Mining ClinicalTrials.gov via CTTI AACT for drug target hypotheses

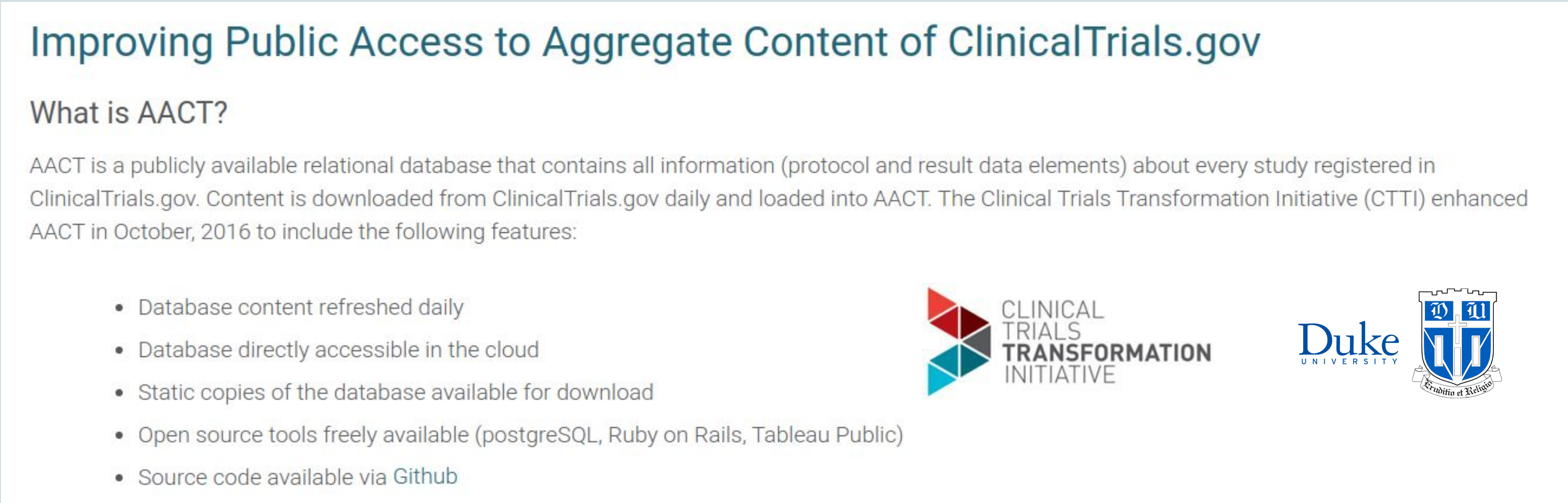
Jeremy Yang[1], Roger Sayle[2], Lars Juhl Jensen[3] and Tudor Oprea[1]

[1]University of New Mexico, Albuquerque, USA; [2]NextMove Scientific Software, Cambridge, UK; [3]Novo Nordisk Foundation Center for Protein Research, Copenhagen, DK

Overview

We mined **ClinicalTrials.gov** using the **CTTI-AACT** db from **Duke University** for drugs/chemicals and diseases/conditions. Named entitiy recognition (NER) requires specialized tools and expertly curated dictionaries for comprehensive and high quality results, hence use of **NextMove Leadmine** for chemical NER and **JensenLab Tagger** for disease NER. Study designs and outcomes can offer new and unique drug target knowledge. Target hypotheses can be inferred indirectly via drugs and diseases, a valuable source of evidence to **Illuminate the Druggable Genome**.

Aggregate Analysis of ClinicalTrials.gov (AACT) db from the Clinical Trials Transformation Initiative (CTTI)



Improving Public Access to Aggregate Content of ClinicalTrials.gov

What is AACT?

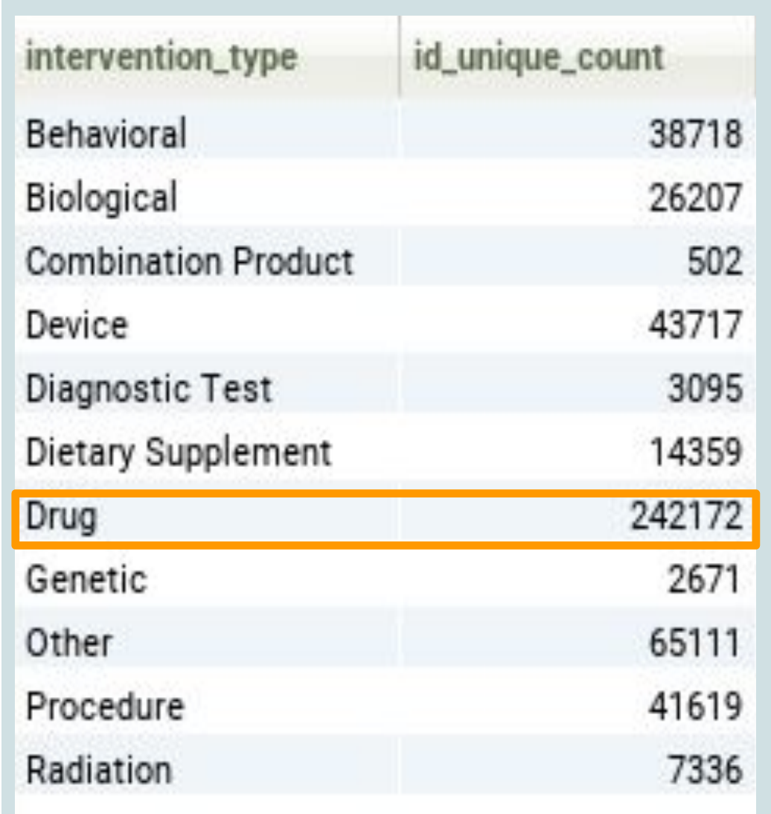
AACT is a publicly available relational database that contains all information (protocol and result data elements) about every study registered in ClinicalTrials.gov. Content is downloaded from ClinicalTrials.gov daily and loaded into AACT. The Clinical Trials Transformation Initiative (CTTI) enhanced AACT in October, 2016 to include the following features:

- Database content refreshed daily
- Database directly accessible in the cloud
- Static copies of the database available for download
- Open source tools freely available (postgreSQL, Ruby on Rails, Tableau Public)
- Source code available via Github

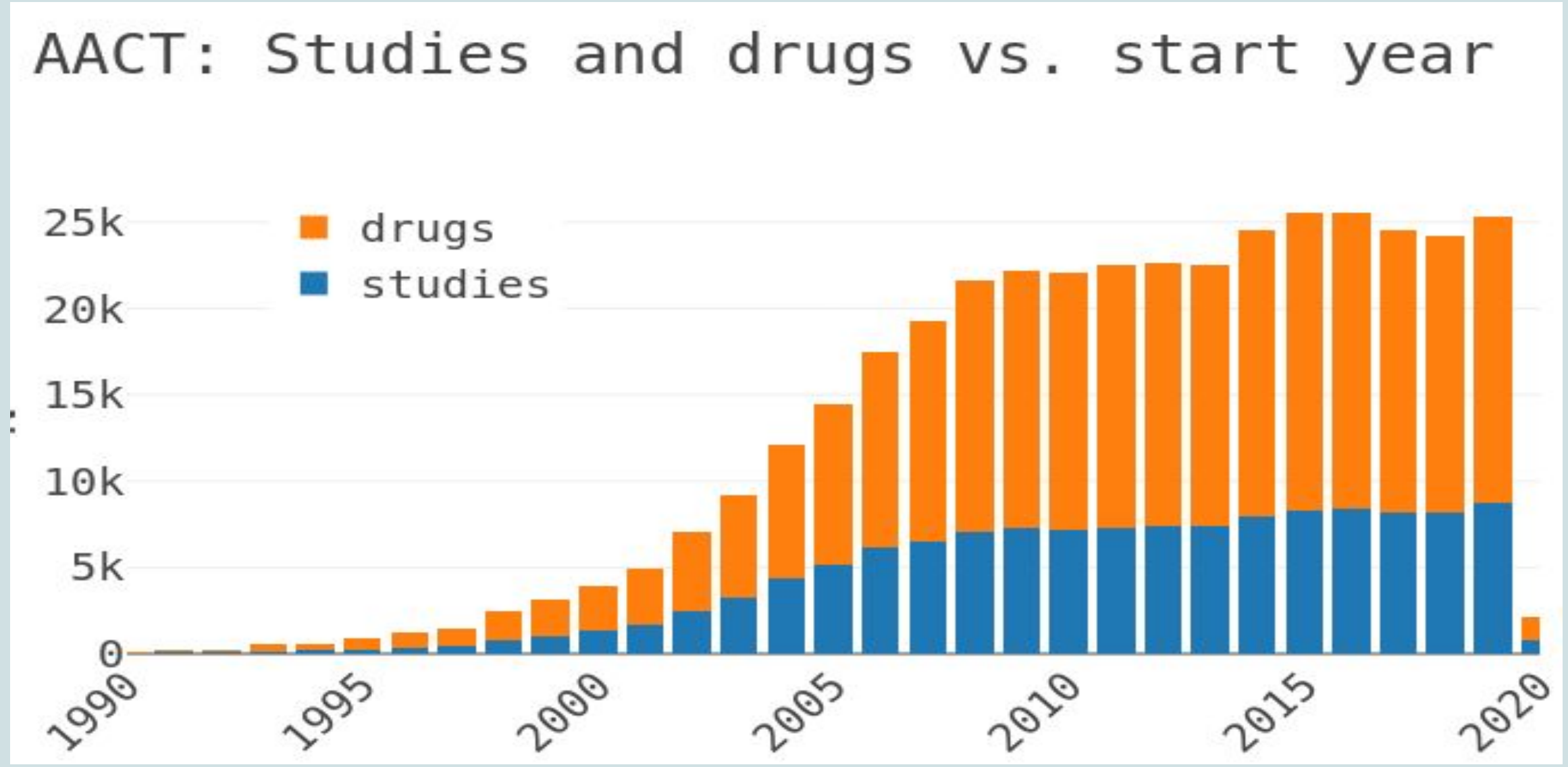
https://aact.ctti-clinicaltrials.org/

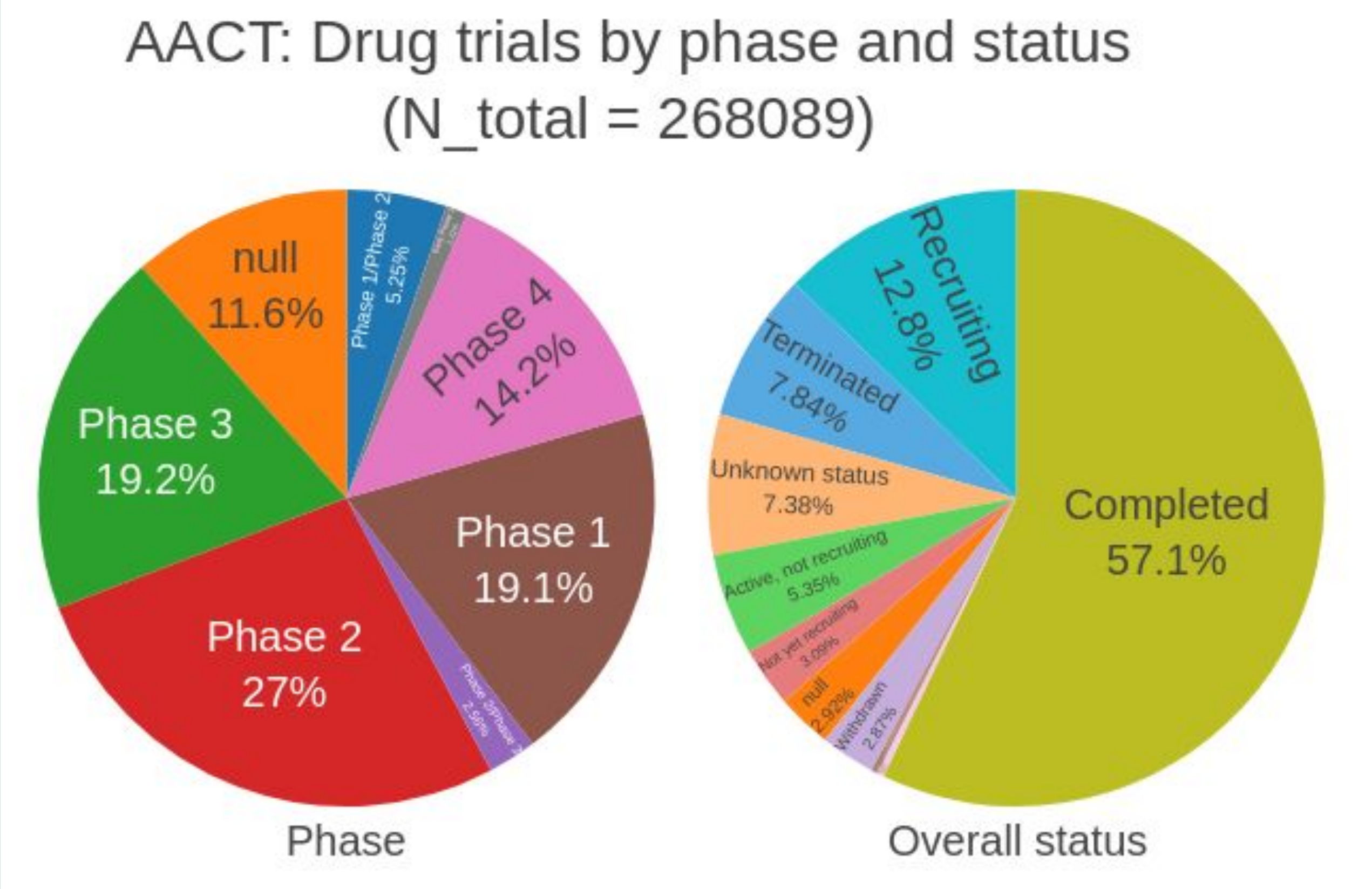
Drug trials produce new and rich experimental data

AACT accessed Dec 3, 2019.



intervention_type	id_unique_count
Behavioral	38718
Biological	26207
Combination Product	502
Device	43717
Diagnostic Test	3095
Dietary Supplement	14359
Drug	242172
Genetic	2671
Other	65111
Procedure	41619
Radiation	7336





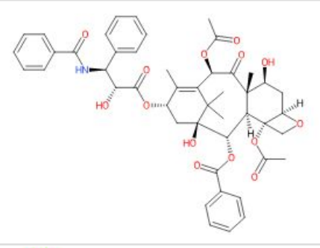
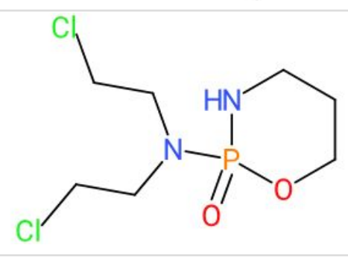
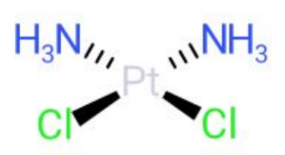
Why not target NER?

Clinical trials not designed to communicate molecular mechanisms to research scientists, but with focus on clinical efficacy and safety. In due diligence we performed target NER, and compared with target NER on arbitrary non-biomedical text: tweets from the Twitter API for #brexit (26 Nov 2019). We find 8.64 target entities per 1000 chars in the tweets (e.g. "TAX", "LIAR", "NHS", "DANGER", "insulin"), vs. 6.63 in the clinical trials descriptions. While not proof this does support prior belief and target inference via chemical NER.

Chemical NER with NextMove LeadMine

https://www.nextmovesoftware.com/

Intervention names and study descriptions mined by LeadMine v3.14.1. Drug trials (interventional): 130740; drug names: 14969; SMILES: 4869. Many non-structures, e.g. "placebo", "test product", "medication", "chemotherapy". Top drugs by total mentions:

CDK_smi2img	N_mentions	names
	2787	Abraxane; PACLITAXEL; Paclitaxel; Taxol; abraxane; paclitaxel; taxol
	2654	CYCLOPHOSPHAMIDE; Ciclophosphamide; Cyclophosphamid; Cyclophosphamide; ciclophosphamide; cyclophosphamide
	2552	CISPLATIN; Cis Platinum; Cis-platinum; Cisplatin; Cisplatine; Cisplatinum; cis Platinum; cis-platinum; cisplatin; cisplatine; cisplatinum

Disease NER with JensenLab Tagger

https://github.com/larsjuhljensen/tagger

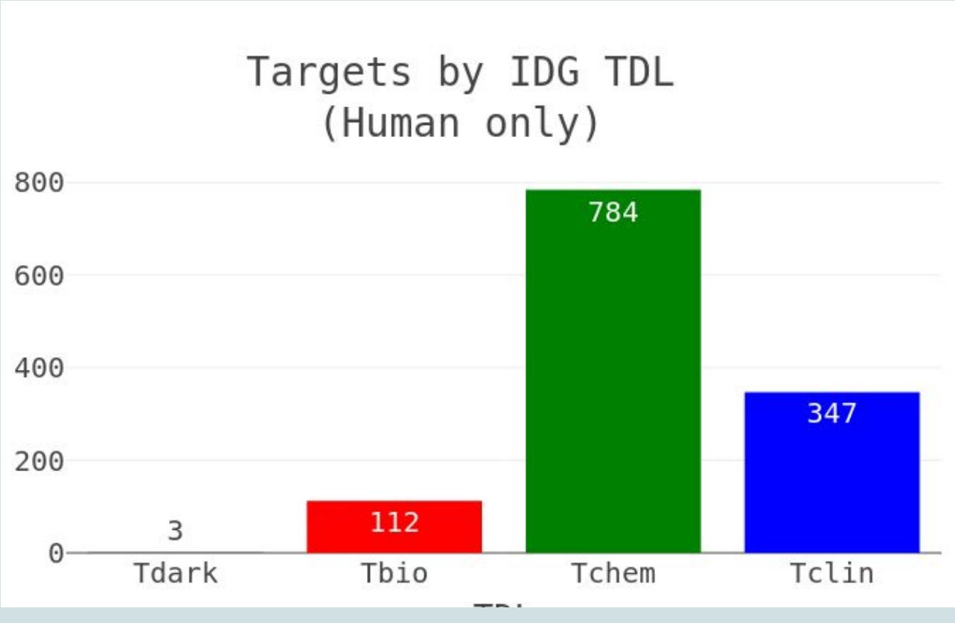
JensenLab diseases dictionary, on detailed descriptions.
Disease mention totals by merging to resolved Disease Ontology term (DOID).

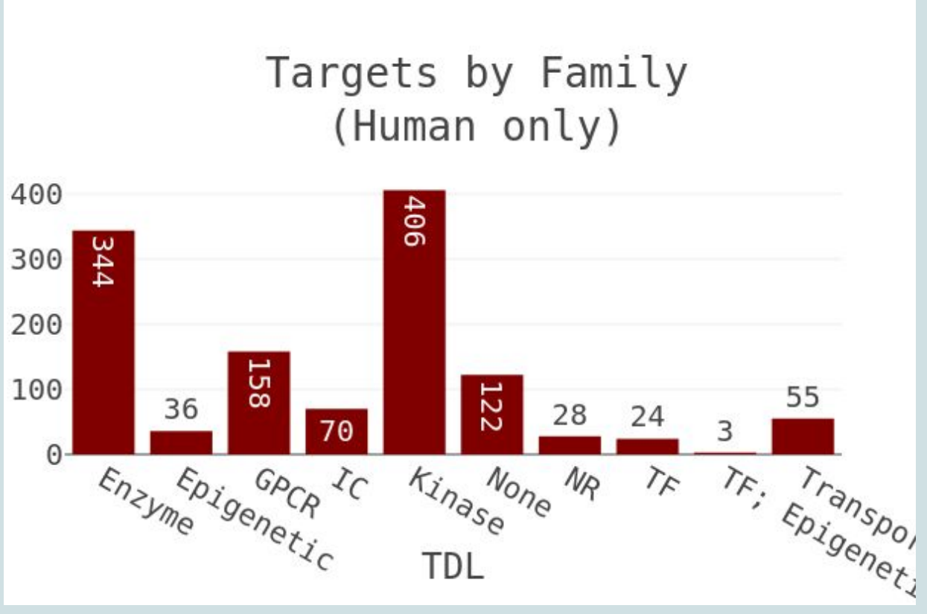
Top diseases by total mentions:

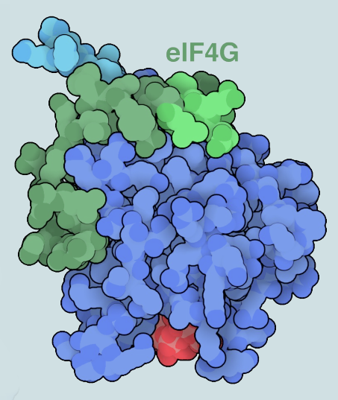
doid	N_mentions	terms
DOID:162	31143	CANCER; CANcer; Cancer; Malignant Tumor; Malignant neoplasm; Malignant tumor; Primary Cancer; Primary cancer; cancer; ...
DOID:9351	18955	DIABETES; DIABETES MELLITUS; DIAbetes; DIabetes; Diabetes; Diabetes Mellitus; Diabetes mellitus; diabetes; diabetes Mellitus...
DOID:6713	18461	CVA; Cerebrovascular Accident; Cerebrovascular Disease; Cerebrovascular accident; Cerebrovascular disease; STROKE; STRokE;…
DOID:2030	13621	ANXIETY; Anxiety; Anxiety Disorder; Anxiety State; Anxiety disorder; Anxiety state; anxiety; anxiety disorder; anxiety state; ...
DOID:1612	11586	BREAST CANCER; BReast CAncer; BReast Cancer; Breast Cancer; Breast cancer; Breast tumor; Breast-cancer; Primary breast cancer;…
DOID:2841	10773	ASTHMA; Asthma; BHR; Bronchial hyper-reactivity; Bronchial hyperreactivity; EIA; Exercise-induced asthma; asthma; bronchial hyper re …
DOID:3083	10726	CHRONIC OBSTRUCTIVE PULMONARY DISEASE; COLD; COPD; COPd; Chronic Obstructive Lung Disease; Chronic Obstructive L…
DOID:9970	10193	OBESITY; OBesity; Obesity; obEsity; obe-sity; obesity
DOID:10763	9816	HBP; HTN; HYPERTENSION; High Blood Pressure; High blood pressure; High-blood pressure; Hypertension; Hypertensive disease; high…

Compound-target mapping via PubChem & ChEMBL

Compounds mapped to: PubChem via PUG REST API, SMILES exact search; ChEMBL via REST API, InChIKey search; targets via ChEMBL bioassays. Targets mapped to IDG-TCRD/Pharos via UniProt ID.







Disease-target associations from drug trial links

Proposed confidence metrics:

- nStudy : Study count for association.
- nStudyNewness : Study count weighted by newness of study (newer better).
- nStudyPhase : Study count weighted by phase of study (completed better).
- nPub : Study publications.
- nPubTypes : Study publications (results type better).
- nDiseaseMention : Disease mention count for disease-target association.
- nDrugMention : Drug mention count for disease-target association.
- nDrug : Drug count for disease-target association.
- nAssay : Assay count for drug-target association.
- nAssayPchembl : Assay count for drug-target association, weighted by pChembl.

1.2M associations; 164K unique disease-gene pairs. Examples:

nct_id	drug_name	cid	disease_term	doid	gene_symbol	uniprot	idgTDL
NCT02600741	Fluphenazine	3372	schizophrenia	DOID:5419	MC5R	P33032	Tchem
NCT00008190	melphalan	460612	acute leukemia	DOID:12603	MMP2	P08253	Tchem
NCT00003012	methotrexate	126941	breast cancer	DOID:1612	KDM4E	B2RXH2	Tchem
NCT01445522	ABT-888	11960529	lymphoma	DOID:0060058	PARP12	Q9H0J9	Tdark
NCT03201250	Cabozantinib	25102847	multiple myeloma	DOID:9538	ANKK1	Q8NFD2	Tbio

This work supported by US National Institutes of Health (grants U54 CA189205 and U24 224370), Illuminating the Druggable Genome Knowledge Management Center (IDG KMC), and by the Novo Nordisk Foundation (grant number NNF14CC0001).

IDG F2F Meeting - 11-12 Feb 2020 - Arlington, VA