Supplementary Text, Figures & Tables:

Systematic analyses of drugs and disease indications in RepurposeDB reveal pharmacological, biological and epidemiological factors influencing drug repositioning

Khader Shameer[1], Benjamin S. Glicksberg[1], Rachel Hodos[1], Marcus A. Badgeley[1], Kipp W Johnson[1], Ben Readhead[1,] Max S. Tomlinson[1], Timothy O'Connor[1,2], Riccardo Miotto[1], Brian A. Kidd[1], Rong Chen[1], Avi Ma'ayan[3] and Joel T. Dudley[1,4*]

[1] Institute of Next Generation Healthcare, Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, Mount Sinai Health System, New York, New York, USA.

[2] Boston College, Morrissey College of Arts and Sciences, Chestnut Hill, MA

[3] Department of Pharmacology and Systems Therapeutics, Icahn School of Medicine at Mount Sinai, Systems Biology Center New York (SBCNY), Mount Sinai Health System, New York, New York, USA.

[4] Department of Population Health Science and Policy, Icahn School of Medicine at Mount Sinai, Mount Sinai Health System, New York, New York, USA.

*Corresponding author: joel.dudley@mssm.edu

Keywords: drug repositioning, translational bioinformatics, drug development, drug discovery

Supplementary Text:

Methods:

RepurposeDB – user interface design and database development:

The backend of the database is developed using the Python based web development framework Flask (http://flask.pocoo.org/). RepurposeDB integrates data visualization and visual analytics tools to let users view, interact and analyze various data elements. These tools enable web-based visualization of small molecules and protein drugs that are repurposed for more than one indication. RepurposeDB provides tools for dynamic visualization of compounds and various chemoinformatics properties. The web interface of RepurposeDB uses JSmol (http://wiki.jmol.org/index.php/JSmol) for visualizing chemical structures of drug compounds and protein drugs, Cytoscape Web and cytoscape.js[1] for visualizing the bipartite network of drug and primary, secondary and orphan indications , and NVD3 (http://nvd3.org/), an extension of d3.js JavaScript library (http://d3js.org/), was used for developing interactive visualization of various physicochemical and chemoinformatics properties. Additional JavaScript libraries including jQuery (https://jquery.com/), Bootstrap (http://getbootstrap.com/), Select2 (https://select2.github.io/), and Selectize (http://brianreavis.github.io/selectize.js/) were used to build the web and responsive mobile device interfaces of RepurposeDB. Biojs (https://github.com/biojs) is used for displaying multiple sequence alignment. Programs for search, display, data aggregation and parsing were developed in Python. Depending on the query and user interaction, dynamic query engine interacts with an nginx (http://nginx.org/) web server in the backend. Annotation data integrated in RepurposeDB is compiled from various databases including DrugBank[2], PubChem BioAssay[3], SIDER[4] , OFFSIDES, TWOSIDES [5] etc. Source code (http://repurposedb.dudleylab.org/code) and the data files (http://repurposedb.dudleylab.org/data) are provided in the public domain.

RepurposeDB – features:

Drug page: The drug page provides several pieces of data about repositioned medications such as the generic name of the medication (Name), a unique medication identifier for RepurposeDB (RxID), DrugBank; FDA; or ChEBI identifiers where applicable (Reference Identifier), primary indication, original indication, and secondary indications, and Drug-Drug interactions and Food-Drug interactions integrated from DrugBank. Additionally available are visualization tools for interacting modeling and bipartite drug repurposing graphs. Chemoinformatics features and annotations for the medications and associated diseases obtained from knowledge mapping to biomedical and healthcare ontologies are easily downloadable. External molecular biology or chemoinformatics databases can be linked to RepurposeDB using the RepurposeDB identifier mapping files.

Disease page: Similar to the drug page, a disease page is indexed unique identifier indicating individual diseases in RepurposeDB (DxID). Each page includes the name of the disease as curated from the original sources (Name), medication name and RxID with visualization tools for visualizing disease-drug network that enables an aggregated visualization of all drugs available for a given diseases, and a bipartite drug-disease network with links to browse the drugs associated with the given disease in RepurposeDB.

Browse: Users can browse RepurposeDB using the list of drugs, diseases or a combined database. Options are also provided to browse using side effects, drug targets and pathways associated with drugs compiled in RepurposeDB. Users can browse the RepurposeDB drugs alphabetically by selecting the "DRUGS" tab from the navigation bar. Selecting a letter from the list shows all RepurposeDB drugs starting with that letter, and clicking on the page numbers shows results not listed on the first page. Clicking on a repositioned drug identifier (RxID, example: Rx000123 is the identifier for idoxuridine) reveals details about the corresponding drug, and clicking on a given indication shows information regarding the indication. Each drug-indication record is linked to relevant reference databases like FDA-

RDRD and PubMed. Users can browse the indications of RepurposeDB drugs alphabetically by selecting the "DISEASE" tab from the navigation bar. Selecting a letter from the list shows all RepurposeDB indications starting with that letter, and clicking on the page number bar to shows results not listed on the first page. In this case, the type of indication is what type of indication the feature is relative to a given drug - for example, progesterone has embryo implantation as a common indication and embryo transfer as a primary indication. Users can get additional details about disease by clicking on a disease identifier (DxID, example: Dx00211 is the identifier for Hypertension) to view details about the corresponding indication, and by clicking on a given drug name to view information regarding it.

Search: RepurposeDB provides tools to search the database using key words, chemical similarity and sequence similarity of the drug targets and protein drugs.

a) ***Key word search****:* Users can search RepurposeDB using keywords against a list of drugs, diseases, drug targets and pathways associated with compounds compiled in RepurposeDB.

b) Chemical similarity search: Chemical similarity search implementation in RepurposeDB uses the Open Babel[6] chemoinformatics file format conversion method to compute a Tanimoto coefficient[7] between a SMILES string given by the user and all the molecules in RepurposeDB. This utility helps drug discovery investigators to quickly search for similarity of new compounds that could be repurposed.

c) Sequence similarity search: A given protein or peptide sequence can be queried against sequence of protein-drugs in RepurposeDB. The Basic Local Alignment Search Tool (BLAST) finds regions of local similarity between sequences. We have implemented our sequence search using BLAST+ to compare protein sequences to sequence databases and compute the statistical significance of matches [8, 9]. BLAST can be used to infer functional and evolutionary relationships using sequence data.

RepurposeDB offers BLASTP-based search utilities to perform sequence search against two sequence databases: a) target-sequence database and b) protein-drug database. This option requires the user to submit a valid protein or peptide sequence in FASTA format (http://www.ncbi.nlm.nih.gov/BLAST/blastcgihelp.shtml) to perform BLAST searches.

Data visualization and visual analytics in RepurposeDB: The drugs, drug target and disease data in RepurposeDB is visualized using various modalities. Atomic resolution structures of small molecules and protein drugs are available as a PNG file and a fully interactive 3D structural model rendering within the web-browser using Jsmol. Converting the SMILES into a 2D and 3D model generates structure files of small molecules in RepurposeDB. The 3D models were then optimized using 500 iterations of local energy minimizations with MMFF94 force field [10]. 3D structure data integrated with Jsmol can be used to map van der Waals surface, molecular surface and can compute various distance calculations. Jsmol can also help users to animate the structure of repurposed drugs. RepurposeDB provides Cystoscape.js (http://js.cytoscape.org/) based interface to visualize the bipartite drug-repurposing graph. The nodes and edges in the bipartite network can be rearranged to generate publication quality figures. RepurposeDB displays various chemoinformatics properties of the drug compounds as scatterplot with dynamic selection of x and y-axes with user defined chemical features. For example users can plot any of the physicochemical features including molecular weight, partition coefficient (logP), solubility (logS), accessible surface area, composition of various functional groups (acidic, basic), composition of atoms, chemical moiety based descriptors (number of bonds, number of rings, number of hydrogen bonds), molar refractivity, geometric properties (diameter, radius, kier shapes, Zagreb group index).

Downloads: Data from RepurposeDB can be downloaded on a drug-basis or as complete tab-delimited files. For an individual drug in RepurposeDB, chemical

structures of the drug molecules are provided in PDB format and SDF format. SDF These files could help in structural studies including visualization and protein-ligand docking or virtual screening experiments. Ontologies mapped using the annotations are also provided to enable enrichment analysis. Bipartite drug-disease networks with demarcated primary and secondary indications are available from each drug page.

List of ADMET Properties:

The following ADMET properties are tested: Human Intestinal Absorption, Blood Brain Barrier, Caco-2 permeable, P-glycoprotein substrate, P-glycoprotein inhibitor I, P-glycoprotein inhibitor II, Renal organic cation transporter, CYP450 2C9 substrate, CYP450 2D6 substrate, CYP450 3A4 substrate, CYP450 1A2 substrate, CYP450 2C19 substrate, CYP450 inhibitory promiscuity, Ames test, Carcinogenicity, Biodegradation, Rat acute toxicity, hERG inhibition (predictor I) and hERG inhibition (predictor II) (Also see Supplemental Table S3)

List of enrichment annotations compiled from DAVID:

GDB, GDB_CLASS, PIR_SEQ_FEATURE, SP_PIR_KEYWORDS, UP_SEQ_FEATURE, PIR_SUMMARY, SP_COMMENT, INTERPRO, PANTHER_FAMILY, PANTHER_SUBFAMILY, PIR_SUPERFAMILY, PROSITE, SMART, SCOP_CLASS, SCOP_FAMILY, SCOP_FOLD, SCOP_SF, GENERIF_SUMMARY, PUBMED_ID, SCOP_FAMILY, SCOP_FOLD, SCOP_SF, UCSC_TFBS, UP_TISSUE (Also See Supplementary Data: RepurposeDB_DAVID.xlsx)

List of enrichment annotations compiled from Enrichr:

Allen_Brain_Atlas_up_table, Cancer_Cell_Line_Encyclopedia_table, CORUM_table, ChEA_table, ENCODE_Histone_Modifications_2015_table, ENCODE_TF_ChIP-seq_2015_table, ESCAPE_table, GO_Biological_Process_table, GO_Cellular_Component_table, GO_Molecular_Function_table, GeneSigDB_table, Genes_Associated_with_NIH_Grants_table, HMDB_Metabolites_table, HomoloGene_table, KEGG_2015_table, Kinase_Perturbations_from_GEO_table,

MGI_Mammalian_Phenotype_Level_3_table, MGI_Mammalian_Phenotype_Level_4_table, MSigDB_Computational_table, PPI_Hub_Proteins_table, Pfam_InterPro_Domains_table, Reactome_2015_table, TF-LOF_Expression_from_GEO_table, TRANSFAC_and_JASPAR_PWMs_table, Transcription_Factor_PPIs_table, WikiPathways_2015_table; (Also See Supplementary Data: RepurposeDB_Enrichr.xlsx)

Limitations

Related resources: It should be noted that RepurposeDB is not the first database on drug repositioning and related knowledge corpus. For example, PROMISCUOUS is a database of network-based drug repositioning and offers some of the similar functionalities of RepurposeDB [11]. The OMICtools database[12] lists 24 tools as of Oct 2016 have some level of data on drug repositioning or related methodologies (https://omictools.com/drug-repositioning-category). Compared to these tools, RepurposeDB is the only resource that encourages a community-based aggregation drug repositioning knowledge and also provides factors driving drug repositioning from systematic analyses of drugs, diseases, and drug targets.

Limitations of text mining and biocuration workflows: The current version of RepurposeDB relies on biocuration of publicly available biomedical databases like PubMed and FDA databases. Multiple indications of drugs are also available for various public (PubMed, ClinicalTrial.gov, DrugBank) and private resources (internal database of failed or shelved compounds at pharmaceutical companies, commercial databases like CiteLine) with or without associated research publications. A complete compendium of repositioning investigations should include all the indications, and we hope future iterations of RepurposeDB will add more multi-indication datasets. The quality of the RepurposeDB content depends on the primary source including literature and various primary databases mined as part of accumulating the initial list of the drugs and disease indications. Potential users should note that text mining data might have false positive associations and we may have missed some of the known entries as part of the process [13-15].

Definition of primary indication and secondary indication also depends on the quality of biocuration from primary data sources [16, 17]. In this current version of RepurposeDB, results from text mining or biocuration are not cross-validated using inter-observer variability measurements. Thus, we strongly recommend additional physician review before considering the data from RepurposeDB to use in the clinical setting. Potential clinical use of the resource would only encourage after completing a comprehensive comparative analyses clinical trial that would potentially compare RepurposeDB with the current standard of care in the clinical setting.

Limitations in biochemical inference based on enrichment analyses: From a biochemical knowledge and inference perspective, repurposed drugs are better studied than other compounds, and the chance of repurposed drug targets acting as a mediator of a pathway crosstalk is biased toward being higher. The subjective bias introduced by the repeatedly studied genes, proteins or pathways are a known limitation of biological enrichment and should be considered as a potential factor while extrapolating results from our study to other biochemical or medicinal chemistry inferences. Annotations are also subject to evolve overtime and difference in results could be obtained depending on the version, release and content of the annotation database used in the study [18-21].

Limitations in the definitions of primary and secondary indications: It should be noted that primary indication is not constant, for example a drug approved for indication may have post-market surveillance based impact including recall for side ffects, toxicity or other adverse events [22-26]. Thus the primary indication may or may not be retained in the primary database we have referenced over a period of time. Users of RepurposeDB should be aware of such market dynamics and encouraged to use the most recent version of the database in conjunction with the latest FDA and other literature reports.

Supplementary Figures:

Figure S1: Sweet-spot (correlation coefficient of MW and LogP) of druggability in DrugBank-F, DrugBank-A, subset of CGEA drugs and small molecules in RepurposeDB

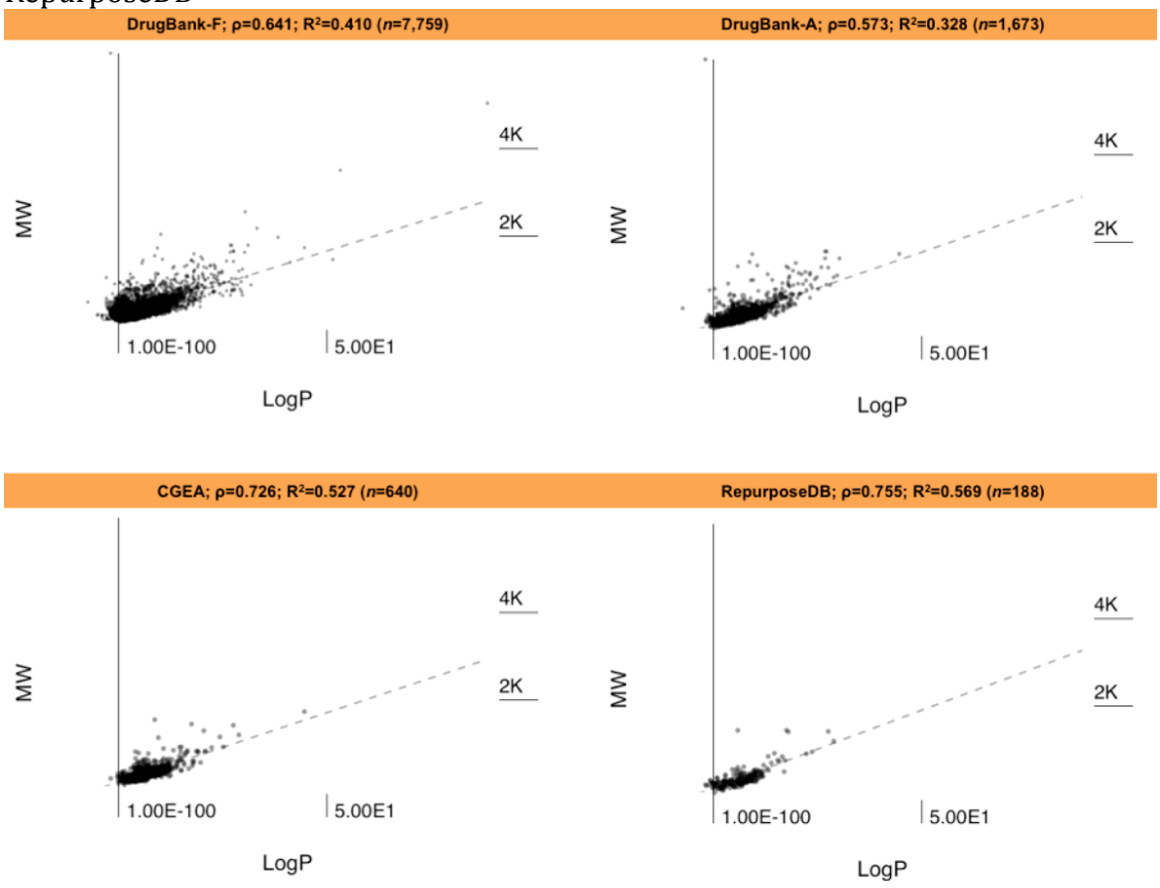
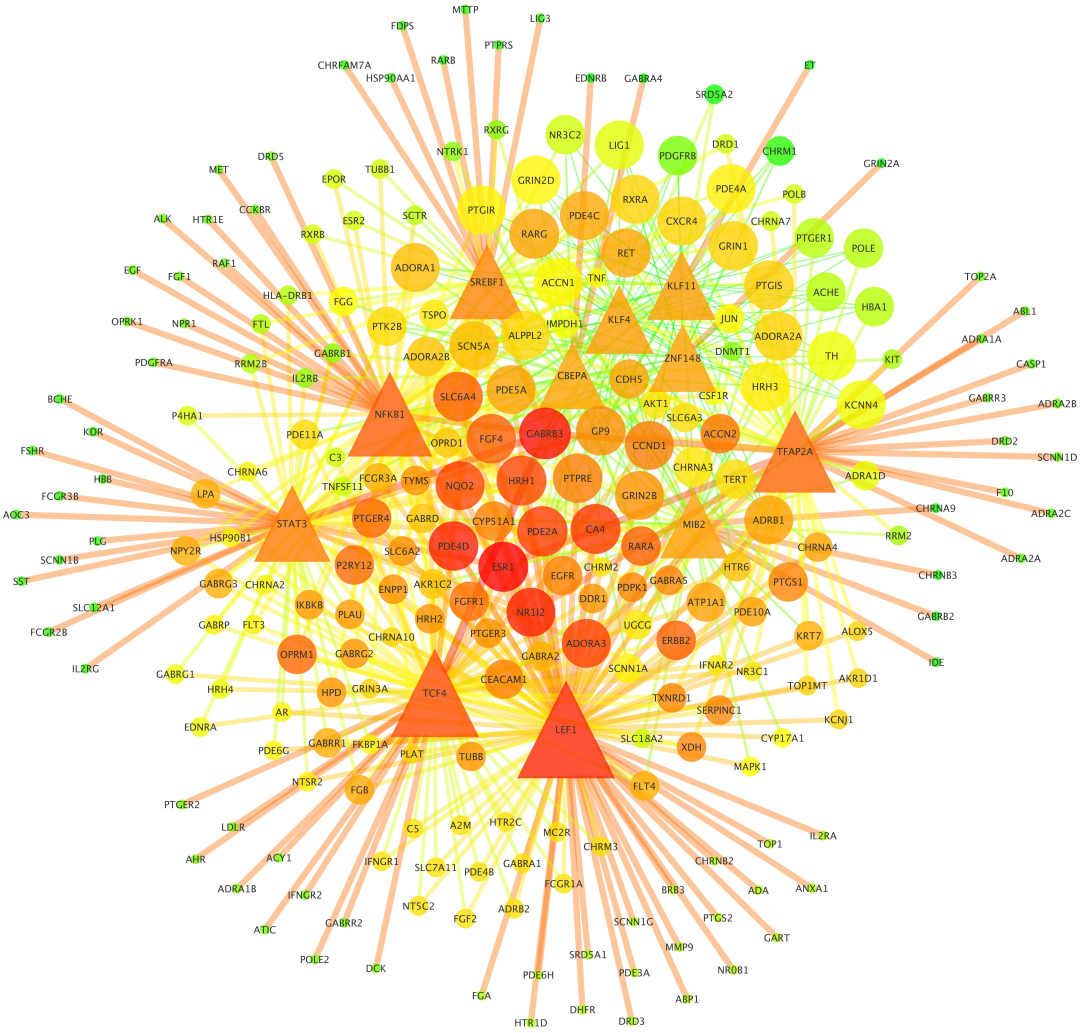


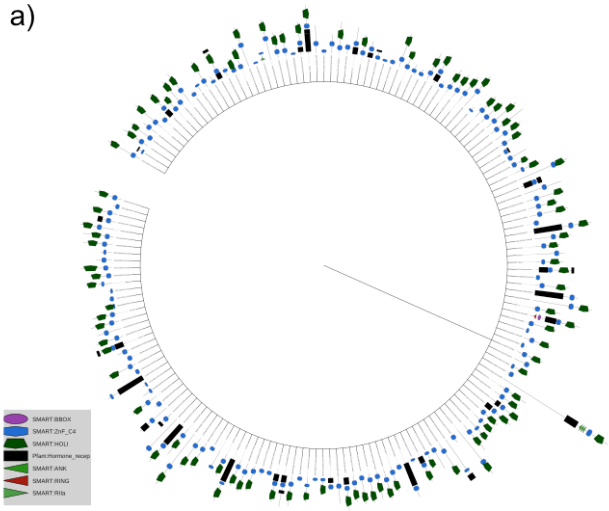
Figure S2: Network of transcription factors regulating drug targets in RepurposeDB



Nodes are genes (Δ=enriched transcription factors associated with drugs targets using Enrichr; O=drug targets from RepurposeDB). Color indicate the edge betweenness (red to green) and the size of the shows the degree of the nodes.

Figure S3: Domain architectures, phylogenetic representation and pathway enrichment of proteins with a) HOLI domains and b) HDc domains

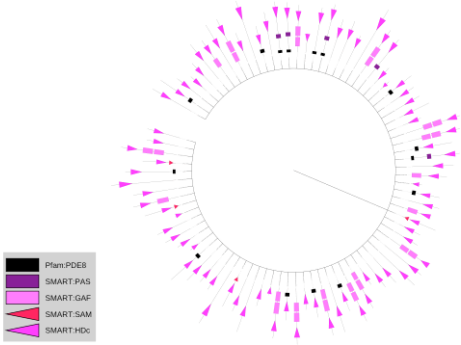
a)



% proteins involved	Description
20.21	PPAR signaling pathway
11.70	Adipocytokine signaling pathway
11.35	Thyroid cancer
10.99	Small cell lung cancer
10.99	Non-small cell lung cancer
8.87	Maturity onset diabetes of the young
8.16	Neuroactive ligand-receptor interaction
4.96	Acute myeloid leukemia
3.19	Wnt signaling pathway
2.84	MAPK signaling pathway
2.84	Prostate cancer
2.48	Progesterone-mediated oocyte maturation
1.42	Circadian rhythm

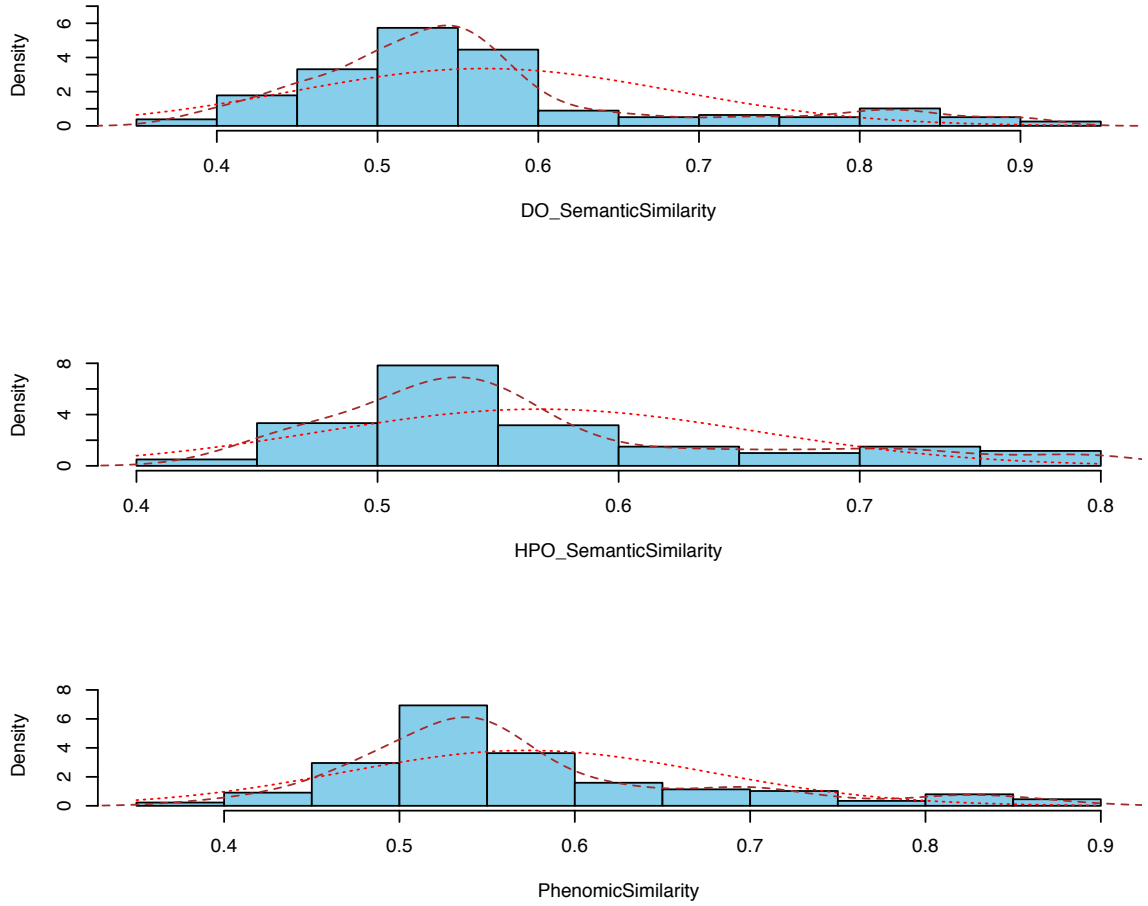
b)

% proteins involved	Description
63.24	Purine metabolism
21.08	Two-component system - General
8.39	Progesterone-mediated oocyte maturation
2.74	Nicotinate and nicotinamide metabolism
1.72	Pyrimidine metabolism
0.39	Glycine, serine and threonine metabolism
0.39	Aminoacyl-tRNA biosynthesis
0.24	Aminosugars metabolism
0.24	Riboflavin metabolism
0.24	Thiamine metabolism
0.24	Fructose and mannose metabolism
0.16	Atrazine degradation
0.08	Tyrosine metabolism
0.08	Glutamate metabolism
0.08	Limonene and pinene degradation
0.08	Butanoate metabolism
0.08	Pyruvate metabolism
0.08	1- and 2-Methylnaphthalene degradation
0.08	Bile acid biosynthesis
0.08	D-Glutamine and D-glutamate metabolism
0.08	Arginine and proline metabolism
0.08	Cyanoamino acid metabolism
0.08	Benzoate degradation via CoA ligation
0.08	Benzoate degradation via hydroxylation



Pathway analyses using human proteome as background and analyses and visualization generated using data from SMART database (http://smart.embl-heidelberg.de/). Pathways mediated by the proteins are shown in the tables using KEGG annotations. Drug molecules hitting these proteins may potentially, highly repurposable. Some of these are function-naïve proteins with uncharacterized experimental function suggests potential targets for target-driven drug repurposing.

Figure S4: Semantic similarity of diseases associated with repositioned drugs



Dotted lines indicate the normal distribution and dashed lines are density estimation. Top histogram is plotted using the semantic similarity scores using Disease Ontology (n=120 drugs); Middle histogram is plotted using the semantic similarity scores using Human Phenotype Ontology (n=157), and the bottom histogram is combined score across the two ontologies (n=176): where the scores are averaged across both ontologies if available, else the singleton scores are considered.

Figure S4: Drug-target interaction network of drugs and target proteins in RepurposeDB

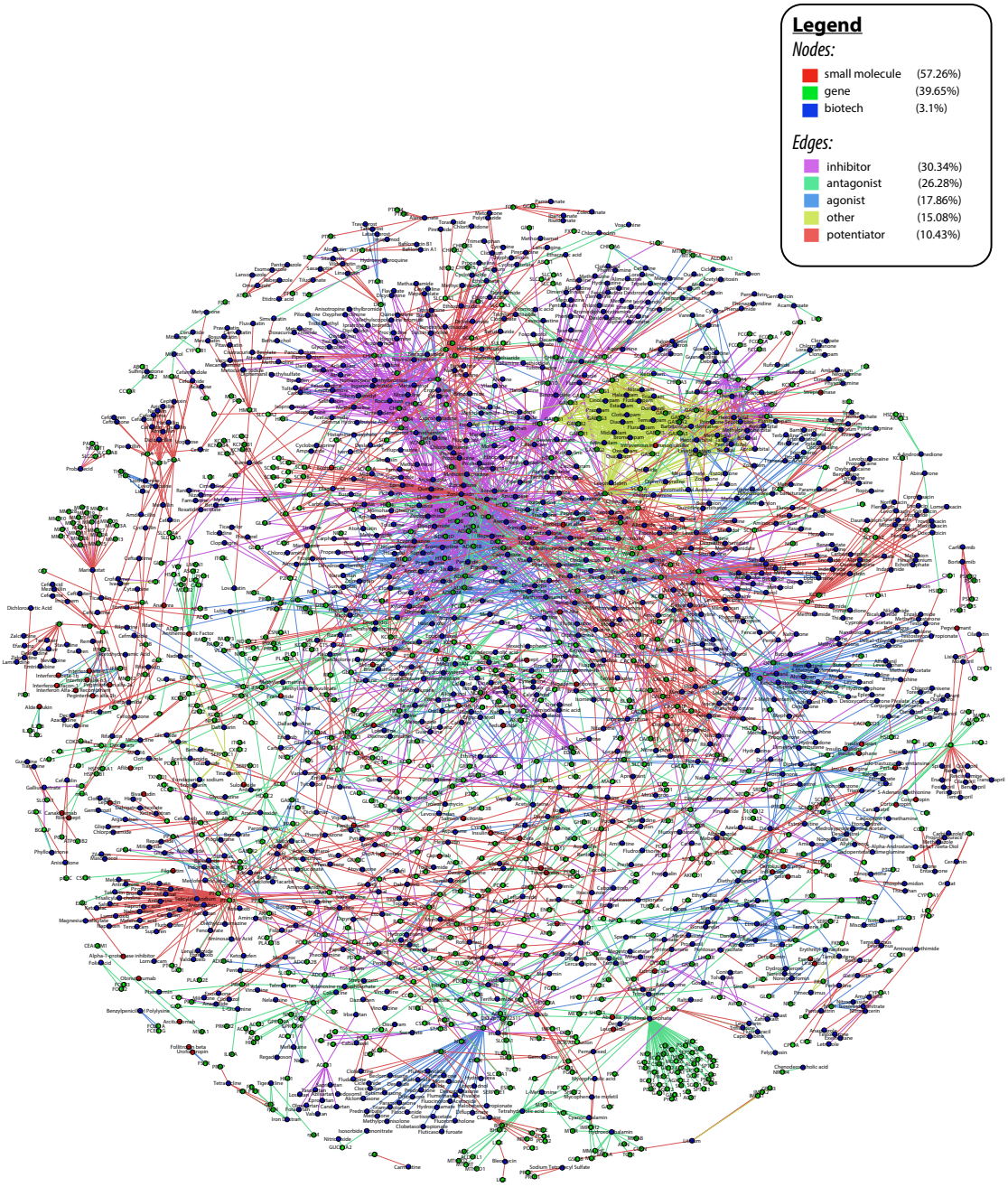
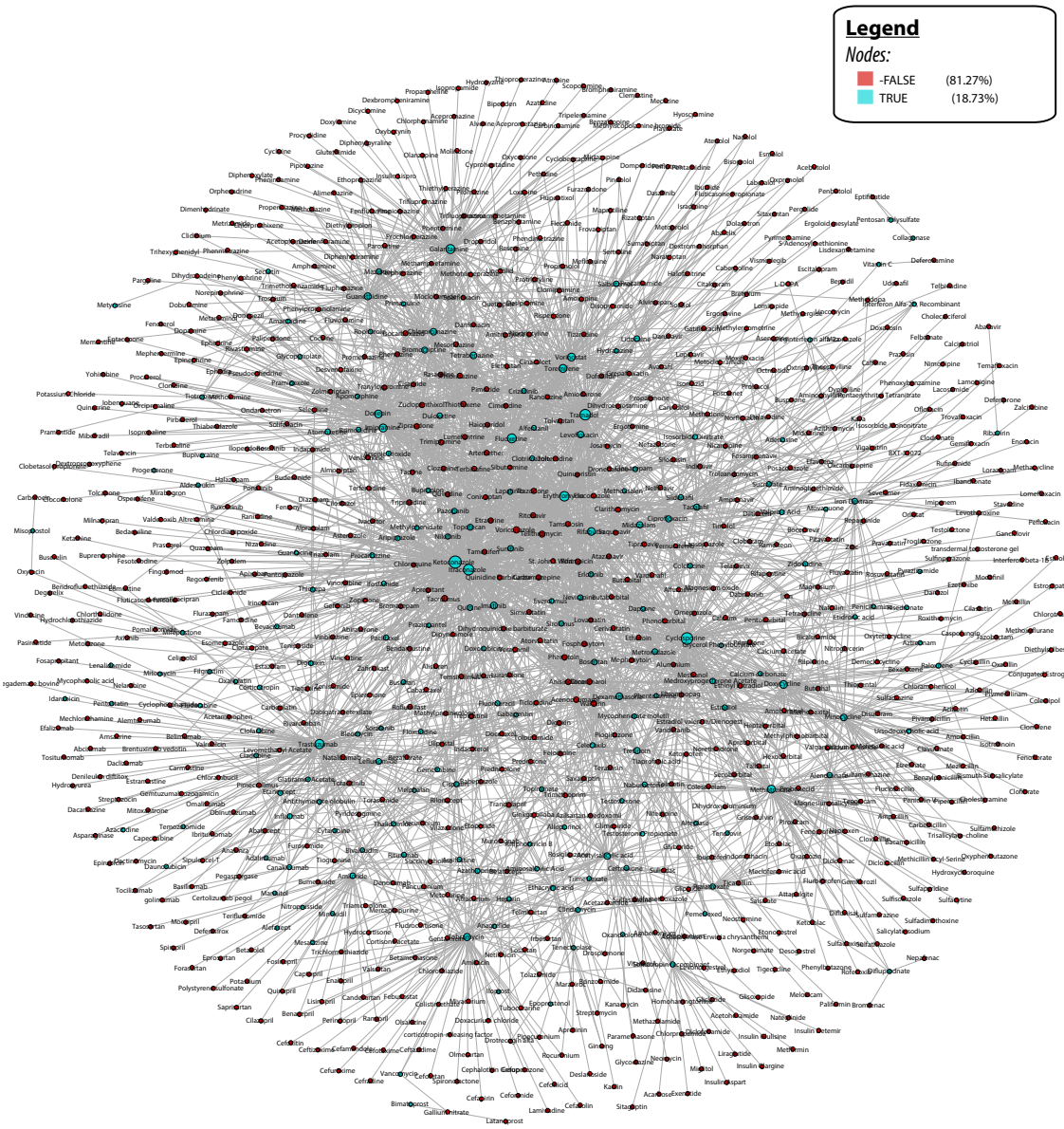


Figure S5: Drug-drug interaction network compiled from RepurposeDB



Blue nodes are drugs with evidence for drug repositioning and present in RepurposeDB. Red nodes doesn't have evidence for repositioning and are not available in RepurposeDB

Figure S6: Seed functional network reconstructed using targets of repurposed drugs in RepurposeDB

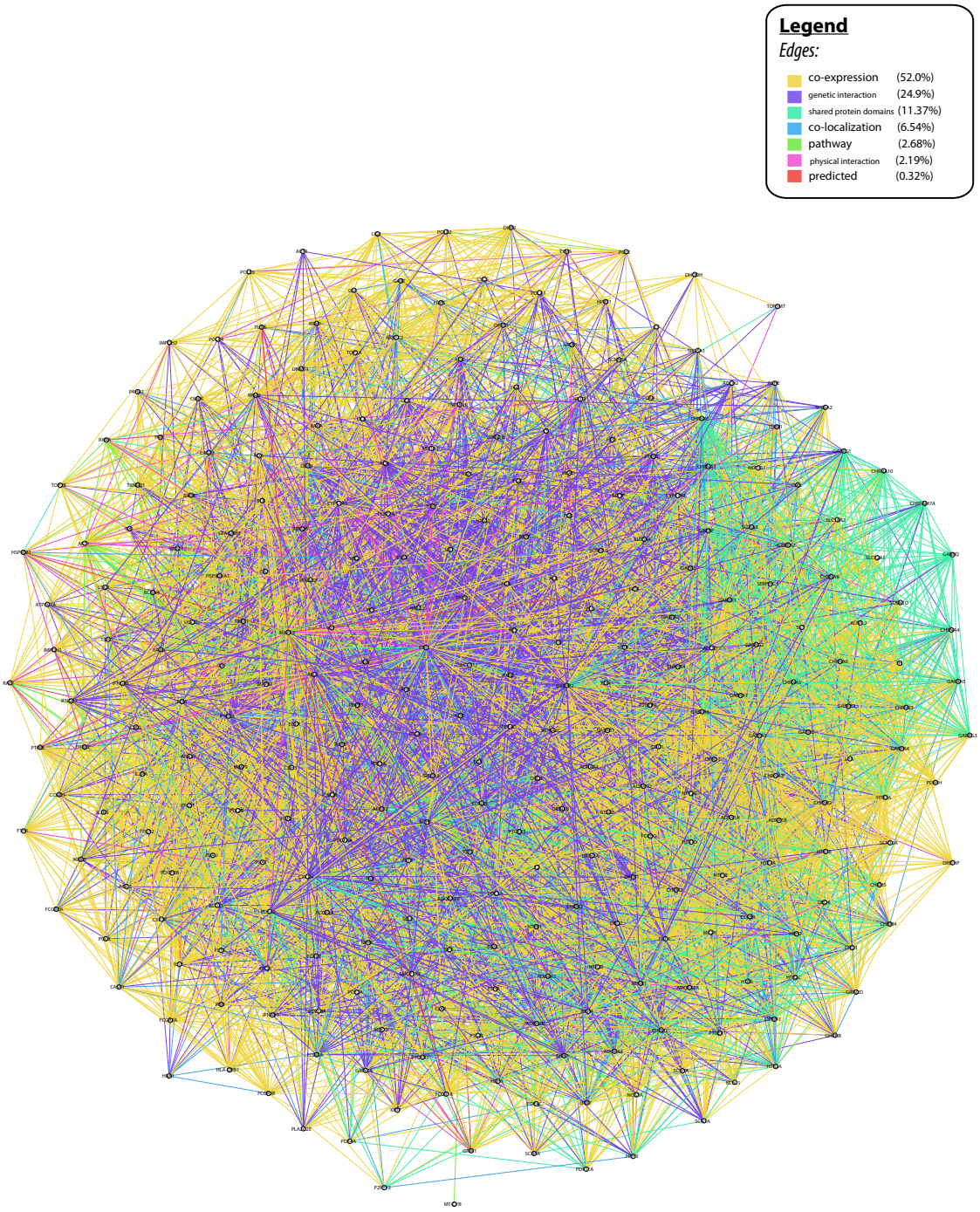
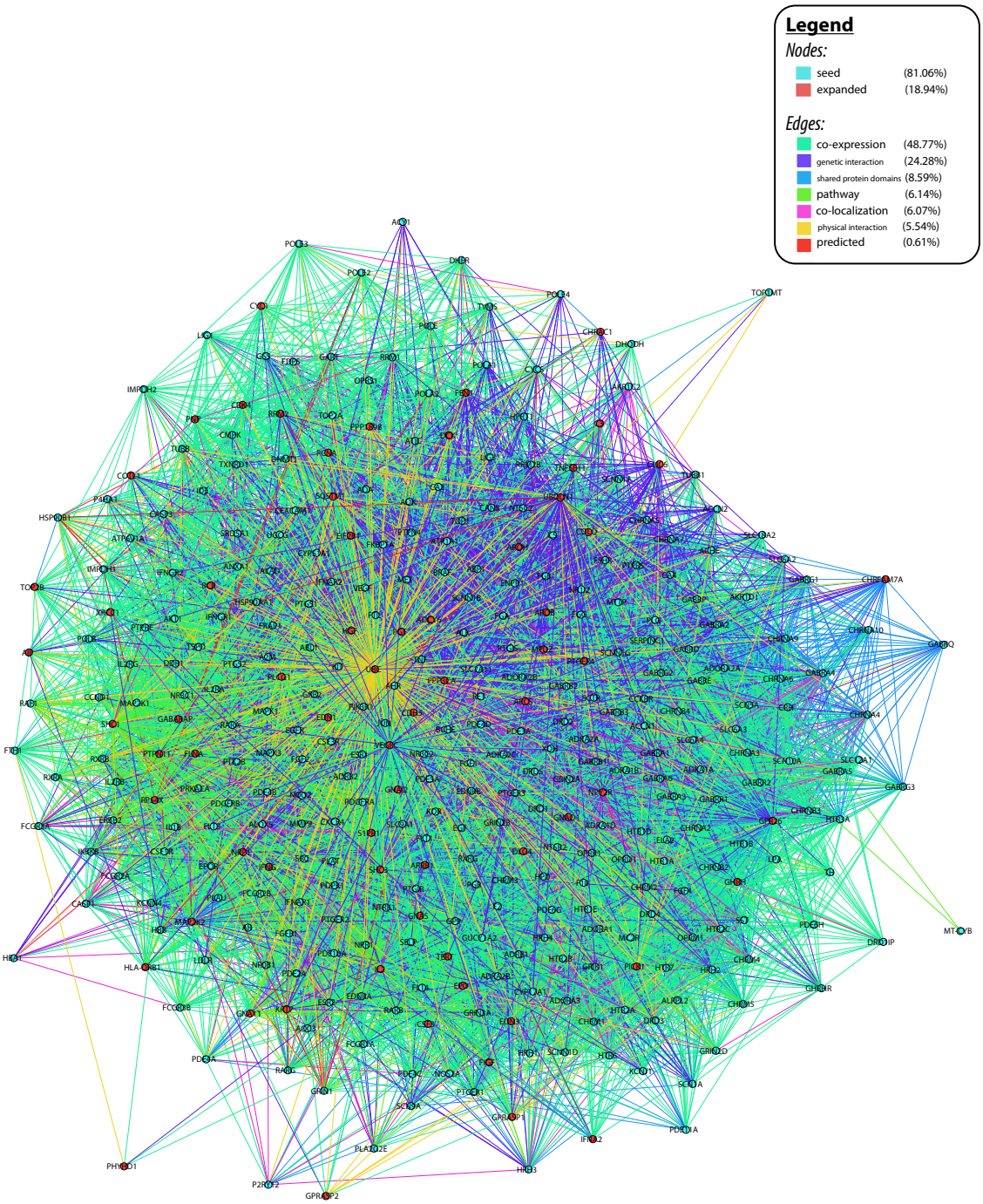


Figure S7: Expanded functional network reconstructed using targets of repurposed drugs in RepurposeDB and their network neighbors



Supplementary Tables:

Table S1: Fields to fill during a new submission to RepurposeDB based on MIADRI guidelines

Field	Description
Generic name of the medication	A common chemical name or the brand name of the repositioned compound or protein-drug
Compound identifiers	RepurposeDB accepts a variety of compound identifiers as part of MIADRI standardization including identifiers from DrugBank, CheBI, PubChem and FDA.
FDA or other regulatory approval status	A link should be to provide to approval status of the drug for rare or common disease indications
Primary indication(s)	Disease name or use the human phenotype ontology (example: HP:0001658 is identifier for Myocardial infarction) or disease ontology (example: DOID: 10264 is identified for mumps) terms can be used to define the primary disease for which the drug is approved.
Secondary indication(s)	Standard clinical terms, HPO terms or DO terms can be used to define the secondary disease for which the drug is repositioned in the current study.
Evidence type	Computational or experimental strategy used to conduct the drug repositioning investigations can be selected here.
Validation	Evidence from validation experiments using clinical trials or EHR-based validation can be provided. If a clinical trial is in progress or submitted, users should provide the ClinicalTrials.gov identifiers.
References	All associated publications can be submitted using the PubMed identifiers (PMIDs)
Additional information	Users can upload additional information that needs to be considered for inclusion in RepurposeDB. The intuitive submission interface can be used to upload documents in PDF format.
Contact details	Users should also submit their name and email address as part of the submission.

Table S2: Statistically significant chemical descriptors associated with repositioned compounds

Descriptors	DL	DrugBank-F	DrugBank-A	RepurposeDB	*P**
Fraction of rotatable bonds in a compound	JL	0.24	0.216	0.198	0.016
Molar refractivity of a compound	JL	110	122.958	116.407	0.005
Molar refractivity of a compound	OB	91.02	101.275	95.934	0.008
Number of acidic groups	JL	0.63	0.381	0.39	0.001
Number of atoms	JL	44.314	49.59	47.829	0.005
Number of basic groups in a compound -	JL	0.301	0.512	0.428	<0.001
Number of bonds in a compound	JL	45.734	51.712	49.481	0.005
Number of carbon atoms in a compound (C)	CM	16.316	18.713	17.332	<0.001
Number of derivatives of acids by replacing a OH group in a compound by OR group (RCOOR)	CM	0.143	0.251	0.198	<0.001
Number of heavy bonds in a compound	JL	25.31	27.681	26.775	0.036
Number of hydrogen atoms in a compound (H)	CM	20.709	24.226	22.909	<0.001
Number of hydrogen bond acceptors (HBA) 1 in a compound	JL	25.828	29.302	28.636	0.009
Number of hydrogen bond Acceptors 1 in a compound (HBA1)	OB	26.961	30.278	29.567	0.014
Number of hydrogen bond donors in a compound (HBD)	OB	2.78	2.306	2.701	0.022
Number of ketone group in a compound (RCOR)	CM	0.128	0.21	0.299	0.007
Number of organo-phosphates or esters of phosphoric acid in a compound (ROPO3)	CM	0.123	0.008	0.016	0.001
Number of phosphorus atoms In a compound	JL	0.163	0.03	0.07	<0.001
Number of phosphorus atoms In a compound	CM	0.163	0.03	0.07	<0.001
Number of single bonds (sbonds)	OB	35.269	41.332	39.947	0.002
Number of tertiary amines in a compound where all 3H - atoms are replaced by alkyl groups (R3N)	CM	0.416	0.639	0.52	<0.001
octanol/water partition coefficient of a compound (LogP)	JL	6.152	7.516	6.714	<0.001
octanol/water partition coefficient of a compound (logP)	OB	2.207	2.651	2.162	0.02
Topological descriptor Zagreb_group_index_1	JL	232.743	265.698	254.193	0.003
Topological descriptor Zagreb_group_index_2	JL	251.096	286.577	274.904	0.004
Total atoms in a compound (atoms)	OB	44.599	50.155	48.032	0.006
Total bonds in a compound (bonds)	OB	46.019	51.908	49.684	0.005

DL=Descriptors derived from the library (CM=Chemminer, JL=JOELib and OB=OpenBabel)
DrugBank-F = DrugBank full; DrugBank-A = Approved subset of DrugBank
*Two-way ANOVA of features across presence in RepurposeDB and approval status

Table S3: Adsorption, distribution, metabolism, excretion and toxicity (ADMET) properties of repositioned compounds

ADMET properties	***P****
Caco-2 permeable	<0.001
P-glycoprotein substrate	<0.001
P-glycoprotein inhibitor I	<0.001
Renal organic cation transporter	0.001
CYP450 2D6 substrate	0.001
Rat acute toxicity	0.005
hERG inhibition (predictor II)	0.006
P-glycoprotein inhibitor II	0.009
Human Intestinal Absorption	0.033
Blood Brain Barrier	>0.05
Ames test	>0.05
Biodegradation	>0.05
CYP450 3A4 substrate	>0.05
Carcinogenicity	>0.05
CYP450 2C19 substrate	>0.05
hERG inhibition (predictor I)	>0.05
CYP450 2C9 substrate	>0.05
CYP450 1A2 substrate	>0.05
CYP450 inhibitory promiscuity	>0.05

Table S4: Methylation patterns of repositioned drug targets

Term	Adjusted *P*	Genes
H3K27me3_cardiac mesoderm_hg19	4.97733E-23	104
H3K27me3_bronchial epithelial cell_hg19	3.81707E-22	78
H3K27me3_H7_hg19	6.36977E-17	68
H3K27me3_kidney epithelial cell_hg19	1.47422E-17	73
H3K27me3_mammary epithelial cell_hg19	1.71711E-17	79
H3K27me3_SK-N-SH_hg19	8.9743E-17	70
H3K27me3_keratinocyte_hg19	1.38397E-15	69
H3K27me3_fibroblast of lung_hg19	8.78499E-16	83
H3K27me3_astrocyte_hg19	1.38397E-15	65
H3K27me3_endothelial cell of umbilical vein_hg19	2.39351E-15	87
H3K27me3_CD14-positive monocyte_hg19	1.71335E-14	86
H3K27me3_fibroblast of dermis_hg19	1.92003E-13	61
H3K9me3_CD14-positive monocyte_hg19	6.36996E-13	60
H3K27me3_BJ_hg19	1.13086E-12	61
H3K27me3_skeletal muscle myoblast_hg19	6.22588E-11	56
H3K27me3_GM12878_hg19	4.78482E-11	75
H3K27me3_mononuclear cell_hg19	1.21605E-08	51
H3K27me3_osteoblast_hg19	3.05851E-08	50
H3K27me3_H1-hESC_hg19	1.21605E-08	51
H3K27me3_K562_hg19	3.92198E-07	66
H3K27me3_A549_hg19	1.04185E-06	59
H3K27me3_NT2-D1_hg19	1.32445E-06	46
H3K27me3_MCF-7_hg19	8.30108E-05	26
H3K9me3_skeletal muscle myoblast_hg19	4.22642E-05	42
H3K4me1_NT2-D1_hg19	8.63418E-05	41
H4K20me1_H1-hESC_hg19	8.63418E-05	41

*Adjusted *P*-values from Enrichr

Supplementary References:

1. Ono, K., Demchak, B. & Ideker, T. Cytoscape tools for the web age: D3.js and Cytoscape.js exporters. *F1000Res* **3**, 143 (2014).
2. Law, V. et al. DrugBank 4.0: shedding new light on drug metabolism. *Nucleic acids research* **42**, D1091-1097 (2014).
3. Wang, Y. et al. PubChem BioAssay: 2014 update. *Nucleic acids research* **42**, D1075-1082 (2014).
4. Kuhn, M., Campillos, M., Letunic, I., Jensen, L.J. & Bork, P. A side effect resource to capture phenotypic effects of drugs. *Molecular systems biology* **6**, 343 (2010).
5. Tatonetti, N.P., Ye, P.P., Daneshjou, R. & Altman, R.B. Data-driven prediction of drug effects and interactions. *Science translational medicine* **4**, 125ra131 (2012).
6. O'Boyle, N.M. et al. Open Babel: An open chemical toolbox. *Journal of cheminformatics* **3**, 33 (2011).
7. Rogers, D.J. & Tanimoto, T.T. A Computer Program for Classifying Plants. *Science* **132**, 1115-1118 (1960).
8. Altschul, S.F. et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research* **25**, 3389-3402 (1997).
9. Camacho, C. et al. BLAST+: architecture and applications. *BMC bioinformatics* **10**, 421 (2009).
10. Cheng, A., Best, S.A., Merz, K.M., Jr. & Reynolds, C.H. GB/SA water model for the Merck molecular force field (MMFF). *Journal of molecular graphics & modelling* **18**, 273-282 (2000).
11. von Eichborn, J. et al. PROMISCUOUS: a database for network-based drug-repositioning. *Nucleic acids research* **39**, D1060-1066 (2011).
12. Henry, V.J., Bandrowski, A.E., Pepin, A.S., Gonzalez, B.J. & Desfeux, A. OMICtools: an informative directory for multi-omic data analysis. *Database : the journal of biological databases and curation* **2014** (2014).
13. Huang, C.C. & Lu, Z. Community challenges in biomedical text mining over 10 years: success, failure and the future. *Briefings in bioinformatics* **17**, 132-144 (2016).
14. Harpaz, R. et al. Text mining for adverse drug events: the promise, challenges, and state of the art. *Drug Saf* **37**, 777-790 (2014).
15. Oda, K. et al. New challenges for text mining: mapping between text and manually curated pathways. *BMC bioinformatics* **9 Suppl 3**, S5 (2008).
16. Gaudet, P. et al. Recent advances in biocuration: meeting report from the Fifth International Biocuration Conference. *Database : the journal of biological databases and curation* **2012**, bas036 (2012).
17. Howe, D. et al. Big data: The future of biocuration. *Nature* **455**, 47-50 (2008).
18. Huntley, R.P., Sawford, T., Martin, M.J. & O'Donovan, C. Understanding how and why the Gene Ontology and its annotations evolve: the GO within UniProt. *Gigascience* **3**, 4 (2014).

19. Draghici, S. et al. A systems biology approach for pathway level analysis. *Genome Res* **17**, 1537-1545 (2007).
20. Khatri, P., Sirota, M. & Butte, A.J. Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS computational biology* **8**, e1002375 (2012).
21. Rhee, S.Y., Wood, V., Dolinski, K. & Draghici, S. Use and misuse of the gene ontology annotations. *Nature reviews. Genetics* **9**, 509-515 (2008).
22. Samp, J.C., Schumock, G.T. & Pickard, A.S. Retracted publications in the drug literature. *Pharmacotherapy* **32**, 586-595 (2012).
23. Freifeld, C.C. et al. Digital drug safety surveillance: monitoring pharmaceutical products in twitter. *Drug Saf* **37**, 343-350 (2014).
24. Kesselheim, A.S. & Gagne, J.J. Strategies for postmarketing surveillance of drugs for rare diseases. *Clinical pharmacology and therapeutics* **95**, 265-268 (2014).
25. Lorberbaum, T. et al. Systems pharmacology augments drug safety surveillance. *Clinical pharmacology and therapeutics* **97**, 151-158 (2015).
26. Nagaich, U. & Sadhna, D. Drug recall: An incubus for pharmaceutical companies and most serious drug recall of history. *Int J Pharm Investig* **5**, 13-19 (2015).