }
  }
}

```

31. Press the play button. This query fetches all Dark targets of interest to the IDG that share associated diseases with STAT3. Returned is the target name, gene symbol, Uniprot id, IDG TDL, and shared associated diseases (Figure 13).

### Entire Relational Database Download Page

32. Navigate to the TCRD website (<http://juniper.health.unm.edu/tcrd/>).
33. Click on the “Downloads” tab on the navigation bar at the top of the page to be redirected to a table of downloadable files; ex: MySQL dump of the full TCRD (latest.sql.gz).

## Basic Protocol 2: Accessing the data in Harmonizome

The Harmonizome resource contains processed datasets detailing functional associations between genes/proteins and their attributes extracted from 66 online resources. The information from the original datasets was distilled into attribute tables that define significant associations between genes and their attributes, where attributes could be other genes, proteins, pathways, cell lines, tissues, experimental perturbations, diseases, phenotypes, drugs, or other entities depending on the dataset. The Harmonizome web application can be accessed from <https://maayanlab.cloud/Harmonizome/> (Rouillard et al., 2016).

### Necessary Resources

#### Hardware

- Desktop or a laptop computer, or a mobile device, with a fast Internet connection

#### Software

- An up-to-date web browser such as Google Chrome (<https://www.google.com/chrome/>), Mozilla Firefox (<https://www.mozilla.org/en-US/firefox/>), Apple Safari (<https://www.apple.com/safari/>), or Microsoft Edge (<https://www.microsoft.com/en-us/edge>).
- Text editor or development environment of choice, such as Visual Studio (<https://visualstudio.microsoft.com/vs/>); most updated version of Python
- (<https://www.python.org/downloads/>); Python module for Harmonizome (<https://maayanlab.cloud/Harmonizome/static/harmonizomeapi.py>)

## Protocol steps and annotations

### Metadata Search

1. Navigate to the Harmonizome website (<https://maayanlab.cloud/Harmonizome/>).
2. The front page features a search bar where keywords of interest can be input. Click the filter button on the left of the search bar to narrow searches to “genes”, “gene sets”, or “datasets” (Figure 14). Type “STAT3” into the search bar and click the submit button. The results page includes a single gene landing page for STAT3 and 75 gene sets with STAT3 as an attribute (Figure 15).
3. Click on the STAT3 “gene” result to be redirected to a single gene landing page (Figure 16). The page includes identifying metadata for the gene, download links for accessing functional associations between STAT3 and other attributes, and links to other gene-related information from ARCHS4 (Lachmann et al., 2018). Additionally, a list of functional associations for STAT3 from the various processed datasets included in Harmonizome is available (Figure 17). Click the “+” button to view associations for STAT3 for any of the datasets.
4. Click on any of the STAT3 “gene set” results. The gene set results page includes metadata for the STAT3 gene set, in this case the gene set includes all target genes of STAT3. All of the genes included in the gene set are found in the “Genes” section (Figure 18). Click on any of the gene symbols to be redirected to a single gene landing page.

### Download Page

1. Click on the “Download” section on the navigation bar at the top of the page to be redirected to a table of all the datasets included in Harmonizome (Figure 19).
2. Click on “Achilles” in the resource column to be redirected to a page with identifying metadata for the resource and a list of all datasets derived from the resource (Figure 20).
3. Click on “Cell Line Gene Essentiality Profiles” in the dataset column to be redirected to a page with identifying metadata for the dataset and links to downloadables contained within this dataset (Figure 21). Further down the page are links to visualizations of the dataset contents and a table of gene sets (Figure 22). Click on any of the gene set names to be redirected to a gene set specific page.

### Visualize

4. Click on the “Visualize” section on the navigation bar at the top of the page and a dropdown menu will appear (Figure 23).
5. Click on “Global Heat Map” within the dropdown menu to be redirected to an interactive clustergram that visualizes the appearances of each gene in Harmonizome. Select different gene classes with the buttons on the left. Switch the ordering of the clustergram between “cluster” and “rank” by clicking the corresponding button (Figure 24).



6. Click on “Dataset Heat Maps” or “Gene Similarity Heat Maps” or “Attribute Similarity Heat Maps” within the dropdown menu to be redirected to a page with a dropdown menu of Harmonizome datasets. Open the dropdown menu and select any dataset to generate a hierarchically clustered heat map visualization of the dataset (Figure 25).
7. Click on “Dataset Pair Heat Maps” within the dropdown menu to be redirected to a page with a dropdown menu of Harmonizome datasets. Open the dropdown menu and select a dataset. A second dropdown menu will appear for selecting a second dataset to compare. Click visualize to generate a hierarchically clustered heat map visualization of the two datasets (Figure 26).
8. Click on “Heat Map with Input Genes” within the dropdown menu to be redirected to a page with a dropdown menu of Harmonizome datasets and a gene list text box. Click the “Example input” button to populate the fields with an example dataset and gene set. Click “Submit” to generate a hierarchically clustered heat map visualization of the associations between the uploaded genes and biological entities in the dataset (Figure 27).

### Predict

9. Click on the “Predict” section on the navigation bar at the top of the page and a dropdown menu will appear (Figure 28). Click “Intro” within the dropdown menu.
10. The intro page contains information about how machine learning studies were devised using the Harmonizome datasets. A table with four separate case studies: “Ion Channel Predictions”, “Mouse Phenotype Predictions”, “GPCR-Ligand Interaction Predictions”, “Kinase-Substrate Interaction Predictions” contains links to view and download tables of predicted associations (Figure 29).

### Using the Harmonizome API

11. These are the entity types supported by the Harmonizome API:

DATASET, GENE, GENE\_SET, ATTRIBUTE, GENE\_FAMILY, NAMING\_AUTHORITY, PROTEIN, RESOURCE

Open a new or existing Python code file. Import the required Harmonizome API Python module at the top of the file:

```
from harmonizomeapi import Harmonizome, Entity
```

The Harmonizome object includes several methods to read, parse, and download data from the Harmonizome API. The Harmonizome object includes `.get().next()` and `.download()` methods. For example, to display the datasets available in Harmonizome run the following code block:

```
entity_list = Harmonizome.get(Entity.DATASET)
more = Harmonizome.next(entity_list)
```

In order to minimize database queries and request times, the Harmonizome API uses a technique called “cursoring” to paginate large result sets. Therefore, the first line in the above code block returns the first 100 datasets, whereas the second line continues from where the previous entity list left off and retrieves the subsequent 14 datasets that are available in Harmonizome. The `Harmonizome.get()` and `Harmonizome.next()` methods can be used for all entity types supported by the Harmonizome API.

12. To download datasets available in Harmonizome to a local directory use the `Harmonizome.download()` generator function. Alternatively `Harmonizome.download_df()` can be used to download files and load them in directly as sparse (with an added `sparse=True` argument) or dense Pandas DataFrames (assumed). The function takes a list of datasets and downloadables as arguments. Leaving the datasets argument empty will download all datasets by default. Leaving the what argument empty will download all downloadables for each dataset by default. In the example code below, the “gene\_attribute\_matrix.txt.gz” downloadable from the “CTD Gene-Chemical Interactions” dataset is downloaded, decompressed, and saved to a local directory named after the dataset if it hasn’t already been processed:

```
dl, = Harmonizome.download(datasets=['CTD Gene-Chemical Interactions'],
what=['gene_attribute_matrix.txt.gz'])
```

*More information regarding the Harmonizome API is available at <https://maayanlab.cloud/Harmonizome/documentation>.*

### Basic Protocol 3: The ARCHS4 Resource

ARCHS4 (Lachmann et al., 2018) is a web resource that provides access to published RNA-seq gene and transcript level data from human and mouse experiments. FASTQ files from RNA-seq experiments deposited in the Gene Expression Omnibus (GEO) were aligned using a cloud-based infrastructure. The ARCHS4 web interface facilitates the exploration of the processed data through querying tools, interactive visualizations, and single gene landing pages that provide average expression of a specific gene across cell lines and tissues, top co-expressed genes, and predicted biological functions and protein–protein interactions for each gene based on prior knowledge combined with co-expression.

#### Necessary Resources

##### Hardware

- Desktop or a laptop computer, or a mobile device, with a fast Internet connection

##### Software

- An up-to-date web browser such as Google Chrome (<https://www.google.com/chrome/>), Mozilla Firefox (<https://www.mozilla.org/en-US/>)

firefox/), Apple Safari (<https://www.apple.com/safari/>), or Microsoft Edge (<https://www.microsoft.com/en-us/edge>).

- Most updated version of R (<https://www.r-project.org/>); R Studio (<https://www.rstudio.com/>); rhdf5 library (<https://www.bioconductor.org/packages/release/bioc/html/rhdf5.html>)

## Protocol steps and annotations

### Metadata Search

1. Navigate to the ARCHS4 web application (<https://maayanlab.cloud/archs4/>).
2. Click the “Get Started” button on the homepage to proceed to the data search and visualization page (Figure 30).
3. The data search and visualization page by default shows an interactive 3D t-SNE scatter plot of all the human gene expression samples found in ARCHS4 (Figure 31). The metadata search field on the left enables querying of specific terms which will be highlighted in the 3D scatter plot. Searching for the term “Pancreatic Islet” and then clicking on the search button results in the highlighting of the relevant samples. The samples that are related to the search term cluster in the scatter plot because the samples contain similar expression profiles (Figure 32).
4. Any submitted search term will be found in its corresponding section within the “Search Result” table below the interactive t-SNE scatter plot visualization. The table contains metadata regarding the organism, number of samples, number of series, as well as a button to download an R script that can be used to retrieve the identified sample files. An X button is also available to delete the query (Figure 33).

### Signature Search

5. Switching to the signature search functionality can be done by clicking on the corresponding tab within the “Search” field on the left (Figure 34). The signature search uses a set of highly and lowly expressed genes from each sample to identify matching samples to the given input.
6. Query the example up and down gene sets by clicking “Try an example”. The corresponding samples are highlighted within the scatter plot and are added to the “Search Result” table (Figure 35). Note that the previous query of “Pancreatic Islet” is still visualized within the scatter plot and listed in the “Search Result” table.

### Enrichment Analysis

7. Switch to the enrichment search by clicking on the corresponding tab within the “Search” field on the left (Figure 36). The enrichment search highlights samples that are enriched in gene sets from eight gene set libraries. Select the gene set library, gene set of interest within the selected library, and a signature direction.

8. Query the example by clicking “Search enriched samples”. The corresponding samples are highlighted within the scatter plot and added to the “Search Result” table along with the previous queries (Figure 37).

### Gene-Centric Visualization

9. Switch to gene-centric searches by clicking on the orange button under the “Species” field in the upper left. Use this field to also switch between human and mouse samples by clicking the corresponding teal button (Figure 38).
10. The page will now contain an interactive t-SNE scatter plot where each point represents a gene instead of a sample (Figure 39).
11. Choose a gene set library and a gene set within the “Search” field on the left (Figure 40). Query the default options by clicking “Search genes”.
12. The corresponding samples are highlighted within the scatter plot and added to the “Search Result” table under the “Genes” section (Figure 41). The table includes the number of genes included in the queried gene set which can be clicked to view the gene symbols in the gene set (Figure 42). Additionally, the gene set can be submitted to Enrichr (Kuleshov et al., 2016) for gene set enrichment analysis by clicking on the Enrichr icon within the table (Figure 43).

### Gene Search

13. Single genes can be queried using the autocomplete field within the “Search” field on the left. Input a gene of interest, for example SOX2, and click the search button (Figure 44).
14. A single gene page is generated for SOX2 (Figure 45). The top of the page includes a description of the gene and links to other resources with identifying metadata for the gene. The “Functional Annotation Prediction” section contains ROC curves and tables of gene sets from six distinct gene set libraries SOX2 is predicted to be a member of based on co-expression. Known associations are marked in teal.
15. The “Most similar genes based on co-expression” section contains a table of the top 100 genes that are most similar to SOX2 based on the Pearson correlation of their expression across all ARCHS4 samples (Figure 46). The most correlated genes from the table can be submitted to Enrichr by clicking the corresponding link in the top right.
16. The “Tissue Expression” section contains a dendrogram of tissue types divided into organs and cell types. The average expression of SOX2 within a specific tissue or a cell type context is visualized as a collection of box plots (Figure 47).
17. The “Cell Line Expression” section contains a dendrogram of various cell lines organized by the tissue of origin. The plot visualizes the average expression of SOX2 across the cell lines based on data from ARCHS4 (Figure 48).

## Downloading Gene Expression Data from ARCHS4

18. As described in previous steps, after submitting a search within the data search and visualization page, the “Search Results” table includes a download link to an R script that can be used to retrieve the selected samples. Click the download icon to download the script.
19. Open R Studio and copy and paste the R script from the downloaded R file into R Studio.
20. Ensure that the “rhdf5” library is installed. Open the console in R Studio and input the following:

```
if (!requireNamespace("BiocManager", quietly = TRUE))
  install.packages("BiocManager")
BiocManager::install("rhdf5")
```

21. Now run the R script downloaded from ARCHS4 to produce an expression matrix for the selected samples that were returned from the search. The expression matrix can be used for further analysis, for example, it can be used to compute the average expression of a gene in a specific disease, cell line, or tissue contexts.

## Basic Protocol 4: Making predictions about gene function with PrismExp

PrismEXP is an Appyter (Clarke et al., 2021; Lachmann et al., 2021) that employs machine learning to predict gene function using gene-gene mRNA co-expression correlations from mRNA-sequencing (RNA-seq) data sourced from ARCHS4, a database composed of human and mouse RNA-seq sample gene counts from GEO (Lachmann et al., 2018). The difference between gene function predictions made by PrismExp and the gene function prediction available from the ARCHS4 website is that the ARCHS4 data is divided first into clusters and then gene-gene correlations are computed for each cluster. 51 correlation matrices are precomputed and stored in the cloud. At runtime, the correlation data is extracted from the cloud storage and a pretrained Random Forest model is applied on the correlation features to rank the level of association of a single gene to all gene sets from a user-specified gene set library.

### Necessary Resources

#### Hardware

- Desktop or a laptop computer, or a mobile device, with a fast Internet connection

#### Software

- An up-to-date web browser such as Google Chrome (<https://www.google.com/chrome/>), Mozilla Firefox (<https://www.mozilla.org/en-US/firefox/>), Apple Safari (<https://www.apple.com/safari/>), or Microsoft Edge (<https://www.microsoft.com/en-us/edge>).

## Protocol steps and annotations

### Navigating the Input Form

1. Navigate to the PrismEXP Appyter (<https://appyters.maayanlab.cloud/PrismEXP/>).
2. The Appyter input form includes a “Gene Selection” section with a field for inputting a gene symbol of interest for which novel functions will be predicted. Additionally, the “GMT Selection” section includes a field for selecting a GMT file from which predictions will be made (Figure 49). Click the “Upload” button within the “GMT Selection” section to upload a custom GMT file (Figure 50).
3. Click submit on the Appyter input form and a Jupyter Notebook with the input parameters will be launched in the cloud.

### Gene Function Predictions

4. A Jupyter Notebook will begin executing in the cloud once the input form is submitted. The notebook includes an option to download the notebook, toggle displaying the code, and running the notebook locally. Additionally, a table of contents exists with clickable elements that link to specific sections within the notebook (Figure 51).
5. Scroll down to the “Load Gene Correlation” section. The Dataframe displays genes that correlate with your query gene in 51 pre-computed correlation matrices from ARCHS4 (Figure 52).
6. Scroll down to the “Avg Correlation Scores” section. This Dataframe displays computed correlation scores to each of the gene set terms from the GMT file based on co-expression values between the query gene and each of the genes included in the gene set (Figure 53).
7. The average correlation score matrices are used as the input features for the PrismEXP model. Scroll down to the “Prediction Validation” section. The ROC curve displayed in this section characterizes how well the known annotations for this gene were recovered by the PrismEXP model (Figure 54).
8. Scroll down to the “Top Predictions” section. The Dataframe displays the top 20 gene set terms that the query gene is predicted to be associated with. The table displays the prediction score from the model, z-score, p-value, and Bonferroni corrected p-value (Figure 55).
9. Scroll down to the “Download Files” section. Click on the appropriate link to download the prediction table or ROC curve in .pdf or .png format (Figure 56).

### Basic Protocol 5: Using Geneshot to illuminate knowledge about under-studied targets

Geneshot is a search engine for querying biomedical terms to retrieve lists of genes most associated with the term from PubMed ID (PMID) co-mentions (Lachmann et al., 2019). To convert search terms to genes, Geneshot uses one of two resources: GeneRIF and AutoRIF. Both GeneRIF and AutoRIF are text files documenting gene-PubMed ID associations. These



associations are used to rank genes for a query term based on the number of co-mentions. Geneshot further prioritizes other related genes based on co-occurrence and co-expression matrices with the genes associated with the term from the literature. Additionally, Geneshot includes a gene function prediction feature that prioritizes novel gene set membership for a query gene based on co-occurrence or co-expression.

## Necessary Resources

### Hardware

- Desktop or a laptop computer, or a mobile device, with a fast Internet connection

### Software

- An up-to-date web browser such as Google Chrome (<https://www.google.com/chrome/>), Mozilla Firefox (<https://www.mozilla.org/en-US/firefox/>), Apple Safari (<https://www.apple.com/safari/>), or Microsoft Edge (<https://www.microsoft.com/en-us/edge>).
- Text editor or development environment of choice, such as Visual Studio (<https://visualstudio.microsoft.com/vs/>); most updated version of Python (<https://www.python.org/downloads/>); Python requests library (<https://requests.readthedocs.io/en/master/user/install/>)

## Protocol steps and annotations

### PubMed Query

1. Navigate to the Geneshot homepage (<https://maayanlab.cloud/geneshot/>).
2. The PubMed Query page includes an input form for submitting search terms (Figure 57). The top search bar is for terms that the search should include, whereas the lower search bar is for terms that should be omitted from the search. Toggle the size of the gene set that will be used to make further predictions with the “Top Associated Genes to Make Predictions” filter. Use the toggle bar to switch between AutoRIF and GeneRIF (Maglott et al., 2011) as the underlying databases for gene-PMID associations. Click “Wound Healing” in the example section of the input form to launch a search (Figure 58).
3. The first output from the search is a scatter plot of all genes associated with “wound healing” (Figure 59). The x-axis of the scatter plot displays the counts of Publications with Search Term, and the y-axis shows the fraction of Publications with Search Term / Total Publications. Hover over any point on this plot to display the gene name and its corresponding X and Y values.
4. Clicking on any of the points in the scatter plot generates a histogram displaying the association of the gene with the search terms based on literature co-mentions over time (Figure 60). The number of publications for the selected gene that do not match the search term is displayed as pink bars, while the number of publications matching the search term and the gene is displayed as blue bars.

5. Scroll down to view the tables of associated genes and predicted genes (Figure 61). The left table includes the top genes associated with “wound healing” ranked by number of PubMed ID co-mentions. The right table shows the top 200 genes predicted to be associated with “wound healing” based on co-expression with the top 20 genes from the associated table. Each of the tables include a row of buttons that, when clicked, filter the genes from each table into a specific gene family. Additionally, the genes from each table can be submitted to Enrichr for gene set enrichment analysis, and each table itself can be downloaded.
6. To recalculate the predictions, use the drop-down menu above the associated table to select a new gene-gene similarity matrix and increase or decrease the associated gene set size using the scroll bar. Click the “Recalculate Predictions” button to update the prediction table (Figure 62).

### Gene Function Predictions

7. Navigate to the Gene Function Prediction page by clicking the corresponding link within the navigation bar at the top of the page. This page includes an input form for selecting a gene of interest, Enrichr gene set library from which gene functions will be sourced from, and a gene-gene similarity matrix from which predictions will be calculated (Figure 63). By using functional prediction by association, the input gene can be predicted to be a member of gene sets. Click the example to launch a query.
8. A table of the top predicted functions and ROC curve of prediction performance are generated (Figure 64). Known associations within the table are highlighted in blue, whereas previously unknown associations are not highlighted. The table is available for download.

### Gene Set Augmentation

9. Navigate to the Gene Set Augmentation page by clicking the corresponding link within the navigation bar at the top of the page. The input form on this page includes a text box for pasting a gene set for augmentation, a drop-down menu of gene-gene similarity matrices from which predictions will be calculated, and a toggle bar for switching between GeneRIF and AutoRIF for retrieving publication counts for each gene (Figure 65).
10. Click on the “mixed genes” example to submit a query. The input genes are first sorted into quantiles based on their novelty in the literature (Figure 66).
11. Scroll to the bottom of the page where there is a table with the submitted genes on the left, and a table of genes predicted to be associated with the input genes based on the selected gene-gene similarity matrix, in this case ARCHS4 co-expression, on the right (Figure 67). The “user upload” table ranks the genes by the amount of PubMed abstracts they are mentioned in, along with their novelty. The predicted genes table ranks genes by their similarity score with the input gene set. Genes from both tables can be submitted to Enrichr for gene set enrichment analysis and each table can be downloaded.

## Geneshot API Example

- Open a new or existing Python code file. Import the JSON and requests libraries at the top of the file as follows.

```
import json
import requests
```

- Call the requests.post method to send a POST request to the URL. The payload variable contains the parameters that are sent to the API endpoint specified in GENESHOT\_URL. In this case the endpoint is /search and the parameters are rif, which specifies whether AutoRIF or GeneRIF is used as the association file, and term, which specifies the query term for the search.

```
GENESHOT_URL = 'https://maayanlab.cloud/geneshot/api/search'
payload = {"rif": "generif", "term": "hair loss"}

response = requests.post(GENESHOT_URL, json=payload)

data = json.loads(response.text)
print(data)
```

- Use the json.loads method to view the response as a JSON object containing all genes related to the query term.

```
{
  "PubMedID_count": 34412,
  "gene_count": {
    "ABCC6P2": [
      1,
      0.25
    ],
    "ABI3": [
      2,
      0.125
    ],
    ...
  },
  "query_time": 1.121943712234497,
  "return_size": 298,
  "search_term": "hair loss"
}
```

For more information on using the various Geneshot API endpoints, please refer to the API documentation (<https://maayanlab.cloud/geneshot/api.html>).

## Basic Protocol 6: Exploring understudied targets with TIN-X

TIN-X (Target Importance and Novelty eXplorer) (Cannon et al., 2017), is an informatics workflow, REST API, and web application used to identify, visualize, and explore protein-disease associations. TIN-X is based on text mining data processed from scientific literature. The TIN-X visualizations plot information for protein-disease associations along two axes, specifically “novelty” and “importance.” Briefly, Novelty is used to estimate the scarcity of publications about a protein target, whereas *Importance* estimates the strength of the association between that protein and a specific disease.

### Necessary Resources

#### Hardware

- Desktop or a laptop computer, or a mobile device, with a fast Internet connection

#### Software

- An up-to-date web browser such as Google Chrome (<https://www.google.com/chrome/>), Mozilla Firefox (<https://www.mozilla.org/en-US/firefox/>), Apple Safari (<https://www.apple.com/safari/>), or Microsoft Edge (<https://www.microsoft.com/en-us/edge>).

### Protocol steps and annotations

#### Browse Diseases

1. Navigate to the TIN-X web app (<https://www.newdrugtargets.org/>).
2. The default TIN-X mode, “Browse Diseases”, (upper-left) starts with the Disease Ontology (Schriml et al., 2019), (DO). The DO hierarchy can then be navigated using the left panel (Figure 68). Given this hierarchical nature, a larger number of target-disease associations can be text-mined from biomedical literature for higher-level terms (e.g., N=13405 for “nervous system disease”), as opposed to child terms (e.g., N=9733 for “neurodegenerative disease”, N=4587 for “Synucleinopathy,” N=4587 for “Parkinson’s Disease”) or for leaf terms (e.g., N=227 for “Early Onset Parkinson’s Disease”).
3. Searching by *disease name* is also supported. Targets with stronger associations (higher Importance) are in the upper part of the plot, while targets with a higher number of publications (lower Novelty) are located on the left side of the plot. Points situated in the upper-right area of the plot (if any) are most likely to be of interest, as they are located at the Pareto frontier, i.e., targets for which a large number of published papers mentioning that target also mention the selected disease.
4. Targets are colored by Target Development Levels, and can be filtered as such (Tclin/Tchem/Tbio/Tdark). They can also be filtered by protein superfamily (e.g. kinases). Upon selecting a protein, links to both Pharos and DrugCentral are provided for that protein (Figure 69); selecting the titles allows the user to

navigate through abstracts or to examine the document of interest in PubMed (additional clicks are required).

5. Once the desired level of granularity for diseases is reached, the user can examine target-disease associations, which are plotted along the Novelty-Importance axes in log-log format. To reach “Parkinson’s Disease”, one must click Disease of anatomical entity → Nervous System Disease → Neurodegenerative disease → Synucleinopathy → Parkinson’s Disease.
6. A highly-ranked gene associated with Parkinson’s Disease is “Synaptogyrin-3” (SYNGR3) and is classified as Tdark (Figure 69). While the exact function of SYNGR3 is unknown, there is recently published evidence that SYNGR3 encodes for a synaptic vesicle protein that interacts with a dopamine transporter (Egaña et al., 2009). The most novel association (lowest Importance) is for “Tripartite motif-containing protein 10” (TRIM10), which is supported by one genome-wide association study (Witoelar et al., 2017) focused on the overlap between Parkinson’s Disease and autoimmune diseases.
7. Both the “Browse Diseases” and the “Browse Targets” exploratory modes support an interactive way to manipulate the number of points displayed on the scatter plot. To change the number of plotted points, simply go to the top right side of the panel, where a vertical bar is placed between a “+” and a “-” sign. Sliding this bar up or down increases or decreases the number of visible points within the plot. By default, 300 or fewer points are plotted. Thresholds are defined by non-dominated solution (NDS) ranking, a.k.a. Pareto frontier, meaning that all hidden points are inferior to those visible in one or both variables.

### Browse Targets

8. From the upper left menu, “Browse Targets” can be selected. The Drug Target Ontology (Lin et al., 2017) hierarchy becomes visible, and can be navigated from the left panel (Figure 70). For each protein, Diseases are plotted with log–log Importance–Novelty axes and color-coded according to the top hierarchical Disease Ontology term (e.g., diseases of anatomical entity, diseases of metabolism, etc.).
9. Searching by *target name* is supported. Diseases with stronger associations (higher Importance) are in the upper part of the plot, while diseases with a higher number of publications (lower Novelty) are on the left side of the plot. Diseases that are likely of most interest are plotted in the upper-right area of the plot (Figure 71).
10. The plot, however, remains target-centric. Upon clicking on a point, the disease name and protein name are displayed, with appropriate links to Pharos and DrugCentral (Figure 72).
11. When selecting a target family (e.g., kinase), the user can drill down to the desired level of granularity, before examining disease associations for a specific

protein. Starting from Kinase, for example, the user must click Protein kinase → CAMK group → TRIO family → Kalirin, before diseases associated with Kalirin (KALRN) are displayed (Figure 70).

12. The top disease (highest Importance, lowest Novelty) associated with KALRN is “disease by infectious agent”, followed by “psychotic disorder”. We recommend repeated scrolling before identifying a leaf term corresponding to the Disease Ontology. For example, next to “psychotic disorder” is “schizophrenia” (a child term); this association is supported by 26 publications, including Miller et al. (Miller et al., 2017). The most novel association (lowest Importance) is for “X-linked nonsyndromic deafness” (Figure 72), supported by Cai et al. (Cai et al., 2014). This association is genuine, as the gene name (KALRN) is mentioned in the abstract, in relation to the rs333332 SNP.

### Sharing and downloading data

13. Whether in “Browse Diseases” or “Browse Targets” mode, the user can share data in two ways. First, for any given plot, the specific URL (universal resource locator) for that visualization can be copied and shared with third-party users. This can be done by clicking on the “Share” button. Second, the data can be exported (in comma-separated value format), and thus archived or post-processed with third-party software. Exported data includes Novelty and Importance scores, in addition to Disease names and identifiers in the “Browse Targets” mode, as well as Target names and identifiers in the “Browse Diseases” mode, respectively.

### Basic Protocol 7: Interacting with the DrugCentral user interface

DrugCentral is an online compendium (Ursu et al., 2017) centered on “active pharmaceutical ingredients” and their link to “pharmaceutical products”. DrugCentral distills relevant information from “pharmaceutical product” (or formulation) package inserts; while these are frequently referred to as “drugs” by patients and medical practitioners, herein we reserve the term “drugs” for “active pharmaceutical ingredients”. All data, including downloads, related to DrugCentral can be accessed at its designated web portal (<https://drugcentral.org/>). DrugCentral provides information on active ingredients, chemical entities, pharmaceutical products, drug mode of action, medical uses (indications, contra-indications and off-label uses), pharmacologic action, as well as adverse events (Ursu et al., 2019). As of 2021, DrugCentral (Avram et al., 2021) separately stores adverse events for women and men, and provides regulatory information extracted from the FDA Orange Book. DrugCentral is current (as of the date of the release) with regulatory approvals from the United States (US FDA), the European Union (EMA), Japan (PDMA) and, more recently, some drugs approved in China and Russia. Limited information on drugs that have been discontinued or withdrawn is available, particularly for drugs approved outside the US when package inserts and relevant information are not in English.



## Necessary Resources

### Hardware

- Desktop or a laptop computer, or a mobile device, with a 100 Mbps or higher (fast) Internet connection.

### Software

- An up-to-date web browser such as Google Chrome (<https://www.google.com/chrome/>), Mozilla Firefox (<https://www.mozilla.org/en-US/firefox/>), Apple Safari (<https://www.apple.com/safari/>), or Microsoft Edge (<https://www.microsoft.com/en-us/edge>).

## Protocol steps and annotations

### Queries Supported by DrugCentral

1. Navigate to the DrugCentral portal (<https://drugcentral.org/>).
2. The main DrugCentral search bar supports three types of queries: drug, target and disease. Each of these will filter and prioritize results according to a 4-level ranking system ordered from highest to lowest, as follows:
  - a. query term matching drug name or synonyms mechanism of action target, or drug indication (see below).
  - b. query term matching disease term in drug contraindications or off-label uses, targets listed in drug bioactivity profiles (not MoA targets), or pharmacologic action descriptions.
  - c. query term matching the short drug description text.
  - d. query term matching full text in the FDA drug labels processed from DailyMed (Figure 73).
3. For example, drug query results are sorted to display active ingredients first (e.g., omeprazole), followed by related ingredients (e.g., esomeprazole) and by other active ingredients that are co-formulated with the queried substance into pharmaceutical products. A query by brand name (e.g., prilosec) includes other antacids such as sodium bicarbonate, antibiotics such as amoxicillin and clarithromycin (co-prescribed with omeprazole to treat stomach ulcers caused by *Helicobacter pylori*) as well as acetyl-salicylic acid, which is combined with omeprazole for the prevention of stroke. (Figure 74)
4. Disease names are mappable to multiple terminologies such as Disease Ontology, MeSH, SNOMED-CT and MedDRA. Disease term queries first retrieve indications, followed by off-label and contra-indications, then other sections (e.g., side effects) that contain medical / disease terms. For example, the query “Parkinson’s disease” (PD) first lists drugs indicated for PD (e.g., ropinirole), followed by drugs indicated in complications of PD (e.g., fludrocortisone is indicated for the PD-associated orthostatic hypotension), then by drugs that list PD as side-effect (e.g., dimenhydrinate) (Figure 75).

5. Target name queries support input as text (e.g., “muscarinic m1”), gene symbol (CHRM1) or UniProt (P11229) and SwissProt (ACM1\_HUMAN) identifiers. It is recommended to use the exact target names adopted by UniProt, though gene/protein identifiers are preferred.

#### Queries Supported by DrugCentral: Redial

5. Given its basic science focus, the machine-learning based REDIAL-2020 platform (Kc et al., 2021), which is also part of DrugCentral, supports queries by drug name (e.g., omeprazole), by PubChem compound identifier (e.g., 4594) or by chemical structure in the SMILES (Weininger, 1988) format (e.g., COc1ccc2nc(S(=O)Cc3ncc(C)c(OC)c3C)[nH]c2c1). Regardless of format, all input queries for REDIAL-2020 are converted to SMILES format in order to predict anti-viral properties (Figure 76). *See also* Protocol nr 8.

#### Queries Supported by DrugCentral: L1000

6. The other search interface available in DrugCentral, implemented in R-Shiny (<https://shiny.rstudio.com/>) supports browsing and searching for drug names for which gene perturbation profiles were recorded across one more of the 81 cell lines collected during the LINCS (Library of Integrated Cellular Signatures) project. Based on the L1000 perturbation profiles for 1613 drugs, the L1000 DrugCentral app allows users to query (via drug names) which drugs have the most similar gene perturbation profiles, ranked by cell lines (Figure 77).

#### DrugCentral Drugcards: A step-by-step content guide

7. At its core, DrugCentral is a drug-centric resource. Thus, all queries are likely to provide information that is displayed in the form of “drug cards”. Data elements identified when searching a drug by name would be thus retrieved in a similar manner when searching by target or by disease, as both queries result in lists of drug cards.
8. Each drug card can be directly accessed (linked out) by observing the following (specific) format:  
<https://drugcentral.org/drugcard/<DrugcentralStruct.ID>>  
where “DrugcentralStruct.ID” is the DrugCentral structure ID number. For example, DrugcentralStruct.ID=824 resolves to dexamethasone. This manner of mining drug cards is not intended for casual users. Rather, this format is intended for programmatic access to DrugCentral content (Figure 78).
9. What follows is a “section by section” guide to drug card content, shown by section title. These are not intended as comprehensive explanations, but rather as brief illustrations of the diverse content available through DrugCentral.
10. “Stem definition” displays International Nonproprietary Names (INN), which are associated with “pharmacologically related groups”; that section also displays

Chemical Abstract Services (CAS) registry numbers, in addition to DrugCentral IDs.

11. “Description” depicts the two-dimensional chemical structure (as well as three separate chemical structure file formats), a number of synonyms and computed chemical descriptors such as Lipinski’s “rule of 5”. (Lipinski et al., 2001) The intellectual property / regulatory status of the drug (if available) is also shown under “Status”, with one of 3 options: OFP - off patent; OFM - off market; and ONP - on patent, respectively (Avram et al., 2020).
12. “Drug dosage” provides a sample (typically, the “maximum dose strength”) of the dosages available for oral / non-oral formulations of the drug.
13. “ADMET Properties” - Absorption, Distribution, Metabolism, Excretion and Toxicity - provides experimental ADMET values, when available. These properties are half-life, systemic clearance, volume of distribution at steady state and fraction unbound, all intravenous pharmacokinetic parameters (Lombardo et al., 2018); the fraction excreted unchanged in urine (extent of metabolism), water solubility and their composite parameter BDDCS, Biopharmaceutical Drug Disposition Classification System, as discussed elsewhere (Benet et al., 2011); and MRTD, the Maximum Recommended Therapeutic Daily Dose (Contrera et al., 2004).
14. “Approvals” shows the date of approval by regulatory agencies (if available).
15. “FDA adverse event reporting system (Female)”, followed by “FDA Adverse Event Reporting System (Male)” lists adverse events, separated by sex, in the decreasing order of the likelihood ratio (Huang et al., 2011).
16. “Pharmacologic action” highlights the drug annotations corresponding to (sometimes multiple) ATC (Anatomical, Therapeutic and Chemical) classification system codes - ATC codes are available at (WHOCC); chemical ontology information from ChEBI (EBI Web Team); FDA terminology; and MeSH (Medical Subject Headings) terms (MeSH Browser).
17. “Drug use” lists indications, off-label use and contra-indications, mapped to SNOMED-CT (Bhattacharyya, 2016) and DOID (Disease ontology - institute for genome sciences @ university of Maryland), where available. Drug indications and contra-indications are mined from package inserts (drug labels), whereas off-label uses are from literature.
18. “Acid dissociation constants calculated using MoKa v3.0.0” shows calculated acid/base dissociation constants, as calculated with the MoKa software (Milletti et al., 2010).
19. “Orange Book patent data (new drug applications)” and “Orange Book exclusivity data (new drug applications)” complement DrugCentral information on marketed pharmaceutical formulations by adding FDA Orange Book (Orange book: Approved drug products with therapeutic equivalence evaluations) for patents, as well as exclusivity data, for new drug applications.

20. “Bioactivity Summary” distills information from multiple bioactivity databases, e.g., ChEMBL (Mendez et al., 2019) and the IUPHAR Guide to Pharmacology (Armstrong et al., 2019), in addition to scientific literature and information from drug labels. Numeric information is converted to the negative log molar of the effective drug concentration at measurement. Mechanism-of-action drug targets (Santos et al., 2017) are marked separately.
21. The “External reference” section contains drug identifiers used by other on-line resources. This section includes identifiers used in medical practice, such as the Veterans Health Administration (e.g., VHA unique identifier, VUID), the National Drug File reference terminology (NDFRT, (National drug file - reference terminology source information, 2016) and RxNorm (RxNorm, 2004), as well as identifiers used by PubChem, ChEBI, DrugBank, etc.
22. Last but not least, the “Pharmaceutical products” section provides direct links to DailyMed (DailyMed, 2015), while incorporating simple meta-data descriptors such as “category” (e.g., prescription vs. over-the-counter), number of ingredients, administration route, etc. This section also includes a clickable container that captures the full text (no images) of the FDA approved package insert.

#### **DrugCentral Target Cards: A step-by-step content guide**

23. In Addition to DrugCentral’s Drugcards, a set of Target Cards can be directly accessed by observing the following (URL) syntax: <https://drugcentral.org/target/<UniprotAccession.ID>>
24. For example, <https://drugcentral.org/target/P23975/> resolves to Sodium-dependent noradrenaline transporter. This method of mining Target Cards is not intended for casual users. Rather, this format is intended for programmatic access to machine readable Target metadata (Figure 79).
25. What follows is a “section by section” guide to Target card content and target metadata.
26. “Description” depicts the Accession ,Swissprot, Organism, Gene & Target class followed by Drug relations where the Drugs Bioactivity mechanism-of-actions are identified and marked.
27. To retrieve all cross-referenced Drug Central Targetcards cards mapped to Uniprot Accession Ids use the following (machine readable) URL syntax (Figure 80): [https://drugcentral.org/static/Drugcentral\\_uniprot\\_Mapping.txt](https://drugcentral.org/static/Drugcentral_uniprot_Mapping.txt)

#### **Additional information**

28. The “Download Database dump 9/18/2020 (Postgres v10.12)” option contains all the information stored in DrugCentral. It requires a new or existing Postgres database setup. Users are directed to consult the [Postgresql documentation](#) on how to install, configure and load database contents. This is also available

via public instance at [drugcentral:unmtid-dbs.net](https://drugcentral.unmtid-dbs.net): 5433, username="drugman", password="dosage", with responsiveness depending on user load.

- 29.** Example queries to extract subsets of data from DrugCentral. Requires a local instance of DrugCentral loaded into a PostgreSQL database. To load the DrugCentral database dump assuming PostgreSQL is up and running and the user has admin privileges, run in PostgreSQL.

```
#create database drugcentral and then run using the OS shell
$gunzip -c drugcentral.dump.06212018.sql.gz | psql drugcentral
#Example 1: Select Off-patent drugs that bind to "Mast/stem
cell growth factor #receptor Kit" as mode-of-action target" in
DrugCentral's Postgres Db.
-select
  distinct(structures.name) as drug_name
  from
  structures
  join act_table_full on structures.id = act_table_full.struct_id
  Where
    structures.status = 'OFP' and
  act_table_full.moa = 1 and
  act_table_full.target_name = 'Mast/stem cell growth factor
receptor Kit'
#Example 2: Select drugs indicated for seasonal allergic rhinitis
that have #the lowest LLR for somnolence in males.
-select
  distinct(structures.name) as drug_name,
  faers_male.*
  from
  structures
  join struct2atc on structures.id = struct2atc.struct_id
  join atc on struct2atc.atc_code = atc.code
  join faers_male on structures.id=faers_male.struct_id
  Where
    atc.l2_name = 'ANTI-HISTAMINES FOR SYSTEMIC USE' and
  faers_male.meddra_name = 'Somnolence' and
  faers_male.llr <= 2*faers_male.llr_threshold
  order by
  faers_male.llr asc
```

- 30.** To download additional example SQL queries for extracting subsets of data from DrugCentral use the following URL: [https://unmtid-shinyapps.net/download/example\\_query.sql](https://unmtid-shinyapps.net/download/example_query.sql)

## Basic Protocol 8: Estimating Anti-SARS-CoV-2 activities with DrugCentral REDIAL-2020

There is currently an urgent need to find effective drugs for treating coronavirus disease 2019 (COVID-19). DrugCentral REDIAL-2020 (Kc et al., 2020), is a suite of machine learning models that forecast activities for live viral infectivity, viral entry, and viral replication specifically for severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), *in vitro* infectivity, and human cell toxicity. This application serves the scientific community when prioritizing compounds for in vitro screening and may ultimately accelerate identifying novel drug candidates for COVID-19 treatment. REDIAL-2020 consists of eleven independently trained machine learning models using high throughput screening data from the NCATS COVID19 portal (<https://opendata.ncats.nih.gov/covid19/index.html>) and includes a similarity search module that queries the underlying experimental dataset for similar compounds. These models were developed using experimental data generated by the following assays: the SARS-CoV-2 cytopathic effect (CPE) assay and its host cell cytotoxicity counterscreen, the Spike–ACE2 protein–protein interaction (AlphaLISA) assay and its TruHit counterscreen, the angiotensin-converting enzyme 2 (ACE2) enzymatic activity assay, the 3C-like (3CL) proteinase enzymatic activity assay, the SARS-CoV pseudotyped particle entry (CoV-PPE) assay and its counterscreen (CoV-PPE\_cs), the Middle-East respiratory syndrome coronavirus (MERS-CoV) pseudotyped particle entry assay (MERS-PPE) and its counterscreen (MERS-PPE\_cs), and the human fibroblast toxicity (hCYTOX) assay (Figure 81).

### Necessary Resources

#### Hardware

- Desktop or a laptop computer, or a mobile device, with a 100 Mbps or higher (fast) Internet connection.

#### Software

- An up-to-date web browser such as Google Chrome (<https://www.google.com/chrome/>), Mozilla Firefox (<https://www.mozilla.org/en-US/firefox/>), Apple Safari (<https://www.apple.com/safari/>), or Microsoft Edge (<https://www.microsoft.com/en-us/edge>)

### Protocol steps and annotations

#### Redial: A step-by-step content guide

1. By accessing REDIAL-2020 (<http://drugcentral.org/Redial>) from any web browser, including mobile devices, the submission page is displayed.
2. The web server accepts SMILES, drug names or PubChem CIDs as input. Regardless of input, the protocol converts drug names (from DrugCentral) or PubChem CIDs into SMILES.
3. The user interface provides a summary of the models, such as model type, which descriptor categories were used for training and the evaluation scores. The user interface depicts the processes of cleaning the chemical structures (encoded as SMILES) before training the machine learning models (Figure 82).



4. As an example, amodiaquine has been shown to have promising anti-SARS-CoV-2 behaviour in several papers (Bocci et al., 2020; Si et al., 2021), but its mechanism of action has not been well established yet. When given as an input to Redial, the webapp opens a new window with the predicted activities.
5. The prediction results table shows that amodiaquine is predicted to be active in cytopathic effect experiments while there are no clues on its mechanism (inactive in AlphaLISA, ACE2, 3CL assays) (Figure 83).
6. REDIAL-2020 links directly to DrugCentral for approved drugs and to PubChem for chemicals (where available), enabling easy access to further information on the query molecule (Figure 84).
7. Using REDIAL-2020 estimates, promising anti-SARS-CoV-2 compounds would ideally be active in the CPE assay while inactive in cytotox and in hCYTOX.

#### Queries Supported by Redial

8. Input queries such as drug name and PubChem CID are converted to SMILES before processing. Each SMILES string input is subject to four different steps, namely, converting the SMILES into canonical SMILES, removing salts (if present), neutralizing formal charges (except permanent ones) and standardizing tautomers. REDIAL-2020 predicts input compound activity across all eleven assays: CPE, cytotox, AlphaLISA, TruHit, ACE2, 3CL, CoV-PPE, CoV-PPE\_cs, MERS-PPE, MERS-PPE\_cs and hCYTOX (Figure 85).

#### Additional information

1. All of the codes and the trained models are available from: <https://doi.org/10.5281/zenodo.4606720>
2. The source code and specific models are available through Github at: (<https://github.com/sirimullalab/redial-2020>), or via Docker Hub (<https://hub.docker.com/r/sirimullalab/redial-2020>) for users preferring a containerized version. All the pre-ML processing and “data cleaning” scripts are here: <https://github.com/sirimullalab/redial-2020/tree/master/data-cleaning>
3. All workflows and procedures were performed using the KNIME platform 10. The NCATS data associated with the aforementioned assays were downloaded from the COVID-19 portal. <https://opendata.ncats.nih.gov/covid19/assays>

#### Basic Protocol 9: Drug Set Enrichment Analysis using Drugmonizome

Drugmonizome (Kropiwnicki et al., 2021) serves processed data extracted from drug and small molecule databases available from a variety of online repositories and data portals. The processed data is provided in the form of drug set libraries which serve as the underlying database for drug set enrichment analysis. Drugmonizome enables users to submit lists of drugs and small molecules as the input query. These drug sets are compared against various drug set libraries that contain known associations between drugs and their attributes, for example, side effects, indications, targets, pathways, induced gene expression

signatures, and other attributes. Additionally, Drugmonizome provides options for querying metadata associated with drug sets to find relevant drugs, small molecules, and drug sets for a given free text query.

## Necessary Resources

### Hardware

- Desktop or a laptop computer, or a mobile device, with a fast Internet connection

### Software

- An up-to-date web browser such as Google Chrome (<https://www.google.com/chrome/>), Mozilla Firefox (<https://www.mozilla.org/en-US/firefox/>), Apple Safari (<https://www.apple.com/safari/>), or Microsoft Edge (<https://www.microsoft.com/en-us/edge>).

## Protocol steps and annotations

### Metadata Search

1. Navigate to the Drugmonizome homepage (<https://maayanlab.cloud/drugmonizome/>). The metadata search is displayed by default. Using the search bar, query terms of interest to identify resources, drug set libraries, drug sets, and small molecules contained in Drugmonizome. Example terms are suggested for each type of metadata search (Figure 86).
2. Alternate between resource, drug set library, drug set, and small molecule metadata searches by clicking the corresponding tab. When performing metadata searches for drug sets, use the filter table to query terms within specific resources, drug set libraries, and association types.
3. Upon submitting a term of interest using the search bar, a list of results that match the term is displayed (Figure 87).
4. Clicking on any term displays a page with identifying metadata for the resource, drug set library, drug set, or small molecule. When perusing drug set metadata, a search bar exists for querying specific small molecules of interest within the set (Figure 88).

### Drug Set Enrichment

1. Navigate to the drug set enrichment page by clicking the corresponding tab on the website header. The drug set enrichment page includes a search box where a list of drugs and small molecules can be pasted. The page also includes several example drug sets that are pasted into the box when clicked (Figure 89). As an example, click the “69 in vitro COVID-19 hits from a drug screen by Ellinger et al.” link to populate the search box with a small molecule set.

Note: Drug and small molecule entities can be queried by name, DrugBank IDs, Broad Institute Accession Numbers (BRD-IDs), SMILES strings, and InChIKeys.

2. Click the “Perform Drug Set Enrichment Analysis” button and a results page of all resources with enriched terms is returned. Each of the resources with enriched drug set libraries are represented as an icon with the number of enriched terms for each resource (Figure 90).
3. Click on any of the resource icons to be redirected to a page with the top enrichment results for each drug set library represented by a toggleable bar graph or scatter plot. The drug set library enrichment results can be expanded by clicking the corresponding button (Figure 91).
4. The expanded page includes the scatter plot, bar graph, and table view of the top enriched terms. The table representation displays the top enriched terms and their p-values, odds ratio, and corrected q-values. Terms of interest can be queried using the search bar above the table. The table is also available for download as a .TSV file (Figure 92).

### Resources Pages

5. Navigate to the resources page by clicking the corresponding tab on the website navigation bar (Figure 93).
6. Each of the drug data resources used to create drug set libraries is catalogued on this page. Click on the DrugBank resource card to view metadata specific to DrugBank, as well as drug set libraries curated from DrugBank (Figure 94).
7. Click on the “DrugBank Small Molecule Targets” library to be redirected to a page with identifying metadata for the drug set library. The metadata for the drug set library includes download links for the .DMT files in drug name or InChIKey format (Figure 95). Additionally, each of the drug sets included in this library are listed below. Clicking on any drug set name redirects to a page with metadata specific to the drug set, as well as the set of associated small molecules.

### Basic Protocol 10: The Drugmonizome-ML Appyter

A wealth of data from a multitude of sources is readily available for thousands of bioactive small molecules in Drugmonizome (Kropiwnicki et al., 2021). The information in Drugmonizome can be harnessed to develop machine learning models that utilize such data to predict the properties of small molecules that are poorly annotated. The Drugmonizome database draws upon a variety of publicly available resources to label each small molecule by its associations with pathways, protein targets, induced gene expression profiles, chemical features, and other attributes. Drugmonizome-ML is an Appyter (Clarke et al., 2021) that executes a machine learning pipeline as a Jupyter notebook using the data curated for creating Drugmonizome. Drugmonizome-ML can be used to make predictions for indications and other attributes such as drug targets or side effects for poorly annotated pre-clinical bioactive small molecules.

### Necessary Resources

#### Hardware

- Desktop or a laptop computer, or a mobile device, with a fast Internet connection

## Software

- An up-to-date web browser such as Google Chrome (<https://www.google.com/chrome/>), Mozilla Firefox (<https://www.mozilla.org/en-US/firefox/>), Apple Safari (<https://www.apple.com/safari/>), or Microsoft Edge (<https://www.microsoft.com/en-us/edge>).

## Protocol steps and annotations

### Input Dataset Selection

1. Navigate to the Drugmonizome-ML Appyter ([https://appymaayanlab.cloud/Drugmonizome\\_ML/](https://appymaayanlab.cloud/Drugmonizome_ML/)). The input form is divided into three sections: input dataset selection, target label selection, and machine learning pipeline.
2. Select datasets from Drugmonizome and SEP-L1000 (Kropiwnicki et al., 2021; Wang et al., 2016) to populate the feature matrix that will be used for learning and classification. Each of the datasets' contents are described using tooltips (Figure 96). For the demonstration, select the "LINCS Gene Expression Signatures" from the "Transcriptomic and Imaging Datasets" subfield and "Morgan Fingerprints" from the "Chemical Fingerprints Generated for Compounds from SEP-L1000" subfield.
3. Additional options for pre-processing the feature matrix are available. If selecting features from various data sources, it is likely that not all compounds will be included across all feature sets, therefore a toggleable option decides whether drugs with missing data are retained or dropped from the feature matrix. Additionally, because some of the available feature sets are binary association matrices, there is the option to apply TF-IDF normalization to account for frequency of common and rare features among the small molecules (Figure 97). In general, the default settings for these options are recommended.

### Target Label Selection

4. In this section, select the positive class label for a binary classification problem. There is the option to select an attribute from any of the Drugmonizome drug set libraries in an autocomplete field where relevant drug-set labels from Drugmonizome are offered as potential class labels (Figure 98). Type any characters into the autocomplete field and matching drug-set labels will be displayed. For the demonstration, type "neuropathy peripheral (from SIDER Side Effects)" into the autocomplete field.
5. Alternatively, upload a newline separated .txt file of compounds to be used as positive examples of a class to predict by selecting the "List" option in the "Target Label Selection" section. Example .txt files are available for download to understand the structure of the file (Figure 99). Choose the drug identifier format (drug name or InChI key) that small molecules within the text file are described by. InChI Keys are the recommended format.

6. The “Include stereoisomers” option decides whether to match compounds from the feature matrix to the target vector using the first 14 characters of the InChIKey (which encodes chemical connectivity) thus including stereoisomers of a particular small molecule, or whether to consider only one form of a molecule and match by the whole InChIKey.

### Machine Learning Pipeline

7. In this section, select data visualization options, machine learning classifiers, machine learning hyperparameters, and methods to evaluate the classifier (Figure 100).
8. Select your preferred data visualization method from the drop-down menu under the “Data Visualization Method” field. The default and recommended method is UMAP.
9. If applicable, select a dimensionality reduction algorithm from the drop-down menu under the “Dimensionality Reduction Algorithm” field.
10. If applicable, select a feature selection method from the drop-down menu under the “Machine Learning Feature Selection” field.
11. The “Machine Learning Algorithm” section includes 9 distinct classifiers that can be chosen by clicking on the corresponding classifier name. Furthermore, each classifier has hyperparameter fields that can be modified. For example, select the “Extra Trees classifier”. Input “1250” in the “n\_estimators” field. Select “entropy” in the “criterion” drop-down menu. Select “log2” in the “max\_features” drop-down menu. All other hyperparameters can be kept as default.
12. Select whether to calibrate algorithm predictions by selecting the appropriate choice in the “Calibrate algorithm predictions” field. This setting will calibrate the predictions output by the chosen model, eliminating model-imparted bias. It is recommended to keep this setting as default.
13. Select a cross-validation method from the drop-down menu under the “Cross-Validation Algorithm” field. The recommended option is Repeated Stratified Group K-Fold because this cross-validation method will maintain class ratios across train and validation splits. Furthermore, choose the number of cross-validation folds and cross-validation repetitions in the subsequent fields. For the demonstration, input “10” into the “Number of Cross-Validation Folds” field and “3” into the “Number of Cross-Validated Repetitions” field.
14. Choose the primary evaluation metric for assessing the performance of the model from the drop-down menu under the “Primary Evaluation Metric” field. The default and recommended metric is “roc\_auc”.
15. Choose any additional evaluation metrics from the drop-down menu under the “Evaluation Metrics” field and these metrics will also be reported for the trained model.

16. Click “Submit” at the bottom of the input form.

### **Navigating the Drugmonizome-ML Appyter Notebook**

17. A Jupyter Notebook will begin executing in the cloud once the input form is submitted. The notebook includes an option to download the notebook, toggle displaying the code, and run the notebook locally. Additionally, a table of contents exists with clickable elements that link to specific sections within the notebook (Figure 101).
18. Scroll down to the “Select Input Datasets and Target Classes” section or click on the corresponding section from the table of contents. The feature matrix that was generated based on the selected features from the input form is displayed. The feature matrix is composed of 19,898 compounds and 3026 features from LINCS Gene Expression Profiles and TF-IDF normalized Morgan Fingerprints (Figure 102).
19. Additionally, information is displayed about how the target array is constructed, how many compounds from the target array are included in the feature matrix, and how many compounds were discarded because they were not included in the feature matrix. Unmatched compounds are available for download.
20. Navigate to the “Dimensionality Reduction and Visualization” section to view the input feature space using the dimensionality reduction and visualization methods that were selected in the input form. Positive class labels are labeled within the visualization to demonstrate how the class of interest is clustered in the feature space (Figure 103).
21. Navigate to the “Machine Learning” section to view the trained classifier and evaluations of the classifier’s performance. The receiver operating characteristic curve (Figure 104), precision-recall curve (Figure 105), and confusion matrix (Figure 106) are displayed. Click the hyperlinks in the figure headers to download the figures.
22. Navigate to the “Examine Predictions” section to view the predictions made by the model in addition to the distributions of mean probability estimates and t-statistics. Figures displaying the distribution of mean cross-validation predictions (Figure 107), distribution of t-statistics (Figure 108), a UMAP visualization of the feature space with overlaid predictions (Figure 109), and a filterable table of the top predicted compounds (Figure 110) are displayed. Click the hyperlinks in the figure and table headers to download the corresponding figure or table.
23. Navigate to the “Feature Importance” section to view the most important features from the input feature matrix that were used to make predictions. A table of the most important features used by the model to make predictions (Figure 111), as well as a figure depicting the distributions of average and cumulative sum of feature importance (Figure 112) are displayed. Click the hyperlinks in the figure and table headers to download the corresponding figure or table.

## Basic Protocol 11: The Harmonizome-ML Appyter

Harmonizome (Rouillard et al., 2016) is a collection of processed datasets that abstract knowledge about genes and proteins. Using the processed data from Harmonizome, Harmonizome-ML enables interactive imputation of knowledge about the function and other properties of genes and proteins using machine learning. Combined with a user-friendly interface of an Appyter (Clarke et al., 2021) – a web-based software application enabling users to execute bioinformatics workflows without coding – the Harmonizome-ML Appyter can be used to build and evaluate machine learning pipelines with Harmonizome data in an accessible way. The Harmonizome-ML Appyter asks users to select or upload attributes for learning as well as specify a target vector to predict. Users also need to select from various machine learning algorithms and performance evaluation methods. Once these options are selected, the workflow is executed, and the results are presented as a Jupyter Notebook that is shareable and downloadable.

### Necessary Resources

#### Hardware

- Desktop or a laptop computer, or a mobile device, with a fast Internet connection

#### Software

- An up-to-date web browser such as Google Chrome (<https://www.google.com/chrome/>), Mozilla Firefox (<https://www.mozilla.org/en-US/firefox/>), Apple Safari (<https://www.apple.com/safari/>), or Microsoft Edge (<https://www.microsoft.com/en-us/edge>).

### Protocol steps and annotations

#### Navigating the input page

1. Navigate to the Harmonizome-ML Appyter ([https://appymt.maayanlab.cloud/#/harmonizome\\_ml](https://appymt.maayanlab.cloud/#/harmonizome_ml)). The input form is divided into two sections: “attribute and prediction class dataset selection” and “settings”.
2. In the “Attribute and Prediction Class Selection” section, select attributes by clicking on the check box to the left of an attribute of choice; a blue check mark indicates that an attribute has been selected. Users may opt to upload a custom attribute dataset using the “Browse” button as well. Target selection can be from Harmonizome or customized; click on the text for the target selection desired and customize the class in the text box below (Figure 113).
3. The “Settings” section includes settings for various algorithms (dimensionality reduction, manifold projection, ML feature selection, cross validation, ML algorithm, hyperparameter search type, evaluation metrics) that can be customized. Simply click on the drop-down menu below an algorithm to view and update the options. For example, clicking on the drop-down menu for “Dimensionality Reduction Algorithm” displays the following options: PCA, truncated SVD, incremental PCA, ICA, and Sparse PCA. Click on the desired algorithm to use it for dimensionality reduction (Figure 114).



4. Once all selections have been made, click on the “Submit” button at the bottom of the page to run the analyses and generate the notebook.

### Navigating the notebook

5. Each notebook generated by the Harmonizome-ML Appyter includes explanations followed by code, data, and figures (both static and interactive). To download the notebook, toggle notebook code, or run the notebook locally, select the appropriate button at the top of the page. The notebook is divided into three sections (which can be accessed through the table of contents on the left side of the page): Inputs, Dimensionality Reduction, and Machine Learning (Figure 115).
6. Navigate to the “Inputs” section to view the feature matrix Dataframe generated from the datasets selected in the input form (Figure 116). Note that some Dataframes contain additional columns that can be explored by scrolling left to right. The first two Dataframes are individual datasets, whereas the final Dataframe displays the concatenated feature matrix that will be used for classification.
7. Scroll down to view the target array created from the dataset containing the class label to be predicted. Genes that are known to be associated with the class label are annotated with a 1, whereas genes not known to be associated with the class label are annotated with a 0 (Figure 117).
8. Navigate to the “Dimensionality Reduction” section. The process of dimensionality reduction involves transforming data from high-dimensional spaces to low-dimensional spaces without losing too much information. The input features are reduced using PCA and visualized in a 3D scatter plot (Figure 118). The reduced features are also projected onto a manifold with T-SNE (Figure 119).
9. Navigate to the “Machine Learning” section which features the machine learning pipeline assembled from the input form submission. A model is generated and trained via the customized pipeline and then used to predict genes that are strongly correlated with the target attribute. General explanations for the model’s performance are provided with ROC curves and a prediction matrix (Figure 120).
10. The prediction results are provided at the end of the pipeline and can be downloaded as a tab-separated (.tsv) file by clicking on “results.tsv” at the end of the notebook (Figure 121).

### Basic Protocol 12: GWAS target illumination with TIGA

Target Illumination GWAS Analytics (TIGA) (Yang et al., 2021) is a web application that facilitates drug target illumination by scoring and ranking protein-coding genes associated with traits from genome-wide association studies (GWAS). Similarly, TIGA can score and rank traits with the same gene-trait association metrics. Rather than a comprehensive analysis of GWAS for all biological implications and insights, this focused application provides a rational method by which GWAS findings can be aggregated and filtered for



applicable, actionable intelligence, with evidence usable by drug discovery scientists to enrich prioritization of target hypotheses. TIGA derives its GWAS summary and metadata solely from the NHGRI-EBI GWAS Catalog and study-associated publications. Thus, TIGA traits are identified by Experimental Factor Ontology (EFO) terms.

## Necessary Resources

### Hardware

- Desktop or a laptop computer, or a mobile device, with a fast Internet connection

### Software

- An up-to-date web browser such as Google Chrome (<https://www.google.com/chrome/>), Mozilla Firefox (<https://www.mozilla.org/en-US/firefox/>), Apple Safari (<https://www.apple.com/safari/>), or Microsoft Edge (<https://www.microsoft.com/en-us/edge>).

## Protocol steps and annotations

### Navigating the input page

1. Navigate to the TIGA web app: (<https://unmtid-shinyapps.net/shiny/tiga/>)

### Trait to gene search

2. A trait query may be specified by browsing and selecting from the Traits (ALL) tab, or via the Trait query field.
3. To find genes associated with the EFO term “worry measurement” (EFO\_0009589), begin typing “worry” in the Trait query field, and autosuggest will assist in selecting the trait, (Figure 122).
4. TIGA results will be displayed via the HitsTable tab and HitsPlot tab (Figure 123).
5. The HitsTable is ranked by meanRankScore as a measure of the strength and confidence of the inferred gene-trait association.
6. The HitsPlot displays hits with meanRankScore on the horizontal axis, and Effect on the vertical axis, either measured by odds ratio (OR) or N\_beta (count of beta values).
7. Hits are annotated, either in the table as columns or as hover-tooltips, with several identifiers, measures, and variables, derived from the aggregated GWAS, or annotated from IDG. Target Development Levels (TDLs) are also color coded for ease of use, facilitating identification of well-known targets (Tclin) and understudied targets (Tdark).
8. From the HitsTable, for a specific gene, the magnifying-glass icon links to the TIGA provenance for the corresponding gene-trait association. The provenance displays studies and publications supporting the association, with GWAS Catalog and PubMed link-outs, respectively (Figure 124).

### Gene to trait search

9. In Gene query mode, TIGA behaves much the same as in Trait query mode, but with traits as hits. Data which pertain to gene-trait associations will be the same, such as provenance, regardless of query mode.
10. TIGA genes are, as in the Catalog, identified by Ensembl Gene IDs. The Gene query field will autosuggest based on gene symbols. Thus, by typing “RAS”, autosuggest will assist in selecting “RASA2”, “Ras GTPase-activating protein 2.”
11. As in Gene query mode, results will be via HitsTable and HitsPlot tabs.

### Basic Protocol 13: Prioritizing kinases for lists of proteins and phosphoproteins using KEA3

Kinase Enrichment Analysis 3 (KEA3) (Kuleshov et al., 2021) is a web-based server application that infers overrepresented upstream kinases whose putative substrates are present in a user-inputted list of differentially-phosphorylated proteins. To infer upstream kinases, KEA3 uses a collection of kinase-substrate libraries created from processing data from several online databases. Kinase enrichment analysis results are provided for each kinase-substrate library, as well as two integrated approaches to integrate all libraries: MeanRank and TopRank. The gene sets from the kinase-substrate libraries are compared to the user-inputted protein list, and the Fisher’s Exact Test is used to compute the significance of the overlap to prioritize kinases. The resulting ranked lists of kinases, as well as visualizations of the significant kinases as networks, are returned to the users as interactive and downloadable figures.

#### Necessary Resources

**Hardware:** Desktop or a laptop computer, or a mobile device, with a fast Internet connection

#### Software

- An up-to-date web browser such as Google Chrome (<https://www.google.com/chrome/>), Mozilla Firefox (<https://www.mozilla.org/en-US/firefox/>), Apple Safari (<https://www.apple.com/safari/>), or Microsoft Edge (<https://www.microsoft.com/en-us/edge>).

Note that there is a tutorial on navigating KEA3 results (<https://maayanlab.cloud/kea3/templates/tutorial.jsp>) from which some of the steps in this protocol have been paraphrased.

#### Protocol steps and annotations

##### Submitting a gene set to KEA3

1. Navigate to the KEA3 homepage (<https://maayanlab.cloud/kea3/>).
2. Gene/protein sets may be submitted to KEA3 in two ways: by uploading the set as a plain text file or by pasting a list, one gene/protein name per line, into a text box. When submitting genes/proteins using the text box, a checklist below the

text box denotes duplicates and confirms valid gene symbols in the input. Once uploaded or inputted, click on the “Submit” button to begin the analysis (Figure 125).

Note that only HGNC-approved gene symbols will be accepted.

### Navigating KEA3 results

3. Scroll down to view the “Integrated results” tab which includes bar charts, tables, subnetwork visualizations, and a clustergrammer visualization of integrated results across all KEA3 libraries using the MeanRank and TopRank methods (Figure 126). The MeanRank method calculates the average rank, whereas the TopRank method calculates the best scaled rank of each kinase across all libraries containing the kinase. The tables can be downloaded in TSV format and visualizations can be downloaded in SVG and PNG format. Use the slider above each visualization to change the number of top results that are displayed.
4. The Tables tab displays interactive tables of ranked kinases for each individual KEA3 library (Figure 127). The tables are organized into kinase-kinase substrate interaction libraries, protein-protein interaction libraries, and libraries with all associations. Each table displays the top 10 ranked kinases using the Fisher’s Exact Test p-value. Click on any of the table headers to re-sort the table. Clicking on any of the kinase names will redirect you to a single gene landing page in Harmonizome. Access the complete list of kinases by downloading any table in TSV format using the download icon.
5. The Networks tab displays global kinase co-regulatory networks generated by applying Weighted Gene Co-expression Network Analysis (WGCNA) (Langfelder and Horvath, 2008) to ARCHS4 (Lachmann et al., 2018), GTEx (Aguet et al., 2020), and TCGA (Tomczak et al., 2015) data in order to visualize the top-ranked kinases in the context of the larger human phosphorylation network; the top-ranked kinases are highlighted in the network (Figure 128). To choose the top-ranked kinases from a specific library, navigate to the “Select a library” drop-down menu and click on the desired library. Download each network as an SVG or PNG file by selecting the corresponding download button.
6. The Subnetworks tab displays kinase co-regulatory network visualizations which have been dynamically generated from the top-ranked kinases in each library (Figure 129). An edge between two kinases indicates an interaction supported by library evidence from either a kinase-substrate interaction library (directed edge) or protein-protein interaction library (undirected edge). Hover over an edge to display the library evidence supporting the interaction. Download each network as an SVG or PNG by clicking the desired file type in the bottom left corner of the graph.
7. The Bar Charts tab provides bar charts showing the  $-\log(p\text{-value})$  of the top-ranked kinases for each individual library (Figure 130). The bar charts are organized into kinase-kinase substrate interaction libraries, protein-protein interaction libraries, and libraries with all associations. Use the slider above each

figure to change the number of top kinases within the figure. Download any given chart as an SVG or PNG by selecting the desired file type in the bottom left-hand corner of the chart.

8. The Clustergrammer tab uses the Clustergrammer (Fernandez et al., 2017) application to provide an interactive clustergram of overlapping substrate targets between the input and the top library results (Figure 131). Share, take a snapshot, download, or crop the clustergram matrix using the icons in the menu bar on the left side of the clustergram. Customize row order and column order by selecting one of the options (alphabetically, cluster, rank by sum, rank by variance) under “Row Order” and “Column Order”, respectively. Search for rows using the text search box. Adjust the dendrogram groups, which show clusters at different hierarchical levels and are represented by grey triangles and trapezoids along the bottom and right axes, using the grey triangular sliders on the right and bottom-left sides of the clustergram.

Note: A tour of Clustergrammer that explains its features in more depth can be found here: [http://maayanlab.github.io/clustergrammer/scrolling\\_tour](http://maayanlab.github.io/clustergrammer/scrolling_tour). More details on interacting with the clustergram can be found in the Clustergrammer documentation: [https://clustergrammer.readthedocs.io/interacting\\_with\\_viz.html](https://clustergrammer.readthedocs.io/interacting_with_viz.html)).

1. Open a new or existing Python code file. Import the JSON and requests libraries at the top of the file.

```
import json
import requests
```

2. Call the requests.post method to send a POST request to the URL. The payload variable contains the parameters that are sent to the API endpoint specified in KEA3\_URL. In this case the endpoint is /enrich and the parameters are query\_name, which specifies the name of the query, and gene\_set, which specifies the query gene list to be enriched.

```
KEA3_URL = 'https://maayanlab.cloud/kea3/api/enrich/'
payload = {"query_name": "myQuery", "gene_set":
["FOXO1", "SMAD9", "MYC", "SMAD3", "STAT1", "STAT3"]}
response = requests.post(KEA3_URL, json=payload)
data = json.loads(response.text)
print(data)
```

3. Use the json.loads method to view the response as a JSON object containing the top enrichment results from various libraries.

```
{
  'Integrated--meanRank':
  [ {'Query Name': 'myQuery',
    'Rank': '1',
    'TF': 'CDK4',
```

```

    'Score': '37.73',
    'Library':
'STRING.bind,20;ChengPPI,2;PhosDA11,39;BioGRID,4;HIPPIE,13;ChengKSIN,29;STRIN
G,107;MINT,59;mentha,2;prePPI,137;PTMsigDB,3',
    'Overlapping_Genes': 'SMAD3,STAT1,MYC,STAT3,SMAD9,FOXM1',
    {'Query Name': 'myQuery',
    'Rank': '2',
    'TF': 'PDGFRA',
    'Score': '48.38',
    'Library':
'STRING.bind,11;ChengPPI,7;PhosDA11,59;BioGRID,110;HIPPIE,2;STRING,61;mentha,
8;prePPI,129',
    'Overlapping_Genes': 'SMAD3,STAT1,MYC,STAT3,SMAD9,FOXM1'},
    ...
}

```

Note: More detailed instructions, as well as examples from the command line and in R, can be found at the following link: <https://maayanlab.cloud/kea3/templates/api.jsp>.

#### Basic Protocol 14: Converting PubMed searches to drug sets with the DrugShot Appyter

PubMed contains millions of publications that co-mention drugs with other biomedical terms such as genes or diseases. DrugShot is an Appyter (Clarke et al., 2021) that enables users to enter any biomedical search term into an input form to receive ranked lists of drugs and small molecules based on their relevance to the search term. DrugShot then deploys a Jupyter Notebook in the cloud to display ranked lists of drugs. To achieve this, DrugShot cross-references returned PubMed IDs with DrugRIF, a curated resource of drug-PMID associations, to produce an associated compound list where each compound is ranked according to the total co-mentions with the search term from shared PubMed IDs. Additionally, lists of compounds predicted to be associated with the search term are generated based on drug-drug co-occurrence in the literature, and drug-drug co-expression correlations computed from L1000 drug-induced gene expression profiles. Through its search functionality and abstraction of drug sets from different sources, DrugShot facilitates hypothesis generation by suggesting small molecules related to any searched biomedical term.

#### Necessary Resources

##### Hardware

- Desktop or a laptop computer, or a mobile device, with a fast Internet connection

##### Software

- An up-to-date web browser such as Google Chrome (<https://www.google.com/chrome/>), Mozilla Firefox (<https://www.mozilla.org/en-US/firefox/>), Apple Safari (<https://www.apple.com/safari/>), or Microsoft Edge (<https://www.microsoft.com/en-us/edge>).

## Protocol steps and annotations

### Query Biomedical Term

1. Navigate to the DrugShot Appyter (<https://appyters.maayanlab.cloud/DrugShot/>). The Appyter input form includes options to query a biomedical term to retrieve a prioritized list of small molecules that is augmented using drug-drug similarity matrices, or to submit a list of small molecules to be augmented using drug-drug similarity matrices.
2. Input a biomedical term into the “Biomedical Term” field. The default string used for this demonstration is “Lung Cancer”. Input an integer ranging from 20 to 200 in the “Associated Drug Set Size” field; this value is used to determine the size of the unweighted drug set that is used to predict related compounds. The larger the value selected, the broader the resulting predictions will be (Figure 132).
3. Click submit on the Appyter input form and a Jupyter Notebook with the input parameters will be launched in the cloud.
4. The first output element of the notebook is a table of “Top Associated Compounds” (Figure 133). This table provides the top-ranked drug and compound names associated with the query term (Index Column), the count of PubMed publications associating each drug with the search term (Column 1), and the fraction of the publications associating the drug and search term divided by the total number of publications related to the drug regardless of search term (Column 2). Click on the hyperlinked filename below the table title to download a .CSV file listing all the associated compounds. This file also includes a Score column containing values that are the product of the first two columns.
5. The second output component of this notebook is a scatter plot (Figure 134) of the values from the table of “Top Associated Compounds”. The X axis displays the integer counts of Publications with Search Term, and the Y axis shows the fraction of Publications with Search Term / Total Publications. Hover over any point on this plot to display the compound’s name and its corresponding X and Y values.
6. An unweighted drug set is created through ranking small molecules from the association table by the product of the total associated publications and their normalized fraction.

### Querying a list of small molecules

7. Alternatively, submit a newline separated .txt file of small molecule names using the input form, thereby omitting steps 2–6. The submitted small molecules will be used as the unweighted drug set that will be used in subsequent steps (Figure 135).

### Literature co-mentions predictions

8. A receiver operating characteristic (ROC) curve that describes the ranking of associated compounds in the DrugRIF literature co-mentions matrix is output (Figure 136). This plot shows the True Positive Rate on the Y axis and the False Positive Rate on the x-axis. The predicted compounds are computed using average co-mention counts of PubMed IDs between the unweighted drug set, and other drugs and small molecules within DrugRIF. The area under the curve (AUC) is shown to the right of the plot and hovering over any point on the curve displays the associated X and Y values.
9. The literature co-mentions prediction matrix is seeded with the unweighted drug set and the top predicted compounds are ranked by their average co-mentions with the small molecules in the unweighted drug set. The “average co-mentions” values are provided in a table that displays the top 20 predicted compounds (Figure 137). Click on the hyperlinked filename below the Table 2 header to download the table as a .CSV file.
10. The top 50 co-occurrence predicted compounds are queried using the DrugEnrichr API for drug set enrichment analysis. The top 10 enriched terms from the downregulated and upregulated GO Biological Processes drug set libraries and the SIDER drug set library are displayed as bar plots (Figure 138). Click the link below the bar plots to be directed to the DrugEnrichr enrichment results page (Figure 139).

### Signature similarity predictions

11. A receiver operating characteristic (ROC) curve that describes the ranking of associated compounds in the L1000 signature similarity matrix is output (Figure 140). This plot shows the True Positive Rate on the Y axis and the False Positive Rate on the x-axis. The predicted compounds are computed using average cosine similarity of drug-induced gene expression signatures between the unweighted drug set, and other drugs and small molecules within the co-expression prediction matrix. The area under the curve (AUC) is shown to the right of the plot and hovering over any point on the curve displays the associated X and Y values.
12. The signature similarity prediction matrix is seeded with the unweighted drug set and the top predicted compounds are ranked by their average cosine similarity to the small molecules in the unweighted drug set. The “average cosine similarity” values are provided in a table that displays the top 20 predicted compounds (Figure 141). Click on the hyperlinked filename below the table header to download the table as a .CSV file.
13. The top 50 signature similarity predicted compounds are queried using the DrugEnrichr API for drug set enrichment analysis. The top 10 enriched terms from the downregulated and upregulated GO Biological Processes drug set libraries and the SIDER drug set library are displayed as bar plots (Figure 142)



Click the link to be directed to the DrugEnrichr enrichment results page (Figure 143).

## COMMENTARY

### Background Information

The IDG consortium has generated several different resources that are available to the research community. These resources include experimental data, tools, and reagents from the Data and Resource Generating Centers (DRGCs) covering the IDG highlighted protein families. These proteins are investigated by compound library screening (*in vitro* and *in silico*), antibody development, function and activation state profiling, and mouse expression profiling. Moreover, illuminating the druggable GPCR-ome is achieved by a two-pronged approach of experimental screening of drugs followed by computational screening against modeled structures of the GPCR to produce optimized lead compounds. This work has led to the discovery of several novel compounds, for example, the small molecule “ogerin” binds to GPR68 (Huang et al., 2015). Much of the success of identifying such novel GPCR binding compounds is due to development of a novel screening assay, PRESTO-Tango (Kroeze et al., 2015), which enables simultaneous investigation of every non-olfactory G protein-coupled receptor in the human genome. Additionally, the DRGCs recently gained insight into new potential therapeutics to help treat circadian rhythm disorders via the melatonin receptors MT1 and MT2 (Stein et al., 2020). The DRGCs also illuminate ion channels by utilizing CRISPR technology to map expression profiles, assess channel activities, develop antibodies, and generate new mouse lines. This work recently elucidated *TMEM16C* and its involvement in thermoregulation and protection from febrile seizures in rodent pups (Wang et al., 2021). Furthermore, discovering the function of the understudied druggable kinome includes using Multiplex Inhibitor Beads (MIB) / Mass Spectrometry (MS) to identify kinase activation status in response to perturbagens. This approach is applied to model cell lines and patient-derived xenografts. These data, along with other data collection efforts, are incorporated into the Dark Kinase Knowledgebase (DKK) that provides gene-by-gene and network-level information on the dark kinome and its interaction with other signal transduction regulatory networks (Berginski et al., 2021). For example, recently, the kinase CDC42BPA/MRCK $\alpha$  has been identified as a potential target for brain, ovarian, and skin cancers (East and Asquith, 2021). Moreover, the Kinase Chemogenomic Set (KCGS) is the most highly annotated set of selective kinase inhibitors available to researchers for use in cell-based screens. Recently, the NIH IDG initiative nominated 162 dark kinases to develop chemical and biological tools to seed research on these understudied proteins. Currently, KCGS contains data of 37 inhibitors from the IDG dark kinases, which may be helpful and improve initial chemical tools to study these kinases (Wells et al., 2021).

Congruently, the IDG Knowledge Management Center (IDG-KMC) develops bioinformatics tools and other digital assets, enabling users to query and visualize the data produced by the DRGCs and other sources. The IDG-KMC gathers knowledge covering the entire human genome and expanding to model systems, including GWAS studies, expression data, compound binding, and patent information via ChEMBL (Mendez et al., 2019). Furthermore, the IDG-KMC incorporates associated information related to human



protein-coding genes, diseases, mouse phenotypes, small molecules and approved drugs (perturbagens) that modulate these proteins/genes and diseases. Utilizing these collected and annotated databases generates opportunities for machine learning ready platforms. For example, using these tools (i.e., combining data on genes, proteins, and RNA molecules from fourteen databases and publications), the IDG-KMC developed a machine learning algorithm that prioritizes targets for human genes associated with 17 unique types of pain and identified thirteen potential IDG family drug targets for migraine drug development and four for rheumatoid arthritis (Jeon et al., 2021). Here we provide a collection of step-by-step get-started protocols to gain initial access to the resources created by the IDG-KMC. We hope that these protocols will facilitate experimental and computational biologists to further engage with the unique opportunities offered by the IDG program toward accelerating drug and target discovery.

### Critical Parameters:

There are several libraries and data sources that IDG-KMC web applications rely on. PubMed (<https://pubmed.ncbi.nlm.nih.gov/>) and DrugCentral (Avram et al., 2021) play an important role in several of the protocols. PubMed and DrugCentral are used by IDG-KMC web applications as both sources of data and also as external references which users can reach from within some IDG-KMC web applications.

The Target Central Resource Database (TCRD) is the central resource behind the Illuminating the Druggable Genome Knowledge Management Center (IDG-KMC) (Sheils et al., 2021). TCRD contains information about human targets and emphasizes four families of targets central to the NIH IDG initiative: GPCRs (note that olfactory GPCRs are treated as a separate family), kinases, and ion channels. A unique aim of the KMC is to classify the development/druggability level of targets via Target Development Level (TDLs). TDLs are currently categorized into four development/druggability levels: **Tclin**, **Tchem**, **Tbio**, and **Tdark**. **Tclin** targets have activities in DrugCentral with a known mechanism of action. **Tchem** targets have activities in ChEMBL (Mendez et al., 2019), Guide to Pharmacology (Armstrong et al., 2019), or DrugCentral that satisfy the activity thresholds, but no approved drugs. **Tbio** targets do not have known drug or small molecule activities that satisfy the activity thresholds and satisfy one or more of the following criteria: target is above the cutoff criteria for Tdark, the target is annotated with a Gene Ontology Molecular Function or Biological Process (The Gene Ontology Consortium, 2019) leaf term(s) with an Experimental Evidence code. **Tdark** targets have limited information or knowledge about them. Moreover, TDark currently includes ~31% of the human proteins that were manually curated at the primary sequence level in UniProt, but do not meet any of the **Tclin**, **Tchem** or **Tbio** criteria.

Each of the datasets in Harmonizome are compiled from various resources that contain information regarding gene-attribute associations. Gene-attribute associations can range from chemical perturbations that induce differential expression in select genes (Subramanian et al., 2017) to specific genes differentially expressed in cell lines (Cowley et al., 2014; Barretina et al., 2012). The evidence for these associations depends on the resource and can be from text mining, high-throughput -omics data, and other methods.

The ARCHS4 resource, and by extension the PrismEXP Appyter, depend on FASTQ files generated from RNA-seq experiments deposited in the Gene Expression Omnibus (GEO) (Edgar et al., 2002).

Geneshot relies on knowledge about under-studied targets from GeneRIF (Osborne et al., 2007) and AutoRIF (Lachmann et al., 2019), association files that catalog gene-PubMed ID co-mentions. AutoRIF is larger and more comprehensive than GeneRIF, but potentially less accurate due to its automated creation. Furthermore, Geneshot generates predictions from gene-gene similarity matrices compiled from AutoRIF, GeneRIF, ARCHS4, Enrichr (Kuleshov et al., 2016), and Tagger (Pletscher-Frankild and Jensen, 2019).

For TIN-X, Drug Target Ontology (Lin et al., 2017) is used to establish associations between drug targets and disease states. TIN-X allows the user to browse diseases based on TDL, IDG Family, as well as user-supplied search terms for drug targets associated with the disease being considered.

Drugmonizome depends upon drug-attribute associations compiled from various resources. These drug-attribute associations are stored as drug set libraries, collections of drug sets that describe relationships between biomedical terms and sets of drugs. The drug set libraries are categorized into distinct categories that include: drug targets and associated genes; side effects, adverse events and phenotypes; gene ontology (GO) and pathway terms; chemical structure and sub-structure motifs; and modes of action.

Several of the protocols (namely Protocols 4, 10, 11, and 15) mention Appyters. Appyters turn Jupyter Notebooks into functional standalone web applications for bioinformatics workflows (Clarke et al., 2021). Each Appyter presents a unique workflow tied to an input form that can be modified by the user. Once the user submits the input form options, a Jupyter Notebook is executed in the cloud and populated with the selected options. These notebooks contain various analyses and publication ready figures that can be shared and downloaded by the research community.

GWAS target illumination depends upon GWAS summary and metadata from the NHGRI-EBI GWAS Catalog with study-associated publications.

TIGA traits are identified by Experimental Factor Ontology (EFO) terms.

The prioritization of kinases for lists of proteins and phosphoproteins with KEA3 makes use of individual libraries generated from kinase-substrate interactions and protein-protein interactions, plus two integrated libraries, MeanRank and TopRank.

When converting PubMed searches to drug sets with the DrugShot Appyter, DrugRIF is used as a background database of drug-PMID associations. Furthermore, drug-drug similarity matrices generated from pairwise drug co-mentions from DrugRIF and pairwise cosine similarity of drug-induced gene expression profiles from SEP-L1000 (Wang et al., 2016) are used to predict novel drug-term associations.

## Acknowledgements

This work was partially supported by NIH grants U24CA224260, U54HL127624, U24CA224370, U24TR002278, U01CA239108 and OT2OD030546.

## Data Availability Statement

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

## Literature Cited

- Aguet F, Anand S, Ardlie KG, Gabriel S, Getz GA, Graubert A, Hadley K, Handsaker RE, Huang KH, Kashin S, et al. 2020. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science*. Available at: <https://science.sciencemag.org/content/369/6509/1318.abstract> [Accessed September 15, 2021].
- Armstrong JF, Faccenda E, Harding SD, Pawson AJ, Southan C, Sharman JL, Campo B, Cavanagh DR, Alexander SPH, Davenport AP, et al. 2019. The IUPHAR/BPS Guide to PHARMACOLOGY in 2020: extending immunopharmacology content and introducing the IUPHAR/MMV Guide to MALARIA PHARMACOLOGY. *Nucleic Acids Research*. Available at: 10.1093/nar/gkz951.
- Avram S, Bologna CG, Holmes J, Bocci G, Wilson TB, Nguyen D-T, Curpan R, Halip L, Bora A, Yang JJ, et al. 2021. DrugCentral 2021 supports drug discovery and repositioning. *Nucleic Acids Research* 49:D1160–D1169. Available at: 10.1093/nar/gkaa997. [PubMed: 33151287]
- Avram S, Curpan R, Halip L, Bora A, and Oprea TI 2020. Off-Patent Drug Repositioning. *Journal of chemical information and modeling* 60:5746–5753. [PubMed: 32877182]
- Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, Wilson CJ, Lehár J, Kryukov GV, Sonkin D, et al. 2012. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* 483:603–607. [PubMed: 22460905]
- Benet LZ, Broccatelli F, and Oprea TI 2011. BDDCS applied to over 900 drugs. *The AAPS journal* 13:519–547. [PubMed: 21818695]
- Berginski ME, Moret N, Liu C, Goldfarb D, Sorger PK, and Gomez SM 2021. The Dark Kinase Knowledgebase: an online compendium of knowledge and experimental results of understudied kinases. *Nucleic acids research* 49:D529–D535. [PubMed: 33079988]
- Bhattacharyya SB 2016. Overview of SNOMED CT. In *Introduction to SNOMED CT* pp. 1–2. Springer Singapore, Singapore.
- Bocci G, Bradfute SB, Ye C, Garcia MJ, Parvathareddy J, Reichard W, Surendranathan S, Bansal S, Bologna CG, Perkins DJ, et al. 2020. Virtual and In Vitro Antiviral Screening Revive Therapeutic Drugs for COVID-19. *ACS pharmacology & translational science* 3:1278–1292. [PubMed: 33330842]
- Cai D-C, Fonteijn H, Guadalupe T, Zwiers M, Wittfeld K, Teumer A, Hoogman M, Arias-Vásquez A, Yang Y, Buitelaar J, et al. 2014. A genome-wide search for quantitative trait loci affecting the cortical surface area and thickness of Heschl's gyrus. *Genes, Brain and Behavior* 13:675–685. Available at: 10.1111/gbb.12157. [PubMed: 25130324]
- Cannon DC, Yang JJ, Mathias SL, Ursu O, Mani S, Waller A, Schürer SC, Jensen LJ, Sklar LA, Bologna CG, et al. 2017. TIN-X: target importance and novelty explorer. *Bioinformatics* 33:2601–2603. [PubMed: 28398460]
- Clarke DJB, Jeon M, Stein DJ, Moiseyev N, Kropiwnicki E, Dai C, Xie Z, Wojciechowicz ML, Litz S, Hom J, et al. 2021. Appytters: Turning Jupyter Notebooks into data-driven web apps. *Patterns (New York, N.Y.)* 2:100213.
- Contrera JF, Matthews EJ, Kruhlak NL, and Benz RD 2004. Estimating the safe starting dose in phase I clinical trials and no observed effect level based on QSAR modeling of the human maximum recommended daily dose. *Regulatory toxicology and pharmacology: RTP* 40:185–206. [PubMed: 15546675]

- Cowley GS, Weir BA, Vazquez F, Tamayo P, Scott JA, Rusin S, East-Seletsky A, Ali LD, Gerath WF, Pantel SE, et al. 2014. Parallel genome-scale loss of function screens in 216 cancer cell lines for the identification of context-specific genetic dependencies. *Scientific data* 1:140035.
- DailyMed 2015. Choice 52:52–5368–52–5368.
- Disease ontology - institute for genome sciences @ university of Maryland Available at: <https://disease-ontology.org/> [Accessed August 31, 2021].
- East MP, and Asquith CRM 2021. CDC42BPA/MRCK $\alpha$ : a kinase target for brain, ovarian and skin cancers. *Nature reviews. Drug discovery* 20:167.
- EBI Web Team ChEBI. Available at: <https://www.ebi.ac.uk/chebi/init.do> [Accessed August 31, 2021].
- Edgar R, Domrachev M, and Lash AE 2002. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic acids research* 30:207–210. [PubMed: 11752295]
- Egaña LA, Cuevas RA, Baust TB, Parra LA, Leak RK, Hochendoner S, Peña K, Quiroz M, Hong WC, Dorostkar MM, et al. 2009. Physical and functional interaction between the dopamine transporter and the synaptic vesicle protein synaptogyrin-3. *Journal of Neuroscience* 29:4592–4604. [PubMed: 19357284]
- Fernandez NF, Gundersen GW, Rahman A, Grimes ML, Rikova K, Hornbeck P, and Ma'ayan A. 2017. Clustergrammer, a web-based heatmap visualization and analysis tool for high-dimensional biological data. *Scientific data* 4:170151.
- Hopkins AL, and Groom CR 2002. The druggable genome. *Nature reviews. Drug discovery* 1:727–730. [PubMed: 12209152]
- Huang L, Zalkikar J, and Tiwari RC 2011. A Likelihood Ratio Test Based Method for Signal Detection With Application to FDA's Drug Safety Data. *Journal of the American Statistical Association* 106:1230–1241. Available at: 10.1198/jasa.2011.ap10243.
- Huang X-P, Karpiak J, Kroeze WK, Zhu H, Chen X, Moy SS, Sadoris KA, Nikolova VD, Farrell MS, Wang S, et al. 2015. Allosteric ligands for the pharmacologically dark receptors GPR68 and GPR65. *Nature* 527:477–483. [PubMed: 26550826]
- Jeon M, Jagodnik KM, Kropiwnicki E, Stein DJ, and Ma'ayan A. 2021. Prioritizing Pain-Associated Targets with Machine Learning. *Biochemistry* 60:1430–1446. [PubMed: 33606503]
- Johns MA, Russ A, and Fu HA 2012. Current drug targets and the druggable genome. *Chemical Genomics*:320–331.
- Kc GB, Bocci G, Verma S, Hassan MM, Holmes J, Yang JJ, Sirimulla S, and Oprea TI 2021. A machine learning platform to estimate anti-SARS-CoV-2 activities. *Nature Machine Intelligence* 3:527–535. Available at: 10.1038/s42256-021-00335-w.
- Kc G, Bocci G, Verma S, Hassan M, Holmes J, Yang J, Sirimulla S, and Oprea TI 2020. REDIAL-2020: A suite of machine learning models to estimate anti-SARS-CoV-2 activities. *ChemRxiv*. Available at: [https://chemrxiv.org/articles/preprint/REDIAL-2020\\_A\\_Suite\\_of\\_Machine\\_Learning\\_Models\\_to\\_Estimate\\_Anti-SARS-CoV-2\\_Activities/12915779](https://chemrxiv.org/articles/preprint/REDIAL-2020_A_Suite_of_Machine_Learning_Models_to_Estimate_Anti-SARS-CoV-2_Activities/12915779).
- Kroeze WK, Sassano MF, Huang X-P, Lansu K, McCorvy JD, Giguère PM, Sciaky N, and Roth BL 2015. PRESTO-Tango as an open-source resource for interrogation of the druggable human GPCRome. *Nature structural & molecular biology* 22:362–369.
- Kropiwnicki E, Evangelista JE, Stein DJ, Clarke DJB, Lachmann A, Kuleshov MV, Jeon M, Jagodnik KM, and Ma'ayan A. 2021. Drugmonizome and Drugmonizome-ML: integration and abstraction of small molecule attributes for drug enrichment analysis and machine learning. *Database: the journal of biological databases and curation* 2021. Available at: 10.1093/database/baab017.
- Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, Koplev S, Jenkins SL, Jagodnik KM, Lachmann A, et al. 2016. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic acids research* 44:W90–7. [PubMed: 27141961]
- Kuleshov MV, Xie Z, London ABK, Yang J, Evangelista JE, Lachmann A, Shu I, Torre D, and Ma'ayan A. 2021. KEA3: improved kinase enrichment analysis via data integration. *Nucleic acids research* 49:W304–W316. [PubMed: 34019655]
- Lachmann A, Rizzo K, Bartal A, Jeon M, and Clarke DJB 2021. PrismExp: Predicting Human Gene Function by Partitioning Massive RNA-seq Co-expression Data. *bioRxiv*. Available at: <https://www.biorxiv.org/content/10.1101/2021.01.20.427528v1.abstract>.

- Lachmann A, Schilder BM, Wojciechowicz ML, Torre D, Kuleshov MV, Keenan AB, and Ma'ayan A. 2019. Geneshot: search engine for ranking genes from arbitrary text queries. *Nucleic acids research* 47:W571–W577. [PubMed: 31114885]
- Lachmann A, Torre D, Keenan AB, Jagodnik KM, Lee HJ, Wang L, Silverstein MC, and Ma'ayan A. 2018. Massive mining of publicly available RNA-seq data from human and mouse. *Nature Communications* 9. Available at: 10.1038/s41467-018-03751-6.
- Langfelder P, and Horvath S. 2008. WGCNA: an R package for weighted correlation network analysis. *BMC bioinformatics* 9:559. [PubMed: 19114008]
- Lin Y, Mehta S, Küçük-McGinty H, Turner JP, Vidovic D, Forlin M, Koleti A, Nguyen D-T, Jensen LJ, Guha R, et al. 2017. Drug target ontology to classify and integrate drug discovery data. *Journal of biomedical semantics* 8:50. [PubMed: 29122012]
- Lipinski CA, Lombardo F, Dominy BW, and Feeney PJ 2001. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings IPII of original article: S0169–409X(96)00423–1. The article was originally published in *Advanced Drug Delivery Reviews* 23 (1997) 3–25. 1. *Advanced Drug Delivery Reviews* 46:3–26. Available at: 10.1016/s0169-409x(00)00129-0. [PubMed: 11259830]
- Lombardo F, Berellini G, and Obach RS 2018. Trend Analysis of a Database of Intravenous Pharmacokinetic Parameters in Humans for 1352 Drug Compounds. *Drug metabolism and disposition: the biological fate of chemicals* 46:1466–1477. [PubMed: 30115648]
- Maglott D, Ostell J, Pruitt KD, and Tatusova T. 2011. Entrez Gene: gene-centered information at NCBI. *Nucleic acids research* 39:D52–7. [PubMed: 21115458]
- Mendez D, Gaulton A, Bento AP, Chambers J, De Veij M, Félix E, Magariños MP, Mosquera JF, Mutowo P, Nowotka M, et al. 2019. ChEMBL: towards direct deposition of bioassay data. *Nucleic acids research* 47:D930–D940. [PubMed: 30398643]
- MeSH Browser Available at: <https://meshb.nlm.nih.gov/record/ui?name=Unified%20Medical%20Language%20System> [Accessed August 31, 2021].
- Miller MB, Yan Y, Machida K, Kiraly DD, Levy AD, Wu YI, Lam TT, Abbott T, Koleske AJ, Eipper BA, et al. 2017. Brain Region and Isoform-Specific Phosphorylation Alters Kalirin SH2 Domain Interaction Sites and Calpain Sensitivity. *ACS chemical neuroscience* 8:1554–1569. [PubMed: 28418645]
- Milletti F, Storchi L, Goracci L, Bendels S, Wagner B, Kansy M, and Cruciani G. 2010. Extending pKa prediction accuracy: high-throughput pKa measurements to understand pKa modulation of new chemical series. *European journal of medicinal chemistry* 45:4270–4279. [PubMed: 20633962]
- National drug file - reference terminology source information 2016. Available at: <https://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/NDFRT/index.html> [Accessed August 31, 2021].
- Nguyen D-T, Mathias S, Bologa C, Brunak S, Fernandez N, Gaulton A, Hersey A, Holmes J, Jensen LJ, Karlsson A, et al. 2017. Pharos: Collating protein information to shed light on the druggable genome. *Nucleic Acids Research* 45:D995–D1002. Available at: 10.1093/nar/gkw1072. [PubMed: 27903890]
- Oprea TI, Bologa CG, Brunak S, Campbell A, Gan GN, Gaulton A, Gomez SM, Guha R, Hersey A, Holmes J, et al. 2018. Erratum: Unexplored therapeutic opportunities in the human genome. *Nature Reviews Drug Discovery* 17:377–377. Available at: 10.1038/nrd.2018.52.
- Orange book: Approved drug products with therapeutic equivalence evaluations Available at: <https://www.accessdata.fda.gov/scripts/cder/ob/index.cfm> [Accessed August 31, 2021].
- Osborne JD, Lin S, Kibbe WA, Zhu L, Danila MI, and Chisholm RL 2007. GeneRIF is a more comprehensive, current and computationally tractable source of gene-disease relationships than OMIM. *Bioinformatics Core, Northwestern University Technical Report*. Available at: <https://www.academia.edu/download/37808069/geneRIFDO16.pdf>.
- Pletscher-Frankild S, and Jensen LJ 2019. Design, implementation, and operation of a rapid, robust named entity recognition web service. *Journal of cheminformatics* 11:19. [PubMed: 30850898]
- Rouillard AD, Gundersen GW, Fernandez NF, Wang Z, Monteiro CD, McDermott MG, and Ma'ayan A. 2016. The harmonizome: a collection of processed datasets gathered to serve and mine



knowledge about genes and proteins. Database: the journal of biological databases and curation 2016. Available at: [10.1093/database/baw100](https://doi.org/10.1093/database/baw100).

- Russ AP, and Lampel S. 2005. The druggable genome: an update. *Drug discovery today* 10:1607–1610. [PubMed: 16376820]
- RxNorm 2004. Available at: <https://www.nlm.nih.gov/research/umls/rxnorm/index.html> [Accessed August 31, 2021].
- Santos R, Ursu O, Gaulton A, Bento AP, Donadi RS, Bologa CG, Karlsson A, Al-Lazikani B, Hersey A, Oprea TI, et al. 2017. A comprehensive map of molecular drug targets. *Nature reviews. Drug discovery* 16:19–34. [PubMed: 27910877]
- Schriml LM, Mitraka E, Munro J, Tauber B, Schor M, Nickle L, Felix V, Jeng L, Bearer C, Lichenstein R, et al. 2019. Human Disease Ontology 2018 update: classification, content and workflow expansion. *Nucleic acids research* 47:D955–D962. [PubMed: 30407550]
- Sheils TK, Mathias SL, Kelleher KJ, Siramshetty VB, Nguyen D-T, Bologa CG, Jensen LJ, Vidovi D, Koleti A, Schürer SC, et al. 2021. TCRD and Pharos 2021: mining the human proteome for disease biology. *Nucleic acids research* 49:D1334–D1346. [PubMed: 33156327]
- Sheils T, Mathias SL, Siramshetty VB, Bocci G, Bologa CG, Yang JJ, Waller A, Southall N, Nguyen D-T, and Oprea TI 2020. How to illuminate the Druggable Genome Using Pharos. *Current protocols in bioinformatics / editorial board, Andreas D. Baxevanis ... [et al.]* 69:e92.
- Si L, Bai H, Rodas M, Cao W, Oh CY, Jiang A, Moller R, Hoagland D, Oishi K, Horiuchi S, et al. 2021. A human-airway-on-a-chip for the rapid identification of candidate antiviral therapeutics and prophylactics. *Nature biomedical engineering* 5:815–829.
- Stein RM, Kang HJ, McCorvy JD, Glatfelter GC, Jones AJ, Che T, Slocum S, Huang X-P, Savych O, Moroz YS, et al. 2020. Virtual discovery of melatonin receptor ligands to modulate circadian rhythms. *Nature* 579:609–614. [PubMed: 32040955]
- Subramanian A, Narayan R, Corsello SM, Peck DD, Natoli TE, Lu X, Gould J, Davis JF, Tubelli AA, Asiedu JK, et al. 2017. A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. *Cell* 171:1437–1452.e17. [PubMed: 29195078]
- The Gene Ontology Consortium 2019. The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic acids research* 47:D330–D338. [PubMed: 30395331]
- Tomczak K, Czerwińska P, and Wiznerowicz M. 2015. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemporary oncology* 19:A68–77. [PubMed: 25691825]
- Ursu O, Holmes J, Bologa CG, Yang JJ, Mathias SL, Stathias V, Nguyen D-T, Schürer S, and Oprea T. 2019. DrugCentral 2018: an update. *Nucleic acids research* 47:D963–D970. [PubMed: 30371892]
- Ursu O, Holmes J, Knockel J, Bologa CG, Yang JJ, Mathias SL, Nelson SJ, and Oprea TI 2017. DrugCentral: online drug compendium. *Nucleic Acids Research* 45:D932–D939. Available at: [10.1093/nar/gkw993](https://doi.org/10.1093/nar/gkw993). [PubMed: 27789690]
- Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, et al. 2001. The sequence of the human genome. *Science* 291:1304–1351. [PubMed: 11181995]
- Wang TA, Chen C, Huang F, Feng S, Tien J, Braz JM, Basbaum AI, Jan YN, and Jan LY 2021. TMEM16C is involved in thermoregulation and protects rodent pups from febrile seizures. *Proceedings of the National Academy of Sciences of the United States of America* 118. Available at: [10.1073/pnas.2023342118](https://doi.org/10.1073/pnas.2023342118).
- Wang Z, Clark NR, and Ma'ayan A. 2016. Drug-induced adverse events prediction with the LINCS L1000 data. *Bioinformatics* 32:2338–2345. [PubMed: 27153606]
- Weininger D. 1988. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of chemical information and computer sciences* 28:31–36.
- Wells CI, Al-Ali H, Andrews DM, Asquith CRM, Axtman AD, Dikic I, Ebner D, Ettmayer P, Fischer C, Frederiksen M, et al. 2021. The Kinase Chemogenomic Set (KCGS): An Open Science Resource for Kinase Vulnerability Identification. *International journal of molecular sciences* 22. Available at: [10.3390/ijms22020566](https://doi.org/10.3390/ijms22020566).
- WHOC WHOCC - ATC/DDD Index. Available at: [https://www.whocc.no/atc\\_ddd\\_index/](https://www.whocc.no/atc_ddd_index/) [Accessed August 31, 2021].

Witoelar A, Jansen IE, Wang Y, Desikan RS, Gibbs JR, Blauwendraat C, Thompson WK, Hernandez DG, Djurovic S, Schork AJ, et al. 2017. International Parkinson's Disease Genomics Consortium NABEC and United Kingdom Brain Expression Consortium I. Genome-wide pleiotropy between parkinson disease and autoimmune diseases. *JAMA neurology* 74:780–792. [PubMed: 28586827]

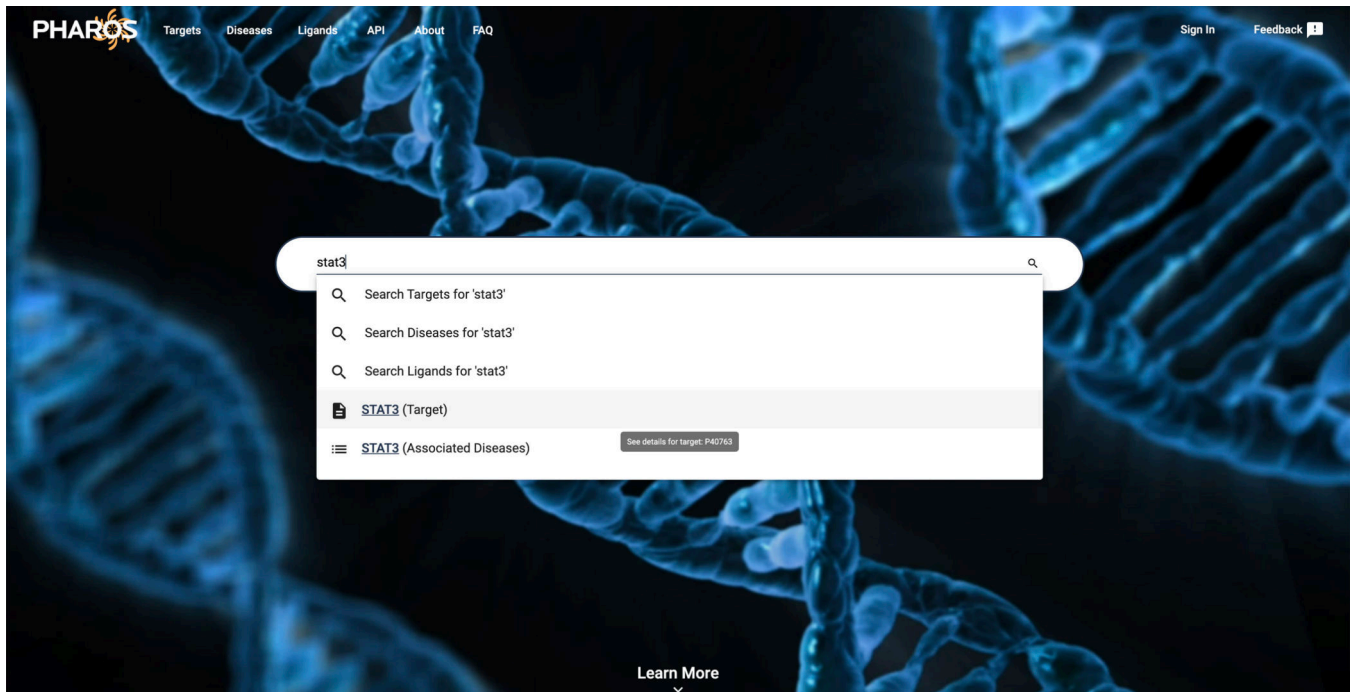
Yang JJ, Grissa D, Lambert CG, Bologa CG, Mathias SL, Waller A, Wild DJ, Jensen LJ, and Oprea TI 2021. TIGA: Target illumination GWAS analytics. *Bioinformatics* . Available at: [10.1093/bioinformatics/btab427](https://doi.org/10.1093/bioinformatics/btab427).

Author Manuscript

Author Manuscript

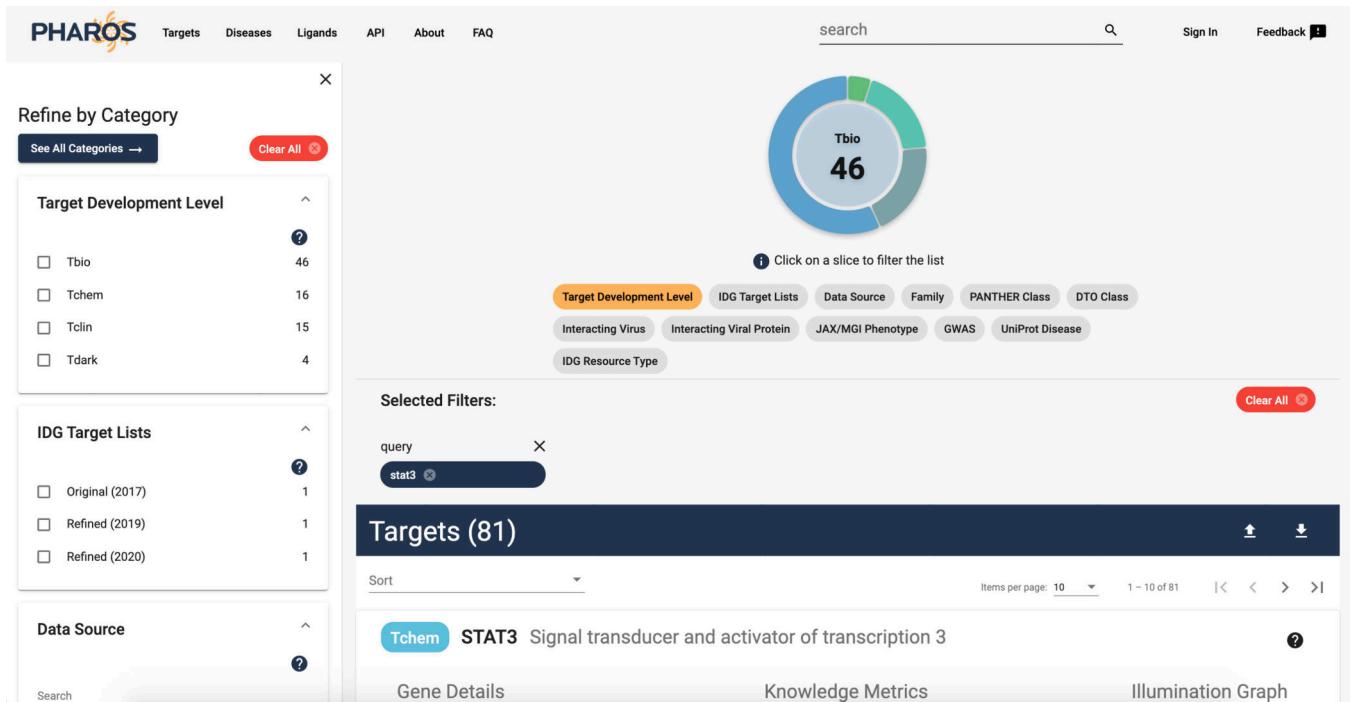
Author Manuscript

Author Manuscript



**Figure 1.**  
Typeahead search results for STAT3 scroll or arrow down to view more options.





**Figure 2.**  
Search Targets for STAT3 search results page.

PHAROS Targets Diseases Ligands API About FAQ search Q Sign In Feedback

Tchem **STAT3** Signal transducer and activator of transcription 3

**Jump to section:**

- Protein Classes
- Protein Summary**
- IDG Development Level Summary
- Approved Drugs
- Active Ligands
- Disease Associations by Source
- GWAS Traits
- PDB Viewer
- Pathways
- Gene Ontology Terms
- Predicted Viral Interactions

### Protein Summary

**Description**

Signal transducer and transcription activator that mediates cellular responses to interleukins, KITLG/SCF, LEP and other growth factors (PubMed:10688651, PubMed:12359225, PubMed:12873986, PubMed:15194700, PubMed:17344214, PubMed:18242580, PubMed:23084476). Once activated, recruits coactivators, such as NCOA1 or MED1, to the promoter region of the target gene (PubMed:17344214). May mediate cellular responses to activated FGFR1, FGFR2, FGFR3 and FGFR4 (PubMed:12873986). Binds to the interleukin-6 (IL-6)-responsive elements identified in the promoters of various acute-phase protein genes (PubMed:12359225). Activated by IL31 through IL31RA (PubMed:15194700). Acts as a regulator of inflammatory response by regulating differentiation of naive CD4(+) T-cells into T-helper Th17 or regulatory T-cells (Treg); deacetylation and oxidation of lysine residues by LOXL3, leads to disrupt STAT3 dimerization and inhibit its transcription activity (PubMed:28065600). Involved in cell cycle regulation by l...more

**Uniprot Accession IDs**  
**P40763** **A8K7B8** **K7ENL3** **O14916**  
**Q9BW54**

**Gene Name**  
**STAT3**

**Ensembl ID**  
ENST00000264657 ENSP00000264657  
ENSG00000168610 ENST00000404395  
ENSP00000384943 ENST00000585517  
ENSP00000467000 ENST00000588969  
ENSP00000467985

**Symbol**  
APRF APRF HIES ADMIO ADMIO1

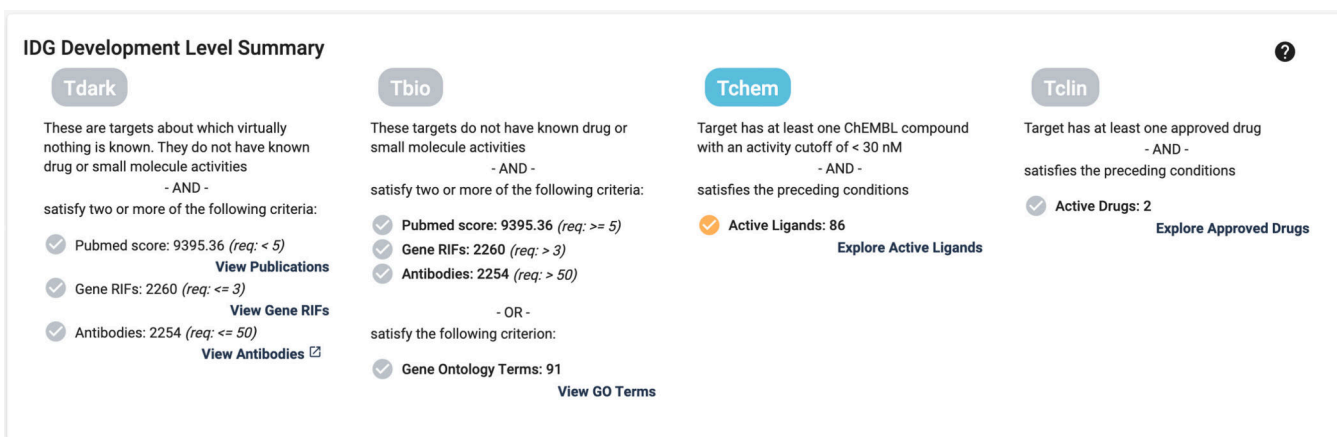
**IDG Partner Tools**

**Illumination Graph**

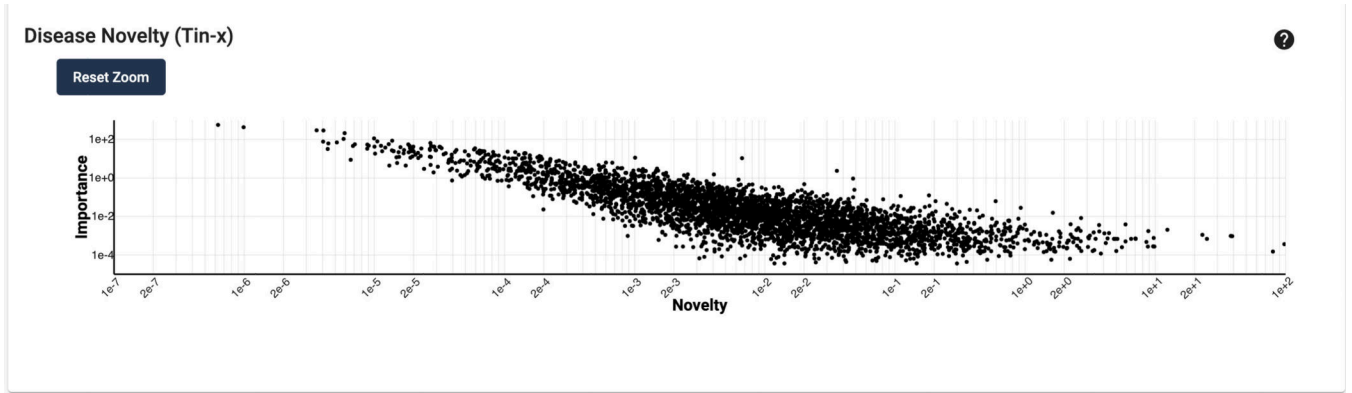
**Knowledge Table**

Most Knowledge About	Knowledge Value (0 to 1 scale)
biological process	1
biological term	1
chemical	1
disease perturbation	1
hub protein	1

**Figure 3.** Target details page for STAT3, the radar chart in the center depicts data from Harmonizome.



**Figure 4.** IDG development level summary section that shows the current development level, and criteria met. Links provide the ability to view either the original source, or the relevant data in Pharos.



**Figure 5.** Scatterplot depicting Tin-x data for STAT3. Hovering over a data point opens up a tooltip, providing novelty and importance data for the disease.

GWAS Traits (9)

• [Explore on Target Illumination GWAS Analytics \(TIGA\)](#)

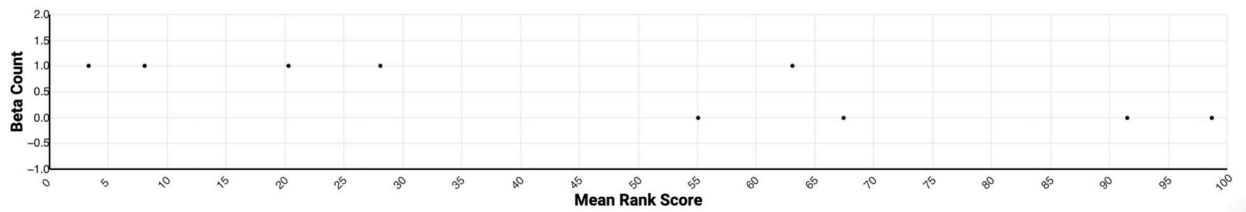
Items per page: 5

1 - 5 of 9

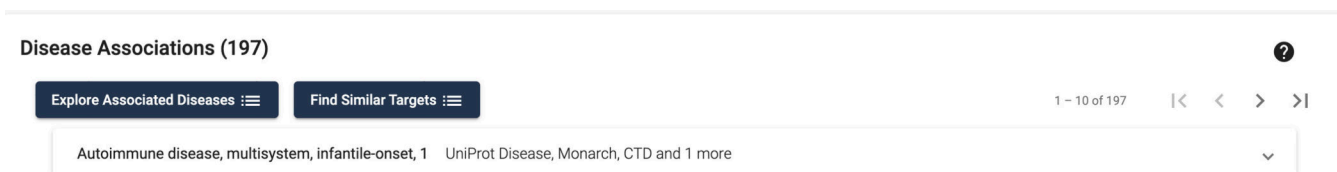
Navigation icons: |< < > >|

GWAS Trait	EFO ID	Study Count	SNP Count	Beta Count	Odds Ratio	Evidence (Mean Rank Score)	Provenance
multiple sclerosis	<a href="#">EFO_0003885</a>	5	5	0	1.1	98.7	<input checked="" type="checkbox"/>
Crohn's disease	<a href="#">EFO_0000384</a>	3	3	0	1.2	91.5	<input checked="" type="checkbox"/>
inflammatory bowel disease	<a href="#">EFO_0003767</a>	1	1	0	1.1	67.4	<input checked="" type="checkbox"/>
C-reactive protein measurement	<a href="#">EFO_0004458</a>	1	1	1		63.1	<input checked="" type="checkbox"/>
ulcerative colitis	<a href="#">EFO_0000729</a>	1	1	0	1.1	55.1	<input checked="" type="checkbox"/>

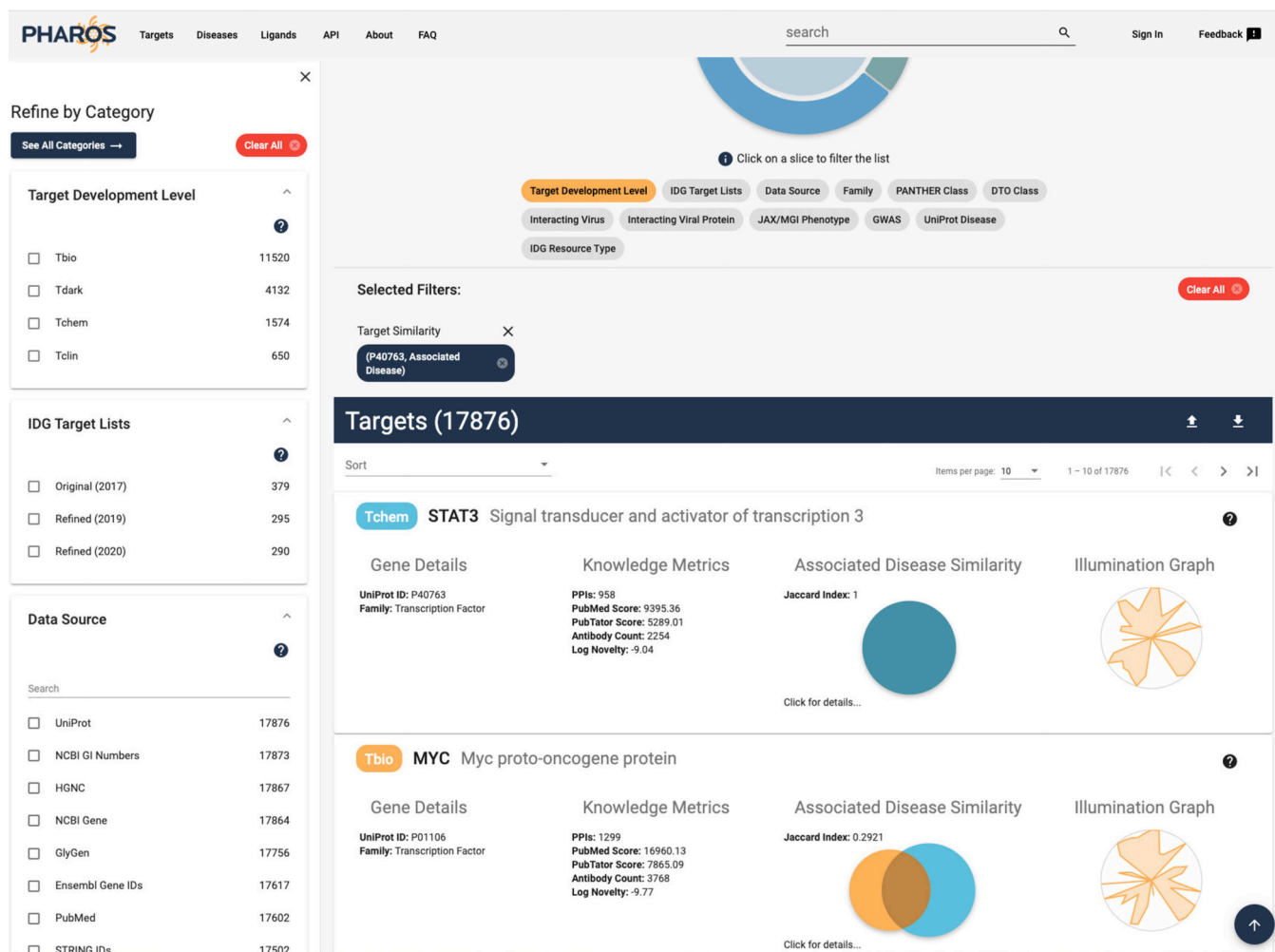
Reset Zoom  Beta Count  Odds Ratio



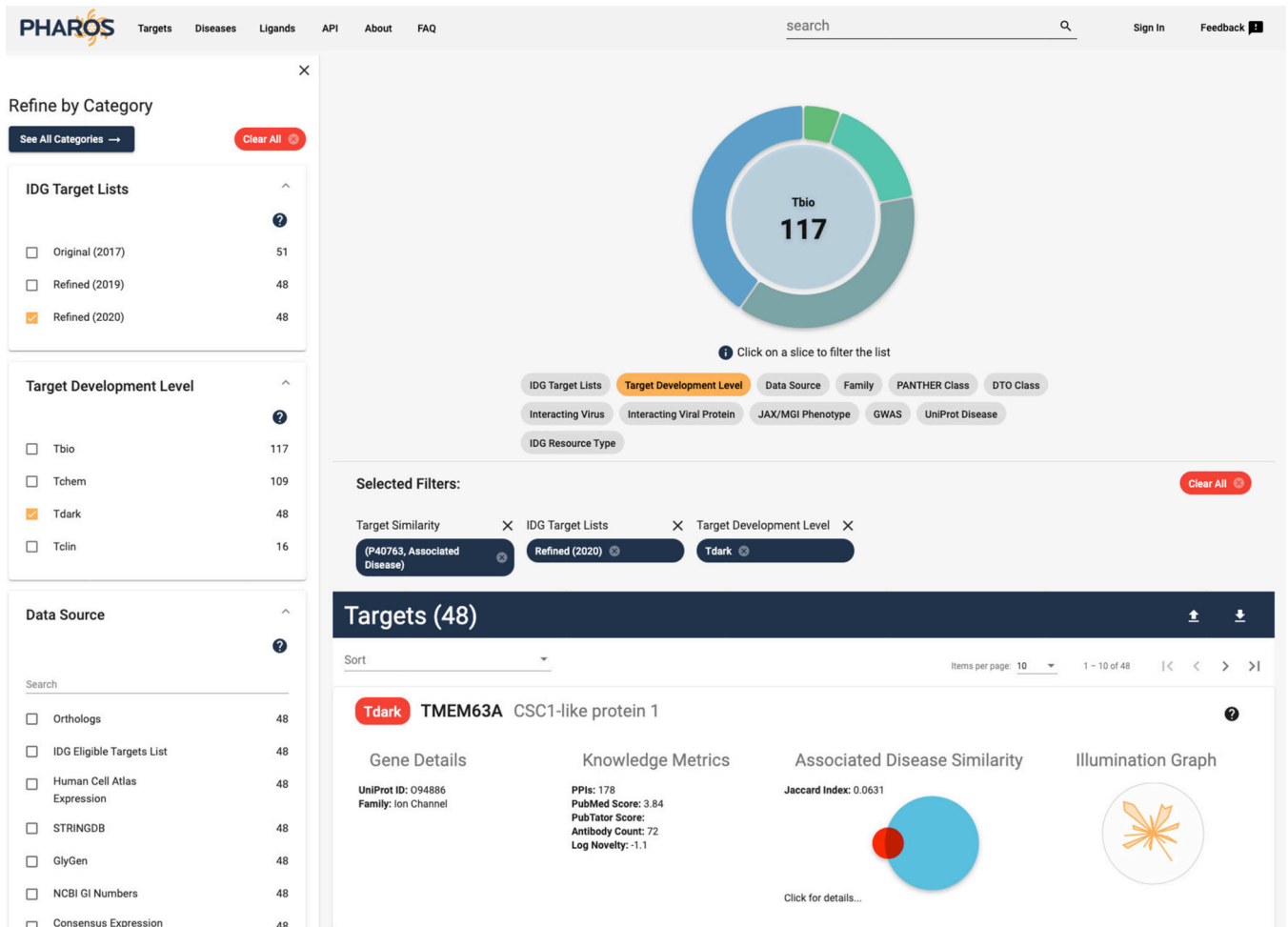
**Figure 6.** GWAS traits, and the associated TIGA scatterplot. For a more in depth exploration of this data, click “Explore on Target Illumination GWAS Analytics”.



**Figure 7.** Additional functions available within Pharos are shown within blue buttons. Users can click to browse filtered lists for targets similar to the current target, or associated diseases or ligands.

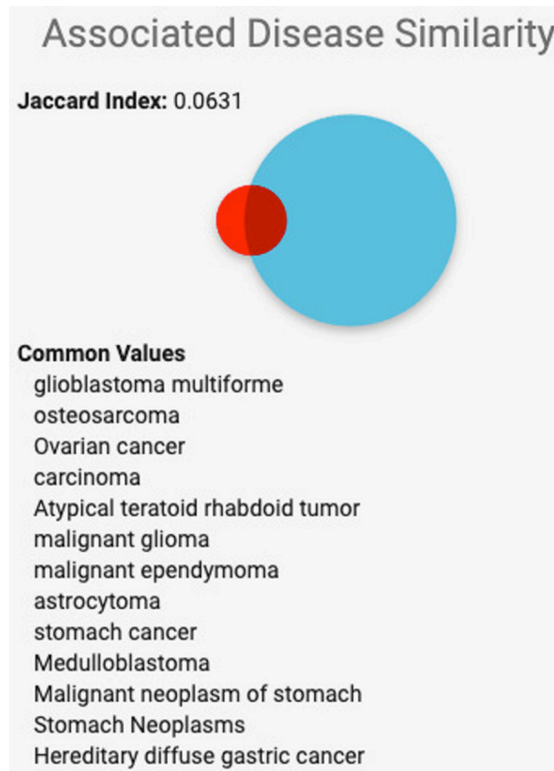


**Figure 8.** List of targets that share associated diseases with STAT3. The Jaccard index is a numerical value of the ratio of overlap between the associated diseases of the target in relation to the original target (STAT3). The Venn diagram is a visual representation of the ratio with the TDL level color coded.



**Figure 9.** The target list from Figure 8 filtered to display Target Development Level of Tdark, and on the Refined(2020) IDG target lists. Click on “Click for details...” to view an expanded list of the overlapping values.





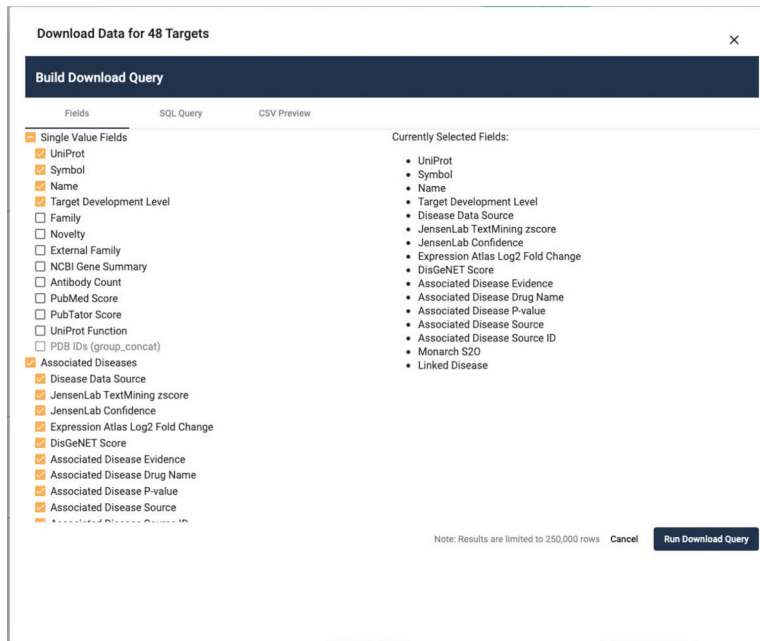
**Figure 10.** Expanded view of the Associated Disease Similarity section of the target card.

## Targets (48)



**Figure 11.**

Target toolbar illustrating the download button on the right side. To the left of the download button is the upload button, which allows for the uploading of custom lists, to explore in the Pharos interface.

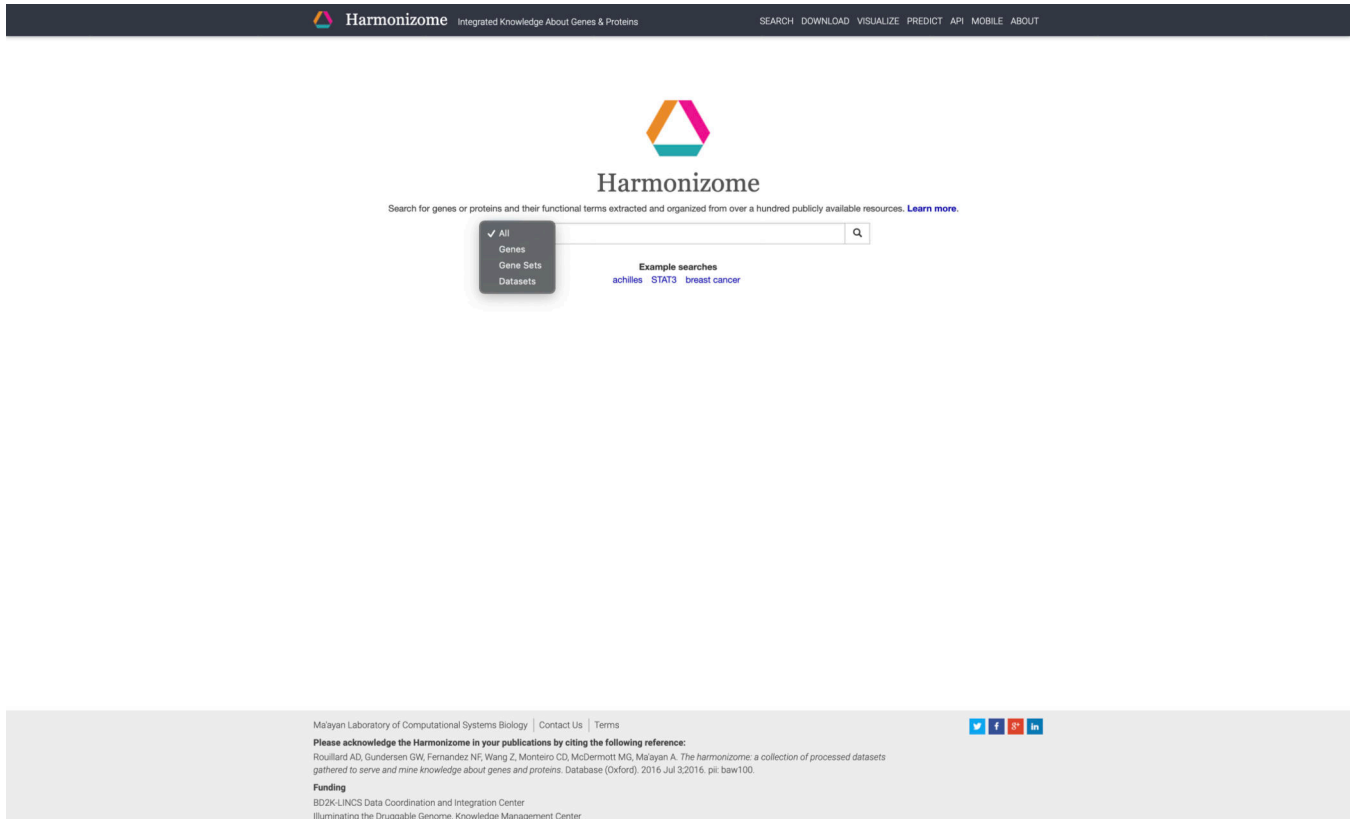


**Figure 12.** Pop-up window featuring the query builder which allows for the download of Pharos list data as a csv file. Subsequent tabs display the raw SQL query used to generate the data, as well as a 10 line preview.

The screenshot shows the PHAROS GraphQL sandbox interface. At the top, there is a navigation bar with links for Targets, Diseases, Ligands, API, About, and FAQ. A search bar is located on the right side of the top bar. Below the navigation bar, the interface is split into three columns:

- Left Column:** Contains the 'Pharos API' header, a welcome message, and a link to 'Learn more about GraphQL: GraphQL.org'. Below this is a section titled 'Example queries' with a list of dropdown menus: Target details, Disease details, Ligand details, Protein-protein interactions, Disease associations, Facets for all targets, Facets for an associated disease, Filtering by facet values, Fetching uncommon facets, and All Facet Names.
- Center Column:** A code editor showing a GraphQL query. The query is a batch query with a filter and a results section. The filter includes facets for 'Target Development Level' and 'IDG Target Lists'. The results section includes a count and a list of targets with similarity details.
- Right Column:** A JSON response showing the results of the query. It includes a 'data' object with a 'batch' containing 'results' and 'targets'. The 'targets' list includes detailed information for two targets, such as 'CSCI-like protein 1' and 'Transmembrane channel-like protein 5', including their gene names, accessions, and similarity details.

**Figure 13.** GraphQL sandbox interface. Examples on the left side, and documentation on the right allow for highly customizable data requests.



**Figure 14.** The Harmonizome homepage. The filter dropdown menu on the left selects between searching for genes, gene sets, and datasets.

Harmonizome All STAT3 Q

SEARCH DOWNLOAD VISUALIZE PREDICT API MOBILE ABOUT

Filters: **Gene** **Gene Set**

Results for "STAT3": 1 gene, 75 gene sets

Show  entries Filter

**STAT3** Gene  
*signal transducer and activator of transcription 3 (acute-phase response factor)*  
 The protein encoded by this gene is a member of the STAT protein family. In response to cytokines and growth factors, STAT family members are phosphorylated by the receptor associated kinases, and then form homo- or heterodimers that translocate to the cell nucleus where they act as transcription activators. This protein is activated through phosphorylation in response to various cytokines and growth factors including IFNs, EGF, IL5, IL6, HGF, LIF and BMP2. This protein mediates the expression of a variety of genes in response to cell stimuli, and thus plays a key role in many cellular processes such as cell growth and apoptosis. The small GTPase Rac1 has been shown to bind and regulate the activity of this protein. PIAS3 protein is a specific inhibitor of this protein. Three alternatively spliced transcript variants encoding distinct isoforms have been described. [provided by RefSeq, Jul 2008]

---

**STAT3** Gene Set  
*From CHEA Transcription Factor Targets*  
 target genes of the STAT3 transcription factor in low- or high-throughput transcription factor functional studies from the CHEA Transcription Factor Targets dataset.

---

**STAT3** Gene Set  
*From ENCODE Transcription Factor Targets*  
 target genes of the STAT3 transcription factor in ChIP-seq datasets from the ENCODE Transcription Factor Targets dataset.

---

**stat3** Gene Set  
*From GeneRIF Biological Term Annotations*  
 genes co-occurring with the biological term **stat3** in literature-supported statements describing functions of genes from the GeneRIF Biological Term Annotations dataset.

---

**STAT3** Gene Set  
*From Hub Proteins Protein-Protein Interactions*  
 interacting proteins for hub protein STAT3 from the curated Hub Proteins Protein-Protein Interactions dataset.

---

**STAT3** Gene Set  
*From JASPAR Predicted Transcription Factor Targets*  
 target genes of the STAT3 transcription factor predicted using known transcription factor binding site motifs from the JASPAR Predicted Transcription Factor Targets dataset.

---

**STAT3** Gene Set  
*From MotifMap Predicted Transcription Factor Targets*  
 target genes of the STAT3 transcription factor predicted using known transcription factor binding site motifs from the MotifMap Predicted Transcription Factor Targets dataset.

---

**STAT3** Gene Set  
*From Pathway Commons Protein-Protein Interactions*  
 interacting proteins for STAT3 from the Pathway Commons Protein-Protein Interactions dataset.

---

**stat3** Gene Set  
*From Phosphosite Textmining Biological Term Annotations*  
 proteins co-occurring with the biological term **stat3** in abstracts of publications describing phosphosites from the Phosphosite Textmining Biological Term Annotations dataset.

**Figure 15.**

Search result page after querying "STAT3". One gene page and 75 gene set pages match the query term "STAT3".

**STAT3** Gene

HGNC Family	<a href="#">SH2 domain containing</a>
Name	Signal transducer and activator of transcription 3 (acute-phase response factor)
Description	The protein encoded by this gene is a member of the STAT protein family. In response to cytokines and growth factors, STAT family members are phosphorylated by the receptor associated kinases, and then form homo- or heterodimers that translocate to the cell nucleus where they act as transcription activators. This protein is activated through phosphorylation in response to various cytokines and growth factors including IFNs, EGF, IL5, IL6, HGF, LIF and BMP2. This protein mediates the expression of a variety of genes in response to cell stimuli, and thus plays a key role in many cellular processes such as cell growth and apoptosis. The small GTPase Rac1 has been shown to bind and regulate the activity of this protein. PIAS3 protein is a specific inhibitor of this protein. Three alternatively spliced transcript variants encoding distinct isoforms have been described. [provided by RefSeq, Jul 2008]
Synonyms	ADMIO1, ADMIO, APRF, HIES
Proteins	<a href="#">STAT3_HUMAN</a>
NCBI Gene ID	<a href="#">6774</a>
API	
Download Associations	
Predicted Functions	<a href="#">ARCHS4</a>
Co-expressed Genes	<a href="#">ARCHS4</a>
Expression in Tissues and Cell Lines	<a href="#">ARCHS4</a>

**Figure 16.** STAT3 single gene landing page that includes identifying metadata for the gene, download links for retrieving functional association data, and gene-related information from ARCHS4.

















## Functional Associations

STAT3 has 14,374 functional associations with biological entities spanning 8 categories (molecular profile, organism, functional term, phrase or reference, disease, phenotype or trait, chemical, structural feature, cell line, cell type or tissue, gene, protein or microRNA) extracted from 100 datasets.

Click the + buttons to view associations for STAT3 from the datasets below.

If available, associations are ranked by **standardized value** 



	Dataset	Summary
	<a href="#">Achilles Cell Line Gene Essentiality Profiles</a>	Cell lines with fitness changed by STAT3 gene knockdown relative to other cell lines from the Achilles Cell Line Gene Essentiality Profiles dataset.
	<a href="#">Allen Brain Atlas Adult Human Brain Tissue Gene Expression Profiles</a>	Tissues with high or low expression of STAT3 gene relative to other tissues from the Allen Brain Atlas Adult Human Brain Tissue Gene Expression Profiles dataset.
	<a href="#">Allen Brain Atlas Adult Mouse Brain Tissue Gene Expression Profiles</a>	Tissues with high or low expression of STAT3 gene relative to other tissues from the Allen Brain Atlas Adult Mouse Brain Tissue Gene Expression Profiles dataset.
	<a href="#">Allen Brain Atlas Developing Human Brain Tissue Gene Expression Profiles by Microarray</a>	Tissue samples with high or low expression of STAT3 gene relative to other tissue samples from the Allen Brain Atlas Developing Human Brain Tissue Gene Expression Profiles by Microarray dataset.
	<a href="#">Allen Brain Atlas Developing Human Brain Tissue Gene Expression Profiles by RNA-seq</a>	Tissue samples with high or low expression of STAT3 gene relative to other tissue samples from the Allen Brain Atlas Developing Human Brain Tissue Gene Expression Profiles by RNA-seq dataset.
	<a href="#">Allen Brain Atlas Prenatal Human Brain Tissue Gene Expression Profiles</a>	Tissues with high or low expression of STAT3 gene relative to other tissues from the Allen Brain Atlas Prenatal Human Brain Tissue Gene Expression Profiles dataset.
	<a href="#">Biocarta Pathways</a>	Pathways involving STAT3 protein from the Biocarta Pathways dataset.
	<a href="#">BioGPS Cell Line Gene Expression Profiles</a>	Cell lines with high or low expression of STAT3 gene relative to other cell lines from the BioGPS Cell Line Gene Expression Profiles dataset.
	<a href="#">BioGPS Human Cell Type and Tissue Gene Expression Profiles</a>	Cell types and tissues with high or low expression of STAT3 gene relative to other cell types and tissues from the BioGPS Human Cell Type and Tissue Gene Expression Profiles dataset.
	<a href="#">BioGPS Mouse Cell Type and Tissue Gene Expression Profiles</a>	Cell types and tissues with high or low expression of STAT3 gene relative to other cell types and tissues from the BioGPS Mouse Cell Type and Tissue Gene Expression Profiles dataset.
	<a href="#">CCLE Cell Line Gene CNV Profiles</a>	Cell lines with high or low copy number of STAT3 gene relative to other cell lines from the CCLE Cell Line Gene CNV Profiles dataset.
	<a href="#">CCLE Cell Line Gene Expression Profiles</a>	Cell lines with high or low expression of STAT3 gene relative to other cell lines from the CCLE Cell Line Gene Expression Profiles dataset.
	<a href="#">CCLE Cell Line Gene Mutation Profiles</a>	Cell lines with STAT3 gene mutations from the CCLE Cell Line Gene Mutation Profiles dataset.
	<a href="#">CHEA Transcription Factor Binding Site Profiles</a>	Transcription factor binding site profiles with transcription factor binding evidence at the promoter of STAT3 gene from the CHEA Transcription Factor Binding Site Profiles dataset.

**Figure 17.**  
Expandable lists of functional associations for STAT3 from each dataset.

Harmonizome All

SEARCH DOWNLOAD VISUALIZE PREDICT API MOBILE ABOUT

## STAT3 Gene Set

Dataset	<a href="#">CHEA Transcription Factor Targets</a>
Category	Genomics
Type	Transcription factor
Description	Signal transducer and activator of transcription 3 (acute-phase response factor)The protein encoded by this gene is a member of the STAT protein family. In response to cytokines and growth factors, STAT family members are phosphorylated by the receptor associated kinases, and then form homo- or heterodimers that translocate to the cell nucleus where they act as transcription activators. This protein is activated through phosphorylation in response to various cytokines and growth factors including IFNs, EGF, IL5, IL6, HGF, LIF and BMP2. This protein mediates the expression of a variety of genes in response to cell stimuli, and thus plays a key role in many cellular processes such as cell growth and apoptosis. The small GTPase Rac1 has been shown to bind and regulate the activity of this protein. PIAS3 protein is a specific inhibitor of this protein. Three alternatively spliced transcript variants encoding distinct isoforms have been described. [provided by RefSeq, Jul 2008] ( <a href="#">NCBI Entrez Gene Database, 6774</a> )
External Link	<a href="http://www.ncbi.nlm.nih.gov/gene/6774">http://www.ncbi.nlm.nih.gov/gene/6774</a>
Similar Terms	<input type="text" value="Q"/>
Downloads & Tools	 

## Genes

6014 target genes of the **STAT3** transcription factor in low- or high-throughput transcription factor functional studies from the CHEA Transcription Factor Targets dataset.

Show  entries Filter

Symbol	Name
<a href="#">A2M</a>	alpha-2-macroglobulin
<a href="#">A4GALT</a>	alpha 1,4-galactosyltransferase
<a href="#">AAAS</a>	achalasia, adrenocortical insufficiency, alacrimia
<a href="#">AACS</a>	acetoacetyl-CoA synthetase
<a href="#">AAK1</a>	AP2 associated kinase 1
<a href="#">AAMP</a>	angio-associated, migratory cell protein
<a href="#">AARS</a>	alanyl-tRNA synthetase
<a href="#">AARS2</a>	alanyl-tRNA synthetase 2, mitochondrial
<a href="#">AASS</a>	aminoadipate-semialdehyde synthase
<a href="#">ABCA1</a>	ATP-binding cassette, sub-family A (ABC1), member 1
<a href="#">ABCA11P</a>	ATP-binding cassette, sub-family A (ABC1), member 11, pseudogene
<a href="#">ABCA12</a>	ATP-binding cassette, sub-family A (ABC1), member 12

**Figure 18.**  
STAT3 gene set page from CHEA Transcription Factor Targets dataset.

**Harmonizome** Integrated Knowledge About Genes & Proteins

SEARCH DOWNLOAD VISUALIZE PREDICT API MOBILE ABOUT

## Downloads

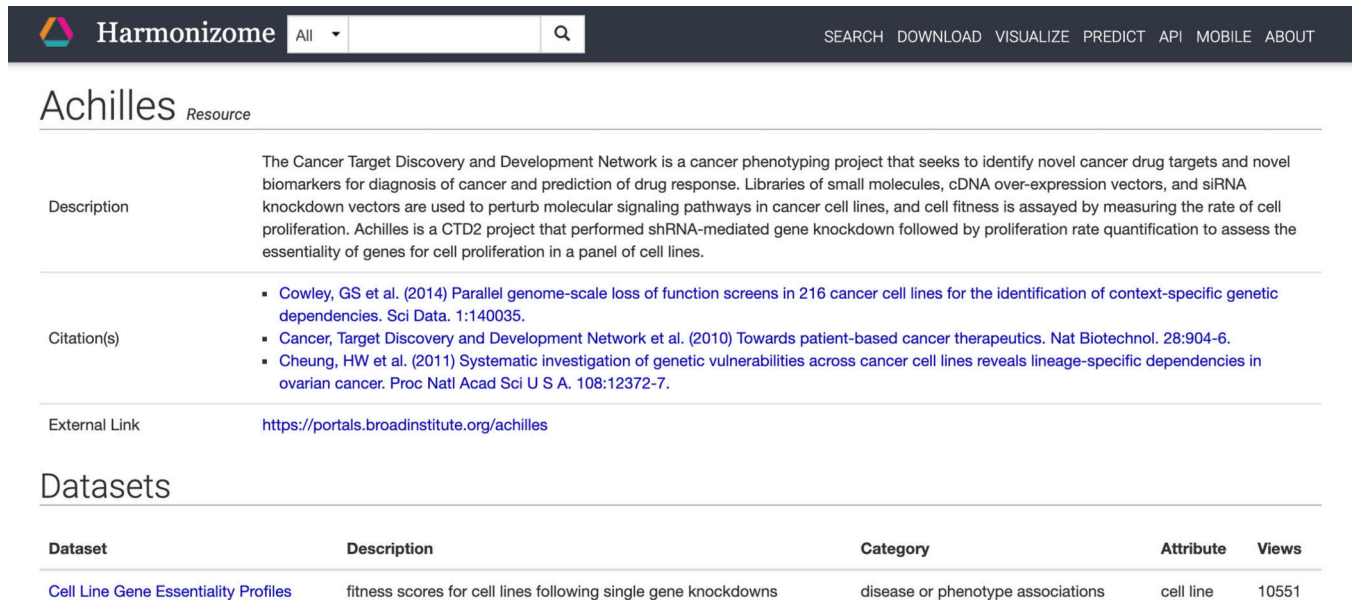
Click on a dataset to access its downloads. Click on a column header to sort the table by that column. Type in the search bar to filter.

To download the data programmatically, use [the API](#) or [this Python script](#).

Show  entries Filter

Resource	Dataset	Description	Category	Attribute	Views
<a href="#">Achilles</a>	<a href="#">Cell Line Gene Essentiality Profiles</a>	fitness scores for cell lines following single gene knockdowns	disease or phenotype associations	cell line	10551
<a href="#">Allen Brain Atlas</a>	<a href="#">Adult Human Brain Tissue Gene Expression Profiles</a>	mRNA expression profiles for 6 adult human brain tissue samples spanning ~300 brain structures	transcriptomics	tissue	15628
<a href="#">Allen Brain Atlas</a>	<a href="#">Adult Mouse Brain Tissue Gene Expression Profiles</a>	mRNA expression profiles for adult mouse brain tissues spanning ~2000 anatomically defined brain structures	transcriptomics	tissue	6762
<a href="#">Allen Brain Atlas</a>	<a href="#">Developing Human Brain Tissue Gene Expression Profiles by Microarray</a>	mRNA expression profiles for human brain tissue samples spanning 27 time points and 26 brain structures	transcriptomics	tissue sample	3684
<a href="#">Allen Brain Atlas</a>	<a href="#">Developing Human Brain Tissue Gene Expression Profiles by RNA-seq</a>	mRNA expression profiles for human brain tissue samples spanning 31 time points and 26 brain structures	transcriptomics	tissue sample	2870
<a href="#">Allen Brain Atlas</a>	<a href="#">Prenatal Human Brain Tissue Gene Expression Profiles</a>	mRNA expression profiles for 4 human prenatal brain tissue samples spanning 4 time points and ~300 brain structures	transcriptomics	tissue	4184
<a href="#">Biocarta</a>	<a href="#">Pathways</a>	sets of proteins participating in pathways	structural or functional annotations	pathway	23666
<a href="#">BioGPS</a>	<a href="#">Cell Line Gene Expression Profiles</a>	mRNA expression profiles for the NCI-60 panel of cancer cell lines	transcriptomics	cell line	19922
<a href="#">BioGPS</a>	<a href="#">Human Cell Type and Tissue Gene Expression Profiles</a>	mRNA expression profiles for human tissues and cell types	transcriptomics	cell type or tissue	9508
<a href="#">BioGPS</a>	<a href="#">Mouse Cell Type and Tissue Gene Expression Profiles</a>	mRNA expression profiles for mouse tissues and cell types	transcriptomics	cell type or tissue	7483
<a href="#">Cancer Cell Line Encyclopedia</a>	<a href="#">Cell Line Gene CNV Profiles</a>	gene-level copy number variation profiles for cancer cell lines	genomics	cell line	16851
<a href="#">Cancer Cell Line Encyclopedia</a>	<a href="#">Cell Line Gene Expression Profiles</a>	mRNA expression profiles for cancer cell lines	transcriptomics	cell line	59025

**Figure 19.**  
Download page for datasets included in Harmonizome.



**Harmonizome** All

SEARCH DOWNLOAD VISUALIZE PREDICT API MOBILE ABOUT

## Achilles Resource

**Description**

The Cancer Target Discovery and Development Network is a cancer phenotyping project that seeks to identify novel cancer drug targets and novel biomarkers for diagnosis of cancer and prediction of drug response. Libraries of small molecules, cDNA over-expression vectors, and siRNA knockdown vectors are used to perturb molecular signaling pathways in cancer cell lines, and cell fitness is assayed by measuring the rate of cell proliferation. Achilles is a CTD2 project that performed shRNA-mediated gene knockdown followed by proliferation rate quantification to assess the essentiality of genes for cell proliferation in a panel of cell lines.

**Citation(s)**

- Cowley, GS et al. (2014) Parallel genome-scale loss of function screens in 216 cancer cell lines for the identification of context-specific genetic dependencies. *Sci Data*. 1:140035.
- Cancer, Target Discovery and Development Network et al. (2010) Towards patient-based cancer therapeutics. *Nat Biotechnol*. 28:904-6.
- Cheung, HW et al. (2011) Systematic investigation of genetic vulnerabilities across cancer cell lines reveals lineage-specific dependencies in ovarian cancer. *Proc Natl Acad Sci U S A*. 108:12372-7.

**External Link** <https://portals.broadinstitute.org/achilles>

## Datasets

Dataset	Description	Category	Attribute	Views
<a href="#">Cell Line Gene Essentiality Profiles</a>	fitness scores for cell lines following single gene knockdowns	disease or phenotype associations	cell line	10551

**Figure 20.**  
Resource page for Achilles with identifying metadata for the Achilles resource.




Harmonizome All ▾

SEARCH DOWNLOAD VISUALIZE PREDICT API MOBILE ABOUT

## Achilles Cell Line Gene Essentiality Profiles Dataset

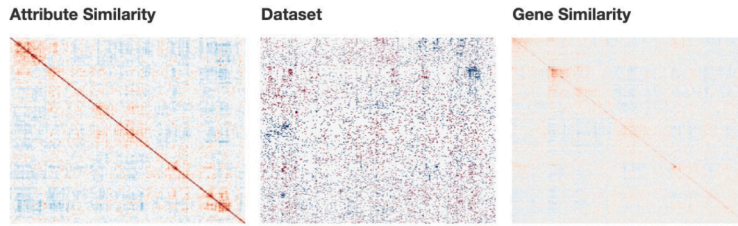
Description	Fitness scores for cell lines following single gene knockdowns
Measurement	Relative cell proliferation
Association	Gene-cell line associations by sensitivity of cell line to gene knockdown
Category	Disease or phenotype associations
Resource	<a href="#">Achilles</a>
Citation(s)	<ul style="list-style-type: none"> <li>▪ <a href="#">Cowley, GS et al. (2014) Parallel genome-scale loss of function screens in 216 cancer cell lines for the identification of context-specific genetic dependencies. Sci Data. 1:140035.</a></li> <li>▪ <a href="#">Cancer, Target Discovery and Development Network et al. (2010) Towards patient-based cancer therapeutics. Nat Biotechnol. 28:904-6.</a></li> <li>▪ <a href="#">Cheung, HW et al. (2011) Systematic investigation of genetic vulnerabilities across cancer cell lines reveals lineage-specific dependencies in ovarian cancer. Proc Natl Acad Sci U S A. 108:12372-7.</a></li> </ul>
Last Updated	
Stats	4831 genes 216 cell lines 104046 gene-cell line associations

### Data Access

API	 
Script	
Downloads	<a href="#">Gene-Attribute Matrix</a> ⓘ <a href="#">Gene-Attribute Edge List</a> ⓘ <a href="#">Up Gene Set Library</a> ⓘ <a href="#">Down Gene Set Library</a> ⓘ <a href="#">Up Attribute Set Library</a> ⓘ <a href="#">Down Attribute Set Library</a> ⓘ <a href="#">Gene Similarity Matrix</a> ⓘ <a href="#">Attribute Similarity Matrix</a> ⓘ <a href="#">Gene List</a> ⓘ <a href="#">Attribute List</a> ⓘ <a href="#">Processing Scripts</a> ⓘ <a href="#">Gene-Attribute Matrix Cleaned</a> ⓘ <a href="#">Gene-Attribute Matrix Standardized</a> ⓘ

**Figure 21.** Dataset page for “Achilles Cell Line Gene Essentiality Profiles” with identifying metadata for the dataset, in addition to download links for files included in this dataset.

## Visualizations



## Cell line Gene Sets

216 sets of gene knockdowns changing fitness of each cell line relative to other cell lines from the Achilles Cell Line Gene Essentiality Profiles dataset.

Show  entries

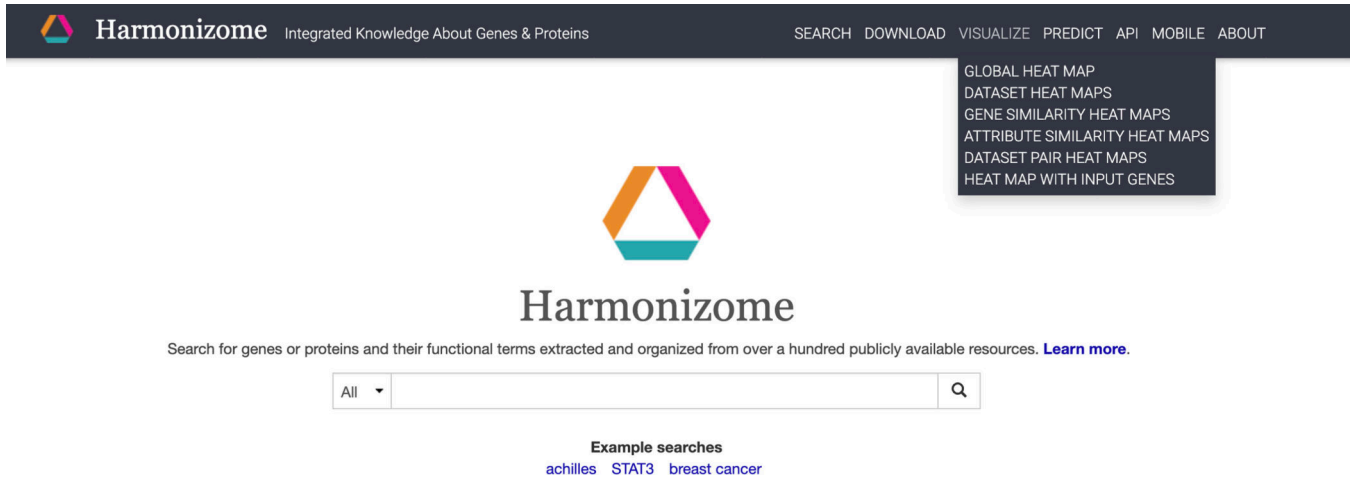
Filter

Gene Set	Description
<a href="#">22RV1</a>	Human prostate carcinoma cell line; derived from a human prostate carcinoma xenograft (CWR22R) that was serially propagated in nude mice after castration-induced regression and relapse of the parental, androgen-dependent CWR22 xenograft.
<a href="#">697</a>	Human B cell precursor leukemia; established from the bone marrow of a 12-year-old boy with acute lymphoblastic leukemia (cALL) at relapse in 1979.
<a href="#">786O</a>	Human renal cell adenocarcinoma cell line, established from a 58 years old caucasian male.
<a href="#">A1207</a>	
<a href="#">A172</a>	A immortal human brain-derived cell line cell that has the characteristics: Human cell line derived from glioblastoma. TKG0183(Deposited from Tohoku Univ.).
<a href="#">A204</a>	
<a href="#">A2058</a>	Human skin melanoma cell line, established from a 43 years adult caucasian male.
<a href="#">A549</a>	Human lung carcinoma established from an explanted lung tumor which was removed from a 58-year-old Caucasian man in 1972; cells were described to induce tumors in athymic mice and to synthesize lecithin.
<a href="#">A673</a>	
<a href="#">ACHN</a>	
<a href="#">AGS</a>	
<a href="#">AM38</a>	
<a href="#">AML193</a>	
<a href="#">ASPC1</a>	
<a href="#">BT20</a>	Human, Caucasian, breast, carcinoma cell line. Morphology: epithelial-like; species: human, Caucasian female 74 years old; tissue: breast; tumor: carcinoma.
<a href="#">BT474</a>	
<a href="#">BXPC3</a>	Human pancreatic adenocarcinoma cell line, established from a 61 year old human female.
<a href="#">C2BBE1</a>	The C2BBE1 (brush border expressing) cell line was cloned in 1988 from the Caco-2 cell line by limiting dilution.
<a href="#">C32</a>	
<a href="#">CADOES1</a>	

### Figure 22.

Links to visualizations of the dataset contents and a table of gene sets. Click any of the gene sets to be redirected to a gene set specific page.





**Figure 23.**  
Dropdown menu of visualization page options.

Author Manuscript

Author Manuscript

Author Manuscript

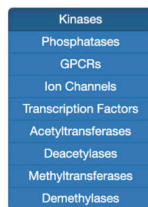
Author Manuscript



## Harmonogram

This interactive clustergram visualizes gene appearance in Harmonizome resources. Genes are shown as rows and resources as columns. The gene occurrence value is normalized relative to the occurrence of other genes in the resource. Resources are grouped into seven categories (see color key) and NIH Grants data is highlighted in blue.

Select different gene classes, e.g. Kinases, by clicking the gene class buttons. Change the clustergram ordering by using the toggle switch or by double-clicking row or column labels.

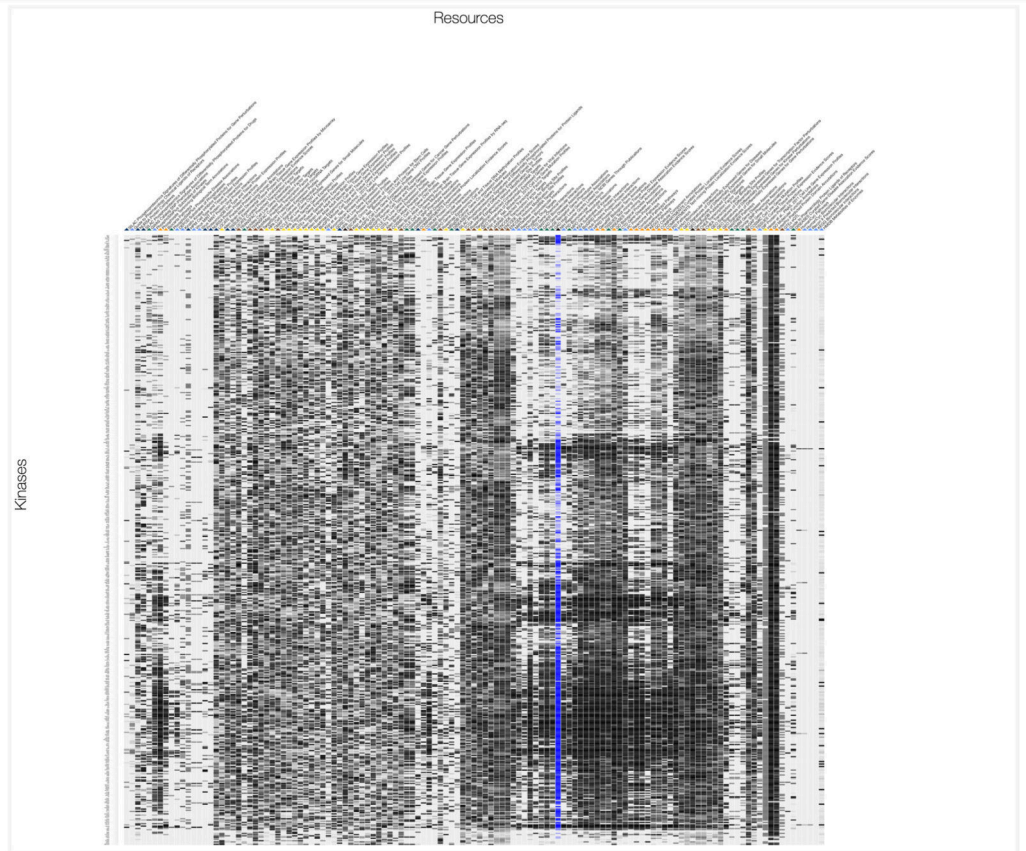


Cluster Rank

### Resource Groups:

- Disease or phenotype associations
- Transcriptomics
- Physical interactions
- Structural or functional annotations
- Genomics
- Proteomics
- Omics

Input Gene Search



**Figure 24.**

Global Heat Map visualization organized by gene families and resources. Switch between gene families using the buttons on the left. Switch between “Cluster” and “Rank” using the toggle on the left. Query a gene of interest using the search bar at the bottom left.

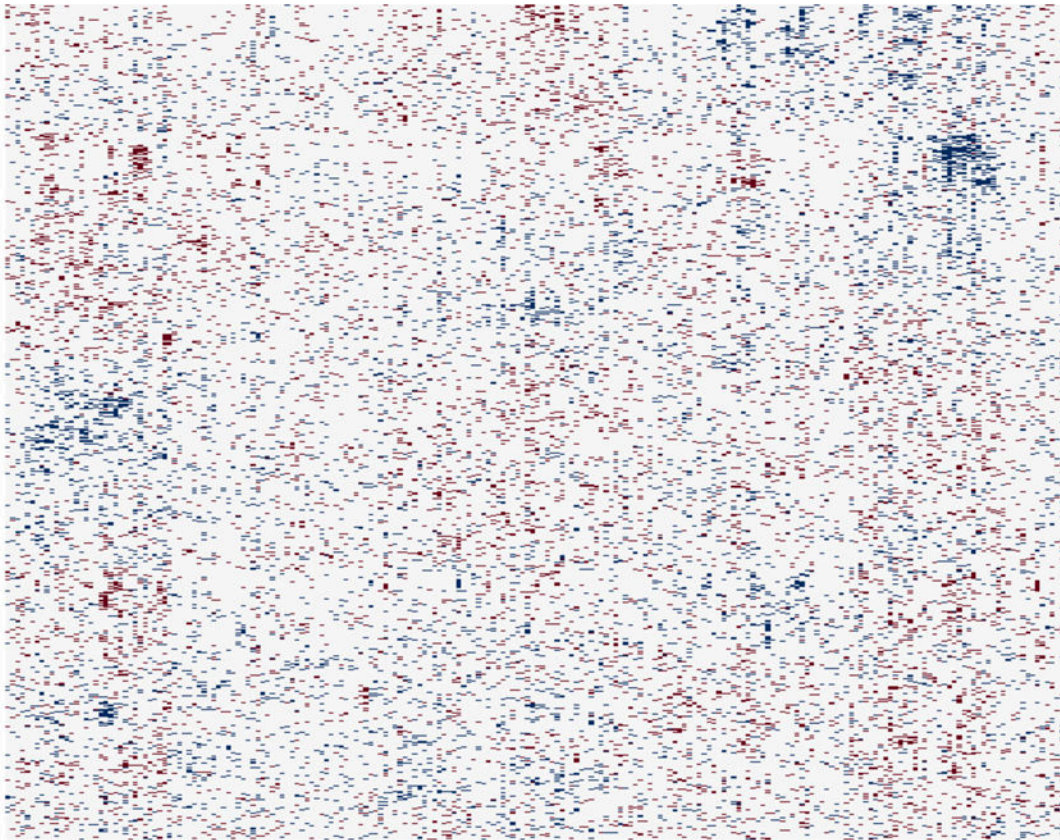
## Dataset Heat Maps

Select from the drop down menu to explore a hierarchically clustered heat map visualization of a dataset. Red tiles indicate positive or unsigned gene-biological entity associations. Blue tiles indicate negative associations.

Dataset

No interactive hierarchical clustering is available at this time.

Genes  
compared with  
Cell lines



**Figure 25.** Dataset Heat Maps page. Select a dataset from the dropdown menu and it will be visualized as a hierarchically clustered heat map.

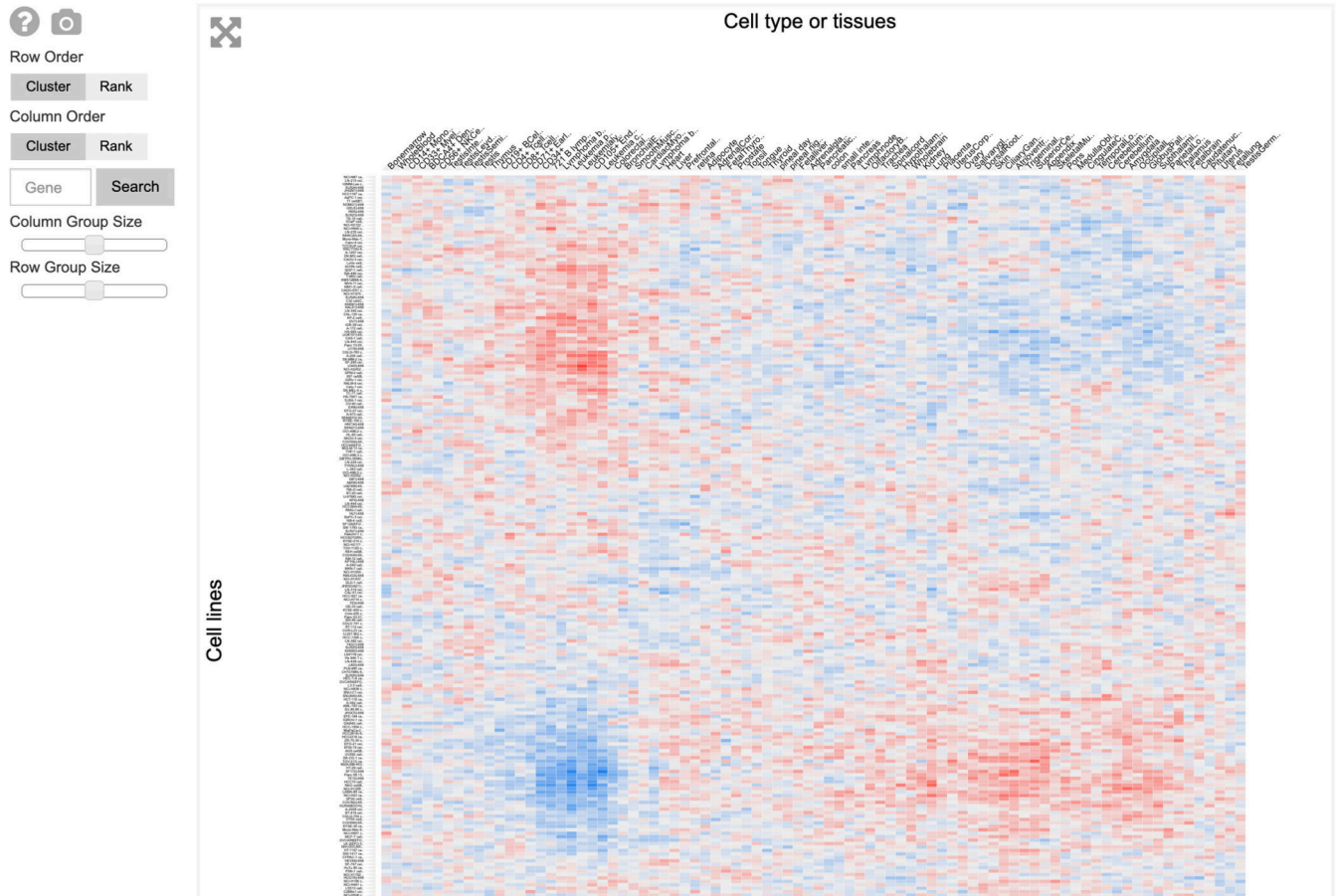


## Dataset Pair Heat Maps

Select from the drop down menu to explore a hierarchically clustered heat map visualization of the similarity of attributes from two datasets. Red tiles indicate pairs of attributes that are similar based on their associations with genes common to both datasets. Blue tiles indicate pairs of attributes that are anti-similar--the two attributes have oppositely signed associations with many of the same genes. White tiles indicate pairs of attributes with few to no overlapping associations.

**Dataset 1**

**Dataset 2**



**Figure 26.** Dataset Pair Heat Maps page. Select two datasets to compare from the dropdown menus and a hierarchically clustered heat map will be generated.

# Heat Map with Input Genes

Select from the drop down menu to choose a dataset. Paste your gene list in the text box. Click submit to build a customized hierarchically clustered heat map visualization of the associations between your uploaded genes and the biological entities in your chosen dataset.

To place an upper bound on the size of the heat maps, which load more slowly as they grow larger, we have two restrictions:

1. Input gene list can have 500 genes maximum.
2. Only datasets with 200 or fewer gene sets can be visualized.

**Dataset** BioGPS Human Cell Type and Tissue Gene Expression Profiles ▾

**Genes**

NSUN3

POLRMT

NLRX1

SFXN5

ZC3H12C

SLC25A39

ARSG

DEFB29

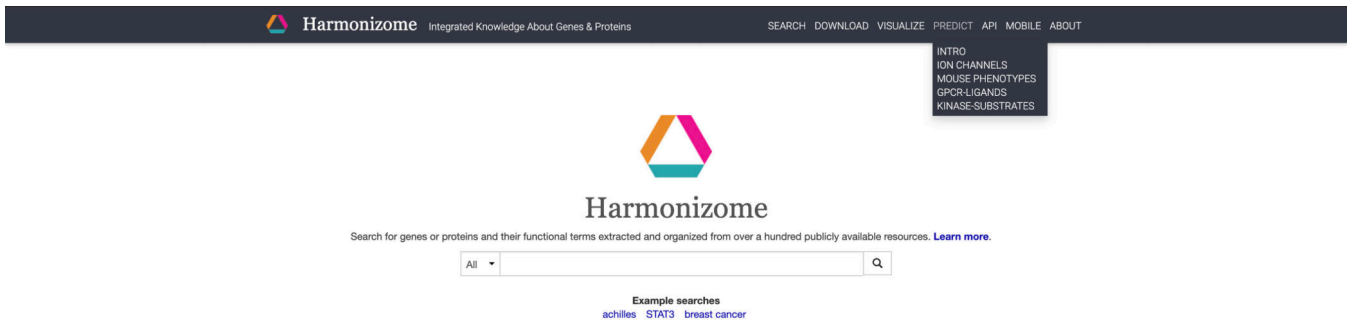
NDUFB6

ZFAND1

Example input Submit



**Figure 27.** Heat Map with Input Genes page. Input a list of maximum 500 genes and select a dataset to build a hierarchically clustered heat map detailing associations between the input genes and biological entities in the dataset.



**Figure 28.**  
Dropdown menu of “Predict” options.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Harmonizome All  Q

SEARCH DOWNLOAD VISUALIZE PREDICT API MOBILE ABOUT

## Machine Learning Case Studies

While the Harmonizome website provides a valuable interface for searching and browsing gene-biological entity associations collected from over 100 datasets, there is also enormous potential for biological discovery by using Harmonizome data to build computational models to predict novel properties of genes or proteins, such as molecular interactions or disease associations. We have made each processed dataset available for [download](#) in several convenient formats to facilitate use of Harmonizome data for computational analysis.

To demonstrate the value of Harmonizome data for computationally-driven hypothesis generation, we developed four supervised machine learning case studies. Our approach was similar for each case study. First we organized gene-biological entity associations from many Harmonizome datasets into a large feature matrix with genes labelling the rows and biological entities (features) labelling the columns. Then we trained a classifier to use the features to distinguish between genes (or pairs of genes) known to have a property of interest and genes (or pairs of genes) unlikely to have that property. Finally, we applied the classifier to make predictions about genes (or pairs of genes) for which knowledge is missing.

Methods and results for the machine learning case studies are described in detail in the [Harmonizome publication](#). Here, we provide brief descriptions of the case studies, interactive tables for browsing the top predictions of the classifiers, and text files that contain the full results.

Case Study	View Table	Download Table
Ion Channel Predictions		
Mouse Phenotype Predictions		
GPCR-Ligand Interaction Predictions		
Kinase-Substrate Interaction Predictions		

**Figure 29.** Machine learning case studies page with details about the case studies were performed. Click on the corresponding buttons to view the tables for each study or download the table of predicted associations.

ARCHS<sup>4</sup> Search Visualize Download Chrome Extension Help  Search

## ARCHS<sup>4</sup>: Massive Mining of Publicly Available RNA-seq Data from Human and Mouse

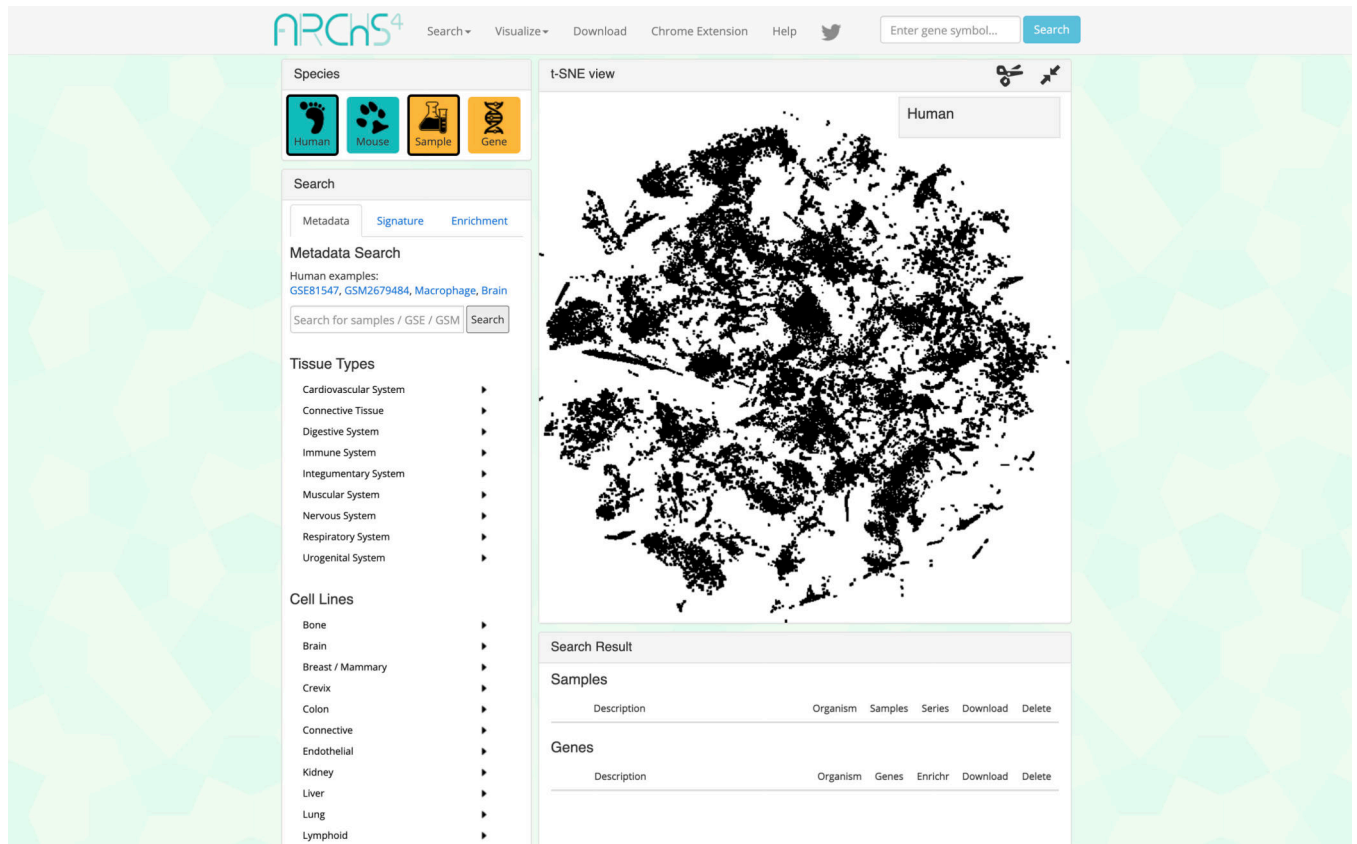
The diagram illustrates the ARCHS4 workflow. It starts with a 'Local server' containing a 'Docker container' with 'RNA-seq quantification' tools: kallisto, libSicklent, STAR tools, Python, wget, R, and Ubuntu. 'Job instructions' are sent from the local server to a 'Cloud Repository' (AWS EC2 instances). The cloud instances are managed by 'Job instructions' and 'SQL'. The workflow is supported by four key features: Process Isolation (Dedicated instances, Docker virtualization), Elasticity (Configuration free, No manual installation), Shared Infrastructure (Shared components), and Host restrictions (Private or public domain). The final output is a 3D visualization of RNA-seq data.

**Get Started**

ARCHS<sup>4</sup> provides access to gene counts from **HiSeq 2000**, **HiSeq 2500** and **NextSeq 500** platforms for human and mouse experiments from GEO and SRA. The website enables downloading of the data in **H5 format** for programmatic access as well as a 3-dimensional view of the sample and gene spaces. Search features allow browsing of the data by meta data annotation, ability to submit your own up and down gene sets, and explore matching samples enriched for annotated gene sets. Selected sample sets can be downloaded into a tab separated text file through auto-generated R scripts for further analysis. Reads are aligned with **Kallisto** using a custom cloud computing platform. Human samples are aligned against the GRCh38 human reference genome, and mouse samples against the GRCh38 mouse reference genome.

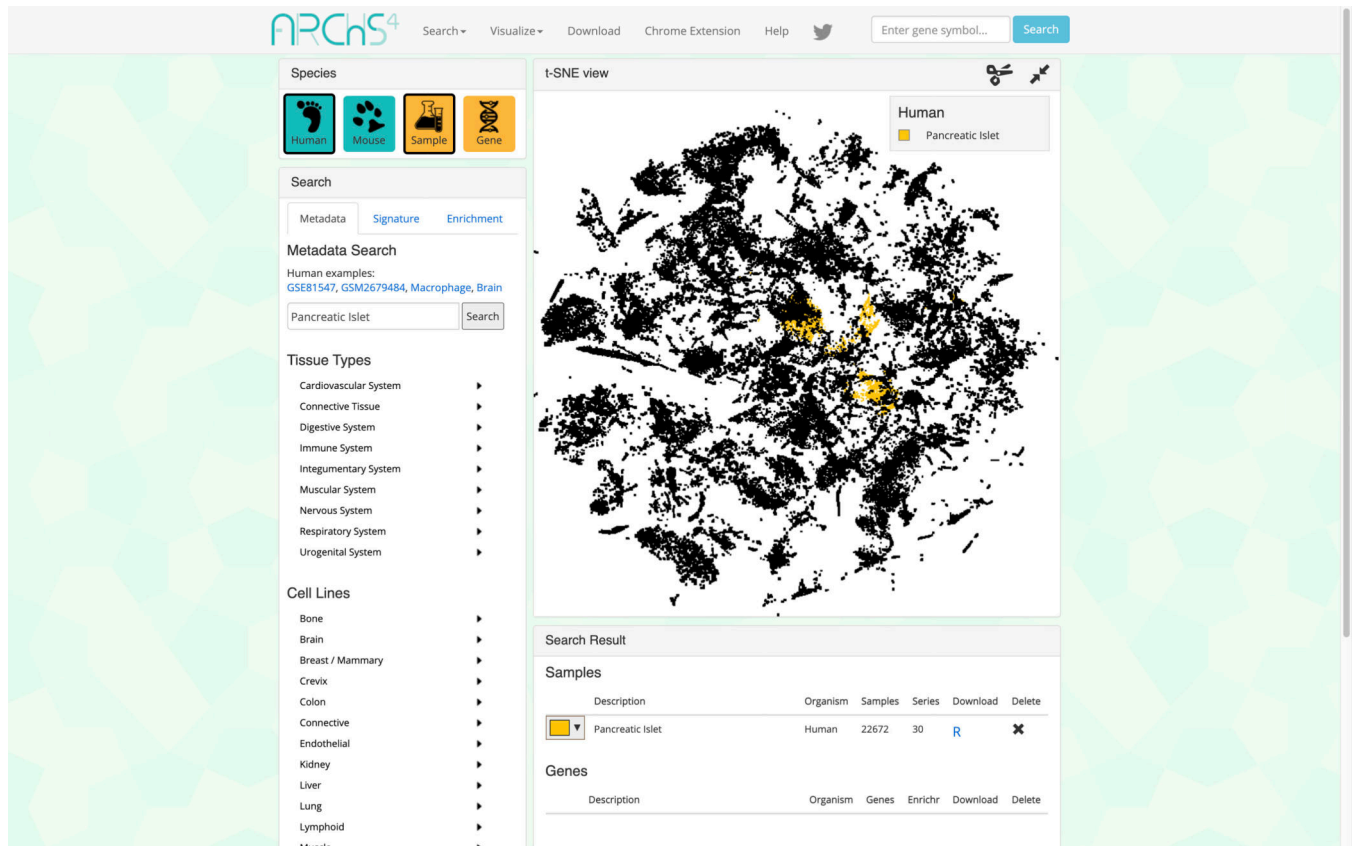
Please acknowledge ARCHS<sup>4</sup> in your publications by citing the following reference:  
 Lachmann A, Torre D, Keenan AB, Jagodnik KM, Lee HJ, Wang L, Silverstein MC, Ma'ayan A. Massive mining of publicly available RNA-seq data from human and mouse. *Nature Communications* 9. Article number: 1366 (2018), doi:10.1038/s41467-018-03751-6

**Figure 30.**  
ARCHS4 Homepage.



**Figure 31.** Data visualization and search page that includes a 3D interactable scatter plot of gene expression data.





**Figure 32.** 3D scatter plot of human gene expression data that includes the term “Pancreatic islet”.

## Search Result

### Samples

Description	Organism	Samples	Series	Download	Delete
 Pancreatic Islet	Human	22672	30	<a href="#">R</a>	<a href="#">✕</a>

### Genes

Description	Organism	Genes	Enrichr	Download	Delete
-------------	----------	-------	---------	----------	--------

**Figure 33.**

Search results table with Pancreatic islet samples listed in their respective section with metadata and options to download an R script to process the samples or delete the query.

Species

Human Mouse Sample Gene

Search

[Metadata](#) [Signature](#) [Enrichment](#)

### Signature Search

Search signatures by up and down gene sets

[Try an example](#)

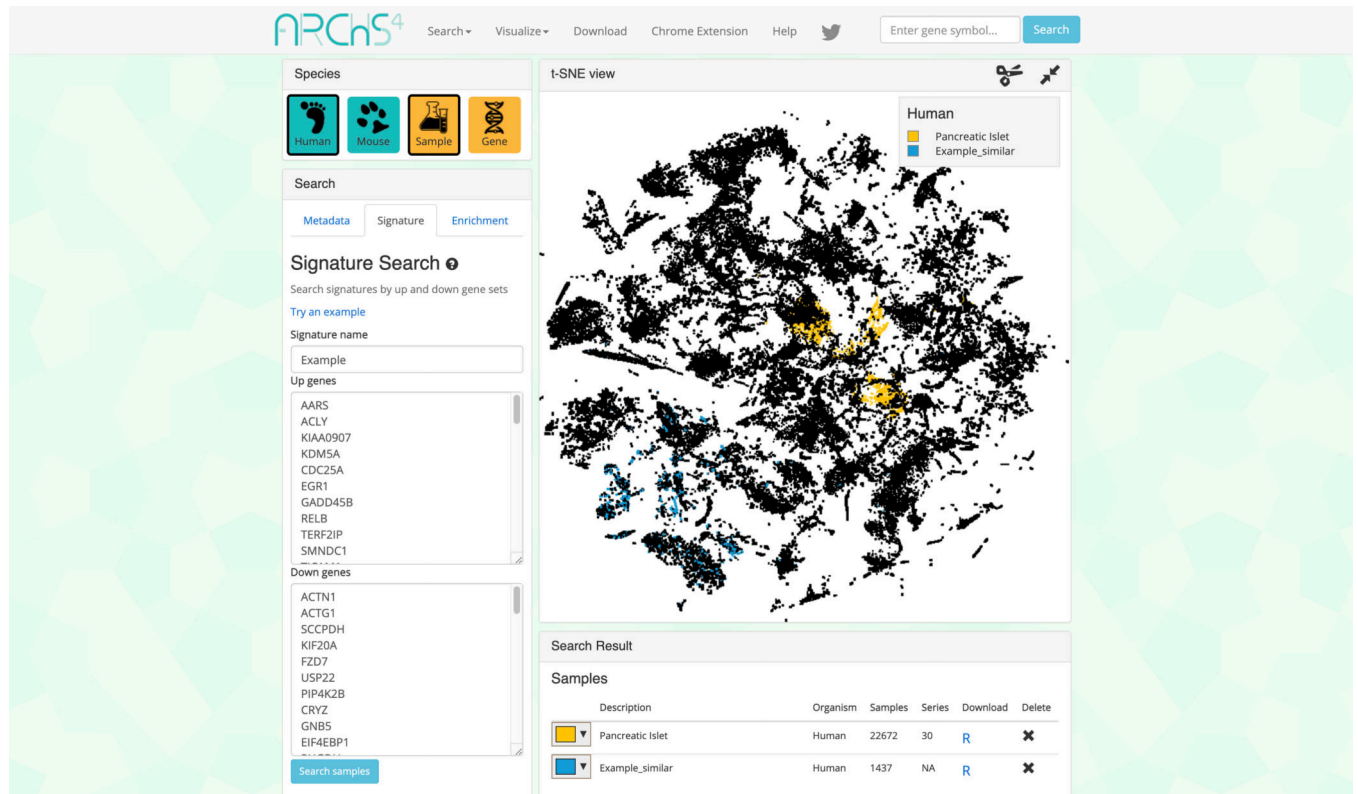
Signature name

Up genes

Down genes

[Search samples](#)

**Figure 34.** Signature search field that allows for querying of up and downregulated genes to identify samples that match the input.



**Figure 35.** Example query from the signature search visualized in the 3D scatter plot. The identified samples are added to the “Search Result” table.

Search

Metadata Signature **Enrichment**

### Enrichment Search ?

Gene Set Library

CHEA 2016 ▼

Search gene set

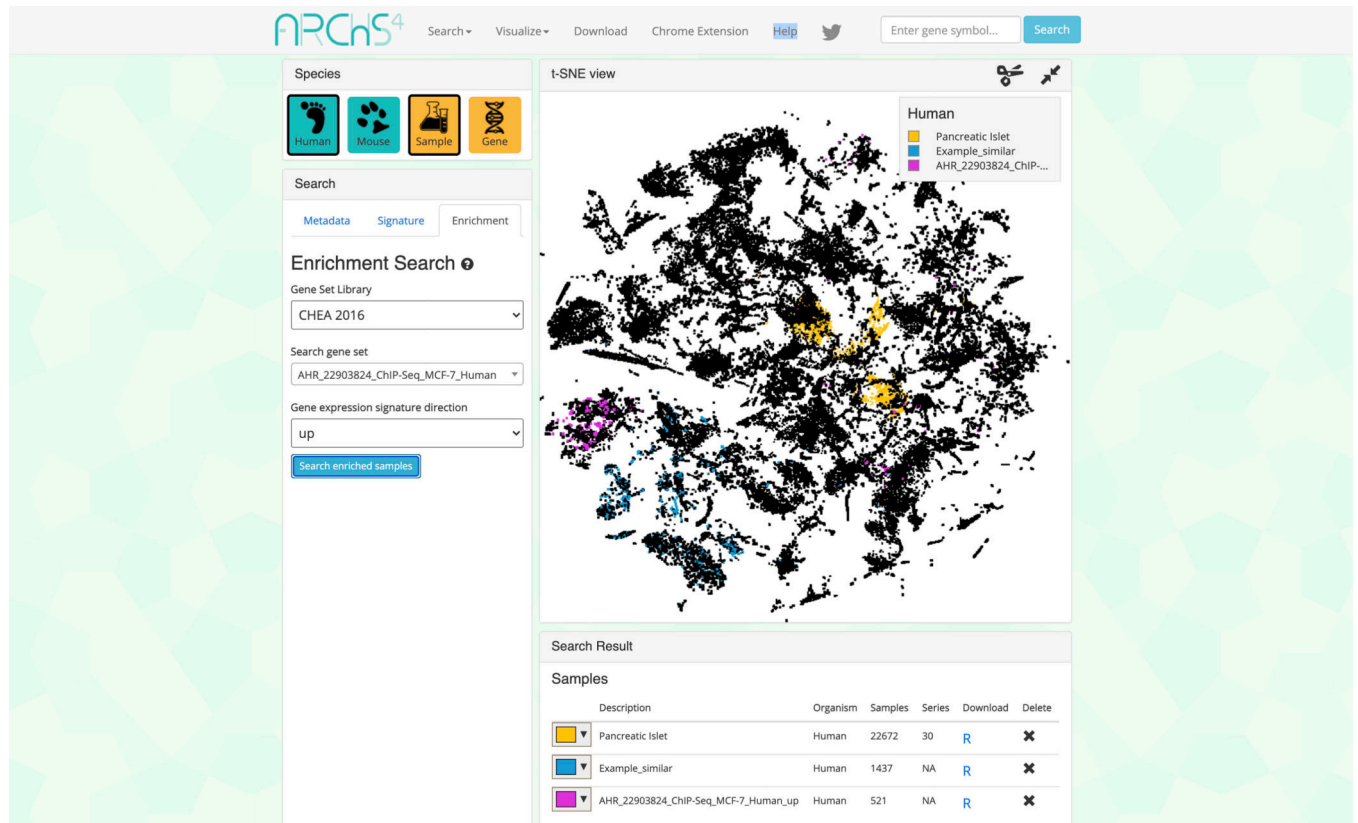
AHR\_22903824\_ChIP-Seq\_MCF-7\_Human ▼

Gene expression signature direction

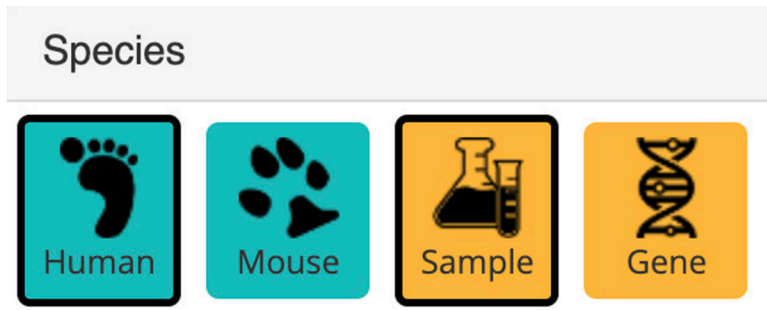
up ▼

Search enriched samples

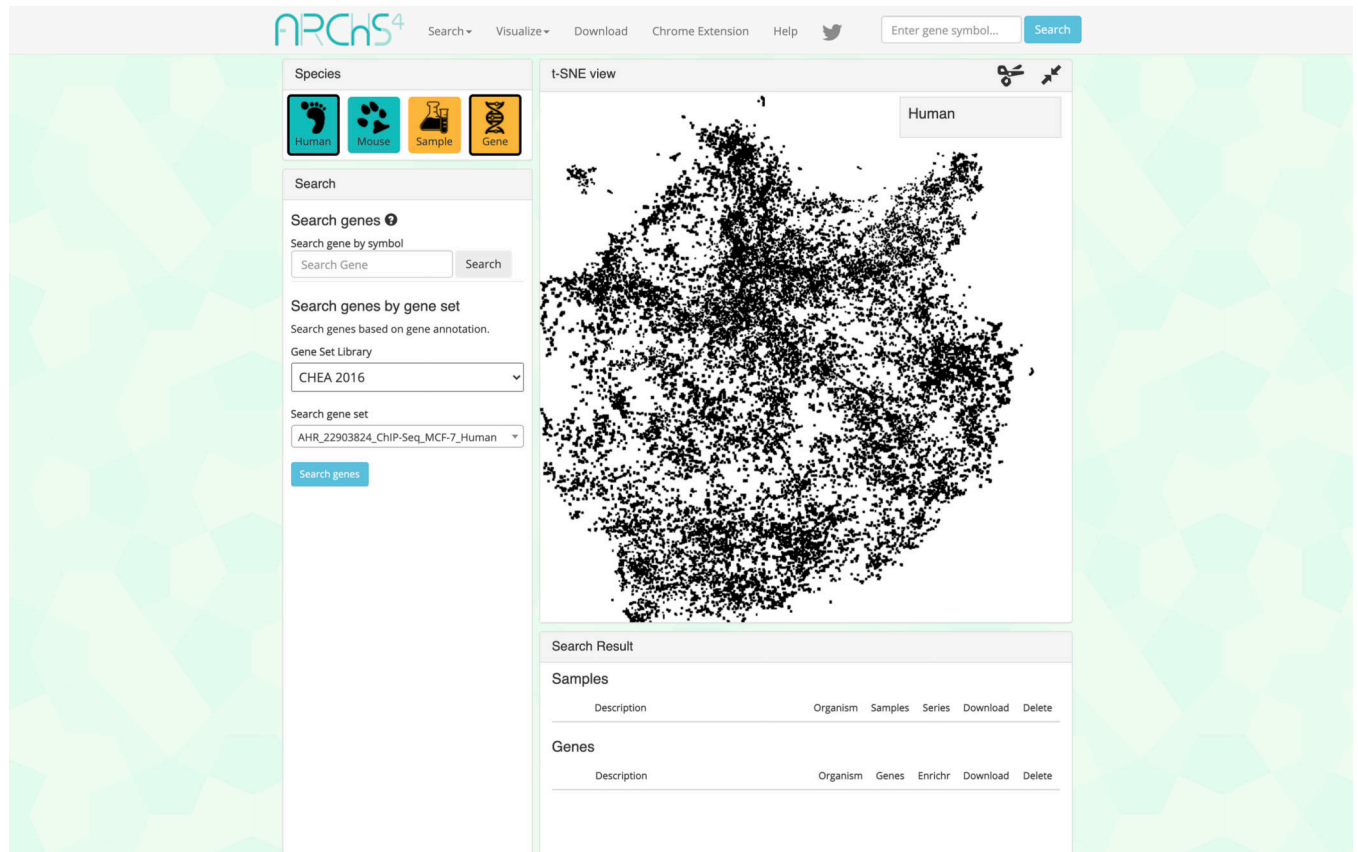
**Figure 36.** Enrichment search field that allows for selection gene set library, gene set within the library, and choice of upregulated or downregulated signatures.



**Figure 37.** Example query from the enrichment search visualized in the 3D scatter plot. The identified samples are added to the “Search Result” table.



**Figure 38.** Selection buttons for switching between human and mouse samples, as well as buttons for switching between sample queries and single gene queries.



**Figure 39.** Scatter plot of single genes instead of samples where the distance between genes quantifies similarity of their expression profiles across all samples in ARCHS4.



## Search genes

Search gene by symbol

## Search genes by gene set

Search genes based on gene annotation.

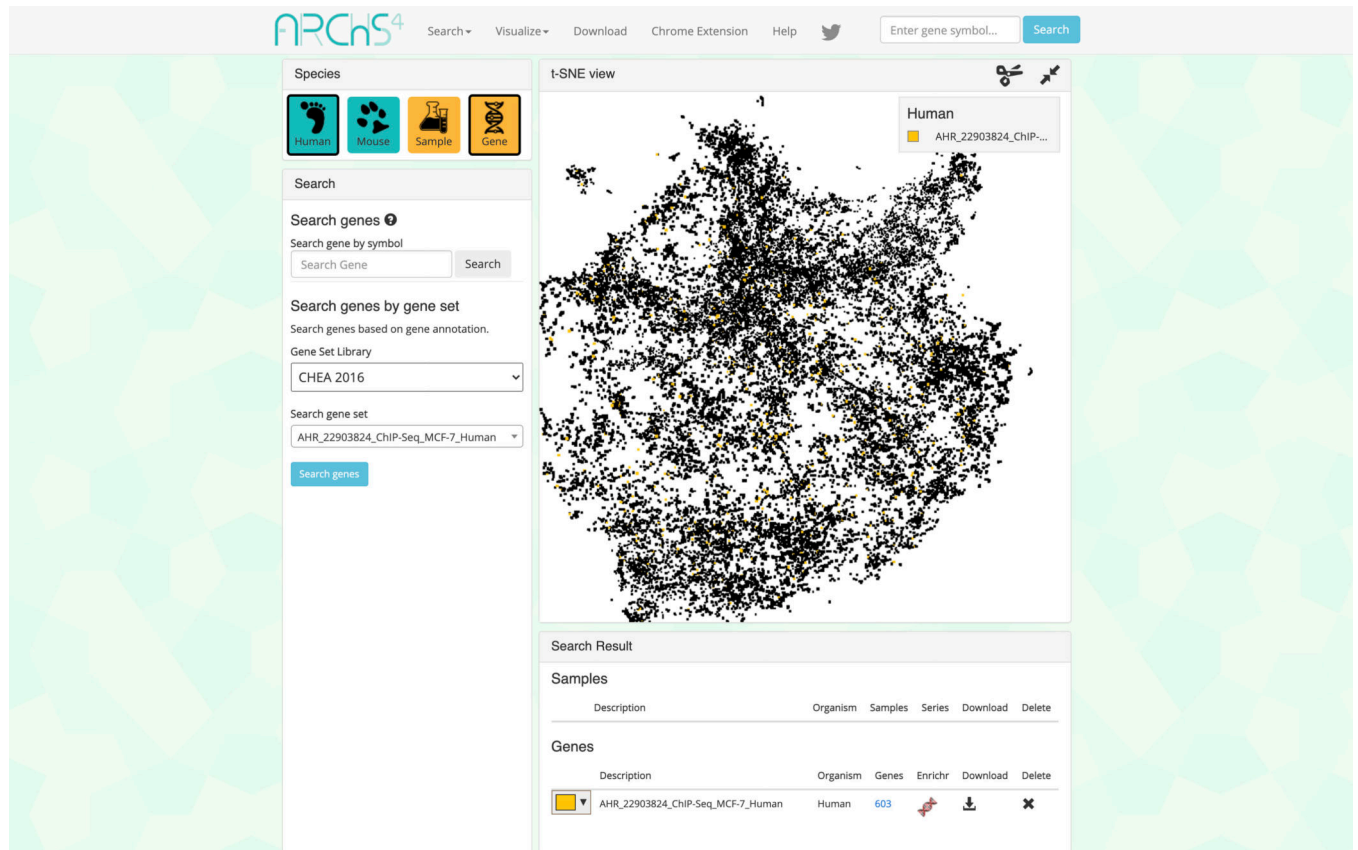
Gene Set Library



Search gene set



**Figure 40.** “Search genes by gene set” field where a gene set library and gene set within the library are selected to be queried.



**Figure 41.** Genes from the selected gene set library and gene set are displayed on the scatter plot. The genes are added to their respective section in the “Search Result” table.

The screenshot displays a search result interface. On the left, there are navigation options for 'Species' (Human, Mouse), 'Search', 'Search genes by symbol', 'Search genes based on', and 'Gene Set Library' (CHEA 2016). The main search bar contains 'AHR\_22903824\_ChIP-Seq\_MCF-7\_Human'. A modal window is open, showing the search results for this query. The title of the modal is 'AHR\_22903824\_ChIP-Seq\_MCF-7\_Human (603)'. Below the title is a long list of gene symbols, including A4GALT, ABCA4, ABCC4, ABHD12B, ABHD2, ABLIM3, ACOXL, ADAM12, ADAMTS9, ADAMTSL3, ADD2, ADK, AGR3, AHDC1, ALDH1A3, ALDH1L1, ALDH3B2, ALS2CL, AMN1, ANKRD13C, ANKRD29, ANO1, ANO4, AP4E1, APBB2, AREG, ARHGAP21, ARHGAP24, ARHGEF10, ARID2, ARL4C, ARL5B, ASAP1, ASB7, ASCC2, ATP1B1, ATP4B, ATP8A2, ATP8B4, ATP9A, ATXN1, ATXN7L1, AUTS2, BARX1, BASP1, BCAS3, BCL2, BCL6, BCOR, BCORL1, BEND5, BMF, BMPR2, BOD1, BRD7, BRWD1, BZW2, C12ORF49, C1ORF100, C1QTNF9B, C5ORF22, C6ORF132, C6ORF141, C8ORF34, C8ORF37, C9ORF3, CABLES1, CACNA1D, CALM2, CALR3, CBX5, CCDC74A, CCND1, CCNJL, CCR7, CD180, CD7, CD99L2, CDH12, CDH4, CDH6, CDKAL1, CDKL4, CDX2, CDYL2, CEACAM6, CELSR1, CHD2, CHN2, CHST11, CHURC1, CITED2, CLEC14A, CLMN, CNBD1, CNTNAP4, COL18A1, COQ4, COTL1, COX6C, CPEB4, CPXM2, CRB2, CRISPLD2, CSNK1G3, CST5, CT62, CTNNA1, CTPS2, CUX1, CWH43, CYP1A1, CYP1A2, CYP1B1, DAB2, DCAF7, DDI4, DEGS1, DHCR7, DHRS2, DICER1, DIRC3, DISC1, DLG2, DLL1, DNAH5, DNAJB6, DNAJB9, DNAJC1, DNMT3, DNMBP, DNMT3L, DNPEP, DPY19L1, DRD1, DSCAM, DST, DYNC1L1, E2F6, EBF1, EDC3, EDIL3, EDN1, EFEMP1, EFNA5, EFR3B, EGFR, EHD4, ELF3, ELF4, EML6, ENPP1, ENTPD3, EPHA4, ERC2, ERCC6, ESR1, EVL, EYA2, EYS, F8A1, FAM101A, FAM102A, FAM102B, FAM105A, FAM120B, FAM151B, FAM153C, FAM155A, FAM174B, FAM178B, FAM65B, FAM65C, FAM84B, FAM92B, FANCC, FAT2, FBLN2, FBP1, FBXO21, FBXO31, FBXW7, FBXW8, FER1L6, FGD2, FGD5, FGF3, FGF2, FHIT, FIGN, FLVCR2, FN1, FNDC3B, FOS, FOSL2, FOXN3, FOXN4, FOXP1, FREM2, FRMD4A, FRMD5, FSCB, FSIP1, FTO, FXR1, GAPVD1, GDF15, GLDN, GLG1, GNA14, GPATCH2, GPC6, GPR45, GPX3, GRB14, GRHL1, GRHL2, GRIA2, GRIK2, GRM4, GSTM3, GSTP1, GULP1, HAAO, HCRTR2, HDACS, HEBP1, HECW2, HES1, HEY1, HEY2, HLCS, HRASLS2, HS6ST3, HSF2BP, HSPD1, HSPD2, HSPD3, HSPD4, HSPD5, HSPD6, HSPD7, HSPD8, HSPD9, HSPD10, HSPD11, HSPD12, HSPD13, HSPD14, HSPD15, HSPD16, HSPD17, HSPD18, HSPD19, HSPD20, HSPD21, HSPD22, HSPD23, HSPD24, HSPD25, HSPD26, HSPD27, HSPD28, HSPD29, HSPD30, HSPD31, HSPD32, HSPD33, HSPD34, HSPD35, HSPD36, HSPD37, HSPD38, HSPD39, HSPD40, HSPD41, HSPD42, HSPD43, HSPD44, HSPD45, HSPD46, HSPD47, HSPD48, HSPD49, HSPD50, HSPD51, HSPD52, HSPD53, HSPD54, HSPD55, HSPD56, HSPD57, HSPD58, HSPD59, HSPD60, HSPD61, HSPD62, HSPD63, HSPD64, HSPD65, HSPD66, HSPD67, HSPD68, HSPD69, HSPD70, HSPD71, HSPD72, HSPD73, HSPD74, HSPD75, HSPD76, HSPD77, HSPD78, HSPD79, HSPD80, HSPD81, HSPD82, HSPD83, HSPD84, HSPD85, HSPD86, HSPD87, HSPD88, HSPD89, HSPD90, HSPD91, HSPD92, HSPD93, HSPD94, HSPD95, HSPD96, HSPD97, HSPD98, HSPD99, HSPD100.

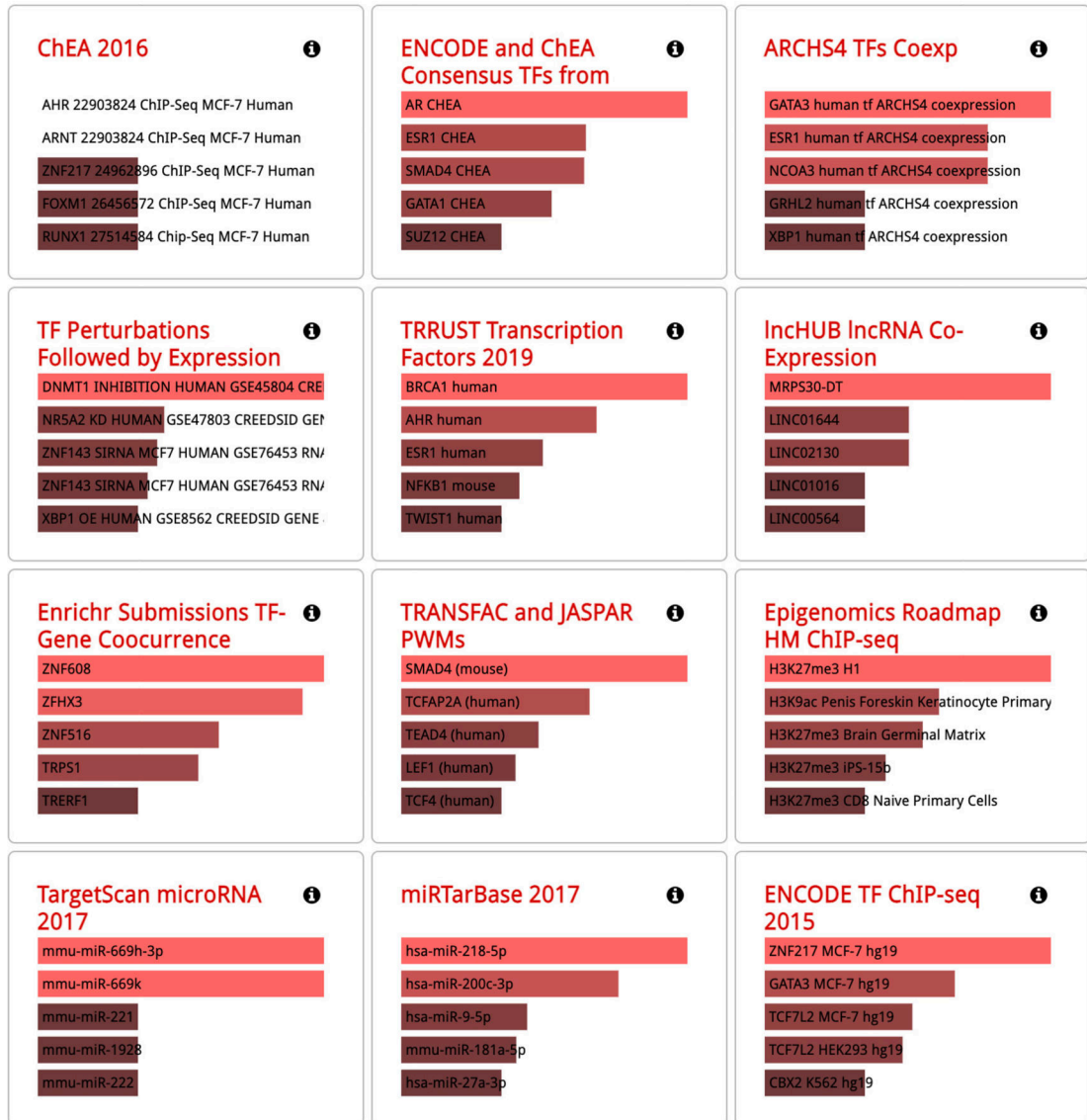
**Figure 42.**

Clicking on the number of genes in the “Search Result” table displays the genes included in the queried gene set.



Transcription Pathways Ontologies Diseases/Drugs Cell Types Misc Legacy Crowd


Description AHR\_22903824\_ChIP-Seq\_MCF-7\_Human (603 genes)



**Figure 43.**

Clicking on the Enrichr icon in the “Search Results” table displays gene set enrichment analysis results for the genes from the queried gene set.

Search

Search genes 

Search gene by symbol

Search genes by gene set

Search genes based on gene annotation.

Gene Set Library

Search gene set

**Figure 44.**  
“Search genes” field populated with the gene symbol “SOX2”.



Predicted functional terms: [GO](#) | [ChEA](#) | [Mouse Phenotype](#) | [Human Phenotype](#) | [KEA](#) | [KEGG](#)  
 Most similar genes based on co-expression: [Pearson correlation](#)  
 Expression levels across tissues and cell lines: [Tissue Expression](#) | [Cell Line Expression](#)

**Description:** This intronless gene encodes a member of the SRY-related HMG-box (SOX) family of transcription factors involved in the regulation of embryonic development and in the determination of cell fate. The product of this gene is required for stem-cell maintenance in the central nervous system, and also regulates gene expression in the stomach. Mutations in this gene have been associated with optic nerve hypoplasia and with syndromic microphthalmia, a severe form of structural eye malformation. This gene lies within an intron of another gene called SOX2 overlapping transcript (SOX2OT). [NCBI Entrez Gene](#) | [GeneCards](#) | [Harmonizome](#)

#### Functional Annotation Prediction

##### Predicted biological processes (GO)

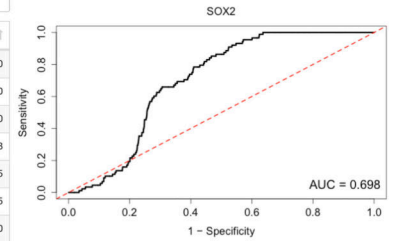
Show 10 entries

Search:

Rank	Gene Set	Z-score
1	regulation of gene silencing by miRNA (GO:0060964)	6.87096240
2	regulation of gene silencing by RNA (GO:0060966)	6.87096240
3	regulation of posttranscriptional gene silencing (GO:0060147)	6.87096240
4	presynaptic membrane assembly (GO:0097105)	5.23303818
5	axon ensheathment in central nervous system (GO:0032291)	5.22702545
6	central nervous system myelination (GO:0022010)	5.22702545
7	neuron fate determination (GO:0048664)	4.81363470
8	regulation of helicase activity (GO:0051095)	4.67302509
9	presynaptic membrane organization (GO:0097090)	4.65147747
10	viral transcription (GO:0019083)	4.51855950

Showing 1 to 10 of 192 entries

Previous 1 2 3 4 5 ... 20 Next

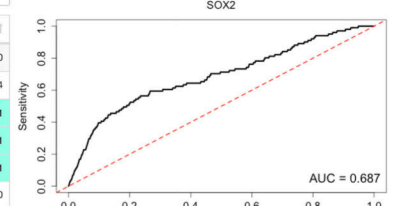


##### Predicted upstream transcription factors (ChEA)

Show 10 entries


Search:

Rank	Gene Set	Z-score
1	E2F7_22180533_Chip-Seq_HELA_Human	7.19845600
2	EZH2_22144423_Chip-Seq_EOC_Human	5.20176004
3	* KLF5_18264089_Chip-ChIP_MESCs_Mouse	4.38094161
4	* KLF4_18264089_Chip-ChIP_MESCs_Mouse	4.38094161
5	* KLF2_18264089_Chip-ChIP_MESCs_Mouse	4.38094161
6	FOXM1_23109430_Chip-Seq_UZOS_Human	3.85513940



**Figure 45.**

Single gene page for SOX2 with identifying metadata at the top of the page. Additionally, tables of predicted functions from various gene set libraries are depicted along with ROC curves to quantify the ability to predict gene sets that SOX2 is a known member of from co-expression data.

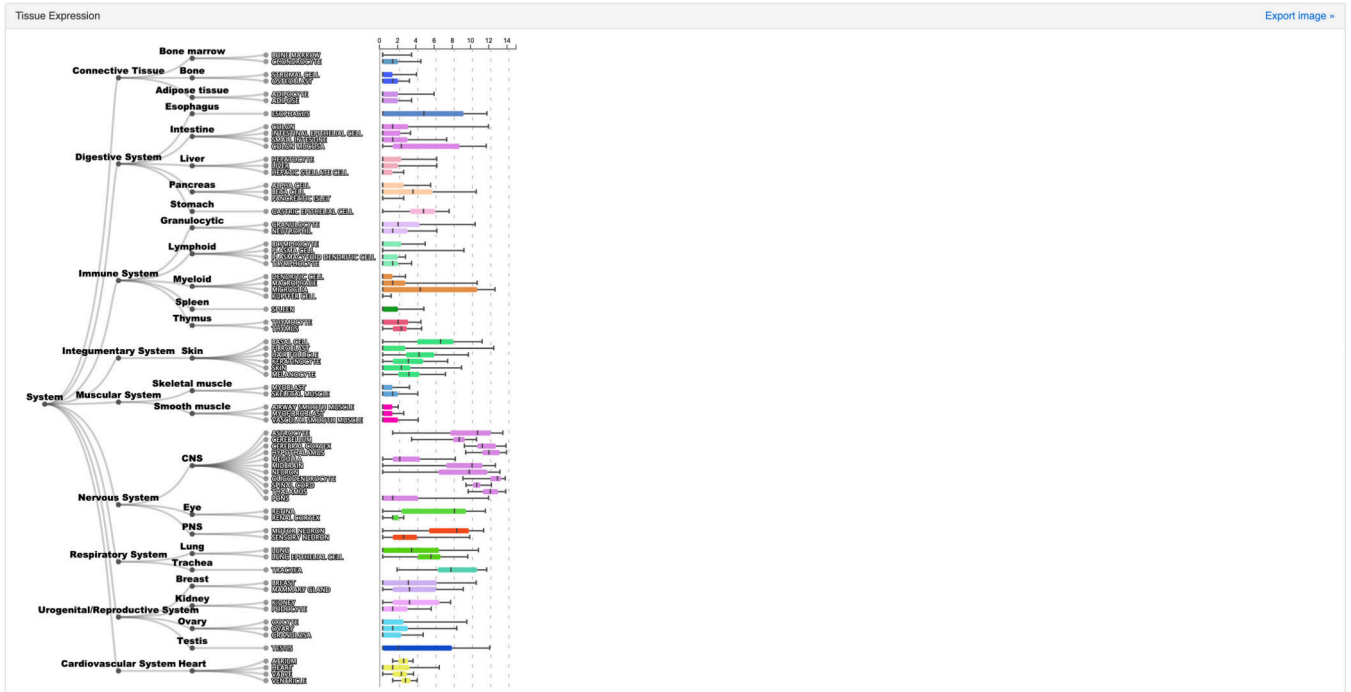
Most similar genes based on co-expression Upload to Enrichr 

Show  entries Search:

Rank	Gene Symbol	Pearson Correlation
1	<a href="#">PTPRZ1</a>	0.6860314011573792
2	<a href="#">TUBB2B</a>	0.6473608016967773
3	<a href="#">SOX21</a>	0.6047778129577637
4	<a href="#">CACNG7</a>	0.5869385600090027
5	<a href="#">VANGL2</a>	0.5742425918579102
6	<a href="#">LRRN1</a>	0.567857027053833
7	<a href="#">SALL3</a>	0.5653490424156189
8	<a href="#">FGFBP3</a>	0.557252824306488
9	<a href="#">SALL2</a>	0.5482832193374634
10	<a href="#">CRABP1</a>	0.5469187498092651

Showing 1 to 10 of 100 entries Previous **1** 2 3 4 5 ... 10 Next

**Figure 46.** Table of the top 100 genes most similar to SOX2 based on co-expression. The genes can be submitted to Enrichr by clicking the “Upload to Enrichr” button.



**Figure 47.** Tissue expression atlas for SOX2 that quantifies the expression of SOX2 in various tissue types.

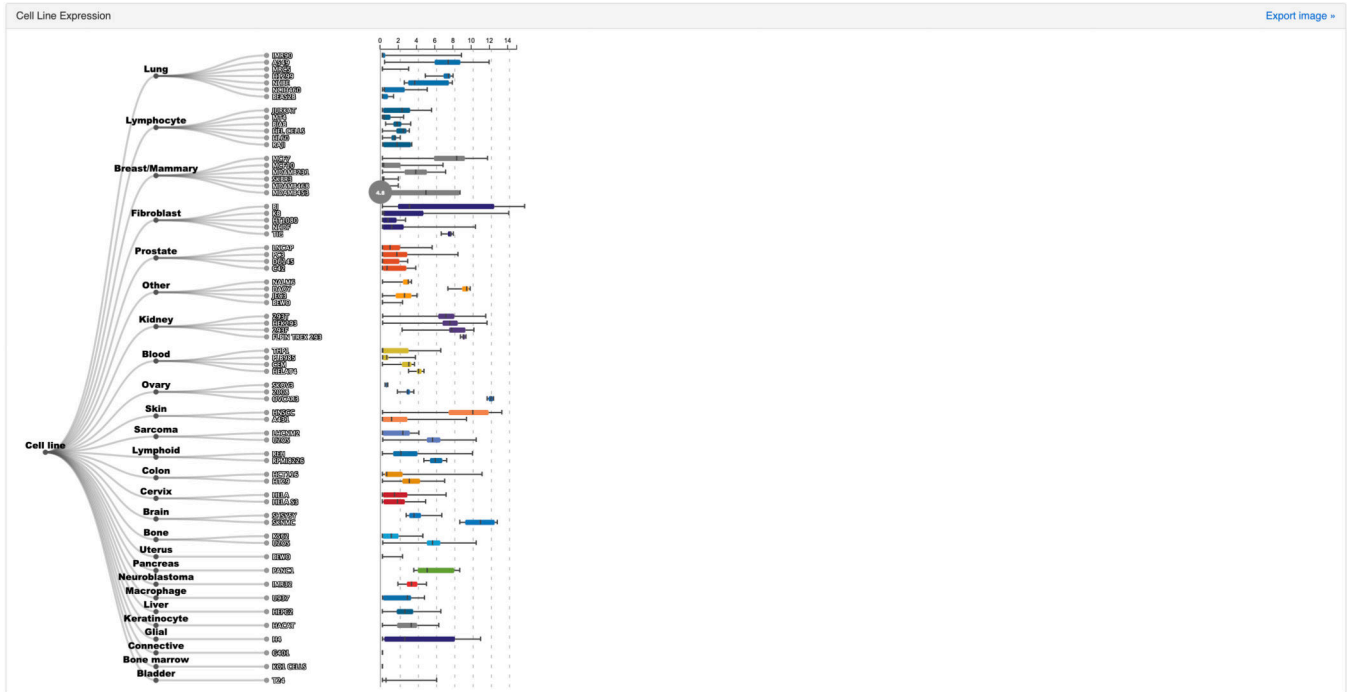
Author Manuscript

Author Manuscript

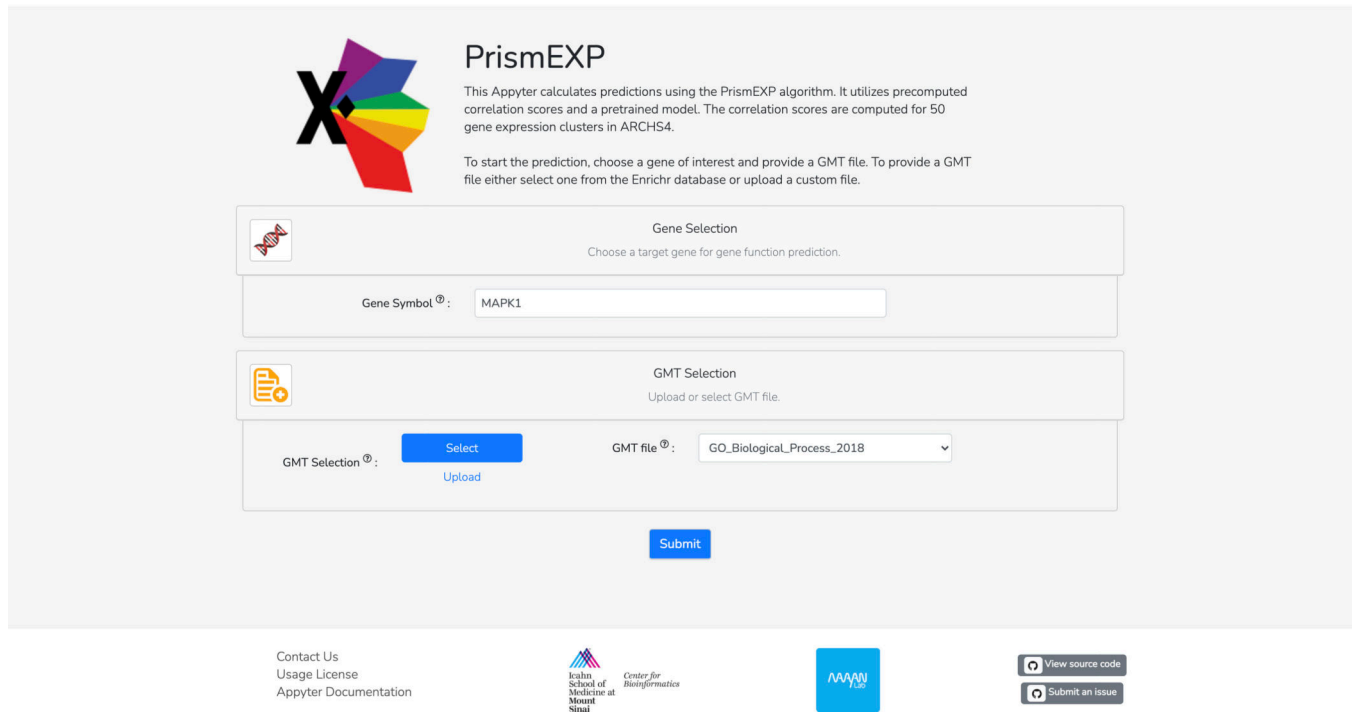
Author Manuscript

Author Manuscript





**Figure 48.** Cell line expression atlas for SOX2 that quantifies the expression of SOX2 in various cell lines.



**PrismEXP**

This Appyter calculates predictions using the PrismEXP algorithm. It utilizes precomputed correlation scores and a pretrained model. The correlation scores are computed for 50 gene expression clusters in ARCHS4.

To start the prediction, choose a gene of interest and provide a GMT file. To provide a GMT file either select one from the Enrichr database or upload a custom file.


**Gene Selection**  
Choose a target gene for gene function prediction.


Gene Symbol <sup>Ⓞ</sup>:

**GMT Selection**  
Upload or select GMT file.

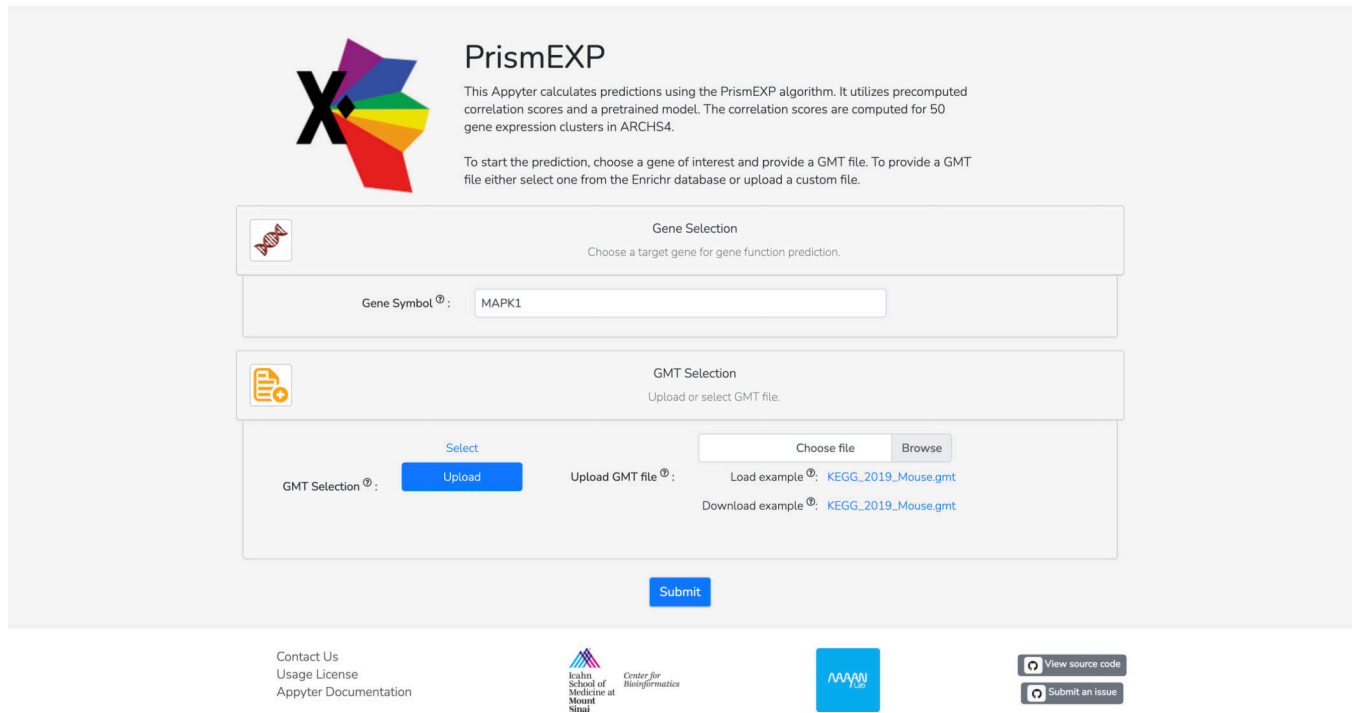
GMT Selection <sup>Ⓞ</sup>:   GMT file <sup>Ⓞ</sup>:

Contact Us  
Usage License  
Appyter Documentation

 Center for Bioinformatics



**Figure 49.** PrismEXP Appyter input form where the user is prompted to input a gene symbol of interest and specify a gene set library (in GMT format) to make predictions from.



**PrismEXP**

This Appyter calculates predictions using the PrismEXP algorithm. It utilizes precomputed correlation scores and a pretrained model. The correlation scores are computed for 50 gene expression clusters in ARCHS4.

To start the prediction, choose a gene of interest and provide a GMT file. To provide a GMT file either select one from the Enrichr database or upload a custom file.

**Gene Selection**  
Choose a target gene for gene function prediction.

Gene Symbol <sup>Ⓞ</sup>:

**GMT Selection**  
Upload or select GMT file.


GMT Selection <sup>Ⓞ</sup>: [Select](#)


Upload GMT file <sup>Ⓞ</sup>:

Load example <sup>Ⓞ</sup>: [KEGG\\_2019\\_Mouse.gmt](#)

Download example <sup>Ⓞ</sup>: [KEGG\\_2019\\_Mouse.gmt](#)

Contact Us  
Usage License  
Appyter Documentation

 Center for Bioinformatics



**Figure 50.**  
Alternative input form option for uploading a custom GMT file.



Download Notebook Toggle Code Run Locally

Table Of Contents

- S3 Data Query
- Load GMT
- Initialize Gene Set Library
- Load Gene Correlation
- Avg Correlation Scores
- Apply PrismEXP
- Prediction Validation
- Top Predictions
- Download Files
- References

## PrismEXP

An appyter interface to compute gene function predictions  
This Appyter calculates predictions using the PrismEXP algorithm. It utilizes precomputed correlation scores and a pretrained model. The correlation scores are computed for 51 gene expression clusters in ARCHS4.

'MAPK1'

'GO\_Biological\_Process\_2018'

In [1]

```
import h5py as h5
import s3fs
import numpy as np
import pandas as pd
from tqdm import tqdm
from typing import List, Dict
import urllib.request
import json
import hashlib
import shutil
import ssl
import os
import sys
import re
import itertools
import pickle
import scipy.stats as st
from sklearn.metrics import roc_auc_score, roc_curve, auc
from matplotlib import pyplot as plt
from IPython.display import display, FileLink, Markdown, HTML
```

In [2]

```
LIBRARY_LIST_URL="https://maayanlab.cloud/speedrichr/api/listlibs"
LIBRARY_DOWNLOAD_URL="https://maayanlab.cloud/Enrichr/geneSetLibrary?mode=text&libraryName="
S3_URL="s3://mssm-prismx/"
GMT_EXAMPLE="https://maayanlab.cloud/Enrichr/geneSetLibrary?mode=text&libraryName=KEGG_2019_Mous
```

**Figure 51.**

The launched Appyter notebook with options to download the notebook, toggle the code, and instructions for running the Appyter locally. Additionally, a table of contents on the left allows for easy traversal between sections of the notebook.

### Load Gene Correlation

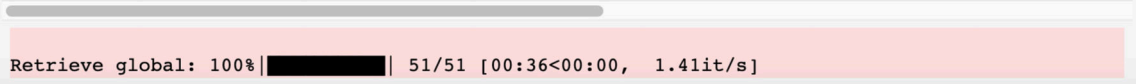
Gene correlation matrices are precalculated for 51 gene expression matrices from ARCHS4. The matrices are stored in AWS S3 and can be queried for target genes of interest. The extraction can take a moment to finish.

In [7]

```
correlation = geneCorrelation(GENE)
correlation
```

	0	1	2	3	4	5	6	7	8	9	...
0610007P14	-0.017517	0.132568	0.005451	-0.004303	0.037903	0.342773	0.001316	0.042419	0.063965	-0.056641	...
0610009B22	-0.007133	0.242798	0.025467	-0.002665	-0.103271	0.074158	0.025192	0.034180	0.073364	0.179321	...
0610009L18	-0.006325	0.249878	0.004566	-0.008400	-0.079651	-0.417236	0.004883	-0.016312	0.090942	0.128540	...
0610009O20	-0.016098	0.017059	0.090881	0.013435	-0.014900	-0.375000	0.055450	-0.033203	0.112915	0.090820	...
0610010F05	0.091187	0.376221	0.103638	0.013855	-0.058136	-0.212646	0.229614	0.016968	0.195679	0.228394	...
...	...	...	...	...	...	...	...	...	...	...	...
MT-ND3	0.167969	0.181641	0.075500	0.008949	-0.002119	0.302979	0.172852	-0.018921	0.194580	0.113586	...
MT-ND4	0.073242	-0.087952	0.049255	0.068481	0.063477	0.298828	0.115479	0.031235	0.134644	-0.048981	...
MT-ND4L	-0.099854	0.088501	0.000661	0.067627	-0.028809	0.287354	0.105286	0.008461	0.160889	-0.083862	...
MT-ND5	0.062866	0.031738	0.012245	0.087524	-0.028717	0.308838	0.119507	0.019714	0.084900	-0.056915	...
MT-ND6	0.078308	-0.254150	0.015701	-0.008034	0.084167	0.209229	0.031860	0.038300	0.032593	-0.045013	...

21553 rows x 51 columns



**Figure 52.**

Dataframe of 51 correlation matrices, each displaying correlation values between the query gene and other mouse genes.

### Avg Correlation Scores

The average correlation scores are equivalent to the predictions performed by [Geneshot](#). The values are used as prediction features of the PrismEXP model.

In [9]

```
avgCor = getAverageCorrelation(correlation, library)
avgCor
```

	0	1	2	3	4	5	6	7	8	
POSITIVE REGULATION OF POSTTRANSCRIPTIONAL GENE SILENCING (GO:0060148)	0.025109	0.116300	0.060048	0.011540	0.012982	0.035267	0.122656	0.014612	0.081640	0.
REGULATION OF CELL CYCLE PROCESS (GO:0010564)	0.019285	0.077702	0.044203	0.009638	0.009380	0.068632	0.084914	0.008132	0.100553	0.
ANGIOTENSIN-ACTIVATED SIGNALING PATHWAY (GO:0038166)	-0.016968	0.137097	0.026035	0.003299	0.066259	-0.136697	0.049524	0.033761	0.048572	-0.
DNA-TEMPLATED TRANSCRIPTION, TERMINATION (GO:0006353)	0.023468	0.049847	0.046839	0.012888	-0.001433	0.030453	0.064282	0.010822	0.094058	0.
REGULATION OF PROTEIN SUMOYLATION (GO:0033233)	-0.011315	-0.025177	0.027910	-0.006089	0.013126	0.058512	0.006767	0.009965	0.003579	0.
...	...	...	...	...	...	...	...	...	...	...
POSITIVE REGULATION OF MITOCHONDRIAL MEMBRANE POTENTIAL (GO:0010918)	-0.009319	0.194031	0.063745	0.031552	0.009374	-0.042017	0.040068	0.007974	0.136731	0.
TRNA AMINOACYLATION FOR PROTEIN TRANSLATION (GO:0006418)	0.012072	0.092235	0.061207	0.013417	-0.008695	-0.018159	0.066635	0.011384	0.125229	0.
TRYPTOPHAN METABOLIC PROCESS (GO:0006568)	-0.009496	0.024302	-0.008865	-0.006510	-0.009717	0.140466	0.011199	0.000402	0.022354	-0.
GUANOSINE-CONTAINING COMPOUND METABOLIC PROCESS (GO:1901068)	0.015196	0.124744	0.050128	0.002735	-0.004368	0.052860	0.075684	0.005699	0.070609	0.
TRNA MODIFICATION (GO:0006400)	0.008834	0.037770	0.045411	0.004414	0.004794	-0.019834	0.027451	-0.002077	0.077026	0.

5103 rows x 51 columns

**Figure 53.**

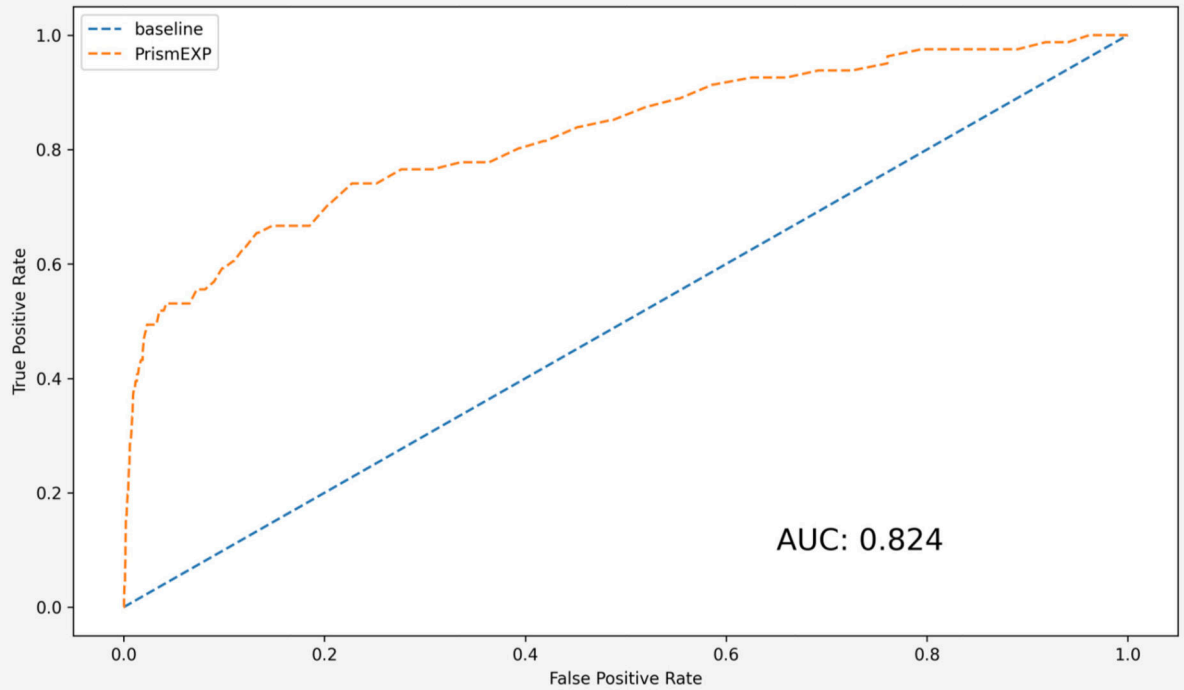
Dataframe of average correlations between each gene set from the specified gene set library and the query gene from the previous 51 correlation matrices.

### Prediction Validation

The ROC curve shows how well previously known annotations for the gene have been recovered by PrismEXP. For this, all gene sets in the library are ranked by the prediction score in descending order. Previously known associations should rank high. The AUC can vary by gene and gene set library.

In [13]

```
calculateGeneAUC(predictions, GENE, rev_library)
```



**Figure 54.**

ROC curve that quantifies the ability of the PrismEXP model to retrieve previously known associations between gene set annotations and the query gene.

## Top Predictions

The top 20 predictions are shown below. The p-values are calculated from the z-scores of the gene set library.

In [14]

top\_predictions

	predictions	z-score	p-value	bonferroni
PROTEIN UBIQUITINATION (GO:0016567)	0.930000	5.496095	1.941468e-08	0.000099
REGULATION OF MITOTIC CELL CYCLE (GO:0007346)	0.930000	5.496095	1.941468e-08	0.000099
STRESS-ACTIVATED PROTEIN KINASE SIGNALING CASCADE (GO:0031098)	0.880000	5.112469	1.589879e-07	0.000811
MODIFICATION-DEPENDENT PROTEIN CATABOLIC PROCESS (GO:0019941)	0.860000	4.959018	3.542517e-07	0.001808
PEPTIDYL-THREONINE PHOSPHORYLATION (GO:0018107)	0.860000	4.959018	3.542517e-07	0.001808
CELLULAR PROTEIN LOCALIZATION (GO:0034613)	0.850000	4.882293	5.242967e-07	0.002675
REGULATION OF CELLULAR RESPONSE TO STRESS (GO:0080135)	0.850000	4.882293	5.242967e-07	0.002675
VESICLE-MEDIATED TRANSPORT (GO:0016192)	0.830000	4.728842	1.129018e-06	0.005761
PROTEIN MODIFICATION BY SMALL PROTEIN REMOVAL (GO:0070646)	0.830000	4.728842	1.129018e-06	0.005761
REGULATION OF TELOMERASE ACTIVITY (GO:0051972)	0.820000	4.652117	1.642722e-06	0.008383
PROTEIN MODIFICATION BY SMALL PROTEIN CONJUGATION (GO:0032446)	0.820000	4.652117	1.642722e-06	0.008383
POSITIVE REGULATION OF TELOMERE MAINTENANCE VIA TELOMERASE (GO:0032212)	0.810000	4.575392	2.376647e-06	0.012128
PROTEIN DEUBIQUITINATION (GO:0016579)	0.800000	4.498667	3.419049e-06	0.017447
PEPTIDYL-THREONINE MODIFICATION (GO:0018210)	0.800000	4.498667	3.419049e-06	0.017447
PROTEIN POLYUBIQUITINATION (GO:0000209)	0.790000	4.421941	4.890900e-06	0.024958
REGULATION OF CELLULAR RESPONSE TO HEAT (GO:1900034)	0.790000	4.421941	4.890900e-06	0.024958
CHEMICAL SYNAPTIC TRANSMISSION (GO:0007268)	0.785833	4.389973	5.668249e-06	0.028925
CELLULAR RESPONSE TO DOPAMINE (GO:1903351)	0.780000	4.345216	6.956930e-06	0.035501
FC RECEPTOR MEDIATED STIMULATORY SIGNALING PATHWAY (GO:0002431)	0.780000	4.345216	6.956930e-06	0.035501
RESPONSE TO DOPAMINE (GO:1903350)	0.780000	4.345216	6.956930e-06	0.035501

**Figure 55.**

Table of top predicted associations for the query gene.



## Download Files

Download full prediction table and high resolution ROC plots. (If plots are opened in the browser the images might require a refresh of the browser page to display)

```
In [15] display(FileLink(GENE+"_"+LIBRARY+"_predictions.tsv", result_html_prefix=str('Download predictio
display(FileLink(GENE+"_"+LIBRARY+"_ROC.pdf", result_html_prefix=str('Download PDF: ')))
display(FileLink(GENE+"_"+LIBRARY+"_ROC.png", result_html_prefix=str('Download PNG: ')))
```

Download prediction table: [MAPK1\\_GO\\_Biological\\_Process\\_2018\\_predictions.tsv](#)

Download PDF: [MAPK1\\_GO\\_Biological\\_Process\\_2018\\_ROC.pdf](#)

Download PNG: [MAPK1\\_GO\\_Biological\\_Process\\_2018\\_ROC.png](#)

### Figure 56.

Download links to prediction table and ROC curve image.

# Geneshot

[PubMed Query](#) | [Gene Function Prediction](#) | [Gene Set Augmentation](#) | [Help](#) | [Download](#) | [API](#)

Submit biomedical terms to receive ranked lists of relevant genes

Search for these terms:

And NOT for these terms:

Examples: [Wound healing](#) | [Hair loss](#) | [Trichostatin A](#) | [Glioblastoma](#) | [Diabetes](#)

Top Associated Genes to Make Predictions:

GeneRIF
  AutoRIF

Submit

Submit any search terms to Geneshot to receive prioritized genes that are most relevant to the search terms. Geneshot finds publications that mention both the search terms and genes. It then prioritizes these genes using various methods: 1) list of genes from publications; 2) predicted genes using gene-gene similarity matrices derived from a variety of resources ([ARCHS4](#) | [Enrichr](#) | [Tagger](#) | [AutoRIF](#) | [GeneRIF](#)).



Please acknowledge Geneshot in your publications by citing the following reference:

Geneshot: search engine for ranking genes from arbitrary text queries

Alexander Lachmann, Brian M. Schilder, Megan L. Wojciechowicz, Denis Torre, Maxim V. Kuleshov, Alexandra B. Keenan, and Avi Ma'ayan

Nucleic Acids Research, gkz393, <https://doi.org/10.1093/nar/gkz393>

Volume 47, Issue W1, 02 July 2019, Pages W571–W577

**Figure 57.**

Geneshot homepage. The search bars allow for querying terms to be included and omitted from the search. Additional options exist for toggling between GeneRIF and AutoRIF and adjusting the gene set size for making predictions.

# Geneshot

[PubMed Query](#) | [Gene Function Prediction](#) | [Gene Set Augmentation](#) | [Help](#) | [Download](#) | [API](#)

Submit biomedical terms to receive ranked lists of relevant genes

Search for these terms:

Wound healing x Search

And NOT for these terms:

Search

Examples: [Wound healing](#) | [Hair loss](#) | [Trichostatin A](#) | [Glioblastoma](#) | [Diabetes](#)

Top Associated Genes to Make Predictions:

GeneRIF
  AutoRIF

Submit

LOADING

Submit any search terms to Geneshot to receive prioritized genes that are most relevant to the search terms. Geneshot finds publications that mention both the search terms and genes. It then prioritizes these genes using various methods: 1) list of genes from publications; 2) predicted genes using gene-gene similarity matrices derived from a variety of resources ( [ARCHS4](#) | [Enrichr](#) | [Tagger](#) | [AutoRIF](#) | [GeneRIF](#) ).



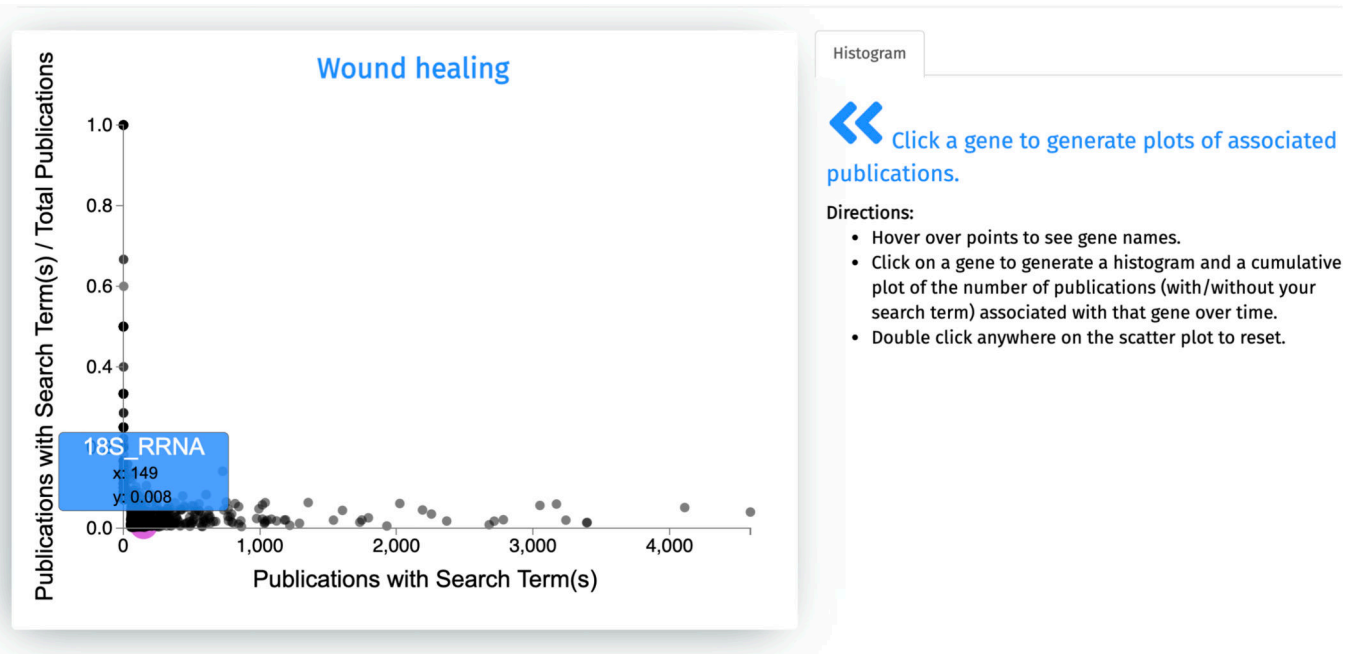
Loading...

Querying PubMed, wait depends on the number of matching publications (up to 1 min).

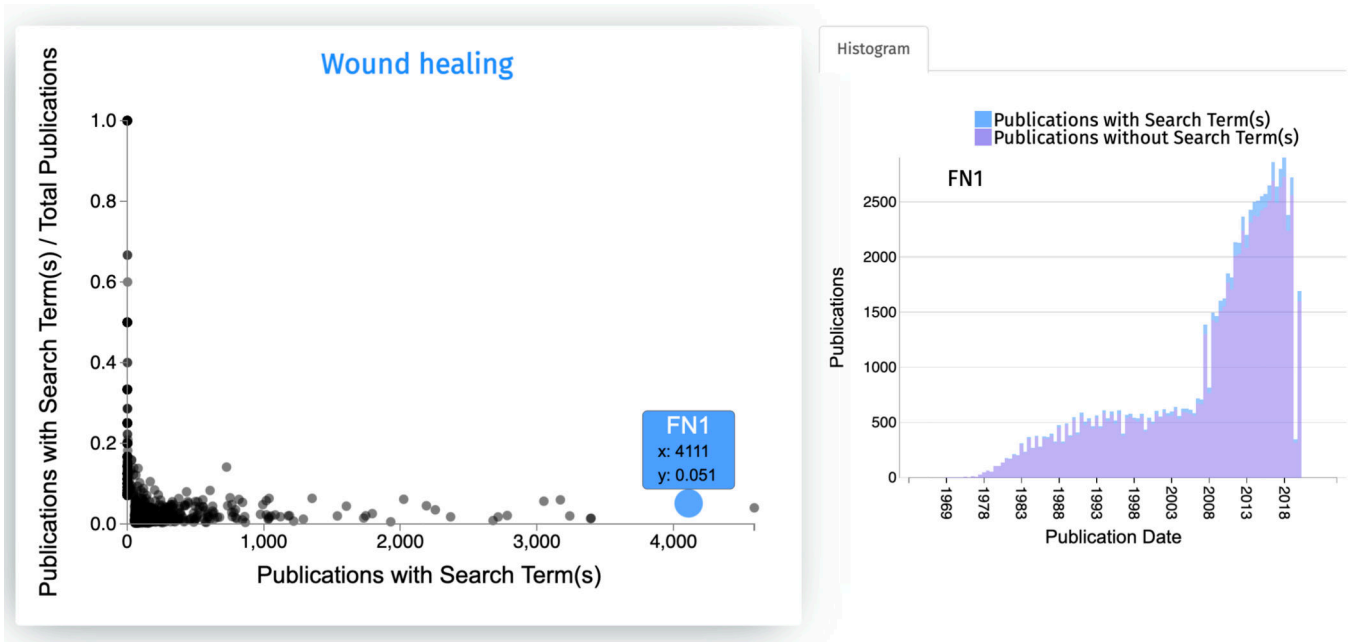


© The Ma'ayan Lab

**Figure 58.**  
Submitted search form populated with the term “Wound healing”.

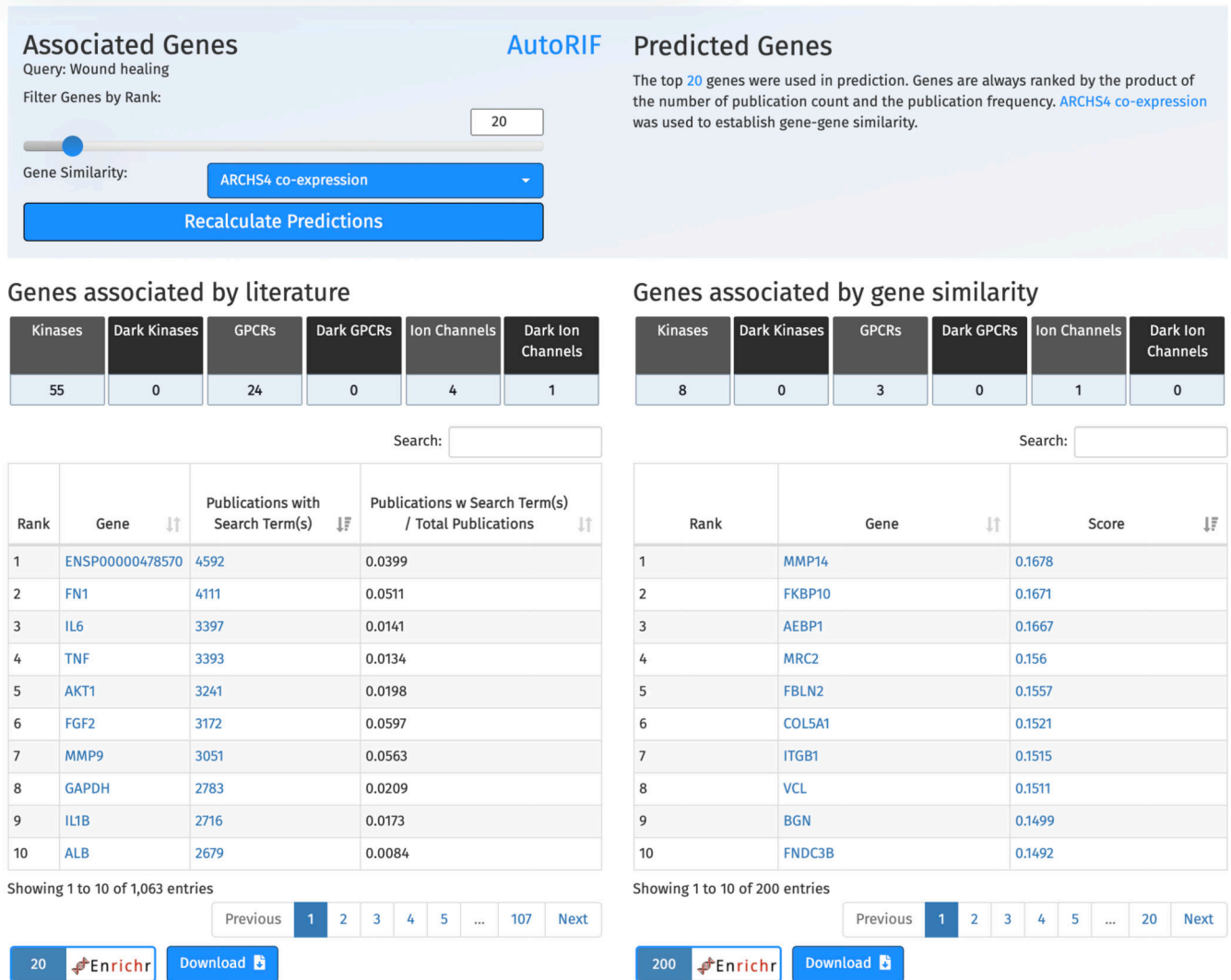


**Figure 59.** Scatter plot of all genes associated with “wound healing”. Each point represents a gene and interacting with any point reveals the gene name, X-axis value, and Y-axis value.



**Figure 60.**

Clicking on any of the points in the scatter plot generates a histogram of associations between the gene and “wound healing” over time. The blue bars represent publications mentioning the gene and search term, whereas purple bars represent publications mentioning just the gene.



**Figure 61.**

Table of top genes associated with “wound healing” ranked by number of publications that mention the gene and search term (left). Table of genes predicted to be associated with “wound healing” based on co-expression with the literature derived genes (right). Both tables can be downloaded and the genes from both tables can be submitted to Enrichr for gene set enrichment analysis.

**Associated Genes** AutoRIF

Query: Wound healing

Filter Genes by Rank:

Gene Similarity: **ARCHS4 co-expression**

Gene Similarity options:

- ARCHS4 co-expression
- GeneRIF co-occurrence
- Enrichr co-occurrence
- Tagger co-occurrence
- ARCHS4 co-expression

Recalculate

Genes associated

**Figure 62.**

The predicted gene table from the “wound healing” search can be recalculated by selecting a different gene-gene similarity matrix for predictions and changing the gene set size derived from the associated gene table.

# Geneshot

[PubMed Query](#) | [Gene Function Prediction](#) | [Gene Set Augmentation](#) | [Help](#) | [Download](#) | [API](#)

Choose a gene and predict properties using gene-gene similarity associations

Gene:

Submit

Library:

WikiPathways

Similarity:

ARCHS4 co-expression

Example: TNF / GO\_Biological\_Process\_2017b / Co-expression

Submit Entrez gene terms to Geneshot and select a gene set library for terms to receive prioritized terms that are most relevant to the searched gene. Geneshot prioritizes the terms using gene-gene similarity matrices derived from a variety of resources ([ARCHS4](#) | [Enrichr](#) | [Tagger](#) | [AutoRIF](#) | [GeneRIF](#)).



**Figure 63.**

Gene function prediction page. The input form allows for the selection of a query gene, a gene set library from which gene sets with functional association terms will be retrieved, and a gene-gene similarity matrix from which predictions will be made.



TNF KEGG Pathways ARCHS4 co-expression

0.89

Show 10 entries

Search:

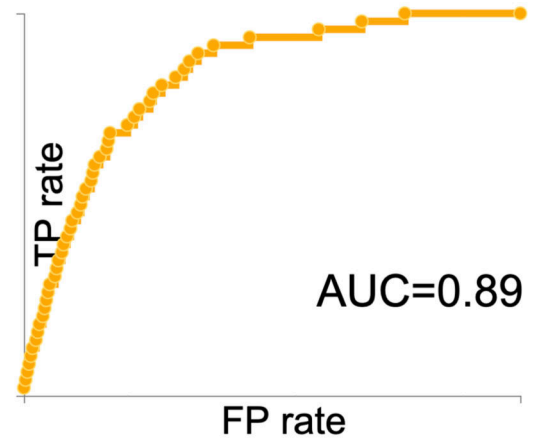
Rank	Property	Score	Known
1	Graft-versus-host disease Homo sapiens hsa05332	0.1605	1
2	Allograft rejection Homo sapiens hsa05330	0.1243	1
3	Leishmaniasis Homo sapiens hsa05140	0.1054	1
4	Inflammatory bowel disease (IBD) Homo sapiens hsa05321	0.1005	1
5	NF-kappa B signaling pathway Homo sapiens hsa04064	0.0998	1
6	NOD-like receptor signaling pathway Homo sapiens hsa04621	0.0965	1
7	Staphylococcus aureus infection Homo sapiens hsa05150	0.0964	0
8	Toll-like receptor signaling pathway Homo sapiens hsa04620	0.0954	1
9	Asthma Homo sapiens hsa05310	0.0932	1
10	Rheumatoid arthritis Homo sapiens hsa05323	0.0916	1

Showing 1 to 10 of 200 entries

Previous 1 2 3 4 5 ... 20

Next

Download



**Figure 64.**

Table of top predicted associations for TNF from the KEGG Pathways gene set library. Known functions are highlighted in blue. The ROC curve quantifies the ability of the prediction method to retrieve functions that TNF is known to be associated with.

# Geneshot

[PubMed Query](#) | [Gene Function Prediction](#) | [Gene Set Augmentation](#) | [Help](#) | [Download](#) | [API](#)

Upload a gene set to receive related genes based on a similarity matrix

Gene Set:

Enter gene symbols...

Examples: [common genes](#) | [mixed genes](#) | [uncommon genes](#)

Similarity:

ARCHS4 co-expression

GeneRIF

AutoRIF

Submit

Submit a gene set to Geneshot and select a gene-gene similarity matrix to receive prioritized genes that are most relevant to the submitted gene set. Geneshot prioritizes the additional genes using gene-gene similarity matrices derived from a variety of resources ( [ARCHS4](#) | [Enrichr](#) | [Tagger](#) | [GeneRIF](#) ).



**Figure 65.**

Gene set augmentation page. The text box accepts a list of gene symbols that will be used as an unweighted gene set to predict related genes based on the selected gene-gene similarity matrix. The source of gene publication data can be changed with a toggle bar between GeneRIF and AutoRIF.

Gene Set:

SYVN1  
CD226  
TP53AIP1  
PABPC1  
OSBPL3  
CHN2  
IL1RAP

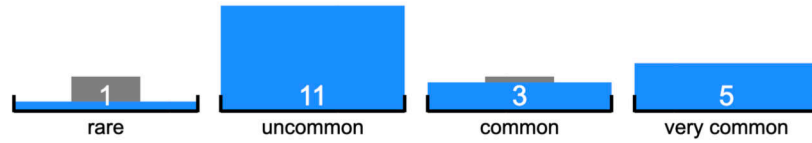
Examples: [common genes](#) | [mixed genes](#) | [uncommon genes](#)

Similarity:  
ARCHS4 co-expression

GeneRIF  AutoRIF

Submit

Submit a gene set to Geneshot and select a gene-gene similarity matrix to receive prioritized genes that are most relevant to the submitted gene set. Geneshot prioritizes the additional genes using gene-gene similarity matrices derived from a variety of resources ( [ARCHS4](#) | [Enrichr](#) | [Tagger](#) | [GeneRIF](#) ).



**Figure 66.**

The “mixed genes” example query with the quantile counts for each of the queried genes.

AutoRIF  ARCHS4 co-expression

## User Upload

Search: 

Rank	Gene	Publication Count	Novelty
1	<a href="#">BIRC3</a>	3641	very common
2	<a href="#">PABPC1</a>	2407	very common
3	<a href="#">CD226</a>	1948	very common
4	<a href="#">RAB27A</a>	1927	very common
5	<a href="#">NCF4</a>	1696	very common
6	<a href="#">CEACAM6</a>	1087	common
7	<a href="#">IL1RAP</a>	798	common
8	<a href="#">SPOP</a>	796	common
9	<a href="#">SYVN1</a>	341	uncommon
10	<a href="#">CHN2</a>	297	uncommon

Showing 1 to 10 of 20 entries

Previous [1](#) [2](#) Next
[20](#)  [Download](#)

## Predicted Genes

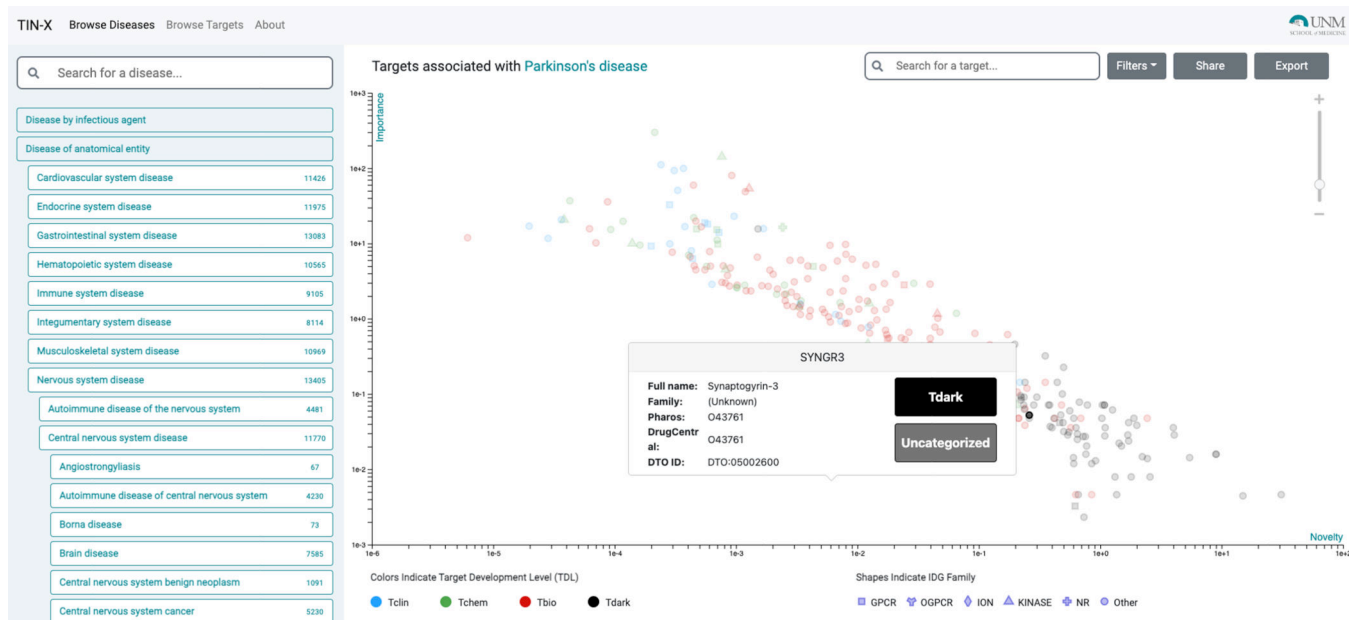
Search: 

Rank	Predicted Gene	Publications	Score
1	<a href="#">RHOG</a>	2134	0.0727
2	<a href="#">PLP1</a>	2580	0.0701
3	<a href="#">ERMN</a>	87	0.0692
4	<a href="#">WIPF1</a>	652	0.0687
5	<a href="#">BC017643</a>	0	0.0685
6	<a href="#">EVI2A</a>	103	0.0672
7	<a href="#">RNF13</a>	52	0.0668
8	<a href="#">CNP</a>	861	0.0662
9	<a href="#">KLRA1</a>	0	0.0659
10	<a href="#">MOG</a>	6634	0.0659

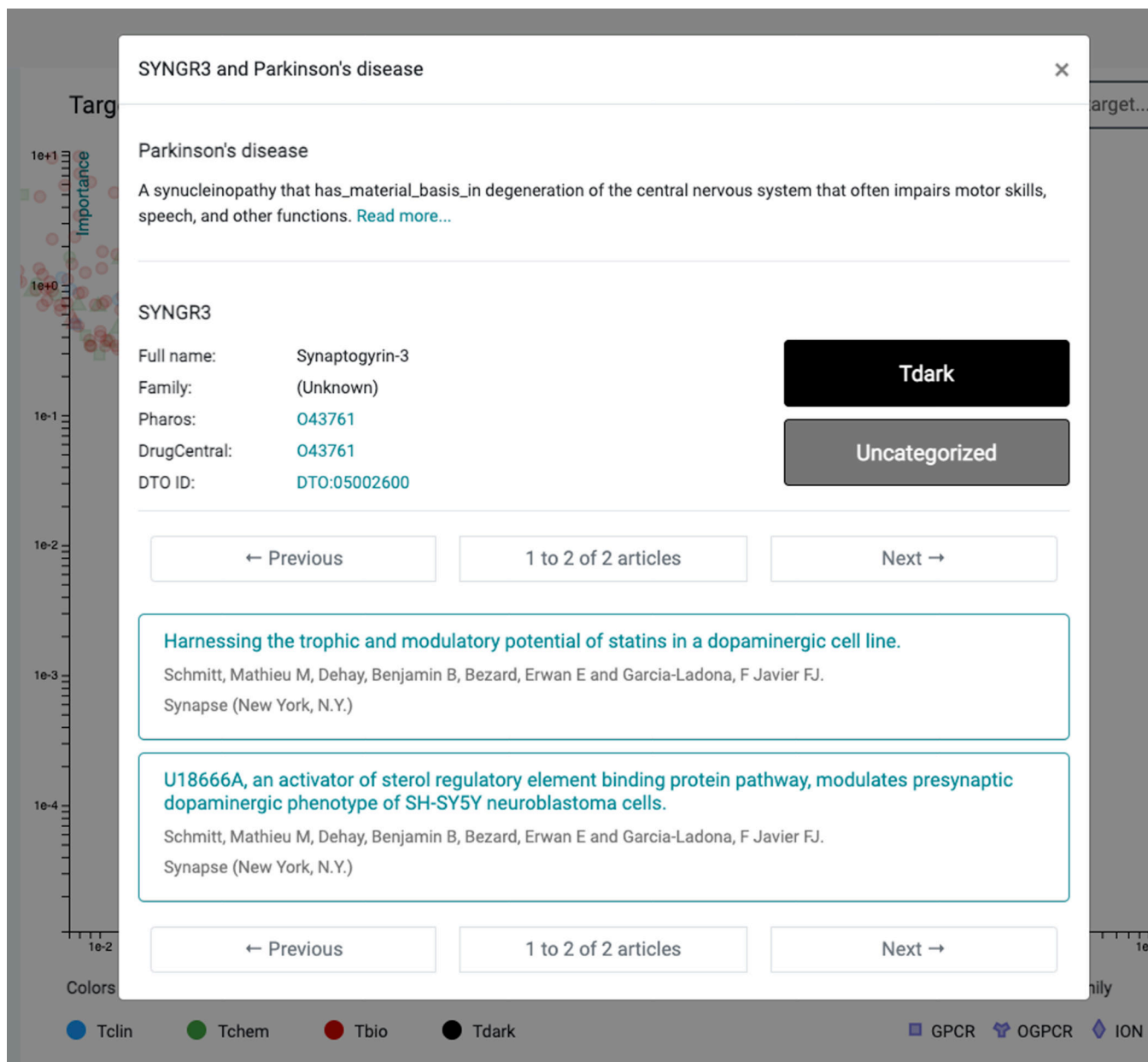
Showing 1 to 10 of 200 entries

Previous [1](#) [2](#) [3](#) [4](#) [5](#) ... [20](#) Next
[200](#)  [Download](#)
**Figure 67.**

Table of queried genes, their publication counts, and novelty (left). Table of top 200 genes predicted to be associated with the query gene set, gene publication counts, and similarity score with the query gene set (right). Each table can be downloaded and the genes from each table can be sent to Enrichr for gene set enrichment analysis.

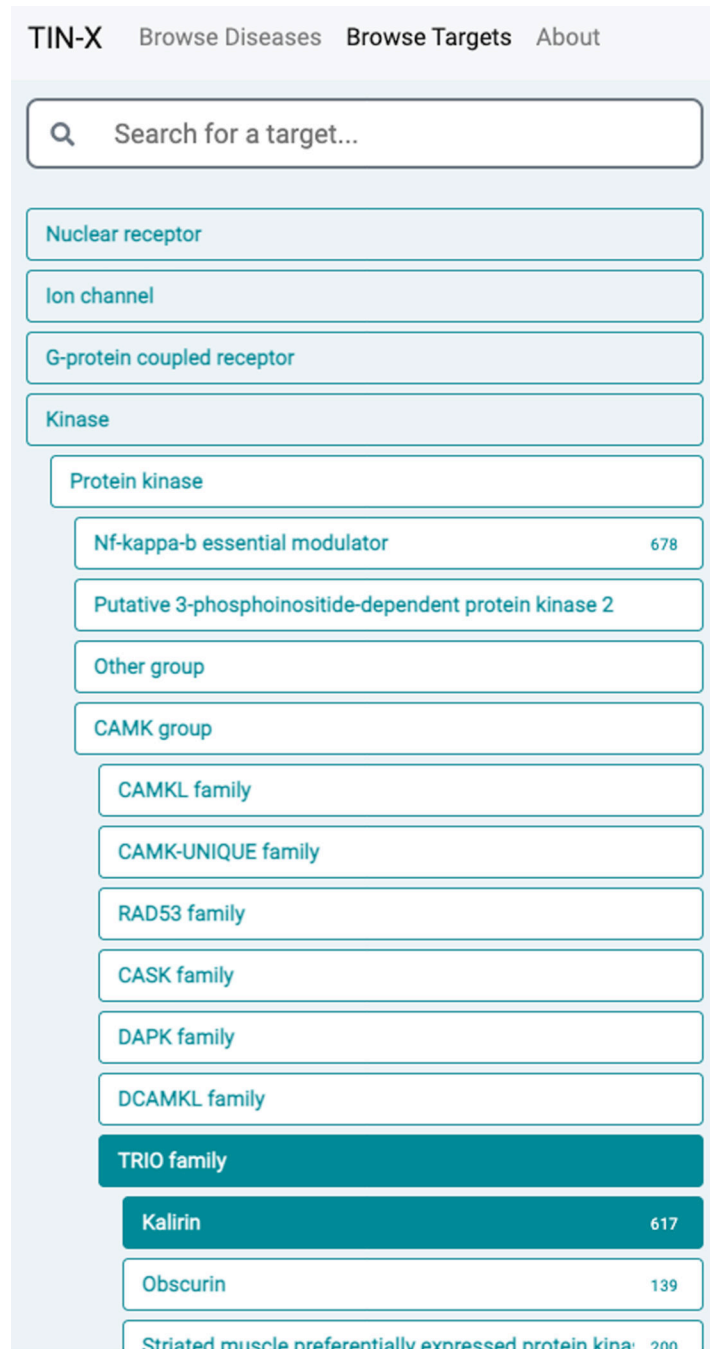


**Figure 68.** The TIN-X “Browse Disease” view (left side) with Parkinson’s Disease selected. Targets associated with Parkinson’s Disease (right side) are plotted on a log scale of Importance vs Novelty, with each data point colored according to its Target Development Level (TDL).



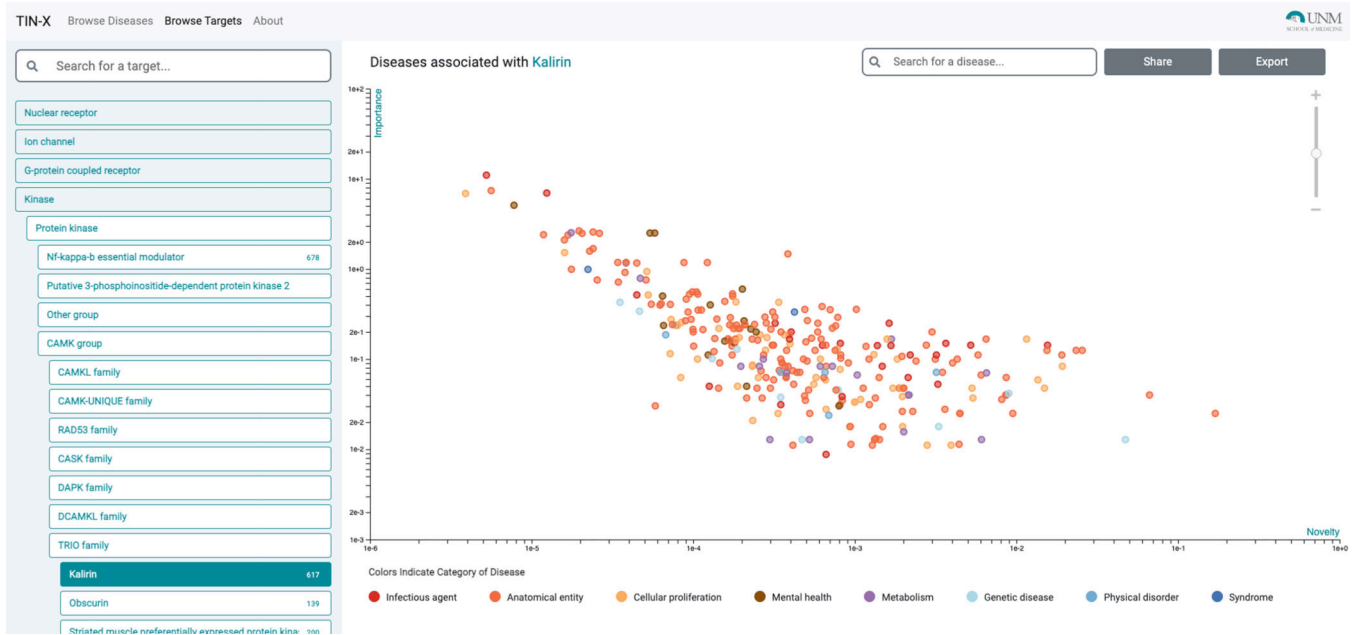
**Figure 69.**

Clicking a target point within the Parkinson's Disease example, "Synaptogyrin-3" (SYNGR3) displays details including the full name and family of the target, Target Development Level (TDL), links to Pharos and DrugCentral, and, importantly, links to the associated two research articles (bottom).

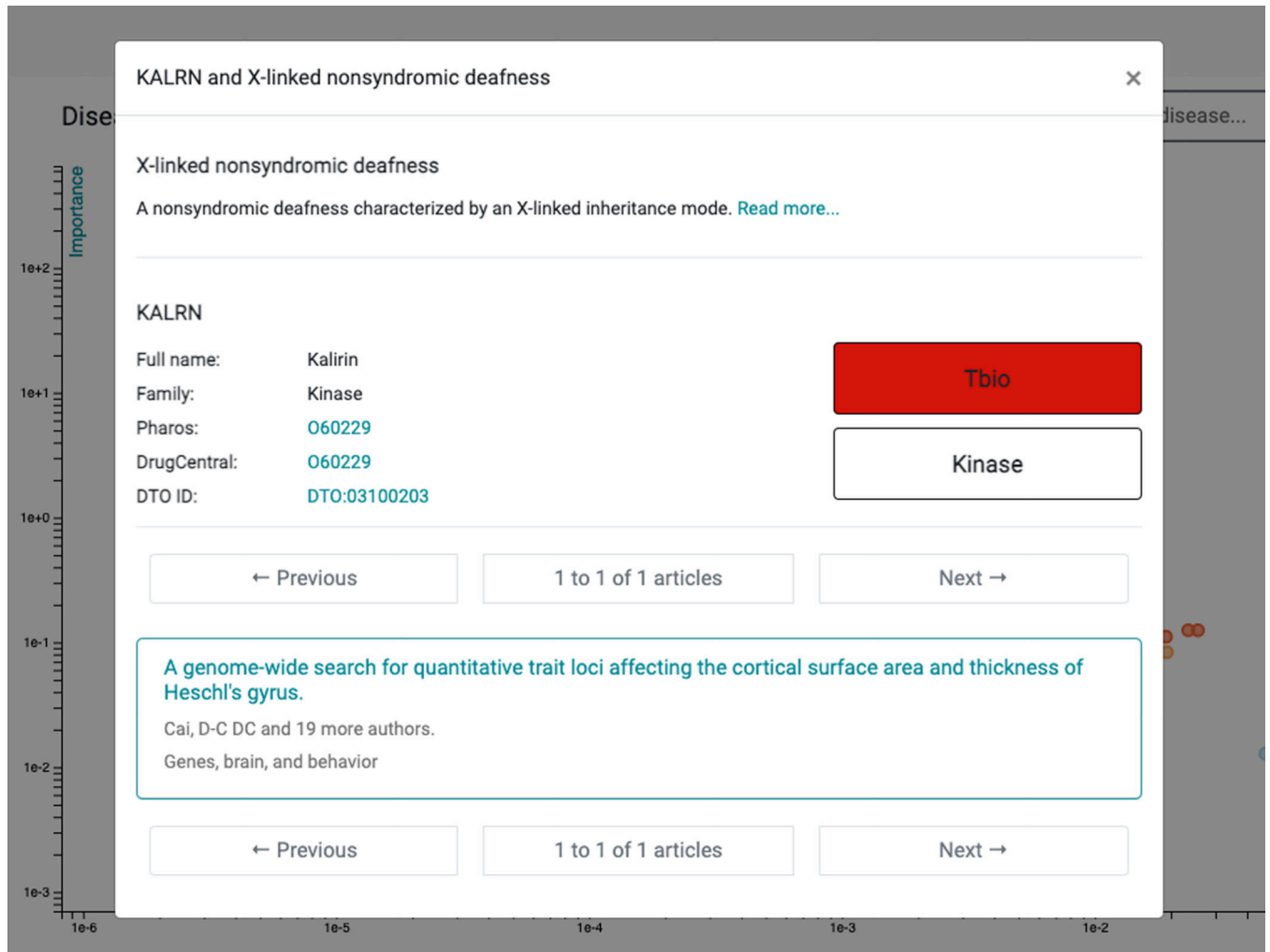


**Figure 70.** Starting with the superfamily Kinase, the user can further refine the selection to Protein kinase → CAMK group → TRIO family → Kalirin by using the left navigation pane within Browse Targets.





**Figure 71.** Within “Browse Targets”, diseases associated with Kalirin (KALRN) are plotted with log–log Importance–Novelty axes, and are colored according to the top hierarchical Disease Ontology term.



**Figure 72.**

For the example target Kalirin (KALRN), the most novel association (lowest Importance) is for “X-linked nonsyndromic deafness”. This detailed view includes the full name and family of the target, links to Pharos and DrugCentral, and in this case, the one article responsible for this association between KALRN and X-linked nonsyndromic deafness.

DrugCentral 2021  
Online drug Compendium

Search Redial About Download L1000 signature FAQ

4,542 Drugs  
112,577 pharmaceuticals

Enter: Drug, Target, Disease

Featured News

[The Latest in Chemistry in Coronavirus Research](#)

Drugs in the News

[Venetoclax Dapagliflozin KEYTRUDA Sacubitril LORBRENA Hydroxychloroquine](#)

DrugCentral Search Overview

Drug search example: Target search example: Pharmacologic action search

Citing DrugCentral Database, November 05, 2020 © 2021. License Designed for Mobile

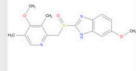
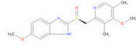
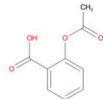
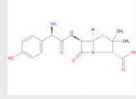
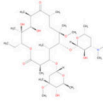
**Figure 73.** DrugCentral homepage. DrugCentral search bar supports three types of queries: drug, target and disease.


DrugCentral 2021  
Online drug Compendium

Search    Redial    About    Download    L1000 signature    FAQ

omeprazole

Results: 7

<a href="#">omeprazole</a>	A 4-methoxy-3,5-dimethylpyridyl, 5-methoxybenzimidazole derivative of timoprazole that is used in the therapy of STOMACH ULCERS and ZOLLINGER-ELLISON SYNDROME. The drug inhibits an H(+)-K(+)-EXCHANGING ATPASE which is found in GASTRIC PARIETAL CELLS.	
<a href="#">esomeprazole</a>	The S-isomer of omeprazole.	
<a href="#">sodium bicarbonate</a>	A white, crystalline powder that is commonly used as a pH buffering agent, an electrolyte replenisher, systemic alkalinizer and in topical cleansing solutions.	
<a href="#">acetylsalicylic acid</a>	The prototypical analgesic used in the treatment of mild to moderate pain. It has anti-inflammatory and antipyretic properties and acts as an inhibitor of cyclooxygenase which results in the inhibition of the biosynthesis of prostaglandins. Aspirin also inhibits platelet aggregation and is used in the prevention of arterial and venous thrombosis. (From Martindale, The Extra Pharmacopoeia, 30th ed, p5)	
<a href="#">amoxicillin</a>	A broad-spectrum semisynthetic antibiotic similar to AMPICILLIN except that its resistance to gastric acid permits higher serum levels with oral administration.	
<a href="#">clarithromycin</a>	A semisynthetic macrolide antibiotic derived from ERYTHROMYCIN that is active against a variety of microorganisms. It can inhibit PROTEIN SYNTHESIS in BACTERIA by reversibly binding to the 50S ribosomal subunits. This inhibits the translocation of aminoacyl transfer-RNA and prevents peptide chain elongation.	

Citing DrugCentral Database, November 05 2020 © 2021. License  Designed for Mobile

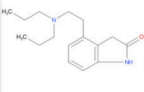
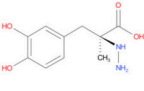
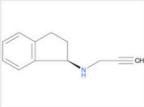
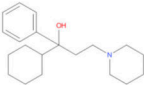
**Figure 74.** DrugCentral search results for “Omeprazole” first lists drugs indicated for “Omeprazole” (e.g., sodium bicarbonate) followed by drugs indicated in complications.

DrugCentral 2021  
Online drug Compendium

Search Redial About Download L1000 signature FAQ

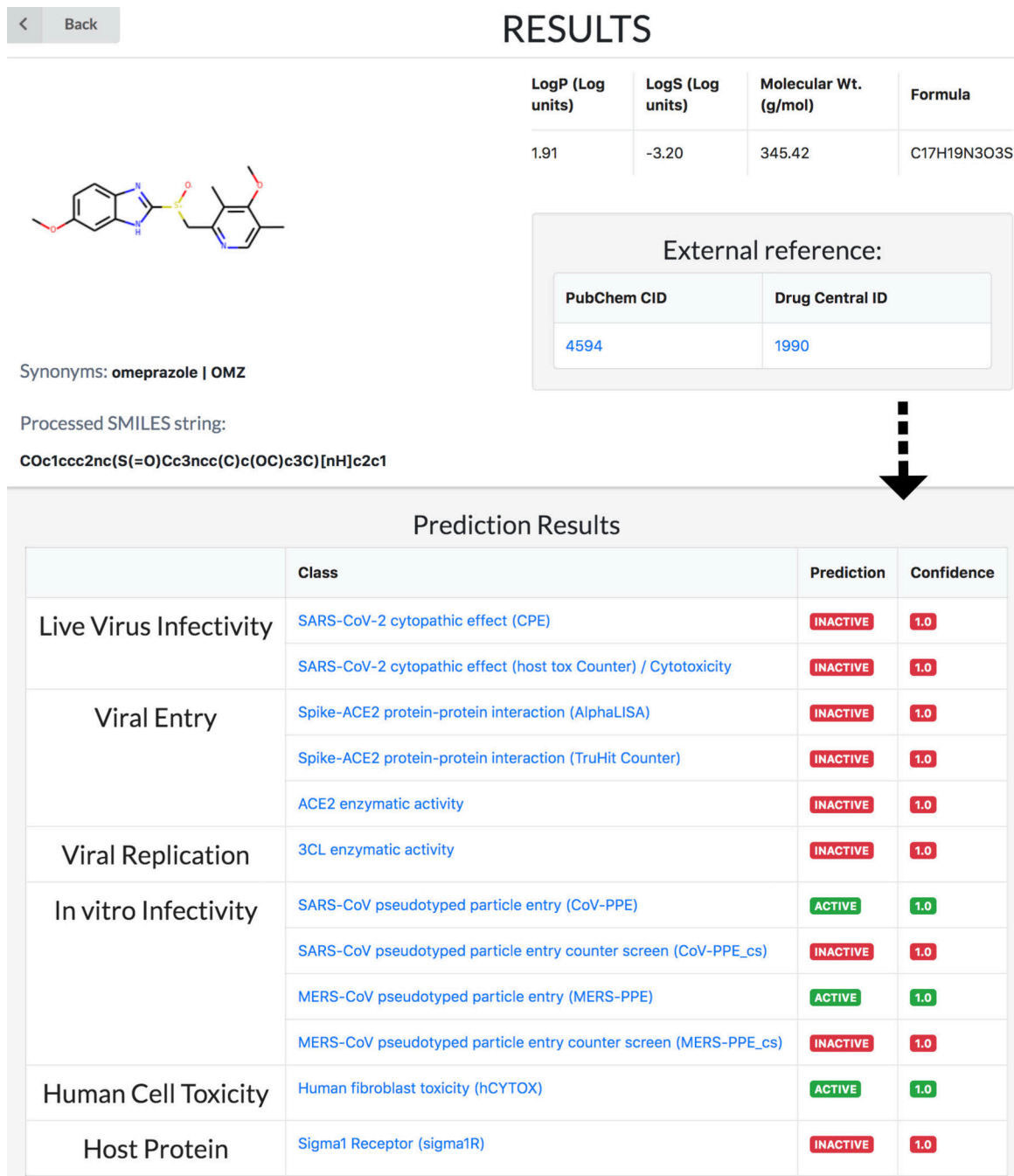
Parkinson's disease

Results: 100

<a href="#">ropinirole</a>	Ropinirole is a non-ergoline dopamine agonist. The precise mechanism of action of ropinirole as a treatment for Parkinson's disease is unknown, although it is thought to be related to its ability to stimulate dopamine D2 receptors within the caudate-putamen in the brain.	
<a href="#">carbidopa</a>	An inhibitor of DOPA DECARBOXYLASE that prevents conversion of LEVODOPA to dopamine. It is used in PARKINSON DISEASE to reduce peripheral adverse effects of LEVODOPA. It has no anti-parkinson activity by itself.	
<a href="#">rasagiline</a>	Rasagiline is a selective, irreversible MAO-B inhibitor indicated for the treatment of idiopathic Parkinson's disease. The results of a clinical trial designed to examine the effects of rasagiline tablets on blood pressure when it is administered with increasing doses of tyramine indicates the functional selectivity can be incomplete when healthy subjects ingest large amounts of tyramine while receiving recommended doses of rasagiline tablets. The selectivity for inhibiting MAO-B diminishes in a dose-related manner. One mechanism is believed to be related to its MAO-B inhibitory activity, which causes an increase in extracellular levels of dopamine in the striatum. The elevated dopamine level and subsequent increased dopaminergic activity are likely to mediate rasagiline's beneficial effects seen in models of dopaminergic motor dysfunction.	
<a href="#">trihexyphenidyl</a>	One of the centrally acting MUSCARINIC ANTAGONISTS used for treatment of PARKINSONIAN DISORDERS and drug-induced extrapyramidal movement disorders and as an antispasmodic.	

**Figure 75.**

Drugcentral query result for "Parkinson's disease" (PD) first lists drugs indicated for PD (e.g., ropinirole), followed by drugs indicated in complications of PD (e.g., fludrocortisone is indicated for the PD-associated orthostatic hypotension), then by drugs that list PD as side-effect (e.g., dimenhydrinate).



**Figure 76.**

DrugCentral Redial query result for Omeprazole. All input queries for REDIAL-2020 are converted to SMILES format in order to predict anti-viral properties.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

The screenshot shows the DrugCentral 2021 website interface. The browser address bar displays 'drugcentral.org/LINCS'. The page title is 'DrugCentral 2021 Online drug Compendium'. The navigation menu includes 'Search', 'Redial', 'About', 'Download', 'L1000 signature', and 'FAQ'. A 'Help' button is located in the top right corner.

The main content area is titled 'Drug gene signature profile similarity'. It features a search input field with 'omeprazole' entered. Below the search field, there are options to 'Rank by': 'Pearson correlation' (selected), 'Normalized RMSD', and 'RMSD'. A dashed arrow points to the 'Normalized RMSD' option.

The results are displayed in a table with the following columns: Drug name, Cell lines count, RMSD, Normalized RMSD, and Pearson correlation. The table shows 10 results, all with a cell lines count of 7. The drugs listed are: diosmin, cinafloxacin, oxandrolone, carbimazole, roxithromycin, talinolol, etoricoxib, carisoprodol, temazepam, and acenocoumarol.

Drug name	Cell lines count	RMSD	Normalized RMSD	Pearson correlation
diosmin	7	2.72	1.05	0.39
cinafloxacin	7	2.78	1.04	0.39
oxandrolone	7	2.72	1.02	0.38
carbimazole	7	2.63	0.99	0.38
roxithromycin	7	2.34	0.95	0.38
talinolol	7	2.76	1.04	0.38
etoricoxib	7	2.59	0.92	0.38
carisoprodol	7	2.68	1.00	0.38
temazepam	7	2.61	1.00	0.37
acenocoumarol	7	2.56	0.99	0.37

Showing 1 to 10 of 1,624 entries. A 'Download' button is located below the table. A pagination control shows 'Previous', '1', '2', '3', '4', '5', '...', '163', and 'Next'.

DrugCentral © 2021. License

**Figure 77.**

The L1000 search input home page. The L1000 DrugCentral app allows users to query (via drug names) which drugs have the most similar gene perturbation profiles, ranked by cell lines.



Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

DrugCentral 2021  
Online drug Compendium

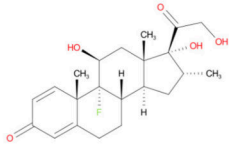
Search Redial About Download L1000 signature FAQ

Enter: Drug, Target, Disease

dexamethasone [Indications/Contra](#) | [FAERs-F](#) | [FAERs-M](#) | [Orange Bk](#) | [BioActivity](#)

Stem definition	Drug id	CAS RN
prednisone and prednisolone derivatives	824	50-02-2

**Description:**

Molecule	Description
 <p><a href="#">Molfile</a> <a href="#">Inchi</a> <a href="#">Smiles</a></p> <p><b>Synonyms:</b></p> <ul style="list-style-type: none"> <li>dexamethasone</li> <li>dexasone</li> <li>dexmethsone</li> </ul>	<p>An anti-inflammatory 9-fluoro-glucocorticoid.</p> <ul style="list-style-type: none"> <li>Molecular weight: 392.47</li> <li>Formula: C<sub>22</sub>H<sub>29</sub>FO<sub>5</sub></li> <li>CLOGP: 1.79</li> <li>LIPINSKI: 0</li> <li>HAC: 5</li> <li>HDO: 3</li> <li>TPSA: 94.83</li> <li>ALOGS: -3.89</li> <li>ROTB: 2</li> </ul> <p>Status: OFP</p> <p><i>Legend:</i> OFP - off patent OFM - off market ONP - on patent</p>

**Figure 78.** DrugCentral Accession “DrugcentralStruct.ID” for cross referencing DrugCentral drug cards.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

DrugCentral 2021  
Online drug Compendium

Search: Redial About Download L1000 signature FAQ

Enter: Drug, Target, Disease

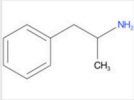
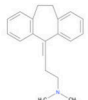
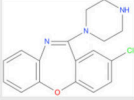
### Sodium-dependent noradrenaline transporter

#### Description:

Description

- Accession: P23975
- Swissprot: SC6A2\_HUMAN
- Organism: Homo sapiens
- Gene: SLC6A2
- Target class: Transporter

#### Drug Relations:

<a href="#">amphetamine</a>	A powerful central nervous system stimulant and sympathomimetic. Amphetamine has multiple mechanisms of action including blocking uptake of adrenergics and dopamine, stimulation of release of monoamines, and inhibiting monoamine oxidase. Amphetamine is also a drug of abuse and a psychotomimetic. The l- and the d,l-forms are included here. The l-form has less central nervous system activity but stronger cardiovascular effects. The d-form is DEXTROAMPHETAMINE.	<a href="#">Bioactivity details</a>	MOA ✓	
<a href="#">amitriptyline</a>	Tricyclic antidepressant with anticholinergic and sedative properties. It appears to prevent the re-uptake of norepinephrine and serotonin at nerve terminals, thus potentiating the action of these neurotransmitters. Amitriptyline also appears to antagonize cholinergic and alpha-1 adrenergic responses to bioactive amines.	<a href="#">Bioactivity details</a>	MOA ✓	
<a href="#">amoxapine</a>	The N-demethylated derivative of the antipsychotic agent LOXAPINE that works by blocking the reuptake of norepinephrine, serotonin, or both; it also blocks dopamine receptors. Amoxapine is used for the treatment of depression.	<a href="#">Bioactivity details</a>	MOA ✓	

**Figure 79.**  
DrugCentral's Target Card. Target card depicts Accession, Swissprot, Organism, Gene & Target class followed by Drug relations where the Drugs Bioactivity mechanism-of-actions are marked.


The screenshot shows the DrugCentral website interface. At the top, there is a navigation bar with the DrugCentral logo and links for Search, Redial, About, Download, L1000 signature, and FAQ. The main content area is titled "DrugCentral Download" and lists several download options, each accompanied by a red pill icon:

- [Download Database dump 9/18/2020](#) (Postgres v10.12) - Contains all information in the DrugCentral database. Requires a new or existing Postgres database setup, please see [Postgresql documentation](#) on how to install, configure and load database contents.
- Also available via public instance at drugcentral:unmtid-dbs.net:5433, username="unmtid", password="dosage", with responsiveness depending on user load.
- [Download-Drug-target interaction data](#) - extracted from literature, drug labels, and external data sources in TSV format
- [SDF file MOL V2000 records](#) - contains chemical structures of drugs, note only V2000 supported records are in this file.
- [SDF file MOL V3000 records](#) - contains chemical structures of drugs, all including records with undefined chemical structure are included here.
- [SQL query examples](#) - Examples of SQL queries to be used on a local instance of Postgres database.
- [SMILES and InChI file](#) - contains structures in SMILES and InChI formats along with INN names, DrugCentral ids, and CAS registry numbers.

Below the download options is a section titled "TCRD Download" with a link: [Files required for TCRD import. Download Files Here.](#) and a note: "Please also see [Target Central](#) on how to install, configure and load database contents."

On the right side of the screenshot, a vertical list of Uniprot Accession IDs is displayed, with a dashed arrow pointing to the "Download" link in the navigation bar. The IDs include: 000444, Q8R4D5, P32246, A3EZ19, D2K2A8, P15388, P49327, P16233, P35236, P09874, P18507, P47869, P47870, Q9HVB9, P50247, P30085, P28907, Q15119, P05181, Q923Y8, P70605, P43681, B1PL86, Q923Y9, Q9RMI3, P18508, P20236, P63138, Q9P1W9, P41231, P05164, P27448, P00915, P50997, Q99640, P46098, Q9Y5Y4, O75676, P06865, P09871, Q52WX2, Q15125, P11802, Q9BVS4, O08858, P30873, P30875, P30935, P49660, Q62968, P43088, Q920D2, P23897, P35918, A3RGC1, Q932Y6, O95271, Q91ZY2, B4URF0, O14578, P28838.

**Figure 80.** Uniprot Accession IDs used for crossreferencing and machine querying DrugCentral Targetcards. [https://drugcentral.org/static/Drugcentral\\_uniprot\\_Mapping.txt](https://drugcentral.org/static/Drugcentral_uniprot_Mapping.txt)



**DrugCentral 2021**  
Online drug Compendium

Search Redial About Download L1000 signature FAQ

## DrugCentral REDIAL 2020

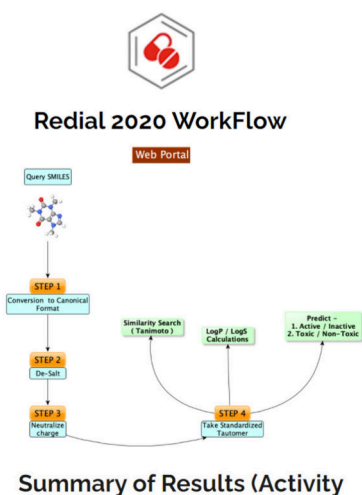
### A portal for estimating Anti-SARS-CoV-2 activities

Provide an Input string:

Some Examples: CC(=O)OC1=CC=CC=C1C(=O)O | Remdesivir | 121304016

**SUBMIT**

Drug central in active development, finalizing frameworks June 29.



**Five filters applied, before building predictive models**

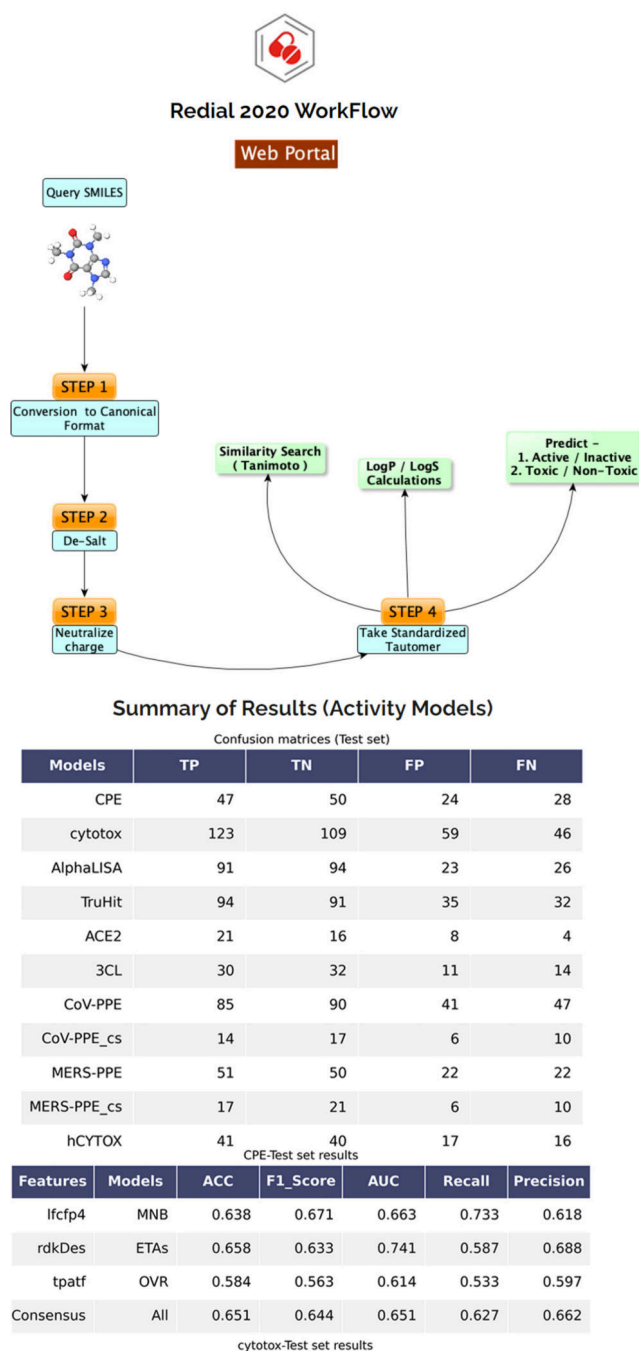
- 1) SMILES were converted into canonical SMILES. Some of the SMILES were not converted into Canonical SMILES, thus discarded.
- 2) RDKit Salt Stripper was implemented to obtain the salt stripped molecules. The

#### Viral Replication

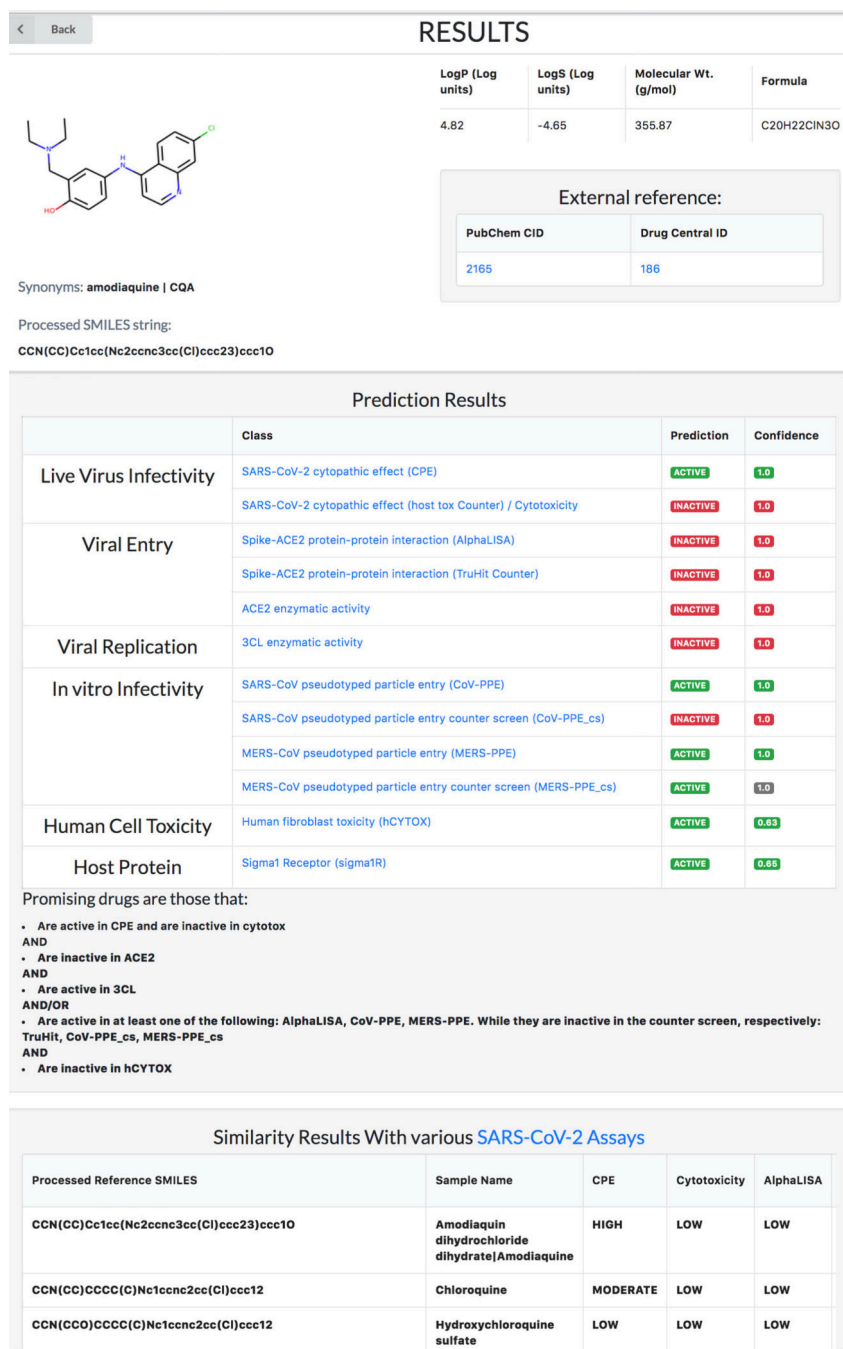
Following entry into the host cell, the main SARS-CoV-2 replication enzyme is 3C-like proteinase (3CL) (also called "main protease" or Mpro) cleaves the two SARS-CoV-2 polyproteins into various proteins that are essential to the viral life cycle (RNA polymerases, helicases, methyltransferases, etc). The inhibition of the 3CL protein causes the disruption of the viral replication process. Hence, 3CL is an attractive drug

Citing Redial, May 03 2021 License  Designed for Mobile

**Figure 81.**  
Redial Home page with Search SMILES, drug names and PubChem CIDs enabled.

**Figure 82.**

Redial interface provides a summary of the models, such as model type, which descriptor categories were used for training and the evaluation scores. The user interface further depicts the processes of cleaning the chemical structures (encoded as SMILES) before training the machine learning models.



**Figure 83.** Redial prediction results table with example search term “amodiaquine”. Amodiaquine is predicted to be active in cytopathic effect experiments while there are no clues on its mechanism (inactive in AlphaLISA, ACE2, and 3CL assays).

LogP (Log units)	LogS (Log units)	Molecular Wt. (g/mol)	Formula
2.20	-2.89	602.59	C27H35N6O8P

External reference:	
PubChem CID	Drug Central ID
<a href="#">121304016</a>	<a href="#">5376</a>

**Figure 84.** REDIAL links directly to DrugCentral for approved drugs and to PubChem for chemicals (where available), enabling easy access to further information on the query molecule.



Prediction Results			
	Class	Prediction	Confidence
Live Virus Infectivity	SARS-CoV-2 cytopathic effect (CPE)	ACTIVE	1.0
	SARS-CoV-2 cytopathic effect (host tox Counter) / Cytotoxicity	INACTIVE	1.0
Viral Entry	Spike-ACE2 protein-protein interaction (AlphaLISA)	ACTIVE	1.0
	Spike-ACE2 protein-protein interaction (TruHit Counter)	INACTIVE	1.0
	ACE2 enzymatic activity	INACTIVE	1.0
Viral Replication	3CL enzymatic activity	INACTIVE	1.0
In vitro Infectivity	SARS-CoV pseudotyped particle entry (CoV-PPE)	ACTIVE	0.69
	SARS-CoV pseudotyped particle entry counter screen (CoV-PPE_cs)	INACTIVE	0.68
	MERS-CoV pseudotyped particle entry (MERS-PPE)	ACTIVE	0.34
	MERS-CoV pseudotyped particle entry counter screen (MERS-PPE_cs)	INACTIVE	0.6
Human Cell Toxicity	Human fibroblast toxicity (hCYTOX)	ACTIVE	0.68
Host Protein	Sigma1 Receptor (sigma1R)	INACTIVE	0.96

Promising drugs are those that:

- Are active in CPE and are inactive in cytotox

AND

- Are inactive in ACE2

AND

- Are active in 3CL

AND/OR

- Are active in at least one of the following: AlphaLISA, CoV-PPE, MERS-PPE. While they are inactive in the counter screen, respectively: TruHit, CoV-PPE\_cs, MERS-PPE\_cs

AND

- Are inactive in hCYTOX

**Figure 85.**

REDIAL-2020 results page predicting compound activity across all eleven assays: CPE, cytotox, AlphaLISA, TruHit, ACE2, 3CL, CoV-PPE, CoV-PPE\_cs, MERS-PPE, MERS-PPE\_cs, and hCYTOX.



Drugmonizome









METADATA SEARCH DRUG SET ENRICHMENT RESOURCES DRUGMONIZOME ML TUTORIAL API ABOUT

Metadata Search Drug Set Enrichment

Search for any term, i.e. drug, side effects, or a disease

Headache / Dry mouth / CXCL10

Resources Drug set libraries Drug sets (110903) Small molecules

	feeling abnormal SIDER Side Effects ← Association Type: side effects	⋮	 Resources
	connective tissue disorder SIDER Side Effects ← Association Type: side effects	⋮	 Drug Set Library
	eye disorder SIDER Side Effects ← Association Type: side effects	⋮	 Association Type
	headache SIDER Side Effects ← Association Type: side effects	⋮	
	pericardial haemorrhage SIDER Side Effects ← Association Type: side effects	⋮	

**Figure 86.**  
Drugmonizome metadata search page with drug set search enabled.

The screenshot shows the Drugmonizome website interface. At the top, there is a blue navigation bar with the logo and menu items: METADATA SEARCH, DRUG SET ENRICHMENT, RESOURCES, DRUGMONIZOME ML, TUTORIAL, API, and ABOUT. Below the navigation bar, there are two tabs: 'Metadata Search' (selected) and 'Drug Set Enrichment'. A search bar contains the term 'Headache'. Below the search bar, there are filters for 'Headache / Dry mouth / CXCL10' and tabs for 'Resources', 'Drug set libraries', 'Drug sets (17)' (selected), and 'Small molecules'. The main content area displays a list of search results:

Icon	Term	Source	Association Type
	tension headache	SIDER Side Effects	side effects
	headache	SIDER Indications	mechanisms of action
	tension headache	PharmGKB OFFSIDES Side Effects	side effects
	cluster headache	SIDER Indications	mechanisms of action
	tension headache	SIDER Indications	mechanisms of action

On the right side of the results list, there are three dropdown menus: 'Resources', 'Drug Set Library', and 'Association Type'.

**Figure 87.** Drugmonizome metadata search page with example term “Headache” queried using the search bar.

Drugmonizome METADATA SEARCH DRUG SET ENRICHMENT RESOURCES DRUGMONIZOME ML TUTORIAL API ABOUT

## tension headache

SIDER Side Effects

<b>Term</b>	<b>Name</b>	tension headache
	<b>Accession</b>	C0033893
<b>Filename</b>	SIDER_side_effects_drugsetlibrary	
<b>Organism</b>	Homo sapiens	
<b>Number of drugs</b>	17	
<b>Association Type</b>	side effects	

tension headache has 17 small molecules in its drug set. 🔍 Search for any term

**tranexamic acid**

🔗 InChI\_key: GYDJEQRTZSCIOI-LJGSYFOKSA-N

🔗 SMILES: NC[C@H]1CC[C@@H](CC1)C(=O)O    🔗 Accession: DB00302

Synonyms:

- 🔗 acide tranéxamique
- 🔗 ácido tranexámico
- 🔗 acidum tranexamicum
- 🔗 tranexamic acid
- 🔗 tranexamsaeure
- 🔗 tranexmic acid
- 🔗 tranhexamic acid
- 🔗 trans amcha
- 🔗 trans-4-(aminomethyl)cyclohexanecarboxylic acid
- 🔗 trans-4-aminomethylcyclohexane-1-carboxylic acid
- 🔗 trans-amcha
- 🔗 trans-tranexamic acid

**Figure 88.**

Drug set page that includes identifying metadata for the drug set and the small molecules included in the drug set. The search bar can be used to query specific drugs or small molecules of interest.

Drugmonizome

Term Search Drug Set Enrichment Resources Drugmonizome ML Tutorial API About

Term Search **Drug Set Enrichment**

Perform Drug Set Enrichment Analysis to Identify Common Properties Shared by Your Drug Set

lopinavir  
 remdesivir  
 mefloquine  
 loratadine  
 almitrine  
 camostat  
 pevonedistat  
 octenidine

50 valid entries  
19 invalid entries  
0 needs review

Perform Drug Set Enrichment Analysis

Top enriched terms from 2 drug set libraries in SIDER

SIDER Side Effects

Result visualizations and downloads

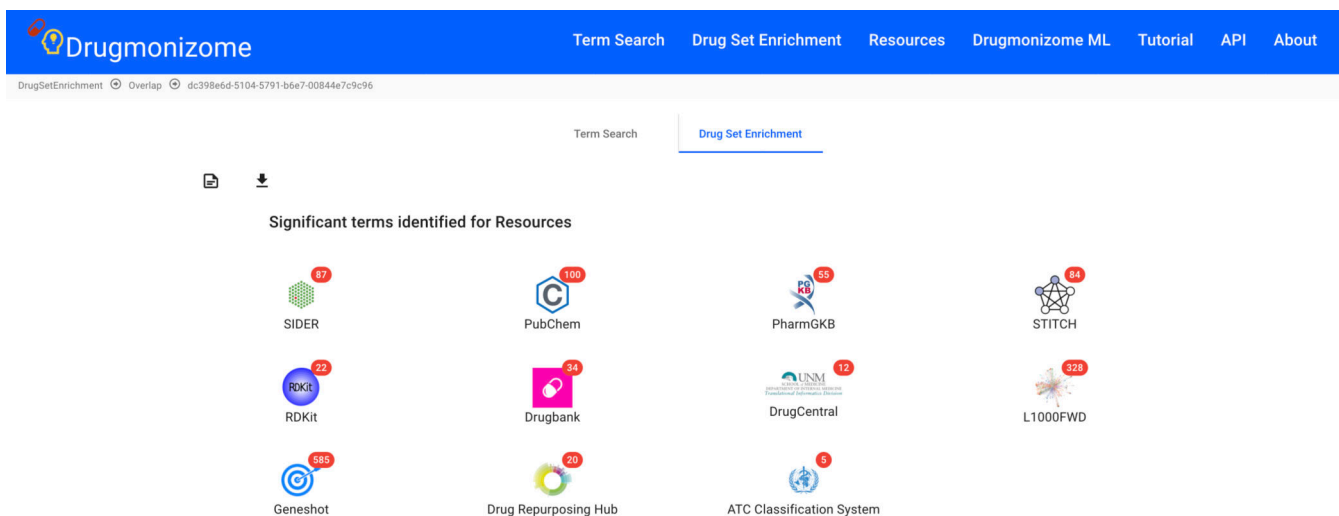
[69 in vitro COVID-19 hits from a drug screen by Ellinger et al.](#)  
[39 in vitro COVID-19 hits from a drug screen by Heiser et al.](#)  
[27 in vitro COVID-19 hits from a drug screen by Jeon et al.](#)

[Terms of Service](#)  
[View Source Code](#)  
[Submit Bugs and Corrections](#)

POWERED BY: Signature Commons

**Figure 89.**

Drug set enrichment page with the “Ellinger et al.” example drug set pasted into the search box.



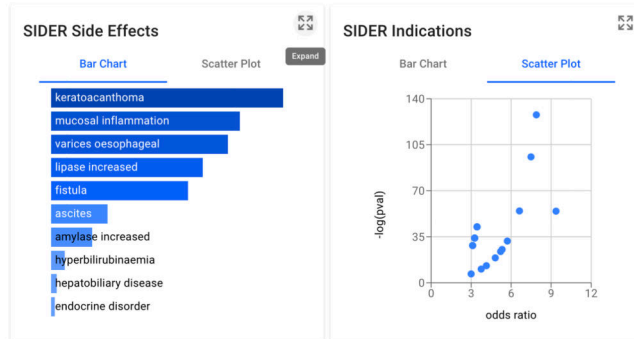
**Figure 90.** Enrichment results page after submitting the “Ellinger et al.” example drug set. Each resource is represented by an icon and the number of enriched drug sets from each resource are displayed above the icon.

Drugmonizome

Term Search Drug Set Enrichment Resources Drugmonizome ML Tutorial API About

Enrichment Overlap dc398e6d-5104-5791-b6e7-00844e7c9c96 Resources 0dd36110-b67b-4546-ac36-974a6432896a

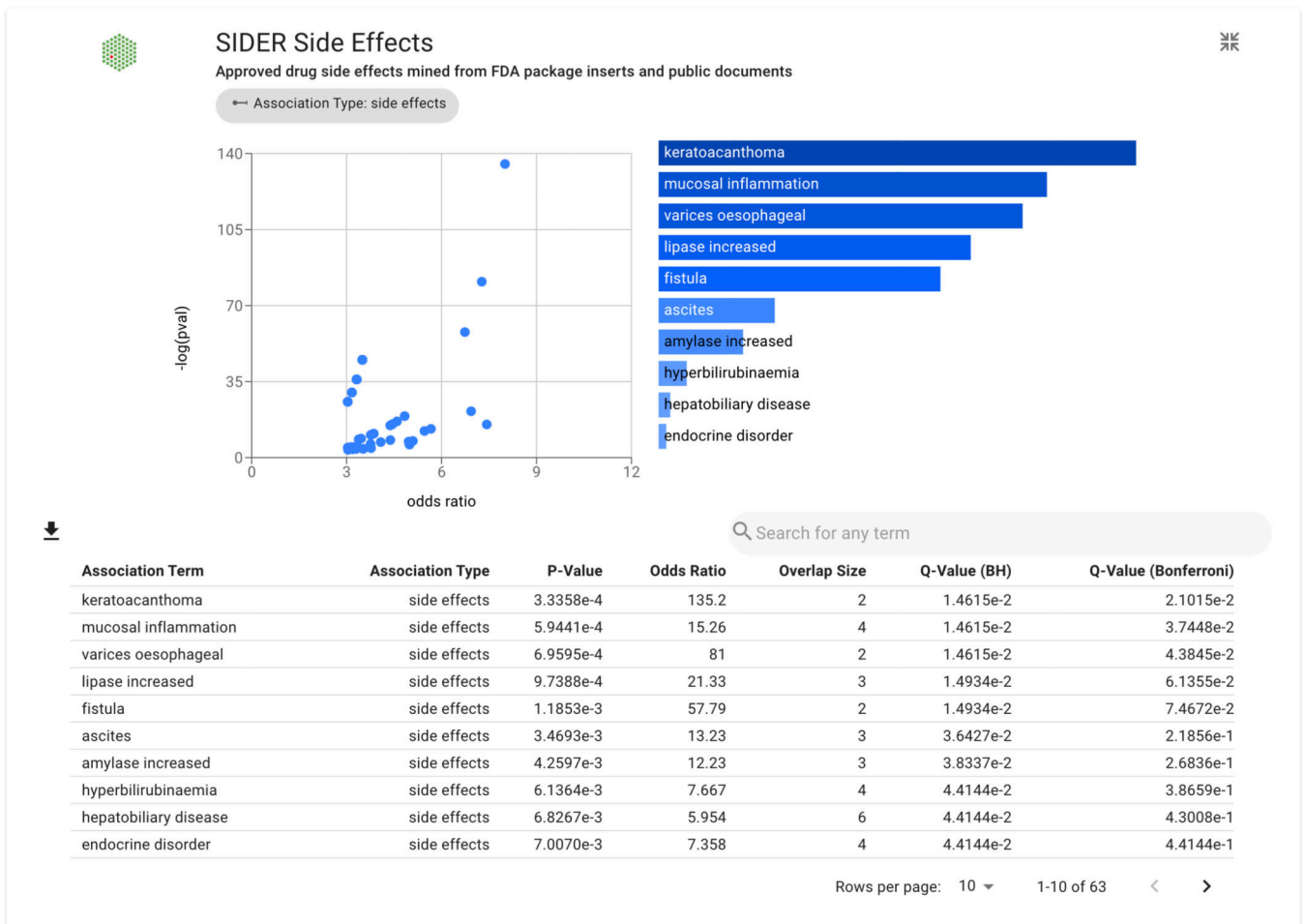
### Top enriched terms from 2 drug set libraries in SIDER



**Figure 91.**

After clicking on the SIDER resource, the top enriched terms from both drug set libraries from SIDER are displayed side by side. Bar charts and scatter plots visualize the top enriched terms. The view for a particular library can be expanded by clicking the “expand” button.

## Top enriched terms from 2 drug set libraries in SIDER



**Figure 92.**

Expanded view for the SIDER Side Effects drug set library. This view includes the bar chart of top enriched terms, scatter plot of top enriched terms, and table of top enriched terms with each of their p-values, odds ratios, overlap sizes, and corrected q-values.

The screenshot displays the Drugmonizome website interface. At the top, a blue navigation bar contains the Drugmonizome logo and several menu items: Term Search, Drug Set Enrichment, Resources, Drugmonizome ML, Tutorial, API, and About. Below the navigation bar, a grid of nine white boxes, each representing a different drug data resource, is presented. Each box includes a logo, the resource name, and a brief description of its function.

Resource Name	Description
DrugCentral	DrugCentral provides information on active ingredients of chemical entities, pharmaceutical products, drug mode of action, indications, and pharmacologic action.
ATC Classification System	A classification system used to organize chemicals by chemical, therapeutic, pharmacological subgroups.
CREEDS	A crowdsourcing resource for the curation and reanalysis of gene expression profiles from GEO that includes drug perturbation signatures.
Drugbank	A bioinformatics and cheminformatics resource that combines detailed drug data with comprehensive drug target information.
L1000FWD	L1000 fireworks display (L1000FWD) is a web application that provides interactive visualization of over 16,000 drug and small-molecule induced gene expression signatures.
PharmGKB	PharmGKB is a pharmacogenomics knowledge resource that includes clinically actionable gene-drug associations and genotype-phenotype relationships.
PubChem	A chemical information resource that collects chemical structures, identifiers, chemical and physical properties, biological activities, patents, health, safety, toxicity
Drug Repurposing Hub	A hand-curated collection of compounds with experimentally confirmed identities and annotations of literature-reported targets.
Geneshot	Submit any search terms to Geneshot to receive prioritized genes that are most relevant to the search terms. Geneshot finds publications that mention both the search

**Figure 93.**  
The resource page listing all drug data resources included in Drugmonizome.



**Drugbank**

**description**  
A bioinformatics and cheminformatics resource that combines detailed drug data with comprehensive drug target information.

**URL**  
<https://www.drugbank.ca>

**PMID**  
29126136

**Resource Name**  
Drugbank

There are 4 drug set libraries in Drugbank.

- Drugbank Small Molecule Targets**  
Drug targets of Drugbank small molecules  
← Association Type: genes
- Drugbank Small Molecule Carriers**  
Genes encoding carriers associated with Drugbank small molecules  
← Association Type: genes
- Drugbank Small Molecule Enzymes**  
Genes encoding enzymes associated with Drugbank small molecules  
← Association Type: genes
- Drugbank Small Molecule Transporters**  
Genes encoding transporters associated with Drugbank small molecules  
← Association Type: genes

Page 1 of 4

**Figure 94.** Expanded view of the DrugBank resource with identifying metadata and drug set libraries curated from DrugBank.

Drugmonizome Term Search Drug Set Enrichment Resources Drugmonizome ML Tutorial API About

## Drugbank Small Molecule Targets

Drugbank

<p><b>Description</b></p> <p><b>DOI</b></p> <p><b>PMID</b></p> <p><b>Year</b></p> <p><b>Version</b></p> <p><b>Filename</b></p> <p><b>Organism</b></p> <p><b>Library name</b></p> <p><b>Download links</b></p> <p><b>Original source</b></p> <p><b>Association Type</b></p> <p><b>Link to resource</b></p> <p><b>Primary Resource</b></p> <p><b>Library created by</b></p> <p><b>Primary Resource Short Version</b></p> <p><b>Link to processing scripts on GitHub</b></p>	<p>Drug targets of Drugbank small molecules</p> <p>10.1093/nar/gkx1037</p> <p>29126136</p> <p>2019</p> <p>2019</p> <p>Drugbank_smallmolecule_target_drugsetlibrary</p> <p>Homo sapiens</p> <p>Drugbank Small Molecule Targets</p> <p><b>InChiKey</b></p> <p><a href="https://maayanlab-public.s3.amazonaws.com/drugmonizome-dmts/Drugbank_smallmolecule_target_drugsetlibrary_inchikey.dmt">https://maayanlab-public.s3.amazonaws.com/drugmonizome-dmts/Drugbank_smallmolecule_target_drugsetlibrary_inchikey.dmt</a></p> <p><b>Drug name</b></p> <p><a href="https://maayanlab-public.s3.amazonaws.com/drugmonizome-dmts/Drugbank_smallmolecule_target_drugsetlibrary_name.dmt">https://maayanlab-public.s3.amazonaws.com/drugmonizome-dmts/Drugbank_smallmolecule_target_drugsetlibrary_name.dmt</a></p> <p>Drugbank</p> <p>genes</p> <p><a href="https://www.drugbank.ca/releases/latest#protein-identifiers">https://www.drugbank.ca/releases/latest#protein-identifiers</a></p> <p>Drugbank</p> <p>MaayanLab</p> <p>Drugbank</p> <p><a href="https://github.com/MaayanLab/Drugmonizome/blob/master/drugsetlibraries/notebooks/Drugbank/Small">https://github.com/MaayanLab/Drugmonizome/blob/master/drugsetlibraries/notebooks/Drugbank/Small</a></p>
---	--

There are 611 drug sets in Drugbank Small Molecule Targets.

**UQCRC2**

Drugbank Small Molecule Targets

← Association Type: genes

⋮

**CYC1**

Drugbank Small Molecule Targets

← Association Type: genes

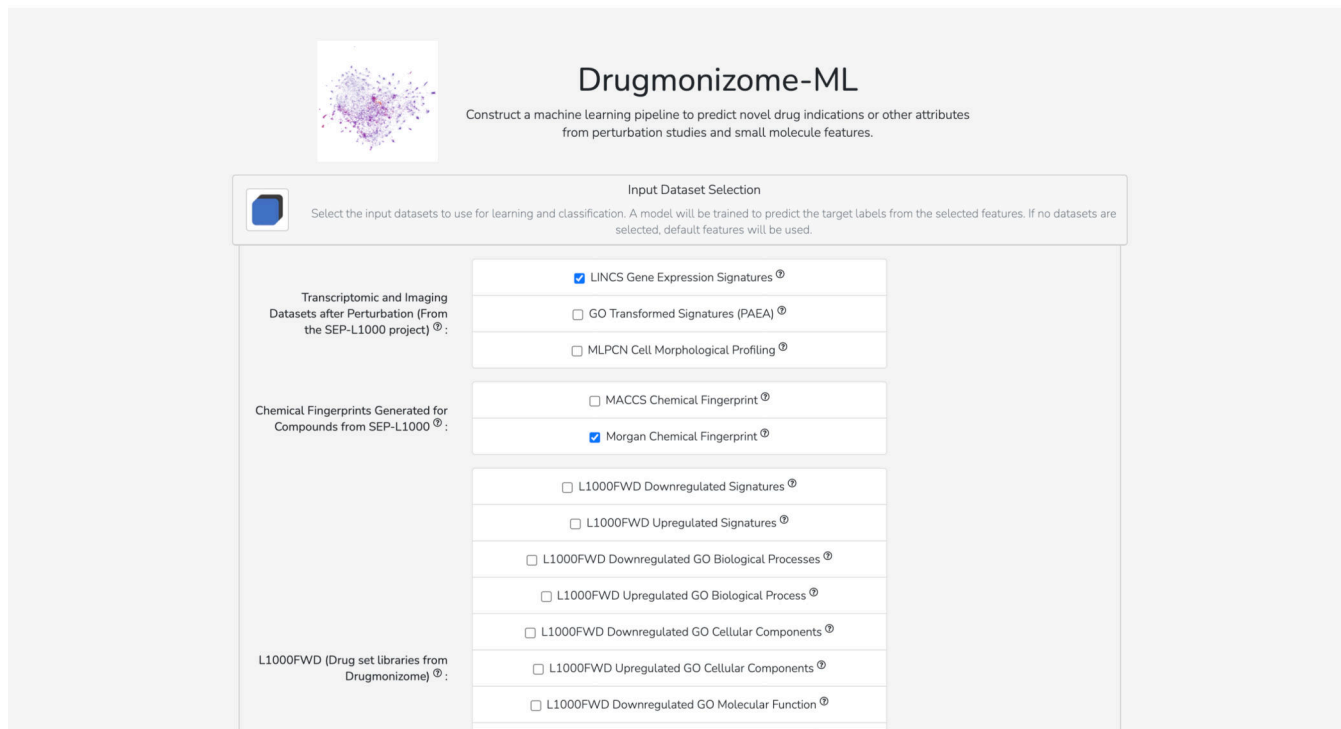
⋮

Search for any term

← Association Type ▾

**Figure 95.**

Expanded view of the DrugBank Small Molecule Targets drug set library with metadata that include download links for the DMT file in drug name and InChiKey formats. All drug sets included in the library are listed below and each drug set can be expanded to view drug set specific metadata and the list of small molecules included in the drug set.



**Drugmonizome-ML**  
Construct a machine learning pipeline to predict novel drug indications or other attributes from perturbation studies and small molecule features.

**Input Dataset Selection**  
Select the input datasets to use for learning and classification. A model will be trained to predict the target labels from the selected features. If no datasets are selected, default features will be used.

**Transcriptomic and Imaging Datasets after Perturbation (From the SEP-L1000 project) <sup>Ⓞ</sup>:**

- LINCX Gene Expression Signatures <sup>Ⓞ</sup>
- GO Transformed Signatures (PAEA) <sup>Ⓞ</sup>
- MLPNCN Cell Morphological Profiling <sup>Ⓞ</sup>

**Chemical Fingerprints Generated for Compounds from SEP-L1000 <sup>Ⓞ</sup>:**

- MACCS Chemical Fingerprint <sup>Ⓞ</sup>
- Morgan Chemical Fingerprint <sup>Ⓞ</sup>

**L1000FWD (Drug set libraries from Drugmonizome) <sup>Ⓞ</sup>:**

- L1000FWD Downregulated Signatures <sup>Ⓞ</sup>
- L1000FWD Upregulated Signatures <sup>Ⓞ</sup>
- L1000FWD Downregulated GO Biological Processes <sup>Ⓞ</sup>
- L1000FWD Upregulated GO Biological Process <sup>Ⓞ</sup>
- L1000FWD Downregulated GO Cellular Components <sup>Ⓞ</sup>
- L1000FWD Upregulated GO Cellular Components <sup>Ⓞ</sup>
- L1000FWD Downregulated GO Molecular Function <sup>Ⓞ</sup>

**Figure 96.** Input dataset selection section of the Drugmonizome-ML Appyter. Each input dataset is annotated with tooltips.

Indications, Modes of Action, and Side Effects (Drug set libraries from Drugmonizome) <sup>Ⓢ</sup>:

- ATC Codes Drugsetlibrary <sup>Ⓢ</sup>
- DrugRepurposingHub Mechanisms of Action <sup>Ⓢ</sup>
- PharmGKB OFFSIDES Side Effects <sup>Ⓢ</sup>
- SIDER Indications <sup>Ⓢ</sup>
- SIDER Side Effects <sup>Ⓢ</sup>

Structural Features (Drug set libraries from Drugmonizome) <sup>Ⓢ</sup>:

- RDKit MACCS Chemical Fingerprints <sup>Ⓢ</sup>
- PubChem Chemical Fingerprints <sup>Ⓢ</sup>

Keep drugs with missing data when joining datasets <sup>Ⓢ</sup>

Apply tf-idf normalization to binary inputs <sup>Ⓢ</sup>

**Figure 97.** Toggleable options for deciding whether to retain or drop drugs with missing data and TF-IDF normalization.

Target Label Selection

Upload a list of compounds or select an attribute from Drugmonizome to be assigned a positive class label for binary classification.

Target Selection <sup>Ⓢ</sup>:  List  Attribute

Attribute <sup>Ⓢ</sup>: neuropathy peripheral (from SIDER Side Effec

Include stereoisomers <sup>Ⓢ</sup>:  Yes  No

**Figure 98.**

Target label selection with “Attribute” selected. The autocomplete field can be populated with search terms that match to drug-set labels in Drugmonizome which will be used as the positive class to predict.

Target Label Selection

Upload a list of compounds or select an attribute from Drugmonizome to be assigned a positive class label for binary classification.

List  
Attribute

Drug Identifier Format: InChI Key

Choose file Browse

Load examples: [COVID19ScreenHits.txt](#)  
[COVID19ScreenHitsInChIKeys.txt](#)

Download examples: [COVID19ScreenHits.txt](#)  
[COVID19ScreenHitsInChIKeys.txt](#)

Target Selection:

Upload List of Compounds:

Include stereoisomers: Yes No

**Figure 99.**

Target label selection with “List” selected. Newline separated .txt files can be uploaded with small molecules that are part of a positive class to predict. The drug identifier format drop-down menu allows specification of how small molecules are catalogued within the uploaded file (names or InChI key).

**Machine Learning Pipeline**

Select from available machine learning algorithms, their unique settings, and methods to use to evaluate the classifier.

Data Visualization Method <sup>Ⓢ</sup>: UMAP

Dimensionality Reduction Algorithm <sup>Ⓢ</sup>: None

Machine Learning Feature Selection <sup>Ⓢ</sup>: None

Machine Learning Algorithm <sup>Ⓢ</sup>:

- GradientBoostingClassifier
- RandomForestClassifier
- AdaBoostClassifier
- ExtraTreesClassifier**
- DecisionTreeClassifier
- KNeighborsClassifier
- RadiusNeighborsClassifier
- MLPClassifier
- SVC

n\_estimators <sup>Ⓢ</sup>: 1250

criterion <sup>Ⓢ</sup>: entropy

min\_samples\_split <sup>Ⓢ</sup>: 2

min\_samples\_leaf <sup>Ⓢ</sup>: 1

max\_features <sup>Ⓢ</sup>: "log2"

min\_impurity\_decrease <sup>Ⓢ</sup>: 0.0

class\_weight <sup>Ⓢ</sup>: "balanced"

ccp\_alpha <sup>Ⓢ</sup>: 0.0

Calibrate algorithm predictions <sup>Ⓢ</sup>:  Yes  No

Cross-Validation Algorithm <sup>Ⓢ</sup>: RepeatedStratifiedGroupKFold

Number of Cross-Validated Folds <sup>Ⓢ</sup>: 10

Number of Cross-Validated Repetitions <sup>Ⓢ</sup>: 3

Primary Evaluation Metric <sup>Ⓢ</sup>: roc\_auc

Evaluation Metrics <sup>Ⓢ</sup>: accuracy, adjusted\_mutual\_info\_score, adjusted\_rand\_score, average\_precision

**Submit**

**Figure 100.** Machine learning pipeline section with methods for data visualization, machine learning classifier selection, hyperparameter settings, and metrics to evaluate the classifier.



1 [What is an Appyter?](#) | [Creating Appyters](#) | [Publishing Appyters](#) | [About](#)

2 [Download Notebook](#) [Toggle Code](#) [Run Locally](#)

Success

3

Table Of Contents

- Select Input Datasets and Target Classes
- Dimensionality Reduction and Visualization
- Machine Learning
- Examine predictions
- Examine feature importances

In [1]

```

%matplotlib inline
# Imports
## Data processing
import pandas as pd
import numpy as np
import scipy as sp
## Machine Learning
import sklearn as sk
from sklearn import (
    calibration,
    decomposition,
    ensemble,
    feature_selection,
    linear_model,
    manifold,
    metrics,
    model_selection,
    multioutput,
    pipeline,
    preprocessing,
    svm,
    tree,
    feature_extraction,
    neural_network,
)
from split import StratifiedGroupKFold, RepeatedStratifiedGroupKFold
import umap
## Plotting
from matplotlib import pyplot as plt

```

**Figure 101.**

(1) To learn more about Appyters, click any of the header tabs to navigate to information pages. (2) Clickable options to download the Jupyter Notebook, toggle code when viewing the notebook, as well as the option to run the notebook locally. (3) Table of contents with clickable elements that link to a specific section within the notebook.



## Select Input Datasets and Target Classes

Selected drug set libraries and phenotypic datasets are downloaded and joined on the compound InChI Key to produce a large input feature matrix. A machine learning model will be trained to predict the specified target labels from these features. This is a binary classification task that can be used to predict compounds that are likely to be associated with the target class.

To construct the input matrix, we download drug set libraries and phenotypic datasets and join them on the InChI Key. Only drugs that are present in all datasets are retained.

### Table 1: Input data

InChI Key	PSME1	ATF1	RHEB	FOXO3	RHOA	IL1B	ASAH1	RALA	ARHGEF12	SOX2	...	Morgan_20
AAALVYBICLMAMA-UHFFFAOYSA-N	0.0476	0.0041	0.0235	0.0047	0.0141	-0.0238	-0.0004	0.0109	-0.0039	-0.0628	...	(
AACFPJSJOWQNBNUHFFFAOYSA-N	0.0031	0.0136	0.0079	0.0030	0.0126	0.0168	0.0134	0.0152	-0.0275	0.0050	...	(
AADCDMQTJNYOSSLBPRGKRZSA-N	-0.0170	0.0269	0.0057	0.0316	-0.0064	0.0191	0.0398	-0.0530	0.0191	0.0041	...	(
AADVJQLQUVDEBPGQIGUUNPSA-N	0.0121	-0.0052	0.0192	-0.0173	0.0878	-0.0101	0.0114	-0.0249	-0.0674	0.0200	...	(
AADVJQLQUVDEBPGUXCAODWSA-N	-0.0050	-0.0087	-0.0084	0.0114	0.0003	0.0279	-0.0178	0.0022	-0.0138	-0.0048	...	(

5 rows x 3026 columns

**Table 1:** *Input data.* The input data contain 19898 compounds and 3026 features per compound, taken from the following datasets: LINCS Gene Expression Signatures, Morgan Chemical Fingerprint.

We produce a target array containing 1 if the compound is associated with the attribute *neuropathy peripheral* in the Drugmonizome resource *SIDER Side Effects* and 0 otherwise.

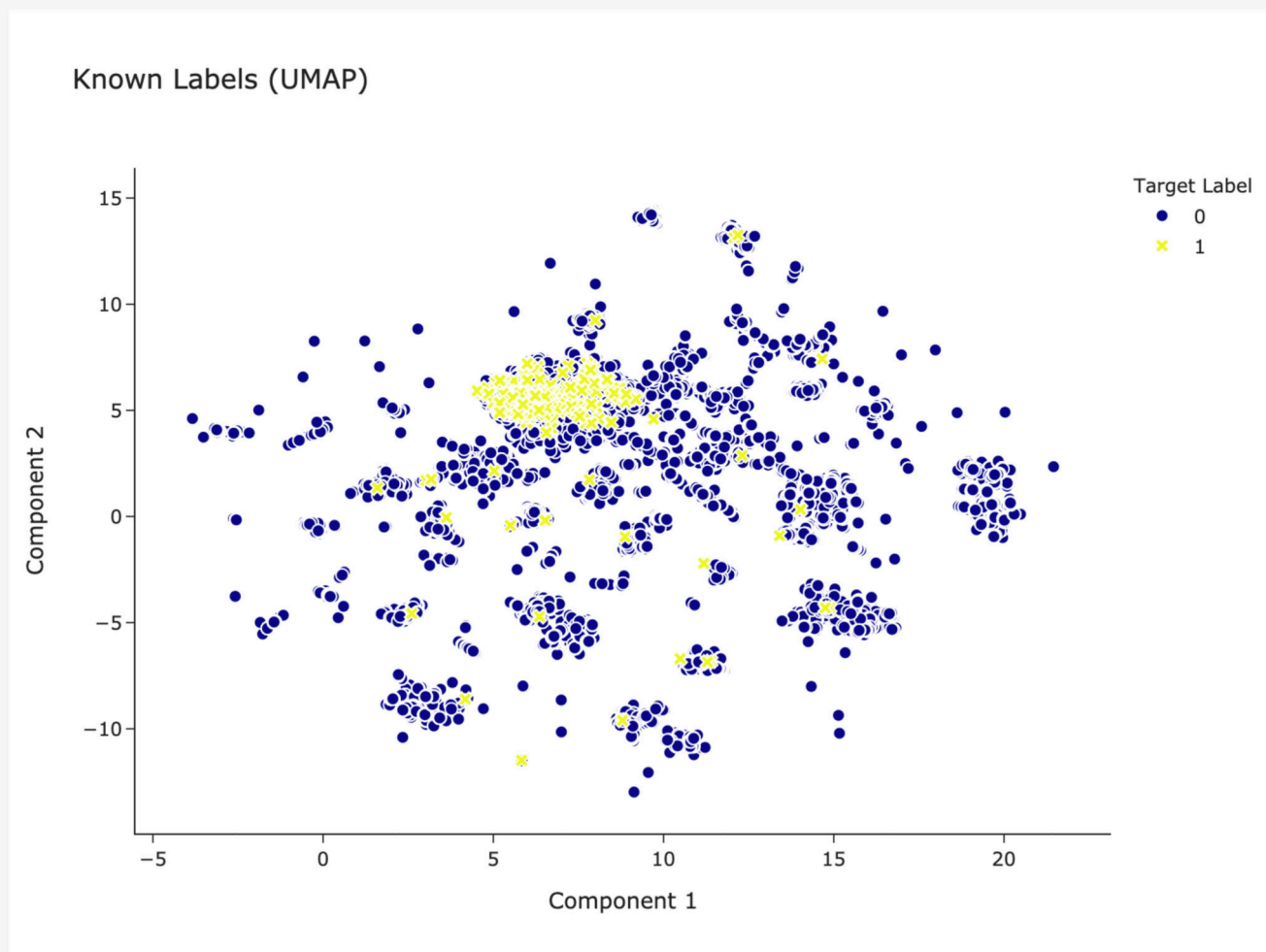
Number of hits matched in input: 220 (1.106 %)

Number of unmatched hits: 115

- File of unmatched InChI keys: [unmatched\\_inchikeys.txt](#)

#### Figure 102.

Input dataset visualized in Dataframe format. The number of matched compounds in the target vector is displayed, along with a downloadable .txt file of unmatched compounds.

**Figure 1:** Input feature space with UMAP dimensionality reduction

**Figure 1:** *Input feature space with UMAP dimensionality reduction.* Each point represents one of 19898 compounds, with 3026 features per compound, taken from the following datasets: LINCS Gene Expression Signatures, Morgan Chemical Fingerprint. Compounds with known positive labels are marked by X's.

**Figure 103.**  
Dimensionality Reduction and Visualization Section with input feature space visualized using UMAP.

## Figure 2: Receiver operating characteristic (ROC) curves across cross-validation splits ([roc.svg](#))

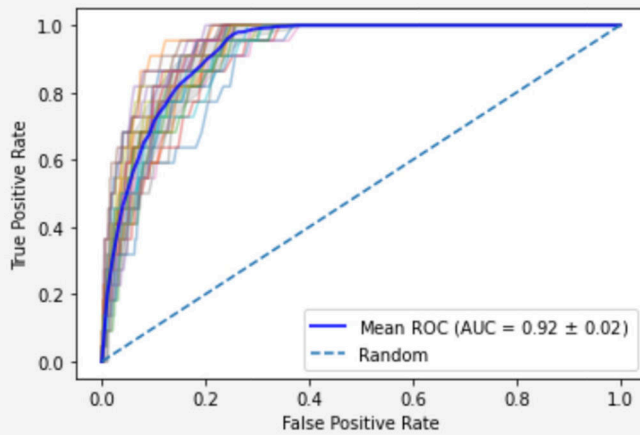


Figure 2: Receiver operating characteristic (ROC) curves across cross-validation splits ([roc.svg](#)). Individual curves are shown for each 10-fold cross-validation split, repeated with 3 different randomizations. Mean ROC shows the average and standard deviation across cross-validation splits.

Confidence interval (95%) (0.8888443897861852, 0.9585480234933569)

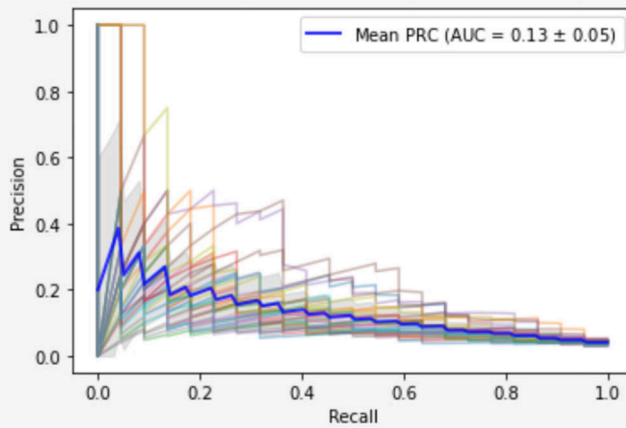
We are 100.000 % confident the model's results are not just chance.

This is statistically significant. These results can be trusted.

### Figure 104.

Receiver Operating Characteristic (ROC) curves of classifier performance after cross-validation splits.

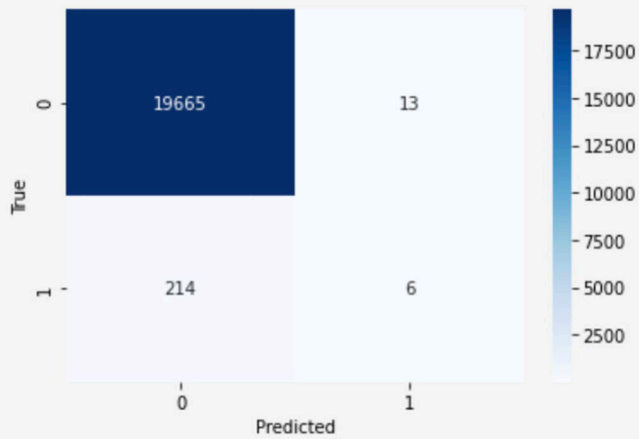
### Figure 3: Precision-recall curves (PRC) across cross-validation splits ([prc.svg](#))



**Figure 3:** Precision-recall curves (PRC) across cross-validation splits ([prc.svg](#)). Individual curves are shown for each 10-fold cross-validation split, repeated with 3 different randomizations. Mean PRC shows the average and standard deviation across cross-validation splits.

**Figure 105.**  
Precision-recall (PR) curves of classifier performance after cross-validation splits.

## Figure 4: Confusion matrix for cross-validation predictions ([confusion\\_matrix.svg](#))



**Figure 4:** Confusion matrix for cross-validation predictions ([confusion\\_matrix.svg](#)). Note that the predicted probabilities can be greatly affected by imbalanced labels and by the model choice. Thus, performance measures such as ROC and PRC, which evaluate performance across a range of prediction thresholds, are more useful than the confusion-matrix, which uses an fixed cutoff of 0.5

### Figure 106.

Confusion matrix for cross-validation predictions from the trained classifier.

## Examine predictions

By examining the validation-set predictions, we can rank the positive compounds and identify additional compounds that were not known to be in the positive class, but nevertheless had high predictions. These may share similar properties with the known compounds.

First, we can compare the distribution of predictions for positive and negative labels.

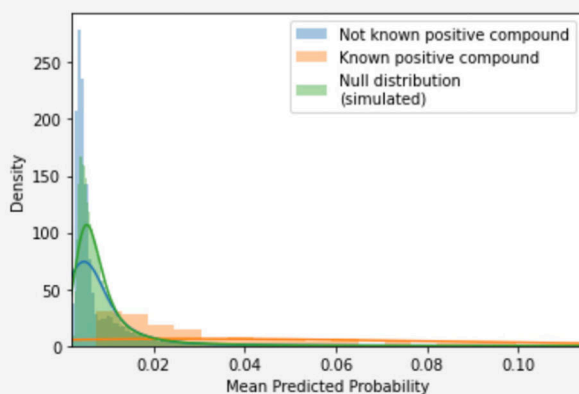
100% | ██████████ | 100000/100000 [00:03<00:00, 30980.45it/s]

100% | ██████████ | 100000/100000 [00:06<00:00, 16225.77it/s]

100% | ██████████ | 19898/19898 [00:03<00:00, 6483.10it/s]

100% | ██████████ | 19898/19898 [00:02<00:00, 6868.30it/s]

## Figure 5: Distribution of mean cross-validation predictions ([mean-prediction-distribution.svg](#))

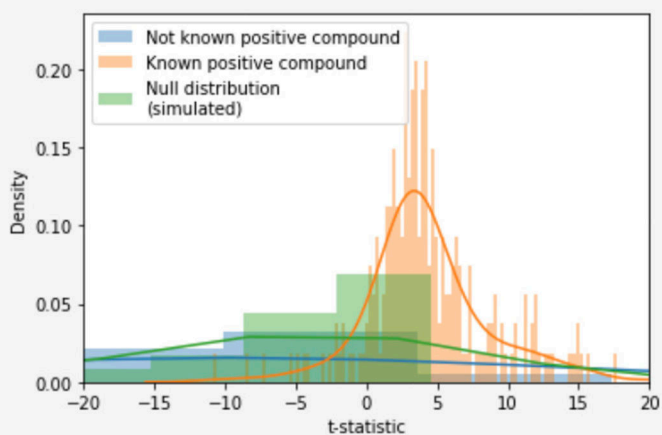


**Figure 5:** Distribution of mean cross-validation predictions ([mean-prediction-distribution.svg](#)). Distribution of mean cross-validation predictions for all 19898 compounds, including both those with known positive labels and other small molecules. The null distribution was simulated by drawing independent samples of predictions with replacement from the distribution of all predictions.

### Figure 107.

Mean probability distribution for classifier predictions including compounds with known positive labels, unknown class labels, and a simulated null distribution.

## Figure 6: Distribution of t-statistics ([t-statistic-distribution.svg](#))



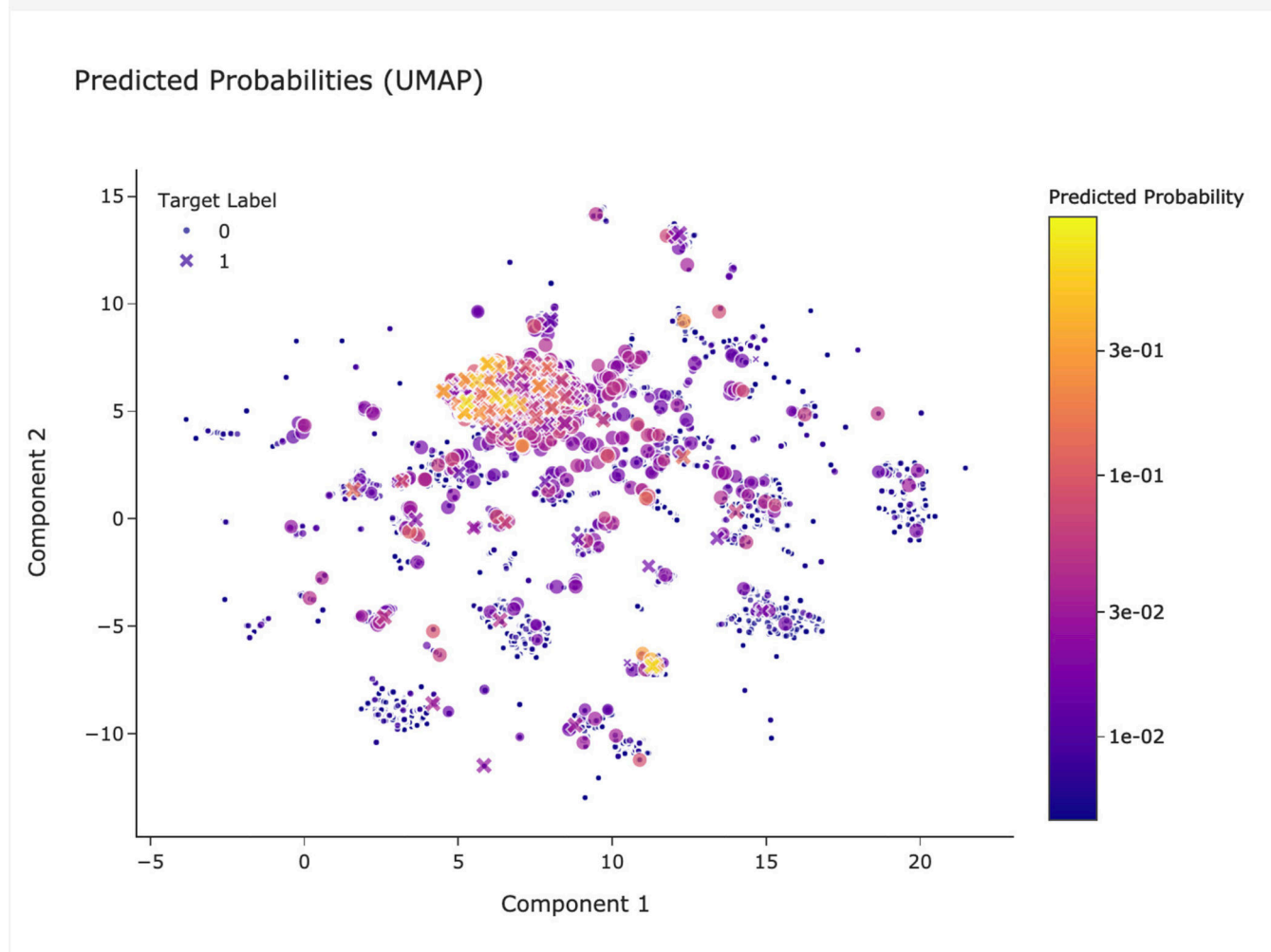
**Figure 6:** *Distribution of t-statistics ([t-statistic-distribution.svg](#)).* Distributions of t-statistics for all 19898 compounds, including both those with known positive labels and other small molecules. The null distribution was simulated by drawing independent samples of predictions with replacement from the distribution of all predictions.

### Figure 108.

T-statistic distribution for classifier predictions including compounds with known positive labels, unknown class labels, and a simulated null distribution.



## Figure 7: UMAP dimensionality reduction of the input feature space overlaid with predictions



**Figure 7:** UMAP dimensionality reduction of the input feature space overlaid with predictions. Each point represents one of 19898 compounds, with 3026 features per compound, taken from the following datasets: LINCS Gene Expression Signatures, Morgan Chemical Fingerprint. Compounds with known positive labels are marked by X's. The color and size of each point correspond to the mean predicted probability and its significance (estimated from the simulated t-statistic null distribution), respectively.

### Figure 109.

UMAP dimensionality reduction of the input feature space with predicted compounds overlaid. The color of each point corresponds to the mean predicted probability, whereas the size of the point corresponds to the significance of the probability.



**Table 2:** Top-predicted compounds ([drug\\_cv\\_predictions.csv](#))

Show  entries

Search:

InChI Key	Name (L1000FWD)	Name (Drugmonizome)	Cross-validation fold	Known	Prediction Probability	Prediction Probability Std. Dev.	t statistic	p value (simulated mean distribution)	p value (one sample t test)	p value (simulated t distribution)
YZDIQTHVDDOVHR-UHFFFAOYSA-N	PLX-4720	n-{3-[(5-chloro-1h-pyrrolo[2...b]pyridin-3-yl)carbon...difluorop...1-sulfonami...	[2, 7, 2]	false	0.976069	0.022983	72.72317	0.00001	0.000095	0.00001
NRUKOCRGYNPUPR-OQMCAFNJSA-N	teniposide	teniposide	[5, 8, 3]	false	0.939583	0.030325	53.0324...	0.00001	0.000178	0.00001
JURKNVYZMNSLP-UHFFFAOYSA-N	cyclobenzaprine	cycloben...	[3, 2, 8]	true	0.859629	0.131679	11.1612...	0.00001	0.003966	0.00014
STQGQHZAVUOBTE-INJOJONLSA-N	daunorubicin	nan	[2, 9, 9]	false	0.836915	0.181208	7.893449	0.00001	0.007837	0.00023
KRMDCWKBEZIMAB-UHFFFAOYSA-N	amitriptyline	amitriptyl...	[1, 1, 1]	true	0.832708	0.091544	15.5452...	0.00001	0.002056	0.00008
VSIKWCGYPAHWDS-FQEVSTJZSA-N	camptothecin	camptoth...	[9, 3, 5]	false	0.775636	0.153799	8.610037	0.00001	0.006611	0.00023
MJIHNNLFOKEZEW-UHFFFAOYSA-N	lansoprazole	lansopraz...	[4, 5, 8]	true	0.760667	0.107511	12.0758...	0.00001	0.003394	0.00013
FPIPGXGPPQFEQ-OVSJKPMPMSA-N	retinol	vitamin a	[1, 2, 3]	false	0.751189	0.112585	11.3858...	0.00001	0.003813	0.00014
HHJUWIANJFBDHT-KOTLKJBCSA-N	vindesine	vindesine	[6, 5, 4]	false	0.744153	0.130926	9.697752	0.00001	0.005233	0.0002
LTMKESNXUBQKBP-UHFFFAOYSA-N	lapatinib	lapatinib	[8, 3, 4]	false	0.743542	0.098505	12.8788...	0.00001	0.002988	0.0001

Showing 1 to 10 of 19,898 entries

Previous  2 3 4 5 ... 1990 Next

**Table 2:** Top-predicted compounds ([drug\\_cv\\_predictions.csv](#)). All 19898 compounds ranked by cross-validation prediction probability. Search 'true' or 'false' to filter compounds with known positive labels or not, respectively. The table can also be sorted by other columns by selecting the column name in the header.

**Figure 110.**  
Table of the top predicted compounds ranked by prediction probability.

**Table 3:** Input features ranked by relative importance ([feature\\_importance.csv](#))Show  entriesSearch: 

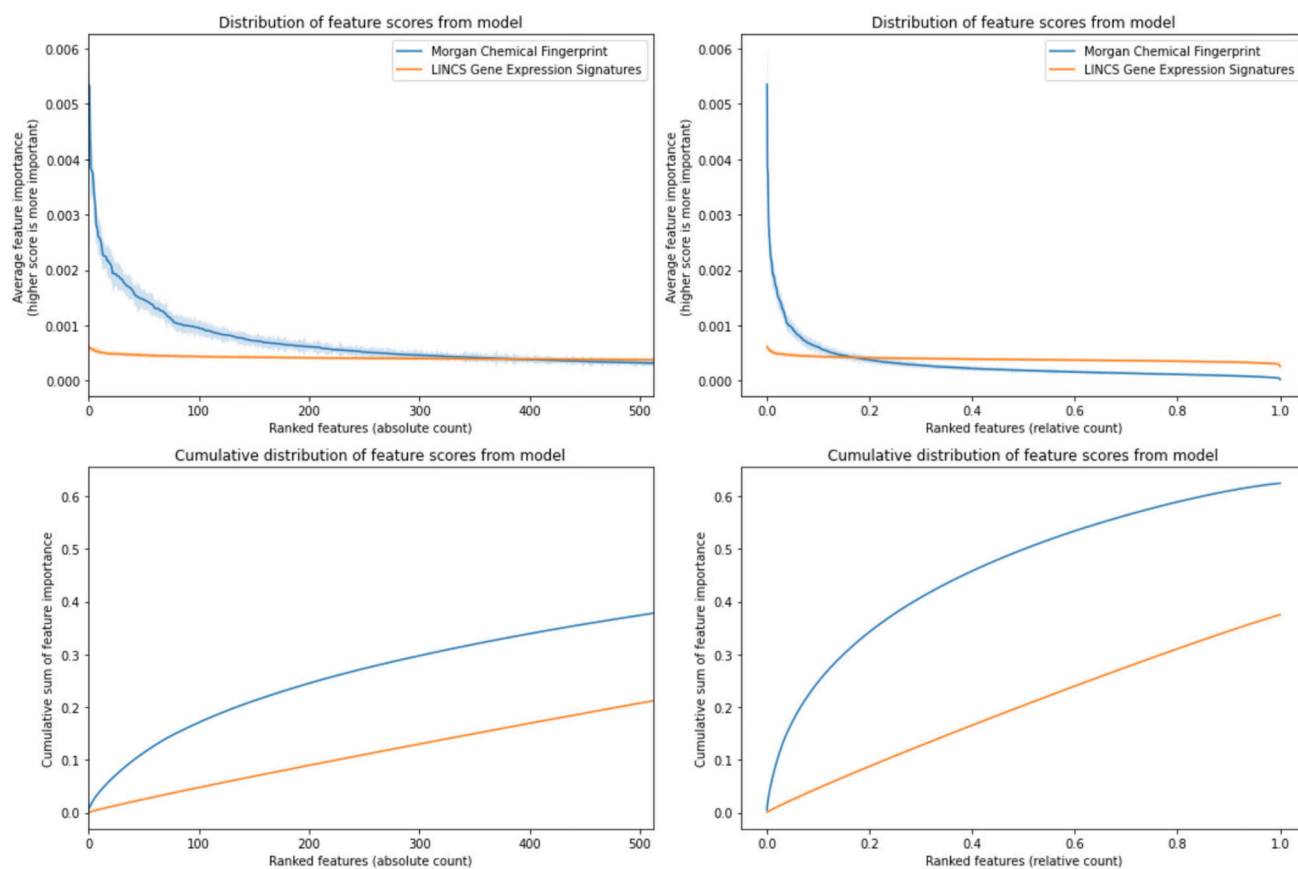
Feature	Dataset	Importance Mean	Importance Std. Dev.
Morgan_1516	Morgan Chemical Fingerprint	0.005353	0.000416
Morgan_222	Morgan Chemical Fingerprint	0.005322	0.000635
Morgan_1905	Morgan Chemical Fingerprint	0.003841	0.000531
Morgan_1116	Morgan Chemical Fingerprint	0.003799	0.000516
Morgan_1019	Morgan Chemical Fingerprint	0.003731	0.000347
Morgan_739	Morgan Chemical Fingerprint	0.003392	0.000368
Morgan_610	Morgan Chemical Fingerprint	0.003215	0.000388
Morgan_834	Morgan Chemical Fingerprint	0.002809	0.000349
Morgan_1716	Morgan Chemical Fingerprint	0.002749	0.00028
Morgan_935	Morgan Chemical Fingerprint	0.002593	0.000327

Showing 1 to 10 of 3,026 entries

Previous  2 3 4 5 ... 303 Next

**Table 3:** *Input features ranked by relative importance* ([feature\\_importance.csv](#)). All 3026 input features are ranked by their relative importance. Tree-based models can be used to calculate impurity-based feature importances (Importance Mean and Std. Dev.). Search a dataset name to filter features from a given dataset. The table can also be sorted by other columns by selecting the column name in the header.

**Figure 111.**  
Feature importance table.

**Figure 8:** Distribution of feature scores from model ([feature\\_importance.svg](#))

**Figure 8:** Distribution of feature scores from model ([feature\\_importance.svg](#)). The distribution of impurity-based feature importances for each dataset. Features with higher scores have greater relative contribution to the overall tree-based model.

**Figure 112.**

Feature importance graphs with distribution scores for each feature and a cumulative distribution score across all features.

**ATTRIBUTE AND PREDICTION CLASS DATASET SELECTION**  
Select the datasets to use for learning and classification.

Attribute Selection (place cursor inside the box to add more datasets) ⓘ:

- CCLE Cell Line Gene Expression Profiles ⓘ Hover over the “?” icon for more information
- ENCODE Transcription Factor Targets ⓘ
- Allen Brain Atlas Adult Human Brain Tissue Gene Expression Profiles ⓘ
- CHEA Transcription Factor Targets ⓘ
- BioGPS Cell Line Gene Expression Profiles ⓘ
- GTEx Tissue Gene Expression Profiles ⓘ

Custom Attribute Dataset (Optional) ⓘ:

Target Selection : Harmonizome  
Custom Harmonizome Class ⓘ:

Click to select an option

**Figure 113.** “Attribute and Prediction Class Dataset Selection” section of the input form. Two datasets are selected to be used as features in the classifier algorithm. Hovering over tool tips displays information about each dataset. There is also an option to upload custom attribute datasets. The Target Selection subsection allows for selection of a class for the classifier to predict.

**SETTINGS**

From here you can select the various available Machine Learning algorithms, their unique settings, and the methods to use to evaluate the classifier.

Dimensionality Reduction Algorithm <sup>Ⓢ</sup> :

✓ PCA
TruncatedSVD
IncrementalPCA
ICA
SparsePCA

Manifold Projection Algorithm <sup>Ⓢ</sup> :

None

Machine Learning Feature Selection:

None

Cross Validation Algorithm:

StratifiedKFold

Machine Learning Algorithm <sup>Ⓢ</sup> :

RandomForestClassifier

Calibrate algorithm predictions <sup>Ⓢ</sup>

Yes

No

Hyper Parameter Search Type <sup>Ⓢ</sup> :

None

Cross-Validated Folds <sup>Ⓢ</sup> :

3

Primary Evaluation Metric <sup>Ⓢ</sup> :

roc\_auc

Evaluation Metrics <sup>Ⓢ</sup> :

neg\_mean\_absolute\_error
neg\_mean\_absolute\_percentage\_error
neg\_mean\_squared\_error
neg\_mean\_squared\_log\_error
neg\_root\_mean\_squared\_error

Drop-down menu example

Select multiple options by holding down "Command" (Mac) or "Control" (PC) while making selections

→

Submit

**Figure 114.**

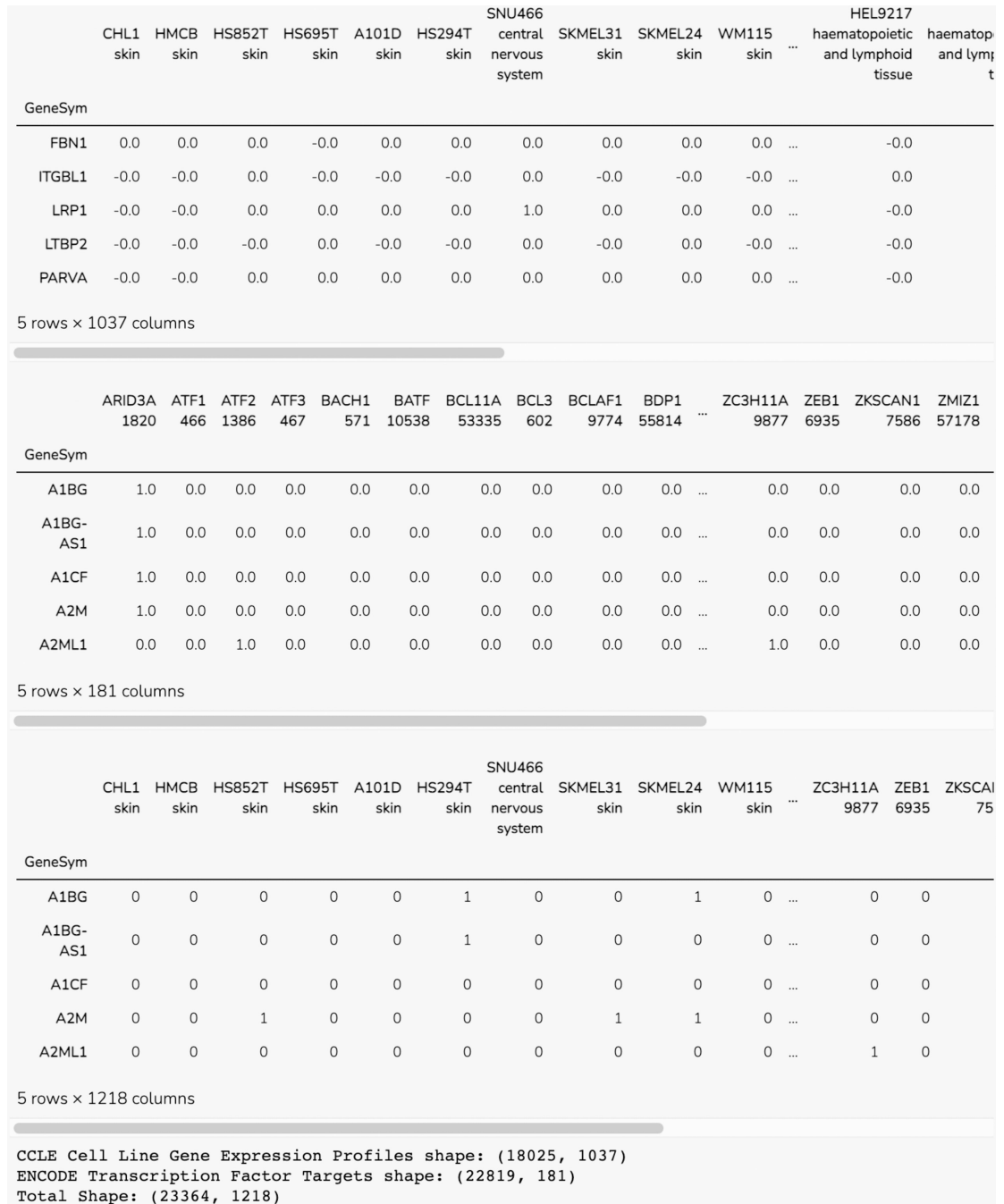
Settings section including a variety of scikit-learn options for building the classifier as well as options for visualizing and evaluating classifier performance and predictions.

The screenshot shows a web browser window with the Appyter interface. The address bar displays the URL: `appymers.maayanlab.cloud/harmonizome_ml/0a916e8c053ae75537afd31d50c1ecc33d9afe72/`. The Appyter logo and tagline "A catalog of appyter notebooks" are visible. A navigation bar contains three buttons: "Download Notebook", "Toggle Code", and "Run Locally". A status bar below the navigation bar shows "Success". On the left, a "Table Of Contents" sidebar lists sections: "Imputing Knowledge about Gene and Protein Function with Machine Learning", "Inputs", "Dimensionality Reduction", and "Machine Learning". The main content area shows a code cell with the following Python code:

```
In [1] # Imports
# Data processing
import pandas as pd
import numpy as np
import scipy as sp
# Machine Learning
import sklearn as sk
from sklearn import (
    calibration, decomposition, ensemble, feature_selection,
    linear_model, manifold, metrics, model_selection, multioutput,
    pipeline, preprocessing, svm, tree, neural_network,
)
# Plotting
import plotly.express as px
from matplotlib import pyplot as plt
# Harmonizome API
from harmonizome import Harmonizome
# Utility
import re
import json
from functools import reduce
from IPython.display import display, Markdown
```

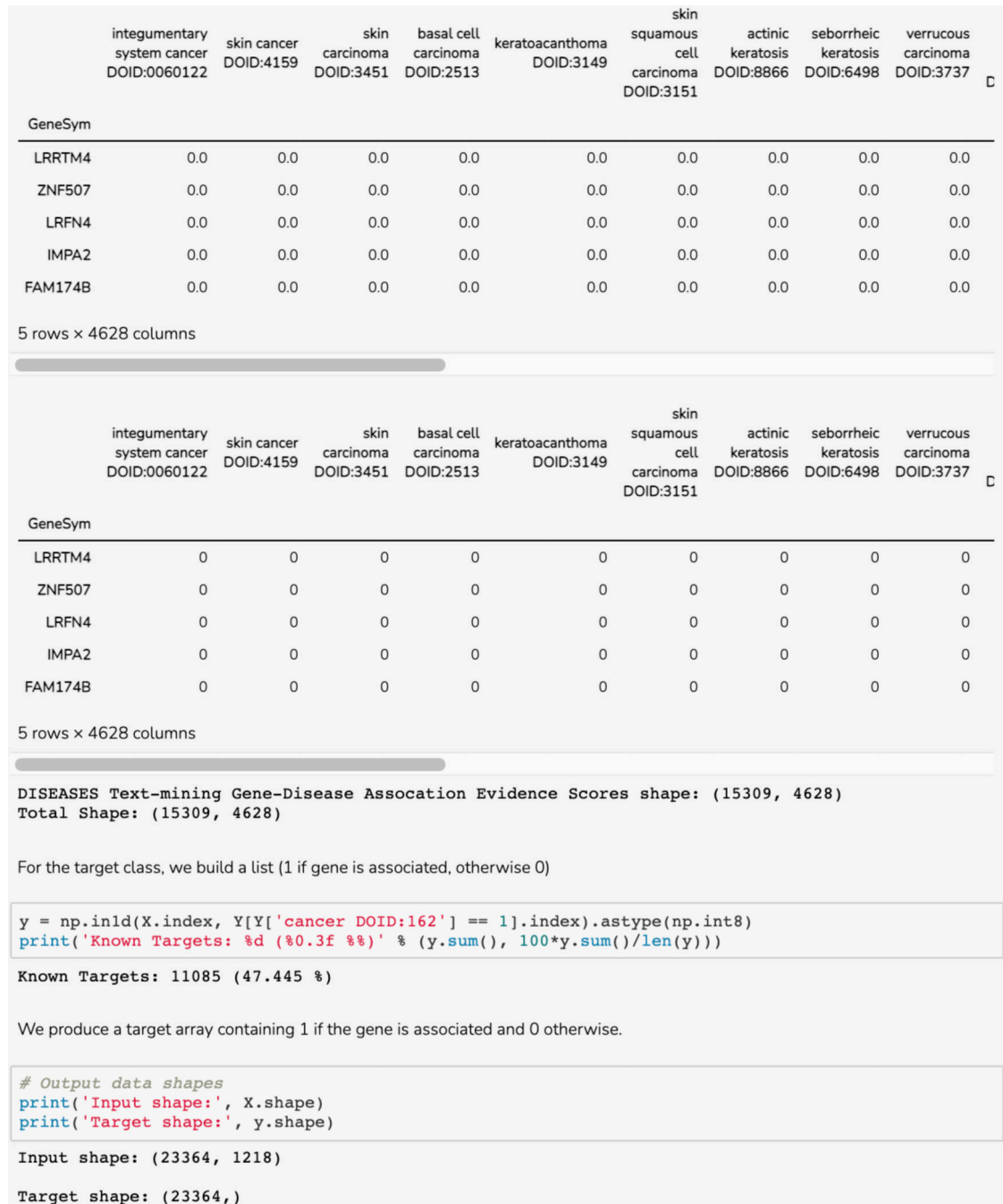
**Figure 115.**

Options to download the appyter notebook, toggle the code, and run the notebook locally. A table of contents on the left allows for navigating the various sections of the notebook.



**Figure 116.**

The input feature datasets visualized as Dataframes. The first and second Dataframes describe the “CCLE Cell Lines Gene Expression Profiles” and “ENCODE Transcription Factors Targets” datasets, respectively. The final Dataframe represents the concatenated feature matrix composed of the previous two datasets.

**Figure 117.**

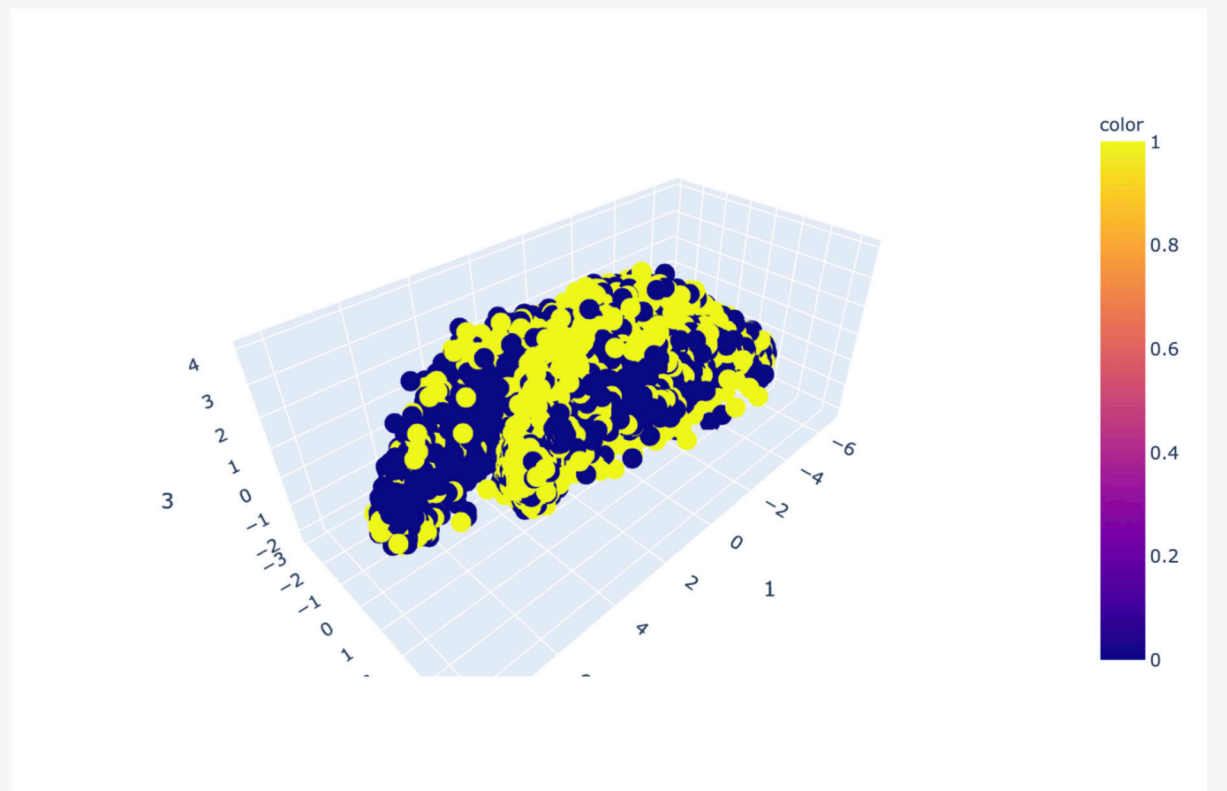
Target array created from the “DISEASES Text-mining Gene-Disease Association Evidence Scores” dataset which contains the class label “cancer DOID:162”. Genes in the target array associated with the class label are marked with a 1, whereas genes that are not known to be associated with the class label are marked with a 0.



## Dimensionality Reduction

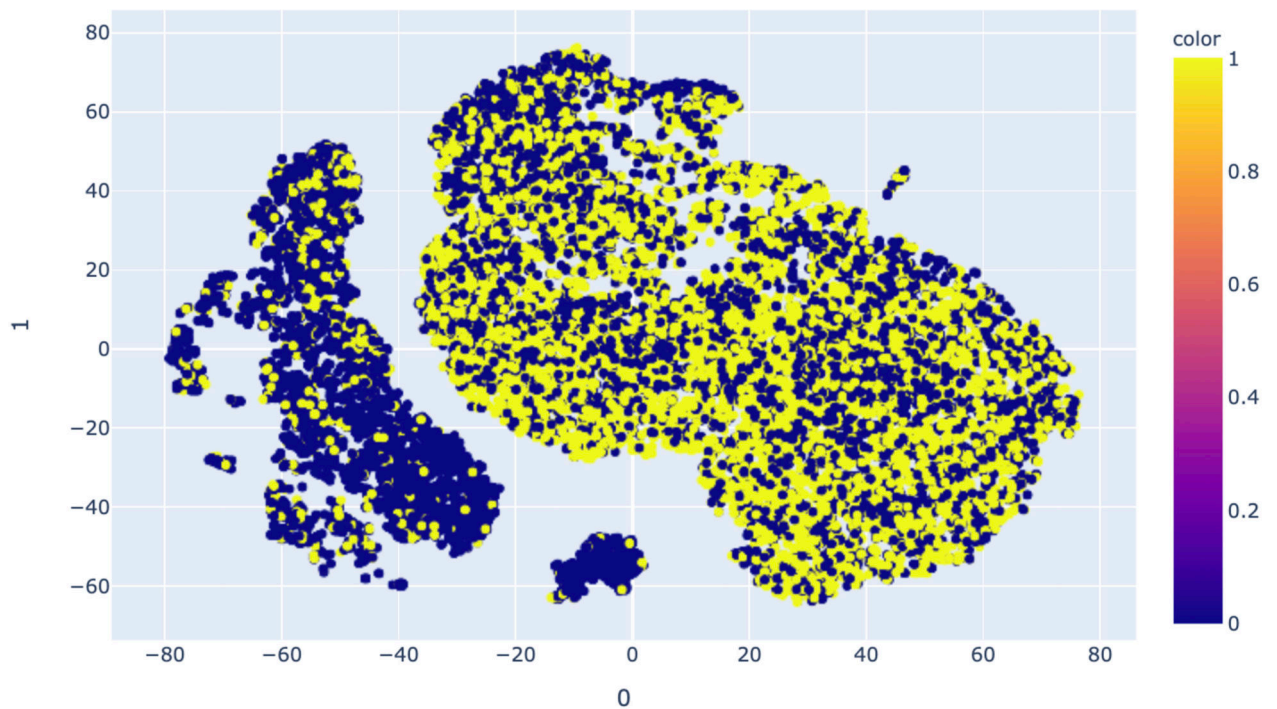
We reduce the dimensionality of our omics feature space with PCA and project it onto a manifold with TSNE.

```
clf_dimensionality_reduction = sk.decomposition.PCA(n_components=64)
X_reduced = pd.DataFrame(
    clf_dimensionality_reduction.fit_transform(X.values),
    index=X.index,
)
display(
    px.scatter_3d(
        X_reduced,
        x=X_reduced.columns[1],
        y=X_reduced.columns[2],
        z=X_reduced.columns[3],
        color=y,
        hover_data=[X_reduced.index],
    )
)
```

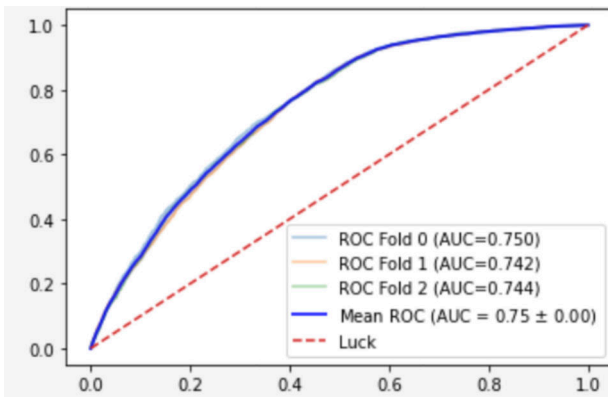


**Figure 118.** 3D scatter plot of PCA reduced input features with genes associated with the target label are colored yellow.

```
proj = sk.manifold.TSNE(n_components=2)
X_transformed = pd.DataFrame(
    proj.fit_transform(X_reduced.iloc[:, :10].values),
    index=X_reduced.index,
)
display(
    px.scatter(
        X_transformed,
        x=X_reduced.columns[0],
        y=X_reduced.columns[1],
        color=y,
        hover_data=[X_transformed.index],
    )
)
```



**Figure 119.**  
T-SNE visualization of the PCA reduced features.



Confidence interval (95%) (0.7385615427343992, 0.7520299883185799)

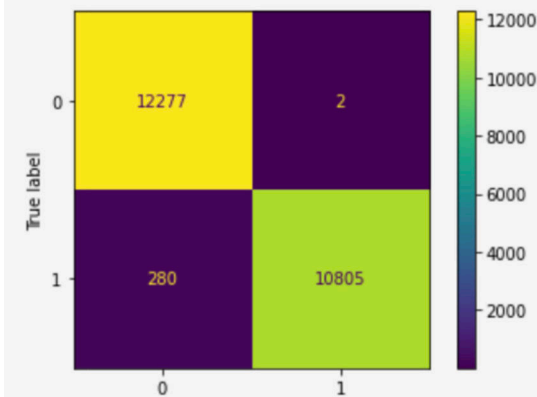
We are 100.000 % confident the model's results are not just chance.

This is statistically significant

This will take a long time as we are evaluating `n_iter` different models `n_splits` different times each computing all the metrics on `product(X.shape)` data points--not to mention the size of each model dictated by the range of parameters specified in the `params` dict.

```
model.fit(X.values, y)
sk.metrics.plot_confusion_matrix(model, X.values, y)
```

<sklearn.metrics.\_plot.confusion\_matrix.ConfusionMatrixDisplay at 0x7f101fe2e250>



**Figure 120.** Receiver operating characteristic (ROC) curves and prediction matrix displaying model performance across cross-validation splits.

```
# Obtain prediction results
y_proba = model.predict_proba(X[:, 1])
results = pd.DataFrame({
    'Known': y,
    'Predicted': (y_proba > 0.5).astype(int),
    'Prediction Probability': y_proba,
}, index=X.index).sort_values(
    'Prediction Probability',
    ascending=False,
)
results[((results['Known'] != results['Predicted']) & (results['Prediction Probability'] > 0.5))
```

GeneSym	Known	Predicted	Prediction Probability
FLVCR1	0	1	0.509489
DIP2C	0	1	0.505093

```
results.to_csv('results.tsv', sep='\t')
display(Markdown('Download model predictions at [results.tsv](./results.tsv)'))
```

Download model predictions at [results.tsv](#)

**Figure 121.**

Table of top genes predicted to be associated with the class label. The results table is available for download by clicking the “results.tsv” link.

IDG  TIGA: Target Illumination GWAS Analytics

EFO\_0009589  
 Gene...  
 Submit Reset

HitsTable (148 genes) HitsPlot (148 genes) Traits (ALL) Genes (ALL) Studies (ALL) Downloads Help

TRAIT: "worry measurement" (EFO\_0009589)

GSYMB	GeneName	idgFam	idgTDL	pVal_mlog	RCRAS	N_snpw	meanRankScore	N_st
MUSTN1	Musculoskeletal embryonic nuclear protein 1		Tdark	13.52	0.00	5.00	80.67	
FCF1	rRNA-processing protein FCF1 homolog		Tdark	13.00	0.00	3.65	75.79	
AREL1	Apoptosis-resistant E3 ubiquitin protein ligase 1	Enzyme	Tdark	13.00	0.00	3.65	75.79	
YLPM1	YLP motif-containing protein 1		Tdark	11.70	0.00	4.95	73.23	
CYP17A1	Steroid 17-alpha-hydroxylase/17,20 lyase	Enzyme	Tclin	11.70	0.00	4.00	72.84	
WBPI1	WW domain binding protein 1-like		Tdark	11.70	0.00	4.00	72.84	
PGF	Placenta growth factor		Tclin	11.70	0.00	3.95	71.76	
RPS6KI1	Ribosomal protein	Kinase	Tdark	11.70	0.00	3.95	71.76	

Hits TDLs  
 Tclin  Tchem  Tbio  Tdark

Plot Y-Axis  
 OR  N\_beta  Auto

Dataset: GWAS Catalog version: 2021-05-06; genes: 18441; traits: 1654; studies: 4725; publications: 2774

Results: "worry measurement" (EFO\_0009589); N\_gene: 148 shown (791 total)

**Figure 122.**  
 TIGA gene plot for trait "worry measurement" (EFO\_0009589).

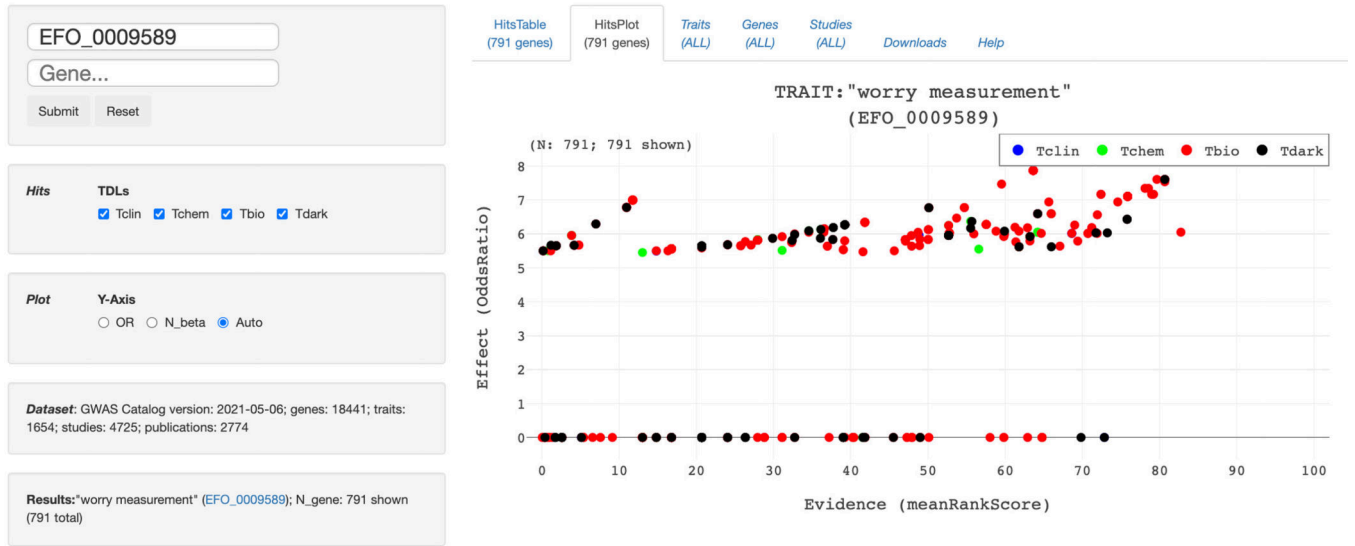
Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

IDG  TIGA: Target Illumination GWAS Analytics



TIGA web app from UNM  and IDG  built from GWAS Catalog [2021-05-06]  and EFO [3.25.0] 

**Figure 123.**  
TIGA gene hitlist for trait “worry measurement” (EFO\_0009589).

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

TRAIT: EFO\_0009589 • **GENE: ENSG00000272573**  
*worry measurement* • **MUSTN1 (Musculoskeletal embryonic nuclear protein 1)**

### Gene-trait association provenance

• **efold:** EFO\_0009589 • **ensemblid:** ENSG00000272573 • **geneFamily:** NA • **n\_beta:** 4 • **n\_snp:** 5 • **n\_snpw:** 5 • **n\_study:** 2 • **or\_median:** 7.605 • **pvalue\_mlog\_max:** 13.523 • **rcras:** 0 • **study\_N\_mean:** 353149 • **TDL:** Tdark • **traitNgene:** 773

Table: Studies with association evidence

Accession	Study	PMID	DatePublished
<a href="#">GCST006478</a>	Meta-analysis of genome-wide association studies for neuroticism in 449,484 individuals identifies novel genetic loci and pathways.	<a href="#">29942085</a>	2018-06-25
<a href="#">GCST006950</a>	Item-level analyses reveal genetic heterogeneity in neuroticism.	<a href="#">29500382</a>	2018-03-02

Showing 1 to 2 of 2 entries

Previous

1

Next

[Download Provenance \(this association\)](#)

#### Figure 124.

TIGA provenance for trait “worry measurement” (EFO\_0009589) associated gene Musculoskeletal embryonic nuclear protein 1 (MUSTN1), with two studies and associated publications, with GWAS Catalog and PubMed link-outs, respectively.

## K3 Kinase Enrichment Analysis version 3

Kinase Enrichment Analysis 3 (KEA3) infers upstream kinases whose putative substrates are overrepresented in a user-inputted list of proteins or differentially phosphorylated proteins.

### Input a list of proteins

KEA3 supports HGNC-approved gene symbols. Symbols that are not included in this set are not analyzed.

0 symbols entered, 0 duplicates, 0 valid symbols ?

Submit

Input an example

Upload proteins list

### The KEA3 workflow

Mean rank			
Rank	Protein	Mean rank	Overlapping Proteins
1	MAPK1	30.36	23 proteins
2	MAPK3	31.55	21 proteins
3	POPK2P	38	6 proteins
4	MAPK2	37.18	15 proteins
5	MAPK14	37.82	20 proteins
6	BRAF	45.5	15 proteins
7	MAP2K1	40.82	20 proteins
8	AKT1	51.73	24 proteins
9	PRKACA	53.55	21 proteins
10	PAK1	56.27	15 proteins

Top rank			
Rank	Protein	Integrated scaled rank	Overlapping Proteins
1	RPS26A1	0.001927	10 proteins
2	POPK1	0.001946	13 proteins
3	AKT1	0.00211	7 proteins
4	MAP2K1	0.00211	8 proteins
5	YES1	0.002375	5 proteins
6	ABL2	0.003854	12 proteins
7	RPS26A3	0.003891	12 proteins
8	MAPK14	0.004219	4 proteins
9	PAK1	0.00463	4 proteins
10	ULK1	0.00578	12 proteins

KEA performs enrichment analysis on the inputted list of differentially expressed proteins using gene set libraries from kinase-substrate interaction databases.

[FAQ](#)

[Tutorial](#)

[API](#)

[Download libraries](#)



Center for  
Biotinformatics



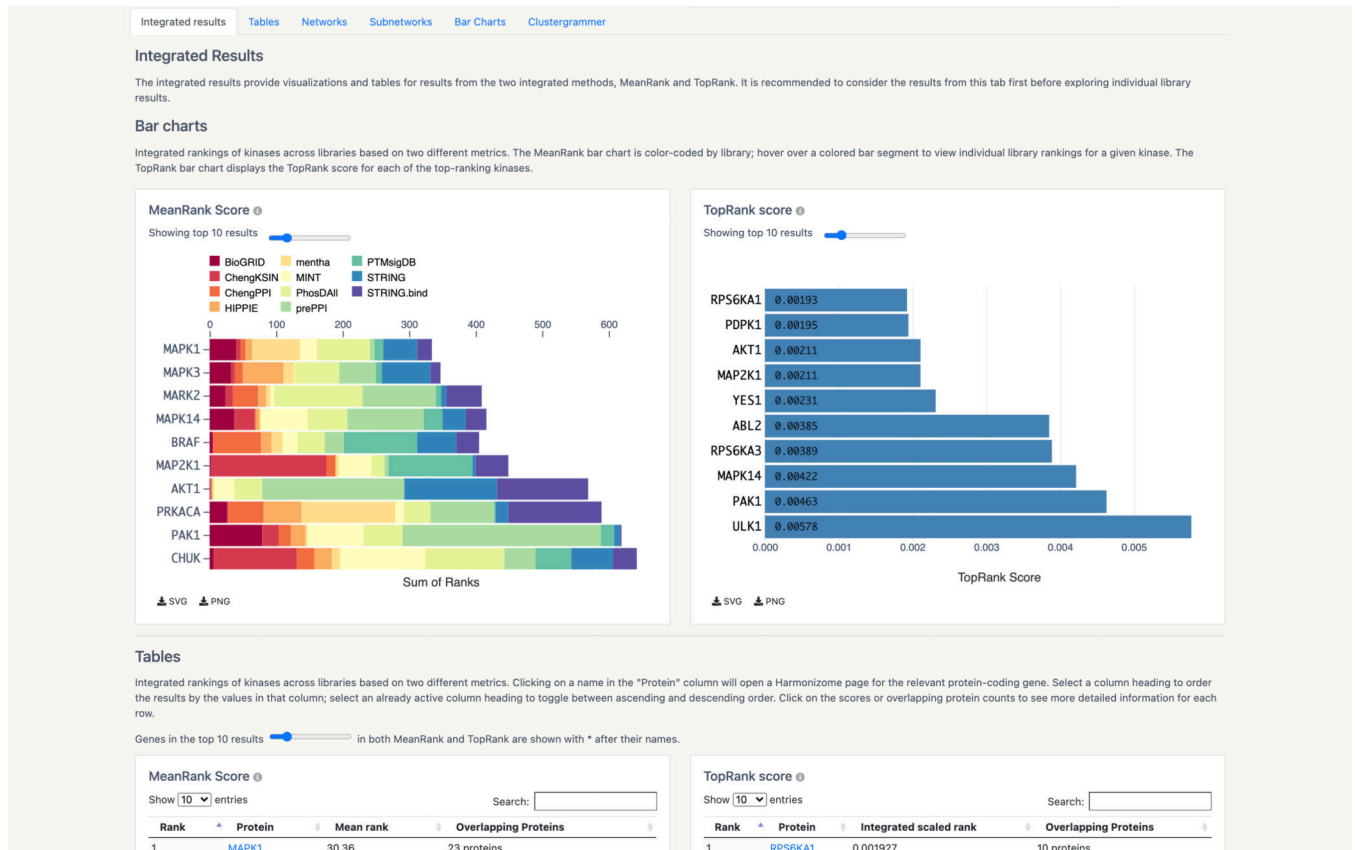
[View source code](#)

[Submit an issue](#)

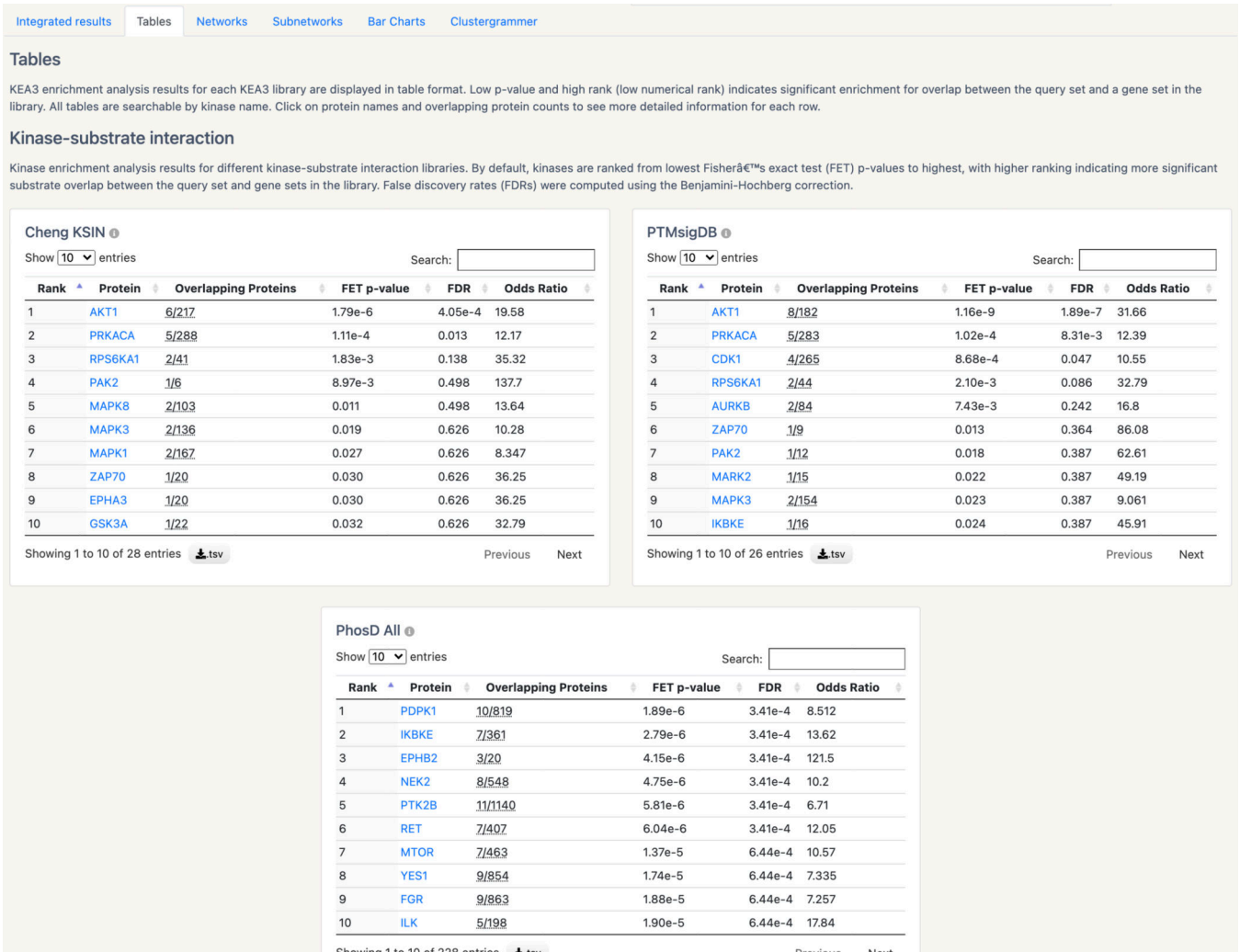
**Figure 125.**

KEA3 homepage with gene input box. HGNC gene symbols can be pasted into the text box or a newline separated .txt file containing the input gene list can be uploaded.

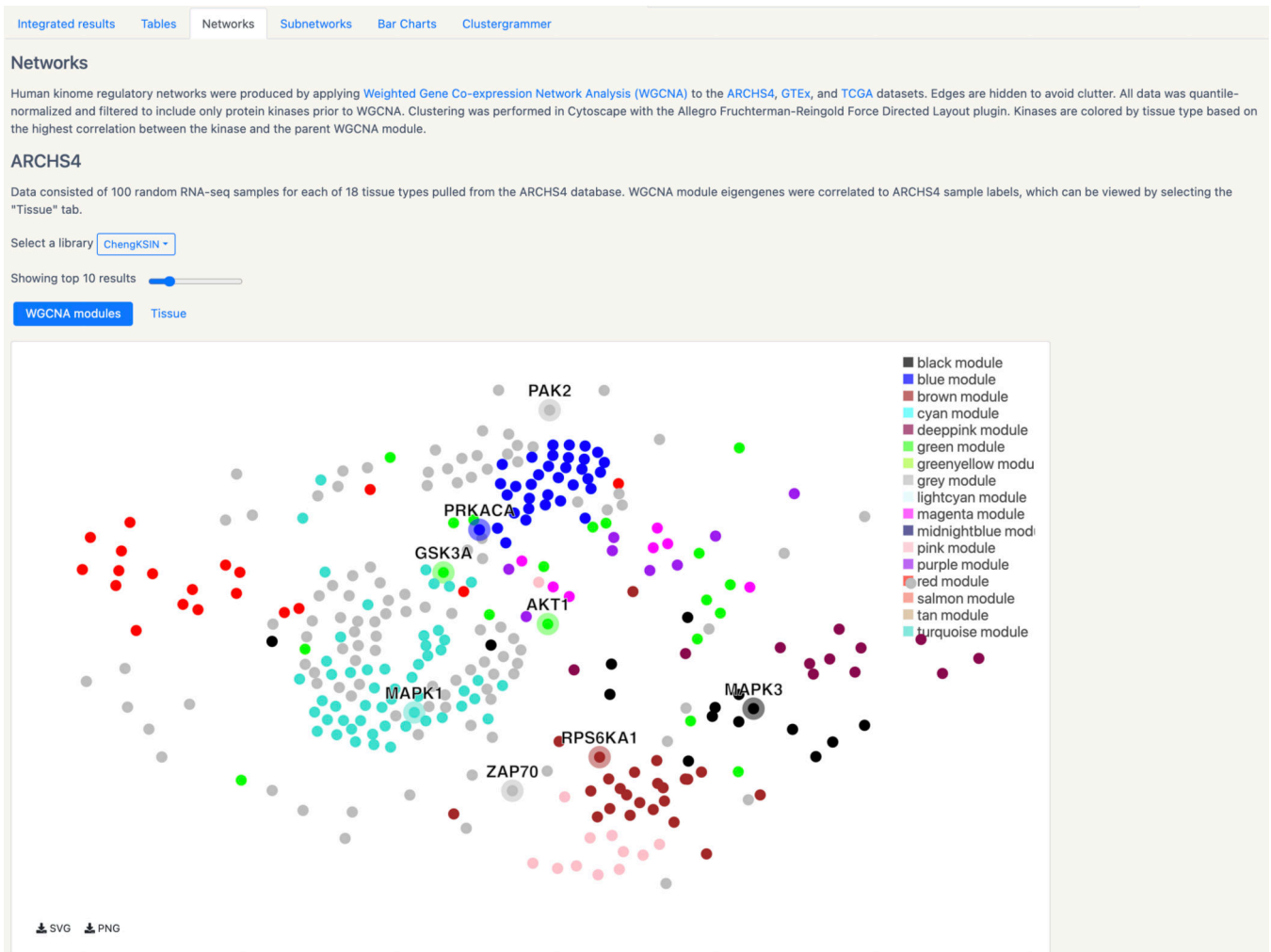




**Figure 126.** Snippet of the integrated results tab showing the top enriched kinases using the MeanRank and TopRank methods through a variety of tables and visualizations.

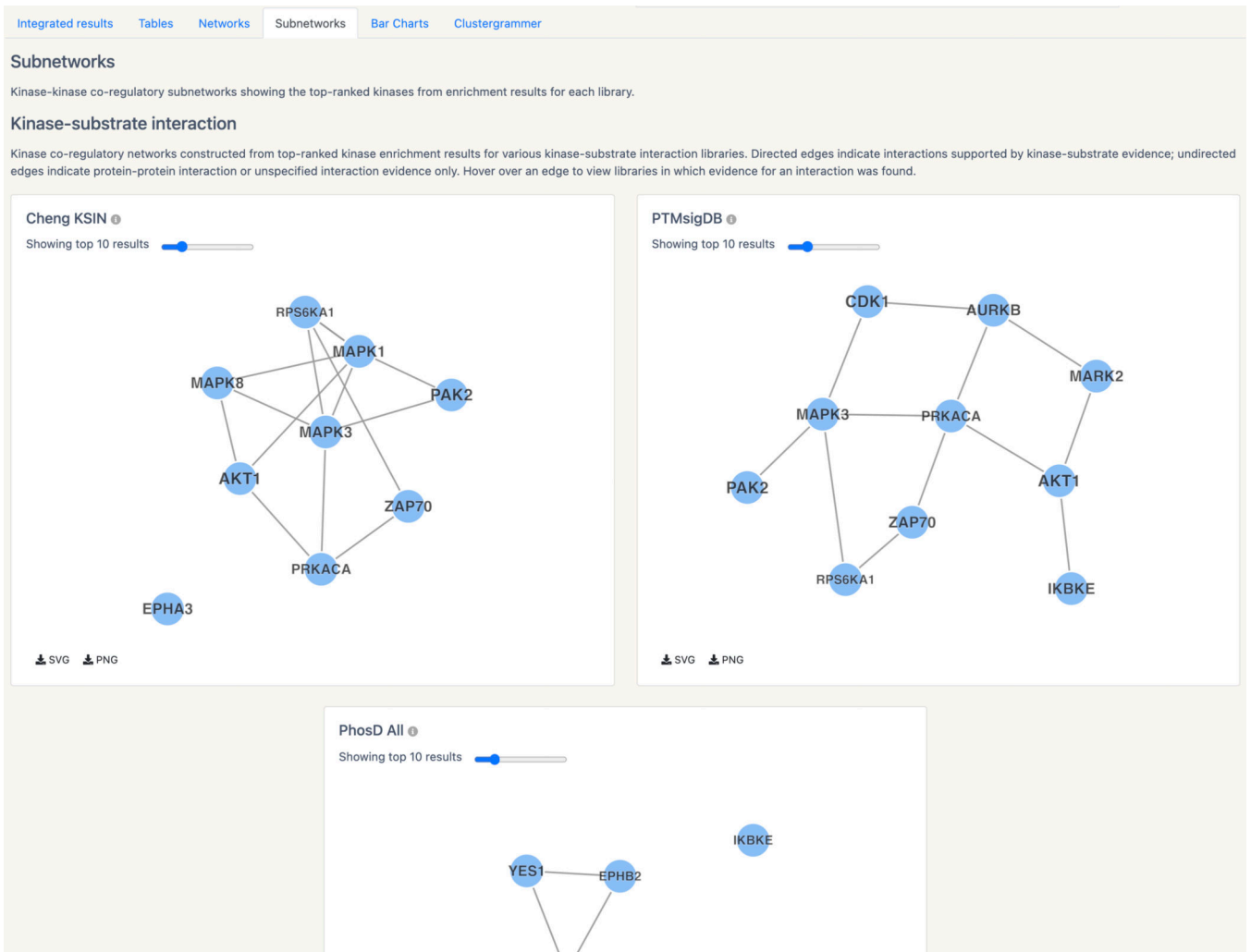


**Figure 127.** Tables tab showing the top enriched kinase results from the kinase-substrate interaction libraries. Each table can be re-sorted by clicking the table headers for each table. Specific terms of interest can be queried in any of the search bars within each table.

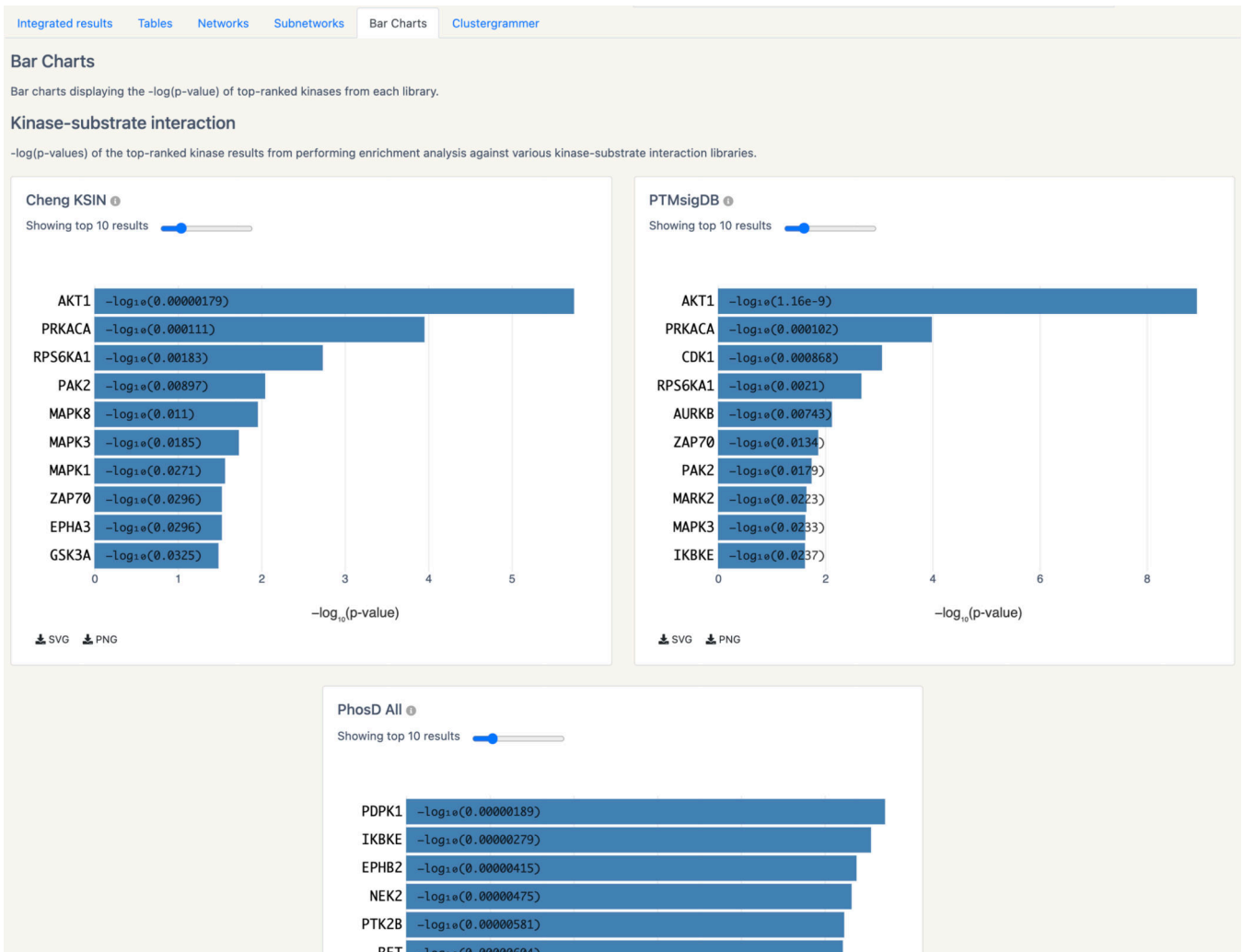


**Figure 128.**

Networks tab displaying human kinome regulatory networks that were produced by applying Weighted Gene Co-expression Network Analysis (WGCNA) to ARCHS4, GTEX, and TCGA datasets. Kinases are colored by tissue type based on the highest correlation between the kinase and parent WGCNA module.



**Figure 129.** Subnetworks tab displaying the kinase-kinase co-regulatory networks showing the top-ranked kinases from enrichment results for kinase-substrate interaction libraries.




**Figure 130.** Bar charts tab displaying the  $-\log(p\text{-value})$  of top-ranked kinases from the kinase-substrate interaction libraries.



**Figure 131.**

This interactive visualization highlights the relationships between the most common kinase-substrate associations detected as overlapping with the input. Each column represents a protein set from a KEA3 library, while the rows are putative substrates from the input list which overlap with proteins within each of the KEA3 library sets. Rows and columns can be sorted by sum to observe the KEA3 sets with the most substrates.

## DrugShot



This Appyter searches PubMed for articles that co-mention any search term that relates to drugs.

If selecting the "Biomedical Term" method: DrugShot finds publications that mention both the search term and drugs. It then prioritizes these drugs using various methods, as well as predicts additional drugs based on shared properties among drugs and other small molecules.

If selecting the "List" method: users can input a list of small molecules that they want to utilize as an unweighted drug set for prioritizing related compounds from co-occurrence and co-expression.

**Method Selection**

Choose between querying a biomedical term of interest to prioritize drugs and small molecules associated with the query term, or uploading a list of small molecules to be augmented using co-occurrence and co-expression matrices.

Method Selection <sup>Ⓢ</sup> :

Biomedical Term

List

Biomedical Term <sup>Ⓢ</sup> :

Associated drug set size <sup>Ⓢ</sup> :

Submit

**Figure 132.** Biomedical Term input form with “Lung Cancer” input in the Biomedical Term field. The associated drug set size is 50, therefore the unweighted drug set will include 50 small molecules.

**Table 1: Top Associated Compounds**  
([Lung\\_Cancer\\_associated\\_drug\\_table.csv](#))

	Publications with Search Term	Publications with Search Term / Total Publications
gefitinib	2620	0.538541
gemcitabine	2393	0.264070
erlotinib	1929	0.535387
etoposide	1693	0.153741
vinorelbine	1367	0.488738
fluorouracil	1118	0.053727
irinotecan	1064	0.160072
camptothecin	883	0.121291
ifosfamide	546	0.165354
tretinoin	430	0.022620
topotecan	400	0.183150
doxorubicin	392	0.059223
paclitaxel	372	0.094225
imatinib	328	0.033031
sorafenib	318	0.093282
tamoxifen	294	0.022407
pidorubicine	259	0.070095
decitabine	234	0.077050
prednisone	227	0.014717
capecitabine	227	0.058370

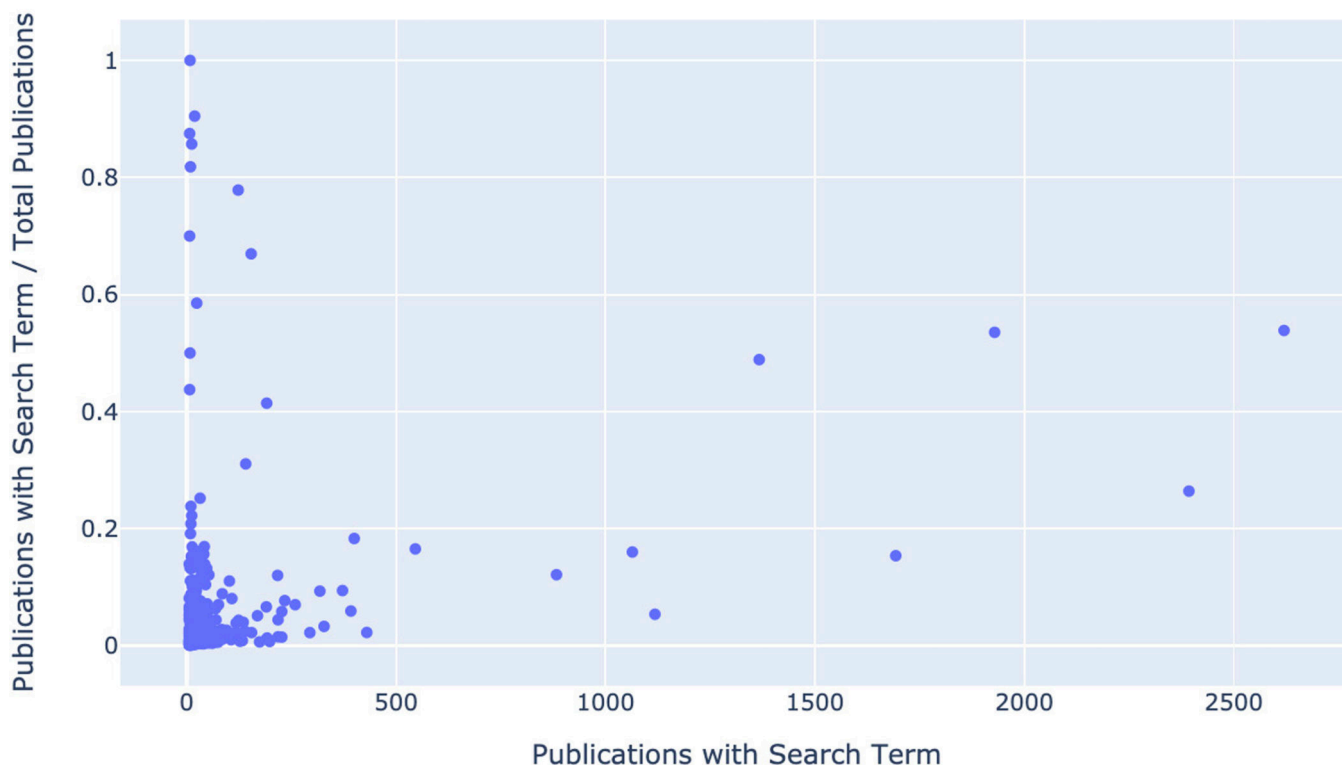
**Table 1: Top 20 Drugs associated with Lung Cancer.**

**Figure 133.**

Table of Top 20 Associated Compounds. This table provides the top-ranked drug and compound names associated with the query term (Column 1); the count of PubMed publications associating each drug with the search term (Column 2); and the fraction of the count from Column 2, divided by the total number of publications related to that drug (Column 3).



## Lung Cancer



**Figure 134.**

Scatter Plot of Drug Frequency in Literature. The X axis displays the integer counts of Publications with Search Term, and the Y axis shows the fraction of Publications with Search Term / Total Publications. Hovering over any point on this plot displays the compound's name and its corresponding X and Y values.

## DrugShot



This Appyter searches PubMed for articles that co-mention any search term that relates to drugs.

If selecting the "Biomedical Term" method: DrugShot finds publications that mention both the search term and drugs. It then prioritizes these drugs using various methods, as well as predicts additional drugs based on shared properties among drugs and other small molecules.

If selecting the "List" method: users can input a list of small molecules that they want to utilize as an unweighted drug set for prioritizing related compounds from co-occurrence and co-expression.

### Method Selection

Choose between querying a biomedical term of interest to prioritize drugs and small molecules associated with the query term, or uploading a list of small molecules to be augmented using co-occurrence and co-expression matrices.

Method Selection <sup>Ⓢ</sup>:

[Biomedical Term](#)

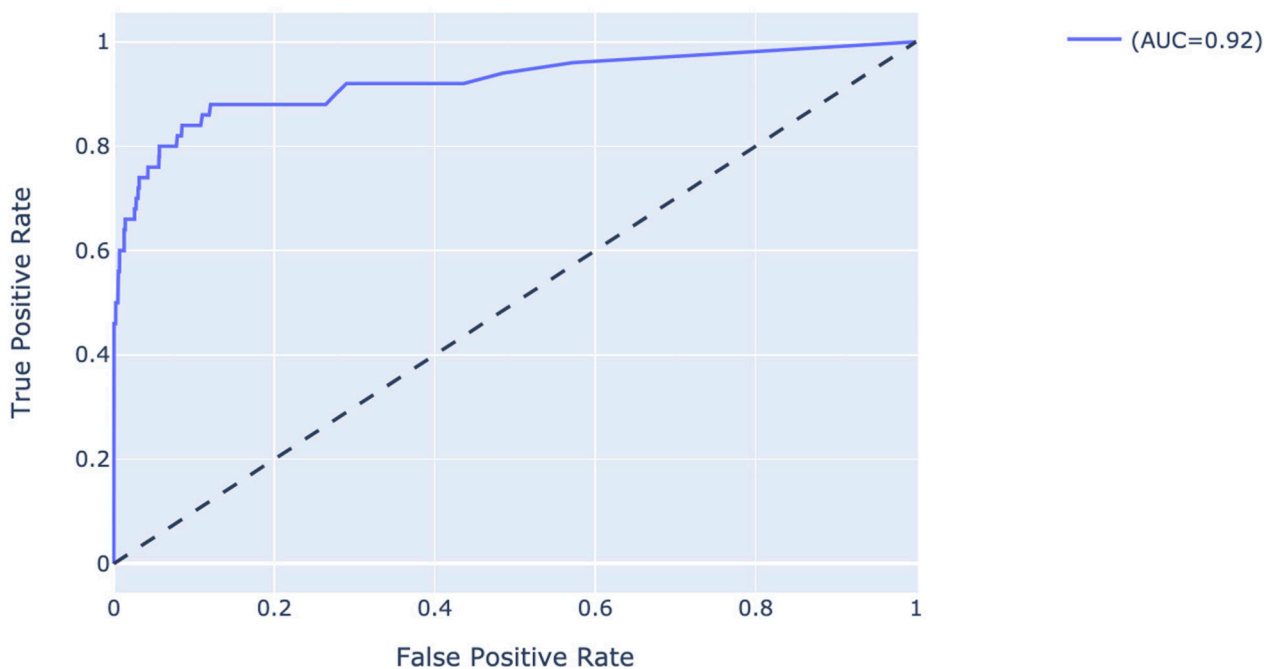
Upload List of Small Molecules <sup>Ⓢ</sup>:

Load example <sup>Ⓢ</sup>: [drug\\_augmentation\\_feature\\_example.txt](#)

Download example <sup>Ⓢ</sup>: [drug\\_augmentation\\_feature\\_example.txt](#)

**Figure 135.**  
List input form where newline separated .txt files of small molecule names are uploaded for drug set augmentation.

## ROC Curve for Associated Compound Rankings in Co-occurrence Prediction Matrix



**Figure 136.**

Receiver operating characteristic curve for rankings of unweighted drug set in co-occurrence matrix. The area under the curve (AUC) is shown to the right of the plot, and hovering over any point on the curve displays the associated X and Y values.

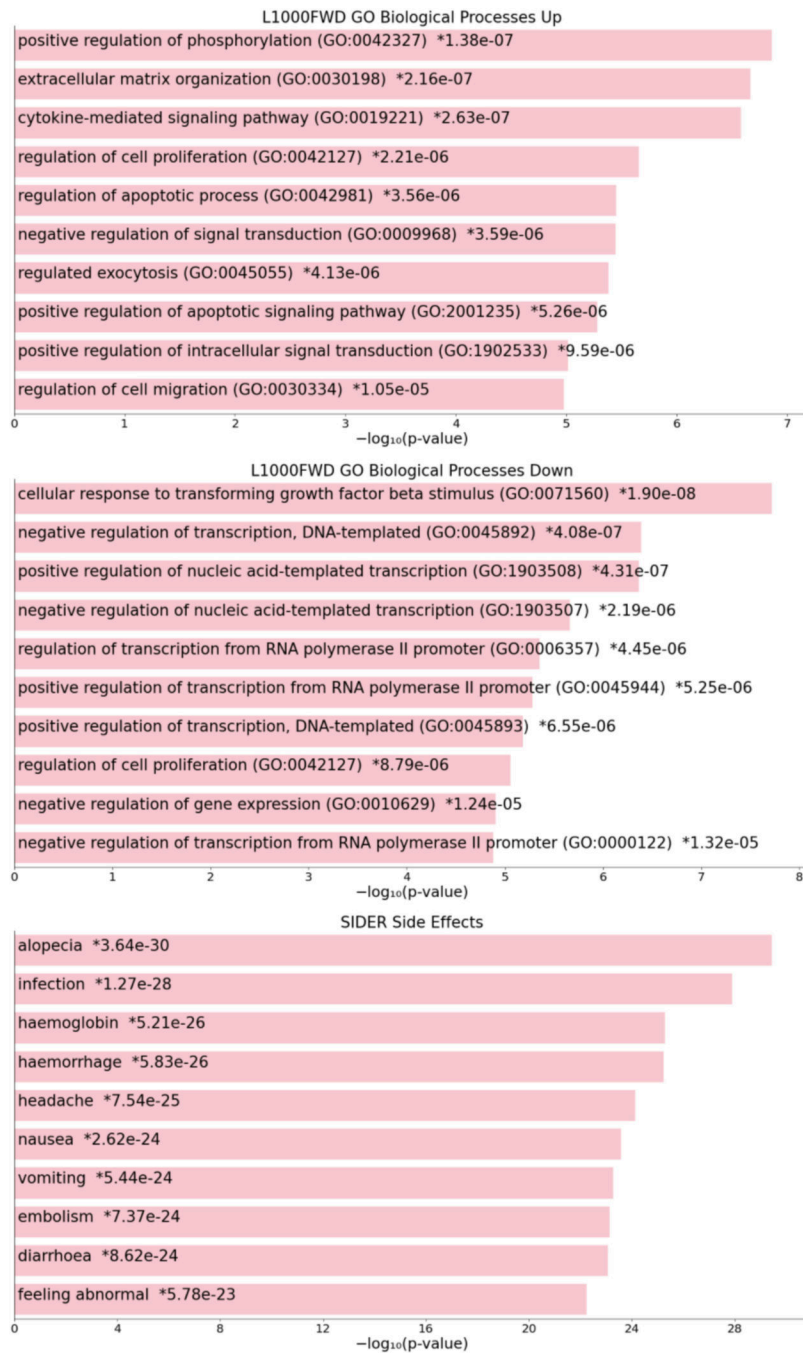
**Table 2: Top Predicted Compounds From Co-occurrence**  
[\(Lung\\_Cancer\\_cooccurrence\\_similarity\\_predicted\\_drug\\_table.csv\)](#)

	Score
cytarabine	81.92
prednisone	63.56
tromethamine	42.30
thalidomide	42.16
dasatinib	39.96
letrozole	39.48
anastrozole	36.08
hydroxyurea	32.44
aspirin	31.72
prednisolone	31.26
rofecoxib	29.54
carmustine	28.64
raloxifene	26.20
hydrocortisone	25.94
valproic-acid	24.50
nilotinib	24.20
idarubicin	23.92
staurosporine	20.60
exemestane	20.36
vitamin e	19.82

**Table 2:** *Top 20 drugs predicted to be associated with Lung Cancer based on DrugRIF co-occurrence.*

**Figure 137.**

Table of top 20 predicted compounds predicted from DrugRIF co-occurrence. Click on the hyperlinked filename below the table header to download a .CSV file listing the complete ranked set of predicted compounds and their associated similarity scores.

**Figure 138.**

Bar plots of top 10 enriched terms across three separate drug set libraries after drug set enrichment analysis of the top 50 co-occurrence predicted drugs using the DrugEnrichr API. Colored bars correspond to terms with significant p-values (<0.05). An asterisk (\*) next to a p-value indicates the term also has a significant adjusted p-value (<0.05).

Link to the complete enrichment analysis results is output below

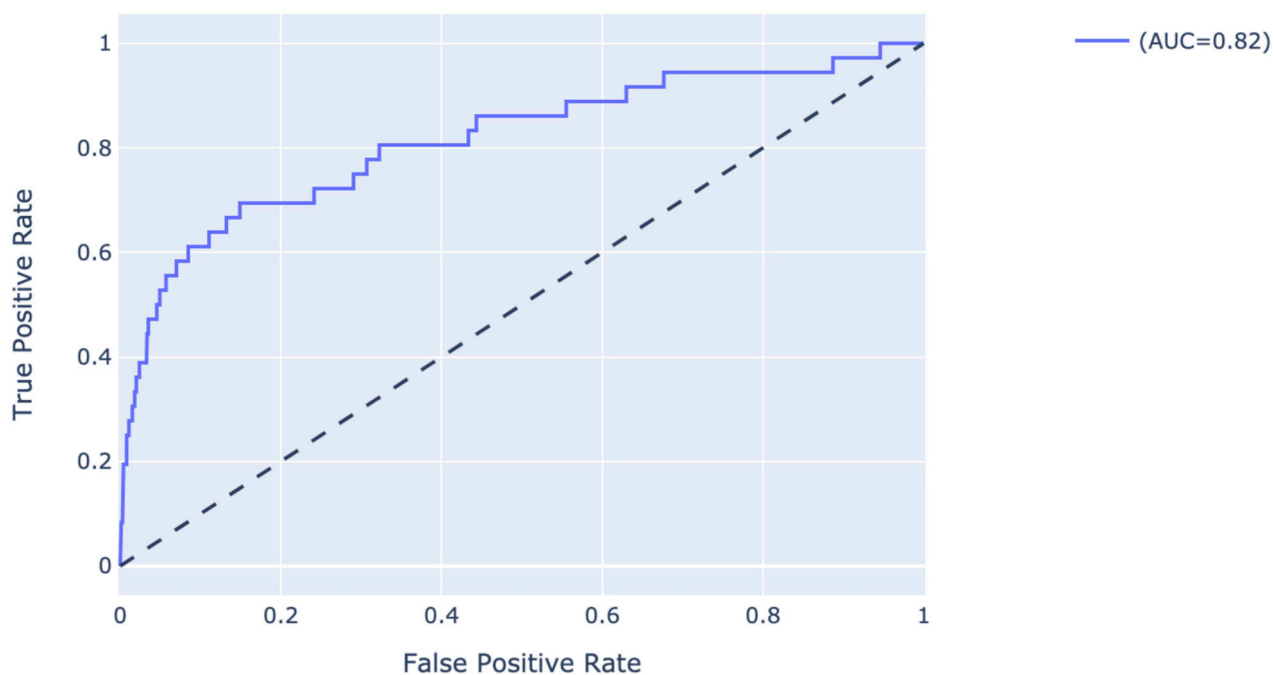
```
drugenrichr_link(short_id)
```

[Access the complete enrichment analysis on the DrugEnrichr website.](#)

**Figure 139.**

DrugEnrichr link to drug enrichment analysis results from querying the top 50 co-occurrence predicted compounds.

## ROC Curve for Associated Compound Rankings in Co-expression Prediction Matrix



**Figure 140.**

Receiver operating characteristic curve for rankings of unweighted drug set in co-expression matrix. The area under the curve (AUC) is shown to the right of the plot, and hovering over any point on the curve displays the associated X and Y values.

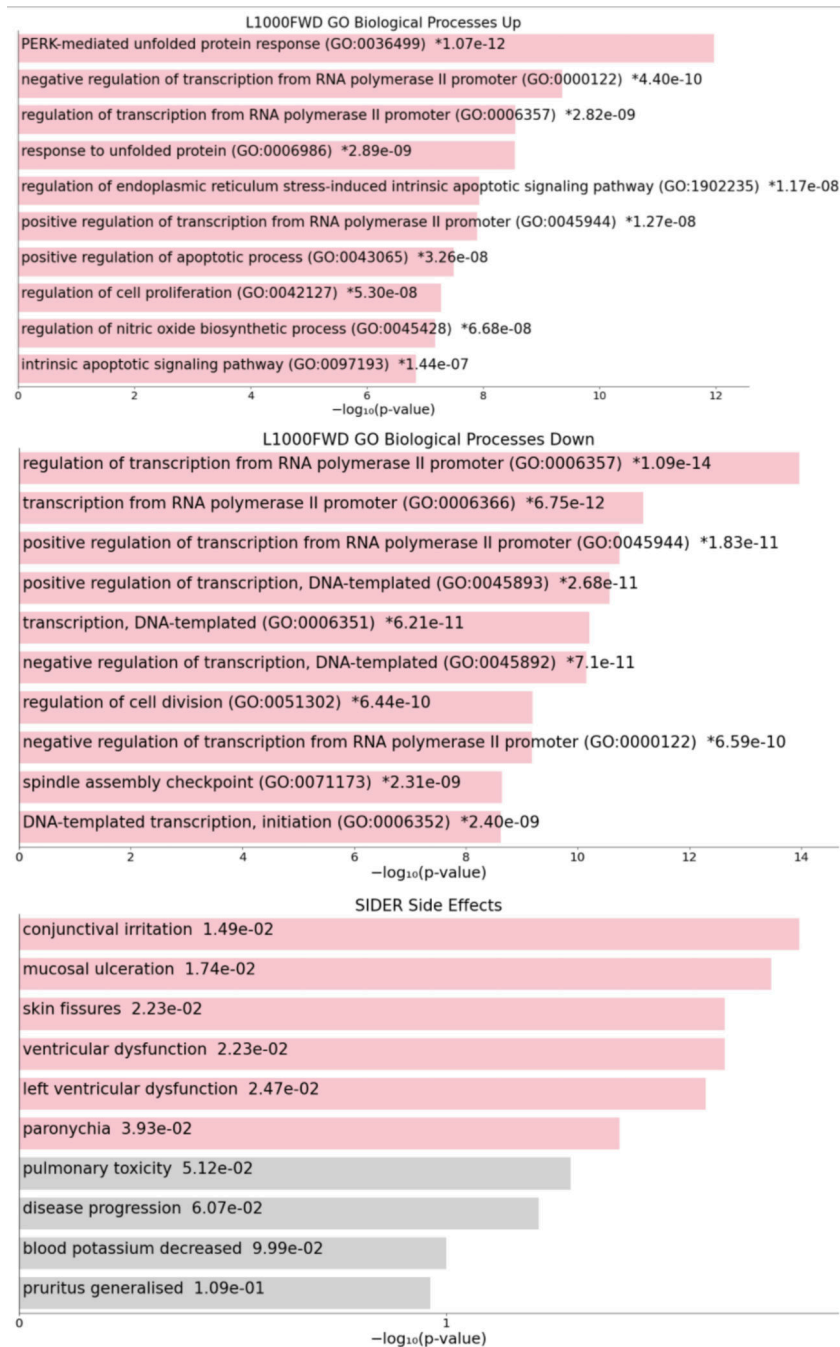
	Score
BIX-01294	0.168949
purvalanol-a	0.166612
SB-590885	0.162408
7b-cis	0.162084
YM-155	0.157794
BRD-K29506255	0.154279
CGK-733	0.151474
afatinib	0.151058
BRD-K77681376	0.150942
avicin-d	0.150777
CAY-10585	0.150705
BRD-K07877311	0.150639
PD-184352	0.150517
KIN001-055	0.150451
VU-0418939-2	0.150310
BRD-K10906552	0.149563
neratinib	0.149104
ispinesib	0.148588
pirarubicin	0.148107
olvanil	0.147889

**Table 3:** *Top 20 drugs predicted to be associated with Lung Cancer based on co-expression.*

**Figure 141.**

Table of top 20 predicted compounds predicted from L1000 co-expression. Click on the hyperlinked filename below the table header to download a .CSV file listing the complete ranked set of predicted compounds and their associated similarity scores.



**Figure 142.**

Bar plots of top 10 enriched terms across three separate drug set libraries after drug set enrichment analysis of the top 50 co-expression predicted drugs using the DrugEnrichr API. Colored bars correspond to terms with significant p-values ( $<0.05$ ). An asterisk (\*) next to a p-value indicates the term also has a significant adjusted p-value ( $<0.05$ ).

Link to the complete enrichment analysis results is output below

```
drugenrichr_link(short_id)
```

[Access the complete enrichment analysis on the DrugEnrichr website.](#)

**Figure 143.**  
DrugEnrichr link to drug enrichment analysis results from querying the top 50 co-expression predicted compounds.