Published online 29 November 2022 *Nucleic Acids Research, 2023, Vol. 51, Database issue* ***D1405–D1416***
https://doi.org/10.1093/nar/gkac1033

Pharos 2023: an integrated resource for the understudied human proteome

Keith J. Kelleher [1], Timothy K. Sheils [1], Stephen L. Mathias [2], Jeremy J. Yang [2], Vincent T. Metzger [2], Vishal B. Siramshetty [1], Dac-Trung Nguyen [1], Lars Juhl Jensen [3], Dušica Vidović [4,5], Stephan C. Schürer [4,5,6], Jayme Holmes [2], Karlie R. Sharma [1], Ajay Pillai [1], Cristian G. Bologa [2], Jeremy S. Edwards [2,7,*], Ewy A. Mathé [1,*] and Tudor I. Oprea [2]

[1]National Center for Advancing Translational Science, 9800 Medical Center Drive, Rockville, MD 20850, USA, [2]Translational Informatics Division, Department of Internal Medicine, University of New Mexico Health Sciences Center, Albuquerque, NM 87131, USA, [3]Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen 2200, Copenhagen, Denmark, [4]Institute for Data Science and Computing, University of Miami, Coral Gables, FL 33146, USA, [5]Department of Molecular and Cellular Pharmacology, Miller School of Medicine, University of Miami, Miami, FL 33136, USA, [6]Sylvester Comprehensive Cancer Center, Miller School of Medicine, University of Miami, Miami, FL 33136, USA and [7]Department of Chemistry and Chemical Biology, University of New Mexico, Albuquerque, NM 87131, USA

Received September 15, 2022; Revised October 12, 2022; Editorial Decision October 14, 2022; Accepted November 28, 2022

ABSTRACT

The Illuminating the Druggable Genome (IDG) project aims to improve our understanding of understudied proteins and our ability to study them in the context of disease biology by perturbing them with small molecules, biologics, or other therapeutic modalities. *Two main products from the IDG effort are the* Target Central Resource Database (TCRD) (http://juniper.health.unm.edu/tcrd/), which curates and aggregates information, and Pharos (https://pharos.nih.gov/), a web interface for fusers to extract and visualize data from TCRD. Since the 2021 release, TCRD/Pharos has focused on developing visualization and analysis tools that help reveal higher-level patterns in the underlying data. The current iterations of TCRD and Pharos enable users to perform enrichment calculations based on subsets of targets, diseases, or ligands and to create interactive heat maps and UpSet charts of many types of annotations. Using several examples, we show how to address disease biology and drug discovery questions through enrichment calculations and UpSet charts.

INTRODUCTION

Biomedical research tends to be dominated by a relatively small number of proteins. By some estimates, only 10% of human proteins receive 75% of research interest (1). One analysis of biomedical research literature and results found that this bias was largely driven by early findings and experimentation conducted in the 1980s and 1990s, rather than the physiological, clinical, or biological relevance of the individual genes (2). To address this bias and incentivize research into understudied proteins, the National Institutes of Health (NIH) initiated the Illuminating the Druggable Genome (IDG) project in 2014. Two main products of the IDG project are the Target Central Resource Database (TCRD) and Pharos, the public, web-accessible interface to the database. TCRD aggregates data from 79 sources and harmonizes the many different (often disjoint) identifiers that the data sources utilize for targets (proteins), diseases, and ligands. Details on the data sources and the processing they undergo are available in previous editions (3,4).

Another outcome for the IDG program is the definition of the Target Development Level (TDL), an annotation that classifies targets based on the amount of data available for them. Very briefly, the TDL is one of four potential values: Tclin, Tchem, Tbio or Tdark. Tclin are targets for which an approved drug exists (5,6), which currently includes 704 human proteins; Tchem are proteins that are not Tclin, but are known to bind small molecules with high potency (currently $N = 1971$); Tbio includes proteins that have Gene Ontology (7) leaf term annotations based on experimental evidence; or meet two of the following three conditions: A fractional publication count (8) >5, three or more Gene RIF, 'Reference Into Function' annotations (https:

*To whom correspondence should be addressed. Tel: +1 301 402 8953; Email: ewy.mathe@nih.gov
Correspondence may also be addressed to Jeremy Edwards. Tel: +1 505 277 6655; Email: jsedward@unm.edu

Figure 1. TDL changes over time • Chart of the TDL changes between an early version (TCRD v3.0), a version at the time of the last update publication (TCRD v6.7), and the current version (TCRD 6.13). The decrease of Tdark and subsequent increase of other development levels shows an overall increase in target illumination.

//www.ncbi.nlm.nih.gov/gene/about-generif), or 50 or more commercial antibodies, as counted in the Antibodypedia portal (9). The fourth category, 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. Figure 1 shows the TDL count changes between version 3 and 6.7 and 6.13.

Since the 2017 NAR Database issue paper (3), additional efforts in development related to TCRD/Pharos have been reported, which include two protocols papers (10,11) and a NAR Database issue update paper (4). Pharos usage now averages 3300 unique users a month, while downloads of the TCRD database average 140 per month. Pharos introduced versioned releases in 2020 and has reached version 3.14.1, which utilizes the latest TCRD version 6.13.4. Pharos typically links to external sites for more information, and to TCRD's primary data sources, when a website is available. Many other sites link to Pharos as well, including: PDB (12), ChEMBL (13), DISEASES (8), DrugCentral (14), MARRVEL (15), Reactome (16), KEGG (17), Guide to Pharmacology (18), GlyGen (19), UniProt (20) and more. Data is accessible by full TCRD download (http://juniper.health.unm.edu/tcrd/), or via the GraphQL API (https://pharos-api.ncats.io/graphql). See https://pharos.nih.gov/api to access the interactive GraphQL IDE that includes a set of example queries to get started.

In the current paper, we describe changes implemented since the 2021 NAR database update paper, such as new and streamlined data sources integrated into TCRD, and new analysis and visualization options available in Pharos. We also present in tabular form several use cases for scientific questions that can be asked of the database, using new list analysis features.

MATERIALS AND METHODS

Throughout this paper, the term 'target' refers to a 'gene or protein of interest', as sometimes attributes are related to genes (e.g. orthologs), and sometimes to proteins. However, TCRD is based on 'reviewed' (manually curated) human protein entries from UniProt (20). We have previously (3,4) described the data aggregation and integration process, as it relates to targets, diseases and ligands. The term 'ligands' includes small molecule and biologic drugs, as well as other therapeutic modalities and is not limited to small molecules that modulate proteins.

New data sources

The newly added data includes disease definitions, hierarchy, and mappings provided by the Mondo Disease Ontology (Mondo) (21,22). This includes synonyms for equivalent terms from 24 ontologies, such as Disease Ontology (DO) (23), Online Mendelian Inheritance in Man (OMIM) (24), Orphanet (25), Unified Medical Language System (UMLS) (26), Genetic and Rare Diseases Information Center (GARD) (21), etc. This allows Pharos to resolve many types of disease IDs to the appropriate details pages, using the link format: https://pharos.nih.gov/diseases/ontology_id. The Mondo dataset also includes hierarchical information about the relationships between diseases. Pharos uses this information to aggregate disease associations, in order to show associations for directly associated diseases, as well as all descendent diseases, as exemplified later in the Results section.

Previously, TCRD incorporated gene–trait relationships directly from the GWAS (Genome-Wide Association Studies) catalog (27). The current TCRD release includes GWAS data processed through the TIGA (Target Illumination GWAS Analytics) analytical pipeline (28). Via TIGA, gene-trait associations are scored according to aggregated evidence from corresponding studies and a citation metric for publications based on the iCite Relative Citation Ratio, RCR (29). A non-parametric mean rank score is used to measure the combined evidence for a gene-trait association, mitigating the noise and uncertainties of individual GWAS publications in order to prioritize targets in an unbiased manner.

Updated data sources

Compound activity data from ChEMBL (13) was updated to version 31, and from DrugCentral (14) updated to the 2023 version (Avram *et al.* 'DrugCentral 2023 extends human clinical data and integrates veterinary drugs' NAR DB, manuscript in preparation). DISEASES (8) and the associated PubMed Scores were updated to the latest versions. Mouse phenotype data was updated to the IMPC Phenotypes (30) version 13.0. TIN-X (31) scores for target-disease associations were last regenerated and reloaded in March 2021, and will be updated again in November 2022 with the upcoming 2022 version of TIN-X. These TIN-X scores are composed of two derived bibliometric statistics: importance and novelty. The novelty of a target or a disease concept is computed based on the relative abundance of associated publication mentions, while the importance score quantifies the relative strength of association between a target and a disease concept.

Additionally, gene and protein expression data was streamlined to reduce the number of data sources and thus better assist users in navigating this data. Five data sources have been selected based on their breadth of target coverage, and complementarity toward each other in the type

Table 1. Enrichment score examples

	List description	How to generate it	Enrichment filter	Question
A	List of targets interacting with a dark target	Follow the link in PPI component on the target details page	Associated disease	What diseases are interacting targets involved in? What diseases might this dark target be involved in?
B	List of targets associated with a disease	Follow the link on the disease details page	PANTHER class/DTO Class	What types of targets might be contributing to this disease?
C	List of targets associated with some phenotype from a GWAS study	Upload custom list	PANTHER class/DTO Class	What types of targets might be contributing to this phenotype?
D	List of ligands that are structurally similar to a novel compound	Structure search page	Target filter/PANTHER Class filter	What targets or target classes have activity for similar compounds?
E	List of ligands from a cell based screening assay	Upload custom list	Target filter	What targets might be causing this phenotype?
F	List of targets with a similar sequence to a non-human gene that has been found to cause disease	Sequence search page	Associated disease	Might there be similar disease mechanisms happening in humans?

A table of examples of interesting questions that can be addressed through the enrichment score functionality in Pharos.
The columns provide a description of the list, a brief description of how to generate such a list, the filter to use for enrichment analysis, and the question that can be addressed by it.

Table 2. UpSet chart example table

	List description	How to generate that list	Which filter to use	Question
Figure 5A	List of active ligands for a target of interest	Link on target details page	Target	What compounds will activate my target, and not some others I want to distinguish between?
Figure 5B	All targets	Go to target list page	NIH target lists	Which targets left the IDG list in 2022?
Figure 5C	All targets	Go to target list page	Data source	Which targets are in the Pro-Kino database, but not Dark Kinase Knowledgebase?
Figure 5D	Any target list	Generate an interesting target list—see Materials and Methods	PANTHER class/DTO class	Which of these targets is involved in RNA binding, but no DNA binding?

A table of examples of interesting questions that can be addressed through the use and interpretation of the UpSet charts in Pharos.
The columns provide a description of the list, a brief description of how to generate such a list, the filter to construct the UpSet plot for, and the question that can be addressed by it.
Images in Figure 5 show the corresponding UpSet plot for these examples.

of expression reported (RNA, protein, or consensus) and the method by which the expression was measured. Protein expression data is provided by Human Protein Atlas (HPA, version 21.1) (32), which measures protein expression through antibody labeling, and Human Proteome Map (HPM, 2014) (33), which measures protein expression through mass spectrometry. Target gene expression is provided by Genotype-Tissue Expression (GTEx, version 8) (34) and Human Protein Atlas (HPA-RNA, version 21.1) (32), which provide high throughput measurements of RNA expression from a wide variety of tissues. Lastly, an aggregated view of expression data from many sources, including text mining, is provided by TISSUES (version 2.0, build 08/28/2022) (35).

Expression data ETL

As expression data was refreshed, the data handling code was migrated into a workflow management tool, Apache Airflow (https://airflow.apache.org/), to help manage and automate the extract, transform, and load (ETL) process. Steps in the ETL process were created to check whether new data exists for each data source, and subsequently, either rebuild the relevant tables based on the new input files or copy from the previous version of TCRD. Moving forward, other data sources, including but not limited to pathways and disease associations, can be incorporated into the ETL in a piecewise fashion to run in parallel. More foundational data sources, like UniProt, would be incorporated as upstream dependencies in the pipeline responsible for triggering a more complete rebuild of the database.

Generating subsets in the UI

One of the core elements for the UI are Pharos details pages, which show primary documentation from many sources for a single target, disease, or ligand. The other core element is list pages, which shows cards or tables for lists of targets, diseases, or ligands. Pairing a method of generating a list of targets (or diseases or ligands) with the subset analysis features can be a powerful tool and is further described below. Tables 1 and 2 define many use cases that start with generating such a list.

The full lists of targets, diseases, and ligands show all entities in TCRD in three corresponding list pages. Using the filters in the left panel of those list pages, these lists can be filtered to include only those entries associated with certain attributes. For example, one can filter for targets as-

sociated with the Gene Ontology, GO. (7,36) Processes annotation 'RNA splicing.' Similarly, on a details page, most annotations include links that will take the user from the details view for a single entity to a list page for entities that share an annotation. Following the same example as above, on a target details page for Synaptic functional regulator *FMR1* (https://pharos.nih.gov/targets/FMR1), the GO Terms component will display 'RNA splicing' as a clickable internal link to the listing of all targets that share that annotation. Data in the Protein-Protein Interactions panel represents all the targets that have been found to interact with a given target. The link to 'Explore Interacting Targets' will generate a list page for that set of targets. Users can pivot from target details pages to disease list pages and ligand list pages via links in the Associated Disease component, and Drugs and Ligands components, respectively.

In addition to filtering target lists down based on annotations, as described above, target lists can be generated based on sequence similarity to a query sequence using the BLAST (blastp) algorithm (37). This is accessible via the 'Sequence Search' link on the target list pages, or via a link on each target details page. The resulting target list includes an interactive component, including a density plot of amino acid residues that match the query sequence. This plot is implemented as an AWS EC2 instance hosting NCBI's dockerized BLAST image (https://github.com/ncbi/docker/tree/master/blast version 2.12), which is used to query the human proteins from the UniProtKB database (20).

The UI now includes a Marvin JS Widget (version 22.11.1) from ChemAxon (https://chemaxon.com/) to load and edit a chemical structure that serves as the input query structure for finding predicted targets. Executing such a search will fetch the list of targets from NCATS Predictor (38), which predicts activity values and a corresponding confidence score to any query structure based on a set of Quantitative structure-activity relationship (QSAR) models. Furthermore, the query structure defined through the Marvin JS Widget can serve as a starting point for finding similar ligands in TCRD. The structure search can be run as a similarity search, or substructure search, to find matching compounds in TCRD based on an Apache Lucene-based structure index of all TCRD ligands. See https://github.com/ncats/structure-indexer for an implementation of this structure search tool. Search results are presented in Pharos as a ligand list that is sorted according to Tanimoto similarity scores (39) against the query structure, calculated using ChemAxon's hashed chemical fingerprints.

All list pages now have buttons to Upload a custom list of targets, diseases, or ligands. This opens up the possibilities to integrate analysis functions, and construct interactive visualizations on a user-defined list. Table 1 has a few examples of using the custom list functionality to import a user's list to help understand their experimental results.

Analysis tools

Pharos list pages now have separate tabs: Table View, which is for viewing a table of results, and List Analysis, which is for performing enrichment analysis and viewing subset-level visualizations like UpSet plots (40) and heat maps. Enrichment scores can be calculated for list pages when they are showing any kind of subset of the full list, i.e. any of the methods (filtering, sequence search, structure search, etc.) mentioned above can be used to generate a subset. *P*-values resulting from a Fisher's exact test (41) are calculated for any of the categorical filters available for the list. *P*-values are then adjusted for multiple comparisons using the Benjamini-Hochberg procedure (42) to limit the False Discovery Rate to $\alpha = 0.05$.

For example, given a subset of targets, users can calculate the degree of enrichment for individual Reactome Pathways, GO Functions, or Associated Diseases that map to targets in the subset, as compared to the full list of targets. Pharos displays the *P*-value, as well as the adjusted *P*-value. Table 1 shows many examples of ways to generate a list in Pharos and calculate enrichment scores to ask various scientific questions.

Heat maps can be constructed on the List Analysis tabs of target, disease, or ligand list pages. Users can construct heat maps of protein-protein interactions colored based on the overall confidence metric that STRING (43) reports for the interaction. Both target list pages and disease list pages will show heat maps of target–disease associations, colored by the number of data sources reporting each association. Similarly, target list pages and ligand list pages will show heat maps of potency values for the target–ligand activities.

The filter panel on the left side of Table View and List Analysis View shows users the *marginal* counts of entries in the list that have each particular filter value, meaning there is no indication as to which filter values have been documented for the same entities. UpSet plots (40), on the other hand, display counts for different combinations of filter values in an intuitive way.

Structured data

Schema.org entities (JSON-LD) are now set on details pages for targets (https://schema.org/Protein), diseases (https://schema.org/MedicalCondition), and ligands (https://schema.org/ChemicalSubstance). The structured data embedded in each details page includes many types of linked data, such as protein-protein interactions (*bioChemInteraction*), pathways (*hasBioChemEntityPart*), GO terms (Processes: *isInvolvedInBiologicalProcess*, functions: *hasMolecularFunction*, component: *isLocatedInSubcellularLocation*), etc. A *rating* element (https://schema.org/Rating) is included for all target details pages to define the *ratingValue* (Tdark, Tbio, etc.), *ratingExplanation*, *reviewAspect*, and the IDG Consortium as the *author* of the TDL classifications. Each use case page is also annotated with a *HowTo* (https://schema.org/HowTo) structured data elements to note that the page reflects a sequence of instructions, where each step in the tutorial is a *HowToStep*.

RESULTS

Pharos continues to improve as a tool for the exploration and analysis of biological data as it relates to targets. Many of these improvements are driven by feedback obtained through frequent demos and user interviews. One key enhancement since our last update (4) is the ability for users to readily download CSV-formatted tables of data from the

website for further analysis or investigation. Users can select which fields to download and a query is made to deliver a table of the requested data. Links to download data are located on all details pages and list pages (where data can be downloaded for all entries in the list). Other new features implemented as a result of discussions with users include the use of ProtVista's integrated sequence and structure viewer, in-page tutorials and use cases to teach new and advanced features, and structure search and new ligand list filters for better support of chemistry workflows (see use cases below).

The structured data elements that are now included on Pharos for all details pages and use case pages will help web crawlers understand and contextualize Pharos' content and improve the visibility of this data within search engine results.

Mondo integration

As of TCRD 6.12 and Pharos 3.10, disease data is aligned using the Mondo disease ontology, which helps aggregate diseases across the many data sources that TCRD ingests. Prior to Mondo integration, there were 26 418 unique disease IDs and 17 989 unique disease names among the set of documented target-disease associations. The alignment of equivalent terms through Mondo mapping resulted in a final count of 13 704 distinct disease terms in Pharos. This mapping remains incomplete, however, since 8121 diseases are mapped to Mondo terms, and 5583 terms are yet to be mapped (primarily terms with UMLS prefixes—83%, and MeSH prefixes - 14%). Going forward, we intend to rectify the gaps in the Mondo mapping, either through proposing extensions to Mondo or developing a fallback for those UMLS and other terms.

As mentioned above, the Mondo ontology includes hierarchy information with parent and child relationships between diseases. For example, when a user browses a list of targets associated with a term that has descendent terms, such as asthma (MONDO:0004979), the list also includes targets that have been found to be associated with child terms, such as allergic asthma (MONDO:0004784) and intrinsic asthma (MONDO:0004765).

Target details pages

UI improvements include a more organized target details page, which now shows an easier-to-understand menu structure. Additionally, Pharos now displays all possible descriptors for targets, regardless of whether or not data is available. These placeholder descriptors provide users with a more obvious indication of missing information about targets. Other changes to the target details pages since the last update include:

- Expression data: A streamlined, up-to-date dataset (see Methods) is displayed, and distinctions are made between data sources that report protein expression (32,33), RNA expression (32,34), and an aggregate score based on a mix of literature and expression data (35). A heat map of tissues expressing the current target vs. the data sources reporting the expression is shown, as well as a circular treemap that groups the tissue expression values according to the hierarchy defined by the UBERON ontology, an anatomy-based ontology (44) (Figure 2A).
- Protein sequence and structure: Users can initiate a blastp search for similar targets when providing protein sequences as input. The integrated ProtVista Viewer (45) is now used, which includes structures from the Protein DataBank (PDB) (12), and predicted structures from AlphaFold (46,47).
- Nearest *Tclin* targets: This is a new component that shows a pageable listing of the nearest Tclin targets (in terms of pathway distance) found in the same KEGG pathways (17).
- Disease novelty: The scatter plot featuring TIN-X (31) data was previously shown as part of the Disease Associations component, but is now a separate component that is accompanied by an interactive circular treemap visualization which groups the associated diseases according to the disease hierarchy. The dynamic highlighting of the scatter plot points helps the user better understand patterns in the types of diseases represented in the associated importance vs novelty scatter plot (Figure 2B). While all regions of the scatter plot might be of interest for various reasons, targets along the upper-right side of the plot (Figure 2B, left panel) are often the most interesting because these targets are poorly understood, but they are still known to be relevant to the disease of interest. These improvements to the TIN-X component makes it easier for users to visually identify data points along the Pareto boundary which exists as the nondominated solution to the optimization of this system containing trade-offs between the two conflicting objectives of importance and novelty.
- GWAS traits: A new component showing GWAS (27) data, as scored and ranked according to the TIGA (28) analysis pipeline described in the methods section (Figure 2C).

Disease details pages

Disease details pages have been expanded since the 2021 update to incorporate a navigation menu on the left panel. New and updated components are summarized here:

- Disease summary: Disease data is now organized based on the Mondo Disease Ontology, and the Mondo Descriptions are shown here, when available. A button to *Explore Associated Targets* can be clicked to generate a target list page with the documented target associations. The resulting list will include targets that are associated with the disease of interest, as well as any child terms in the Mondo disease hierarchy.
- GWAS targets: A new component similar to the *GWAS Traits* component in the target details pages, showing associated targets, scored and ranked according to the TIGA (28) analysis pipeline.
- Disease hierarchy: A set of links to parent and child terms to the disease of interest, according to the Mondo hierarchy.

Ligand details pages

Ligand details pages look mostly the same, except for two new buttons that link to new functionality. The *Ligand Summary* component includes a new button to initiate a structural search using the current compound as a starting query structure. The *Target Activities* component shows a new button labeled *Explore Associated and Predicted Targets*, which navigates to a target list containing known or predicted activity to the current compound. Both features will be included in the use case for characterizing a novel chemical compound below. Both additions can also be accessed through the main Structure Search page (Figure 3).

Use cases

All use cases described in our previous work (4,10) are still possible, and some have been incorporated into the Use Cases detailed as tutorials on Pharos (https://pharos.nih.gov/usecases). Here, we focus on new functionality and how a user's capacity to perform tasks has increased.

The ability to initiate a search based on a chemical structure has added support for a number of different workflows (Figure 3). Users investigating the potential role of a novel chemical compound can input the structure on the Structure Search page (https://pharos.nih.gov/structure), by using a SMILES string or by drawing the structure using the Marvin JS widget. Information about which targets might be activated by such a compound can be generated by using the Find Predicted Targets button, which generates a target list based on the output of a series of QSAR models from NCATS Predictor. Similarly, the user can generate a list of similar compounds found in TCRD (Table 1D), and look for patterns in the targets and target classes that have activity against those similar compounds.

Another useful tool in the user's toolbox is the ability to generate a custom list of entities the user introduces from their own experiments or investigations. Users can upload a list of targets from a GWAS study for a particular trait, or a list of compounds obtained as hits in a cell-based assay (Table 1E). Pharos will resolve the list entries, display the table of results (and any of the visualizations that are shown for lists) and allow the user to run enrichment calculations to determine which annotations in a user's custom list are over-represented. Ligand lists can further be filtered according to the specificity and potency of compounds in the list or can be queried as to the specific target activities they have using the UpSet plot (Table 2A, Figure 5A).

There are several new use cases supported through the addition of analysis capabilities on Pharos list pages. This expansion of functionality has led us to split those pages into two tabs supporting two distinct workflows. Table View is for browsing some high-level information about each target, perhaps to find a particular target details page to explore further. List Analysis View is for exploring patterns in annotations for entries in the list, such as the counts or combinations of filter values that elements in the list have, or constructing heat maps of elements in the list.

Common in modern scientific literature, enrichment calculations are often used to characterize which pathways involve a set of targets (48). In Pharos, enrichment calculations can be made for any target list, for pathways, as well as for any other categorical filter available on the target list page. These calculations can be made from the associated filter panel, or by selecting an option from the drop-down element on the *Filter Value Enrichment* panel. Figure 4 illustrates the problem addressed by enrichment analysis as it shows the filter value counts for the *Associated Disease* filter, for the unfiltered target list, alongside the filter value counts for a list of targets that interact with the D (2) Dopamine Receptor (DRD2). The sorted lists are dominated by values that are very common in the full population of targets. In the case of *Associated Diseases*, the most common filter values tend to be cancers, which are well studied and have wide-ranging effects on cell function, and so have a large number of associated targets. When the enrichment scores are calculated, and the results are sorted according to p-value, the top values in this ranking reveal a different set of *Associated Diseases*, including many examples that people typically associate with DRD2. This example uncovers some expected results for a well-studied target but helps to validate the workflow where a hypothesis is made for the potential role of a dark target in disease, by analyzing the associated diseases of targets that interact with the dark target in question. Table 1 includes this use case and several other examples where enrichment analysis can be used on a subset to examine the patterns of data that may not be obvious at first.

UpSet plots (40) have been trending in recent years, as they help users understand the different sets of values that are present in a list. Pharos shows UpSet plots (40) on the List Analysis tab for all categorical filters that can have multiple values. For filters that can have only one value, such as TDL, a donut chart is shown. At its core, the UpSet plot is a column chart, where column height represents the number of entries in the list that have each combination of filter values. The filled and open circles at the bottom of the chart represent which combination of filter values each column corresponds to. Pharos UpSet plots are interactive and can be used to filter the list based on different combinations of filter values. This allows users to filter their lists with more complex boolean logic: e.g. targets that have been documented with GO Functions for 'DNA Binding' AND NOT 'RNA Binding.' Table 2 and Figure 5 include a set of examples where UpSet plots can be used to answer interesting questions about the data.

DISCUSSION

Both TCRD and Pharos are continuously being developed and since the last 2021 NAR Database update, new implementations align closely with the broader IDG goal of illuminating understudied proteins with the ability to generate hypotheses when data is sparse. New developments include a streamlined process for integrating sources into TCRD and new options for analyses and visualization in Pharos. We'll continue to build on the 79 data sources contributing to Pharos, to further increase the depth of knowledge for each target and expand the ways users can look for patterns that may help to fill the gaps in knowledge and understanding. One addition we are currently working on is a 'Rare Disease' annotation from GARD, so that users will know which diseases in a list have been annotated as such. It is our

Figure 2 Target details updates. (**A**) Expression data from five sources is color coded according to expression type (red: aggregate score, blue: RNA expression, yellow: protein expression) and displayed as a heat map on the left panel. Cells in the heat map display details about the expression data for each tissue when clicked. The right panel contains tabs to show either a circular treemap, where tissues are grouped according to the UBERON hierarchy, and an anatomogram, where tissues on a human form are shaded according to their expression level. The circular treemap is interactive and can be used to filter the heat map. (**B**) Text-mined target-disease associations from TIN-X are shown in an interactive scatter plot of importance vs novelty adjacent to a circular treemap. The circular treemap groups the associations based on the hierarchy defined by the Mondo Disease Ontology. Selecting a circle in the right panel highlights corresponding points on the scatter plot for those diseases. This dynamic visualization helps users find classes of diseases which tend to be both high in the importance and novelty metrics. (**C**) GWAS traits associated with a single target, scored and ranked according to the TIGA data processing pipeline. More reliable associations tend to have a higher Mean Rank Score and a higher Beta Count.

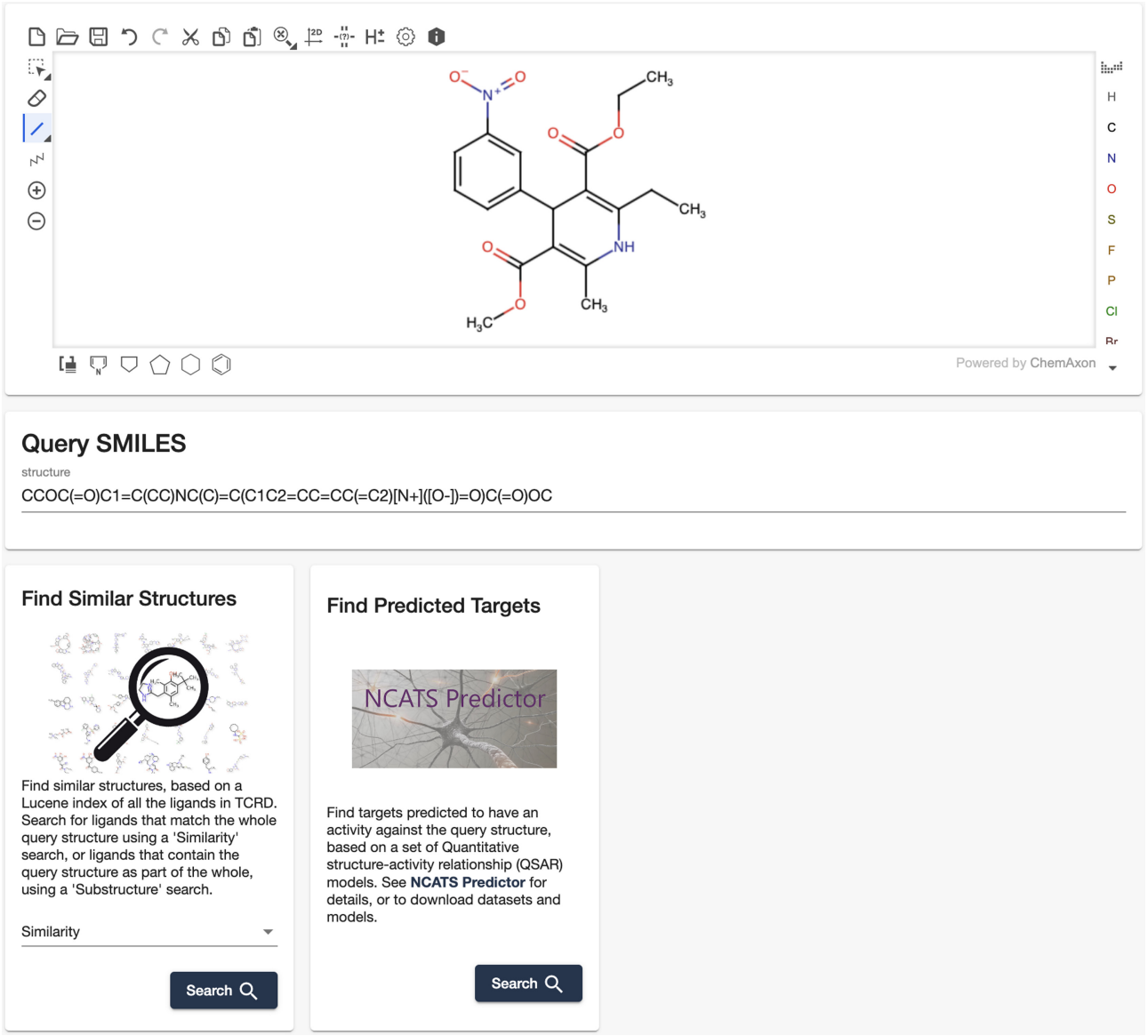

Figure 3. Structure search. The component contains a Marvin JS widget showing an editable query structure. Structures can also be edited via the Query SMILES input or resolved using an external tool (not pictured). This serves as a starting point for performing a similarity, or substructure, search of compounds in TCRD, or a search for predicted targets, using the QSAR models in NCATS Predictor.

vision to become a 'one-stop-shop' for protein-disease biology associations, specifically for understudied targets and rare diseases.

One key new feature is the ability to expand a search for a dark target to retrieve information on related targets. The pivot from a target details page to a list of related targets via amino acid sequence, interacting targets, or other known common attributes allows users to shift their line of inquiry from 'What does this target do?' to 'What do related targets tend to do?'. This shift helps formulate hypotheses as to the role of target(s) in the context of the biological processes they are a part of and the diseases they are associated with.

Another recent addition to Pharos' repertoire is the circular treemap plots that are shown in the Expression and Disease Novelty components. These can help users understand patterns in the data, such as which branches of the UBERON hierarchy are more likely to express a target, or which types of diseases are more often associated with a target (Figure 2). This visualization can also be utilized in a more general way by including those plots on list pages. Users can then generate lists of targets, diseases, or ligands, by any of the methods available, and construct visualizations to help understand patterns in any of the hierarchical annotations. For example, users could not only tell which diseases are associated with targets in a list, but the ancestry of those disease terms, helping them find common root terms that may not be obvious.

In the future, Pharos will continue to expand its supported data types and to enable easy incorporation of new data sources from external groups. Example data under consideration include machine learning predictions and additional experimental data (including but not limited

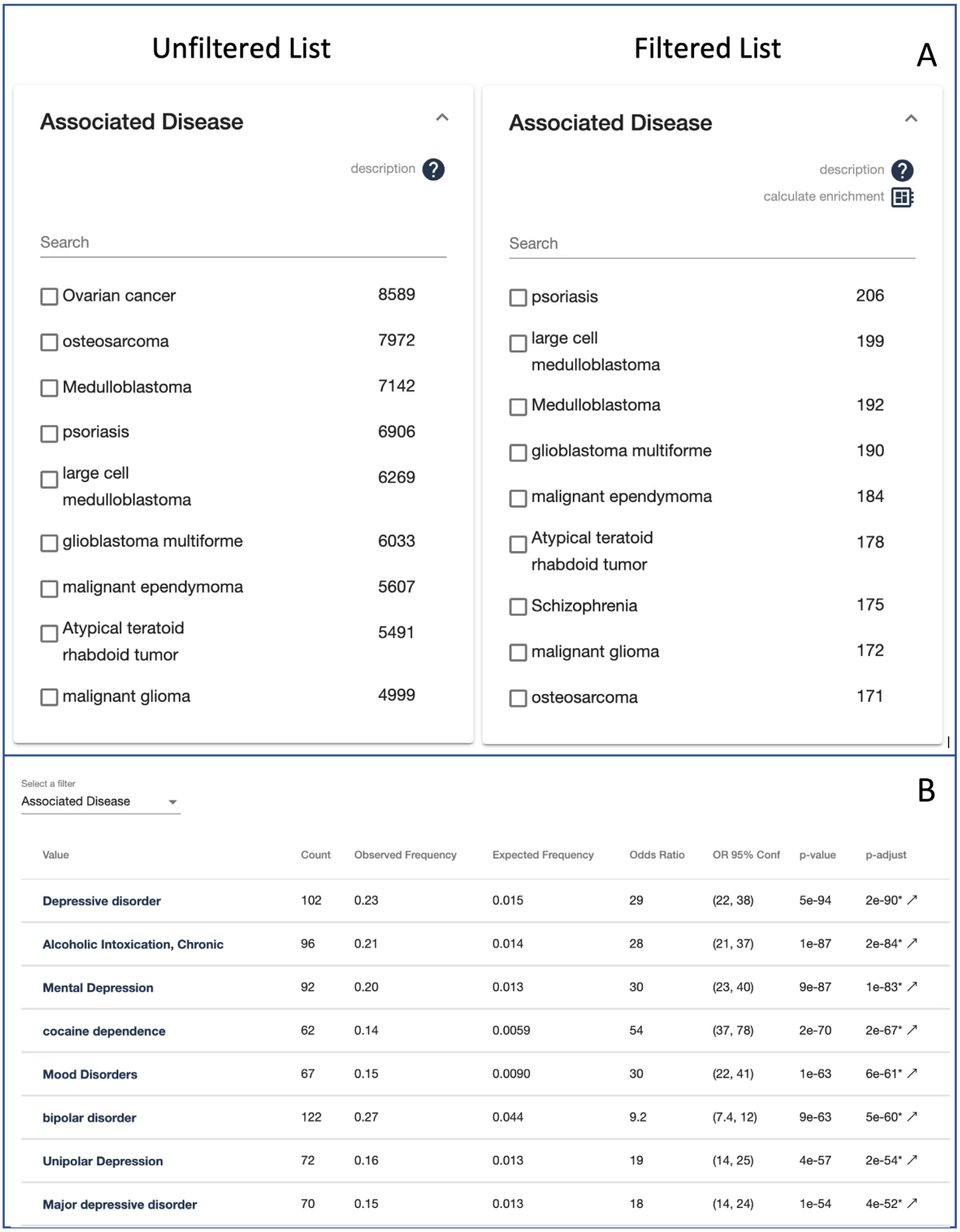

Figure 4. Enrichment analysis. (**A**) Left shows the number of targets associated with different diseases in a full target list. Right shows the number of targets associated with different diseases in a target list consisting of targets with a documented protein-protein interaction with DRD2. Note how there is a lot of overlap between the entries in this list, when the list is sorted by the naive counts. (**B**) Enrichment score results after performing Fisher's Exact Test on the filtered list. Note how the top entries in the list comprises a different set of diseases, including a lot of neurological disorders, and substance dependence disorders, diseases which are more commonly known to be related to DRD2.

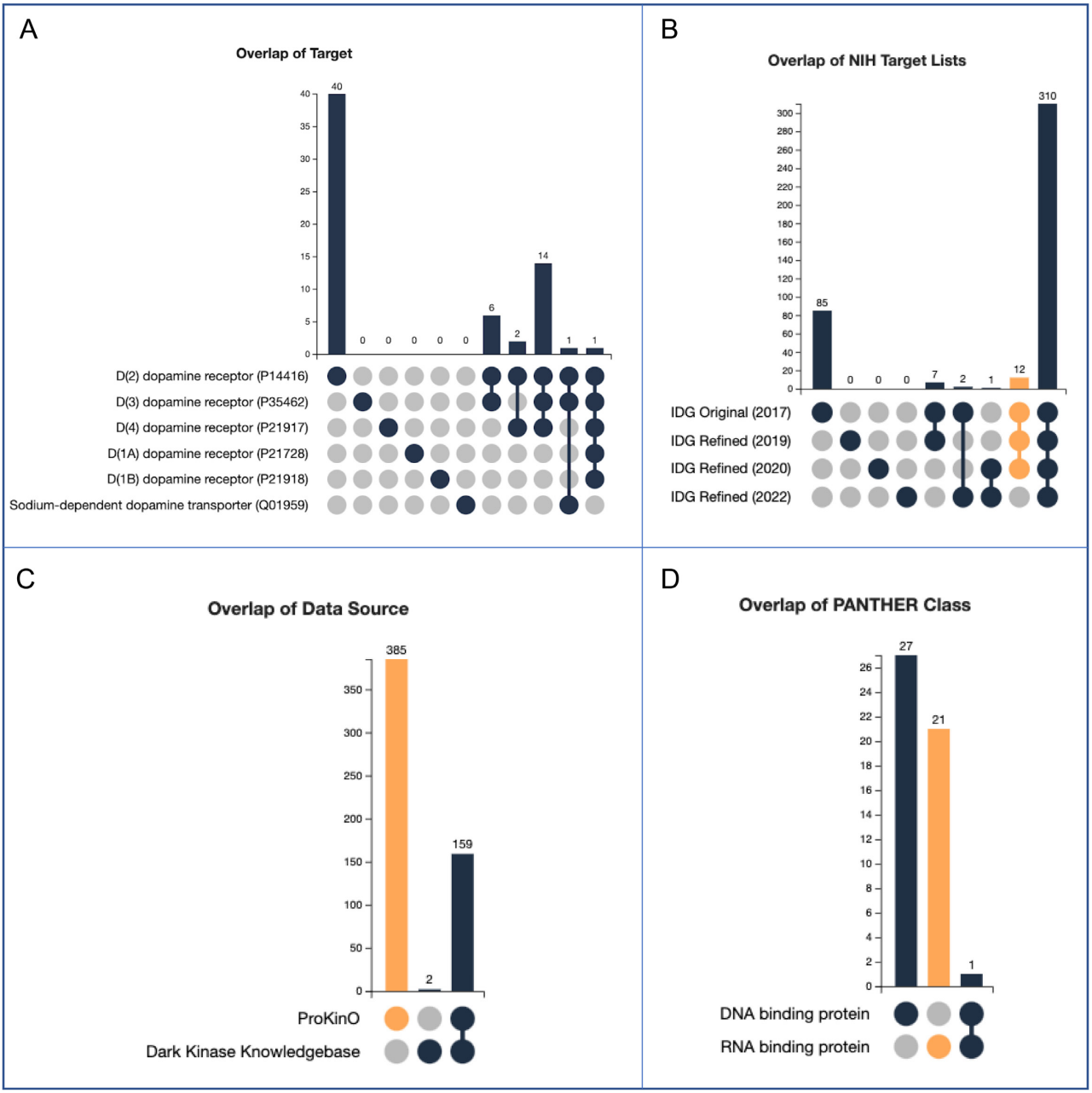

Figure 5. UpSet chart examples. (**A**) For a list of ligands, this plot shows different combinations of targets against which each ligand has been shown to be active. (**B**) For a target list, this plot shows how many targets in the list were on each combination of NIH lists. Gold highlighting shows filters that are currently applied to the list, so this list would consist of 12 targets that were on the IDG list from 2017 to 2020, and removed in 2022. (**C**) In this case, the corresponding target list would show 385 targets that have data from ProKinO, but not from Dark Kinase Knowledgebase. (**D**) In this case, the corresponding target list would show 21 targets annotated by PANTHER to be a RNA binding protein, but not a DNA binding protein.

to phenotypic and expression data on ion channels, G-protein-coupled receptors, and kinases) generated through the IDG (https://druggablegenome.net/ProteinTimeLine), or other external sources. Data may be incorporated directly into TCRD, or accessible through Pharos through external APIs. In the latter case, external groups would create an API (or work with us to set one up) that would return schema.org structured data, such as a list of @Protein objects representing a set of predicted protein-protein interactions, or a list of @MedicalCondition objects representing a set of predicted disease associations. Pharos would fetch data from these APIs for display within Pharos, including a clearly visible link to the original data resource (Supplemental Figure 1). Given the increasing traffic of Pharos, IDG and external groups could then increase the visibility of their data within Pharos and directly share their data through the platform amongst colleagues. Contributing groups would be able to view and analyze their data in light of all the other data in Pharos, empowering them to generate meaningful hypotheses for further research. Sup-

plemental Figure 1 shows the prototype UI component that fetches structured data from an external API. This first implementation returns data from Ravenmehr *et al.* (49), which predicts a relationship between specific cancers and kinases. Incorporating data to Pharos allows scientists to have access to all the visualizations and subset analysis tools that Pharos offers, for their own datasets. In this example (details can be found in the description of Supplemental Figure 1), users could browse the list of kinases predicted to have an effect on a certain cancer and download data for the entire list or perform enrichment calculations to determine what might be common amongst the targets in the list.

Additional future enhancements will include improving navigation of data submitted to and/or contained within Pharos. To accomplish this, we plan to expand functionalities for data analytics and visualization to further explore users' own datasets or custom filtered lists in the context of other studies. Expansions will include additional scoring and statistics outputs to provide ranking and facilitate the discovery of relevant data patterns (e.g. putative mechanisms of actions, drug repurposing efforts). In addition, the Pharos team will continue to embed novel visualizations from other sites via an API and enable users greater flexibility in creating charts that are interactive (e.g. on-the-fly filtering and labeling).

Lastly, in the future we will refer to TCRD/Pharos simply as 'Pharos.' We will also continue to guide its development by use cases and user feedback. By focusing on users and what they could learn from Pharos, we aim to continue expanding and improving the availability of knowledge about the dark proteome (50), thereby continuing to illuminate the druggable genome.

DATA AVAILABILITY

TCRD is an open source database that can be accessed at: http://juniper.health.unm.edu/tcrd/.

Pharos is an open source web platform that can be accessed at: https://pharos.nih.gov/.

The Pharos resources have been split into frontend and backend repositories.

The front end code can be found on Github: https://github.com/ncats/pharos_frontend.

The backend GraphQL implementation code can be found on Github: https://github.com/ncats/pharos-graphql-server.

GraphQL resource documentation can be found on Pharos: https://pharos.nih.gov/api.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

Mitch Miller for his work in developing and maintaining the ligand resolver service used by Pharos to resolve compounds for custom ligand lists, and for structure search. Alexey V. Zakharov for his work in developing and maintaining the NCATS Predictor service used to predict target activity for input structures.

FUNDING

National Institutes of Health (NIH) Common Fund [U24 CA224370 to S.L.M., C.G.B., L.J.J., J.J.Y., J.H., V.T.M., J.E., T.I.O.]; National Institutes of Health (NIH) Common Fund, NCATS [U24 TR002278 to D.V., S.C.S.]; Novo Nordisk Foundation [NNF14CC0001 to L.J.J.]; Intramural Research Program, NCATS [to D.T.N., K.K., T.S., V.S., K.S., A.P., E.M.]. Funding for open access charge: NIH Grant [U24 CA224370]; Intramural/Extramural research program of the NCATS, NIH: Knowledge Management Center for Illuminating the Druggable Genome (Pharos) [ZIA TR000057-08, in part].

Conflict of interest statement. L.J.J. is co-founder and scientific advisory board member of Intomics A/S. T.I.O. and C.G.B. are full-time employees of Roivant Sciences Inc. T.I.O. has received honoraria or consulted for Abbott, AstraZeneca, Chiron, Genentech, Infinity Pharmaceuticals, Merz Pharmaceuticals, Merck Darmstadt, Mitsubishi Tanabe, Novartis, Ono Pharmaceuticals, Pfizer, Roche, Sanofi and Wyeth. He served on the scientific advisory board of ChemDiv Inc. and InSilico Medicine. D.T.N. is now employed at Pfizer.

REFERENCES

1. Edwards,A.M., Isserlin,R., Bader,G.D., Frye,S.V., Willson,T.M. and Yu,F.H. (2011) Too many roads not taken. *Nature*, **470**, 163–165.
2. Stoeger,T., Gerlach,M., Morimoto,R.I. and Nunes Amaral,L.A. (2018) Large-scale investigation of the reasons why potentially important genes are ignored. *PLoS Biol.*, **16**, e2006643.
3. Nguyen,D.-T., Mathias,S., Bologa,C., Brunak,S., Fernandez,N., Gaulton,A., Hersey,A., Holmes,J., Jensen,L.J., Karlsson,A. *et al.* (2017) Pharos: collating protein information to shed light on the druggable genome. *Nucleic Acids Res.*, **45**, D995–D1002.
4. Sheils,T.K., Mathias,S.L., Kelleher,K.J., Siramshetty,V.B., Nguyen,D.-T., Bologa,C.G., Jensen,L.J., Vidović,D., Koleti,A., Schürer,S.C. *et al.* (2021) TCRD and pharos 2021: mining the human proteome for disease biology. *Nucleic Acids Res.*, **49**, D1334–D1346.
5. Santos,R., Ursu,O., Gaulton,A., Bento,A.P., Donadi,R.S., Bologa,C.G., Karlsson,A., Al-Lazikani,B., Hersey,A., Oprea,T.I. *et al.* (2017) A comprehensive map of molecular drug targets. *Nat. Rev. Drug Discov.*, **16**, 19–34.
6. Avram,S., Halip,L., Curpan,R. and Oprea,T.I. (2022) Novel drug targets in 2021. *Nat. Rev. Drug Discov.*, **21**, 328.
7. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene ontology: tool for the unification of biology. The gene ontology consortium. *Nat. Genet.*, **25**, 25–29.
8. Grissa,D., Junge,A., Oprea,T.I. and Jensen,L.J. (2022) Diseases 2.0: a weekly updated database of disease–gene associations from text mining and data integration. *Database*, **2022**, baac019.
9. Björling,E. and Uhlén,M. (2008) Antibodypedia, a portal for sharing antibody and antigen validation data. *Mol. Cell. Proteomics*, **7**, 2028–2037.
10. Sheils,T., Mathias,S.L., Siramshetty,V.B., Bocci,G., Bologa,C.G., Yang,J.J., Waller,A., Southall,N., Nguyen,D.-T. and Oprea,T.I. (2020) How to illuminate the druggable genome using pharos. *Curr. Protoc. Bioinformatics*, **69**, e92.
11. Kropiwnicki,E., Binder,J.L., Yang,J.J., Holmes,J., Lachmann,A., Clarke,D.J.B., Sheils,T., Kelleher,K.J., Metzger,V.T., Bologa,C.G. *et al.* (2022) *Getting started with the IDG KMC datasets tools. Curr. Protoc.*, **2**, e355.
12. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The protein data bank. *Nucleic Acids Res.*, **28**, 235–242.
13. Gaulton,A., Hersey,A., Nowotka,M., Bento,A.P., Chambers,J., Mendez,D., Mutowo,P., Atkinson,F., Bellis,L.J., Cibrián-Uhalte,E.

et al. (2017) The ChEMBL database in 2017. *Nucleic Acids Res.*, **45**, D945–D954.
14. Ursu,O., Holmes,J., Bologa,C.G., Yang,J.J., Mathias,S.L., Stathias,V., Nguyen,D.-T., Schürer,S. and Oprea,T. (2019) DrugCentral 2018: an update. *Nucleic Acids Res.*, **47**, D963–D970.
15. Wang,J., Al-Ouran,R., Hu,Y., Kim,S.-Y., Wan,Y.-W., Wangler,M.F., Yamamoto,S., Chao,H.-T., Comjean,A., Mohr,S.E. *et al.* (2017) MARRVEL: integration of human and model organism genetic resources to facilitate functional annotation of the human genome. *Am. J. Hum. Genet.*, **100**, 843–853.
16. Jassal,B., Matthews,L., Viteri,G., Gong,C., Lorente,P., Fabregat,A., Sidiropoulos,K., Cook,J., Gillespie,M., Haw,R. *et al.* (2020) The reactome pathway knowledgebase. *Nucleic Acids Res.*, **48**, D498–D503.
17. Kanehisa,M., Furumichi,M., Sato,Y., Ishiguro-Watanabe,M. and Tanabe,M. (2021) KEGG: integrating viruses and cellular organisms. *Nucleic Acids Res.*, **49**, D545–D551.
18. Harding,S.D., Armstrong,J.F., Faccenda,E., Southan,C., Alexander,S.P.H., Davenport,A.P., Pawson,A.J., Spedding,M., Davies,J.A. and NC-IUPHARNC-IUPHAR (2022) The IUPHAR/BPS guide to PHARMACOLOGY in 2022: curating pharmacology for COVID-19, malaria and antibacterials. *Nucleic Acids Res.*, **50**, D1282–D1294.
19. York,W.S., Mazumder,R., Ranzinger,R., Edwards,N., Kahsay,R., Aoki-Kinoshita,K.F., Campbell,M.P., Cummings,R.D., Feizi,T., Martin,M. *et al.* (2020) GlyGen: computational and informatics resources for glycoscience. *Glycobiology*, **30**, 72–73.
20. Bateman,A., Martin,M., Orchard,S., Magrane,M., Agivetova,R., Ahmad,S., Alpi,E., Bowler-Barnett,E.H., Britto,R., Bursteinas,B. *et al.* (2021) UniProt: the universal protein knowledgebase in 2021 (2021). *Nucleic Acids Res.*, **49**, D480–D489.
21. Mungall,C.J., McMurry,J.A., Köhler,S., Balhoff,J.P., Borromeo,C., Brush,M., Carbon,S., Conlin,T., Dunn,N., Engelstad,M. *et al.* (2017) The monarch initiative: an integrative data and analytic platform connecting phenotypes to genotypes across species. *Nucleic Acids Res.*, **45**, D712–D722.
22. Vasilevsky,N.A., Matentzoglu,N.A., Toro,S., Flack,J.E. IV, Hegde,H., Unni,D.R., Alyea,G.F., Amberger,J.S., Babb,L., Balhoff,J.P. *et al.* (2022) Mondo: unifying diseases for the world, by the world. medRxiv doi: https://doi.org/10.1101/2022.04.13.22273750, 16 April 2022, preprint: not peer reviewed.
23. Schriml,L.M., Arze,C., Nadendla,S., Chang,Y.-W.W., Mazaitis,M., Felix,V., Feng,G. and Kibbe,W.A. (2012) Disease ontology: a backbone for disease semantic integration. *Nucleic Acids Res.*, **40**, D940–D6.
24. McKusick,V.A. (1998) In: *Mendelian Inheritance in Man: A Catalog of Human Genes and Genetic Disorders*. JHU Press.
25. Weinreich,S.S., Mangon,R., Sikkens,J.J., Teeuw,M.E. and Cornel,M.C. (2008) Orphanet: a european database for rare diseases. *Ned. Tijdschr. Geneeskd.*, **152**, 518–519.
26. Bodenreider,O. (2004) The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Res.*, **32**, D267–D70.
27. Buniello,A., MacArthur,J.A.L., Cerezo,M., Harris,L.W., Hayhurst,J., Malangone,C., McMahon,A., Morales,J., Mountjoy,E., Sollis,E. *et al.* (2019) The NHGRI-EBI GWAS catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.*, **47**, D1005–D1012.
28. Yang,J.J., Grissa,D., Lambert,C.G., Bologa,C.G., Mathias,S.L., Waller,A., Wild,D.J., Jensen,L.J. and Oprea,T.I. (2021) TIGA: target illumination GWAS analytics. *Bioinformatics*, **37**, 3865–3873.
29. Hutchins,B.I., Yuan,X., Anderson,J.M. and Santangelo,G.M. (2016) Relative citation ratio (RCR): a new metric that uses citation rates to measure influence at the article level. *PLoS Biol.*, **14**, e1002541.
30. Dickinson,M.E., Flenniken,A.M., Ji,X., Teboul,L., Wong,M.D., White,J.K., Meehan,T.F., Weninger,W.J., Westerberg,H., Adissu,H. *et al.* (2016) High-throughput discovery of novel developmental phenotypes. *Nature*, **537**, 508–514.
31. Cannon,D.C., Yang,J.J., Mathias,S.L., Ursu,O., Mani,S., Waller,A., Schürer,S.C., Jensen,L.J., Sklar,L.A., Bologa,C.G. *et al.* (2017) TIN-X: target importance and novelty explorer. *Bioinformatics*, **33**, 2601–2603.
32. Thul,P.J. and Lindskog,C. (2018) The human protein atlas: a spatial map of the human proteome. *Protein Sci.*, **27**, 233–244.
33. Kim,M.-S., Pinto,S.M., Getnet,D., Nirujogi,R.S., Manda,S.S., Chaerkady,R., Madugundu,A.K., Kelkar,D.S., Isserlin,R., Jain,S. *et al.* (2014) A draft map of the human proteome. *Nature*, **509**, 575–581.
34. GTEx Consortium (2020) The GTEx consortium atlas of genetic regulatory effects across human tissues. *Science*, **369**, 1318–1330.
35. Palasca,O., Santos,A., Stolte,C., Gorodkin,J. and Jensen,L.J. (2018) TISSUES 2.0: an integrative web resource on mammalian tissue expression. *Database*, **2018**, bay003.
36. Gene Ontology Consortium (2021) The gene ontology resource: enriching a GOld mine. *Nucleic Acids Res.*, **49**, D325–D334.
37. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
38. Zakharov,A.V., Zhao,T., Nguyen,D.-T., Peryea,T., Sheils,T., Yasgar,A., Huang,R., Southall,N. and Simeonov,A. (2019) Novel consensus architecture to improve performance of large-scale multitask deep learning QSAR models. *J. Chem. Inf. Model.*, **59**, 4613–4624.
39. Bajusz,D., Rácz,A. and Héberger,K. (2015) Why is tanimoto index an appropriate choice for fingerprint-based similarity calculations?*J. Cheminform.*, **7**, 20.
40. Lex,A., Gehlenborg,N., Strobelt,H., Vuillemot,R. and Pfister,H. (2014) UpSet: visualization of intersecting sets. *IEEE Trans. Vis. Comput. Graph.*, **20**, 1983–1992.
41. Fisher,S.R.A. (1934) In: *Statistical Methods for Research Workers … Fifth Edition - Revised and Enlarged Edinburgh*, London.
42. Yekutieli,D. and Benjamini,Y. (1999) Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. *J. Stat. Plan. Inference*, **82**, 171–196.
43. Szklarczyk,D., Gable,A.L., Lyon,D., Junge,A., Wyder,S., Huerta-Cepas,J., Simonovic,M., Doncheva,N.T., Morris,J.H., Bork,P. *et al.* (2019) STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.*, **47**, D607–D613.
44. Mungall,C.J., Torniai,C., Gkoutos,G.V., Lewis,S.E. and Haendel,M.A. (2012) Uberon, an integrative multi-species anatomy ontology. *Genome Biol.*, **13**, R5.
45. Watkins,X., Garcia,L.J., Pundir,S., Martin,M.J. and Consortium,UniProt (2017) ProtVista: visualization of protein sequence annotations. *Bioinformatics*, **33**, 2040–2041.
46. Jumper,J., Evans,R., Pritzel,A., Green,T., Figurnov,M., Ronneberger,O., Tunyasuvunakool,K., Bates,R., Žídek,A., Potapenko,A. *et al.* (2021) Highly accurate protein structure prediction with alphafold. *Nature*, **596**, 583–589.
47. Varadi,M., Anyango,S., Deshpande,M., Nair,S., Natassia,C., Yordanova,G., Yuan,D., Stroe,O., Wood,G., Laydon,A. *et al.* (2022) AlphaFold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.*, **50**, D439–D444.
48. Huang,D.W., Sherman,B.T. and Lempicki,R.A. (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.*, **37**, 1–13.
49. Ravanmehr,V., Blau,H., Cappelletti,L., Fontana,T., Carmody,L., Coleman,B., George,J., Reese,J., Joachimiak,M., Bocci,G. *et al.* (2021) Supervised learning with word embeddings derived from pubmed captures latent knowledge about protein kinases and cancer. *NAR Genom Bioinform*, **3**, lqab113.
50. Perdigão,N., Heinrich,J., Stolte,C., Sabir,K.S., Buckley,M.J., Tabor,B., Signal,B., Gloss,B.S., Hammang,C.J., Rost,B. *et al.* (2015) Unexpected features of the dark proteome. *Proc. Natl. Acad. Sci. U.S.A.*, **112**, 15898–15903.