D1334–D1346 *Nucleic Acids Research, 2021, Vol. 49, Database issue* *Published online 6 November 2020*
doi: 10.1093/nar/gkaa993

TCRD and Pharos 2021: mining the human proteome for disease biology

Timothy K. Sheils[1], Stephen L. Mathias[2], Keith J. Kelleher[1], Vishal B. Siramshetty[1], Dac-Trung Nguyen[1], Cristian G. Bologa[2], Lars Juhl Jensen[3], Dušica Vidović[4,5], Amar Koleti[4], Stephan C. Schürer[4,5,6], Anna Waller[7], Jeremy J. Yang[2], Jayme Holmes[2], Giovanni Bocci[2], Noel Southall[1], Poorva Dharkar[1], Ewy Mathé[1], Anton Simeonov[1] and Tudor I. Oprea[2,3,8,9,*]

[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, [7]UNM Center for Molecular Discovery, University of New Mexico Health Sciences Center, Albuquerque, NM 87131, USA, [8]UNM Comprehensive Cancer Center, University of New Mexico Health Sciences Center, Albuquerque, NM 87131, USA and [9]Department of Rheumatology and Inflammation Research, Institute of Medicine, Sahlgrenska Academy at University of Gothenburg, 40530 Gothenburg, Sweden

Received September 15, 2020; Revised October 09, 2020; Editorial Decision October 12, 2020; Accepted October 14, 2020

ABSTRACT

In 2014, the National Institutes of Health (NIH) initiated the Illuminating the Druggable Genome (IDG) program to identify and improve our understanding of poorly characterized proteins that can potentially be modulated using small molecules or biologics. Two resources produced from these efforts are: The Target Central Resource Database (TCRD) (http://juniper.health.unm.edu/tcrd/) and Pharos (https://pharos.nih.gov/), a web interface to browse the TCRD. The ultimate goal of these resources is to highlight and facilitate research into currently understudied proteins, by aggregating a multitude of data sources, and ranking targets based on the amount of data available, and presenting data in machine learning ready format. Since the 2017 release, both TCRD and Pharos have produced two major releases, which have incorporated or expanded an additional 25 data sources. Recently incorporated data types include human and viral-human protein–protein interactions, protein–disease and protein–phenotype associations, and drug-induced gene signatures, among others. These aggregated data have enabled us to generate new visualizations and content sections in Pharos, in order to empower users to find new areas of study in the druggable genome.

INTRODUCTION

It is widely acknowledged that only a subset of the human genome that is considered ‘druggable’ is subject to scientific inquiry (1). Since 2014, the National Institutes of Health (NIH) initiative Illuminating the Druggable Genome (IDG), has been working toward the goal of shedding light on understudied protein targets that could be potentially modulated by small molecules or biologics. Pharos and the Target Central Resource Database (TCRD) are open-access resources developed as a part of the IDG program and jointly serve as the knowledge hub for over 20 000 human protein targets (2,3). First introduced and published in the 2017 NAR database issue (4), TCRD collates information from several gene/protein data sources and Pharos serves as a web interface that presents this information to users. Since its initial launch, the Pharos paper was cited in more than 110 times *cf.* Google Scholar, and the portal is accessed on average by ∼1600 new visitors monthly, with a total of ∼ 425 000 pageviews and >20 000 full TCRD database downloads (as of 3 September

[*]To whom correspondence should be addressed. Tel: +1 505 925 7529; Fax: +1 505 925 7625; Email: toprea@salud.unm.edu

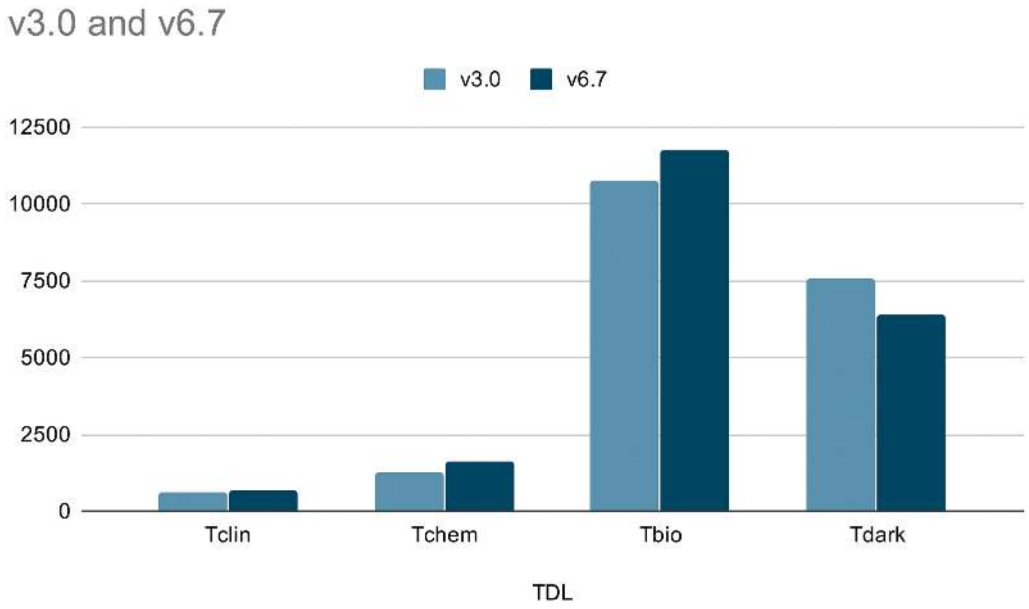
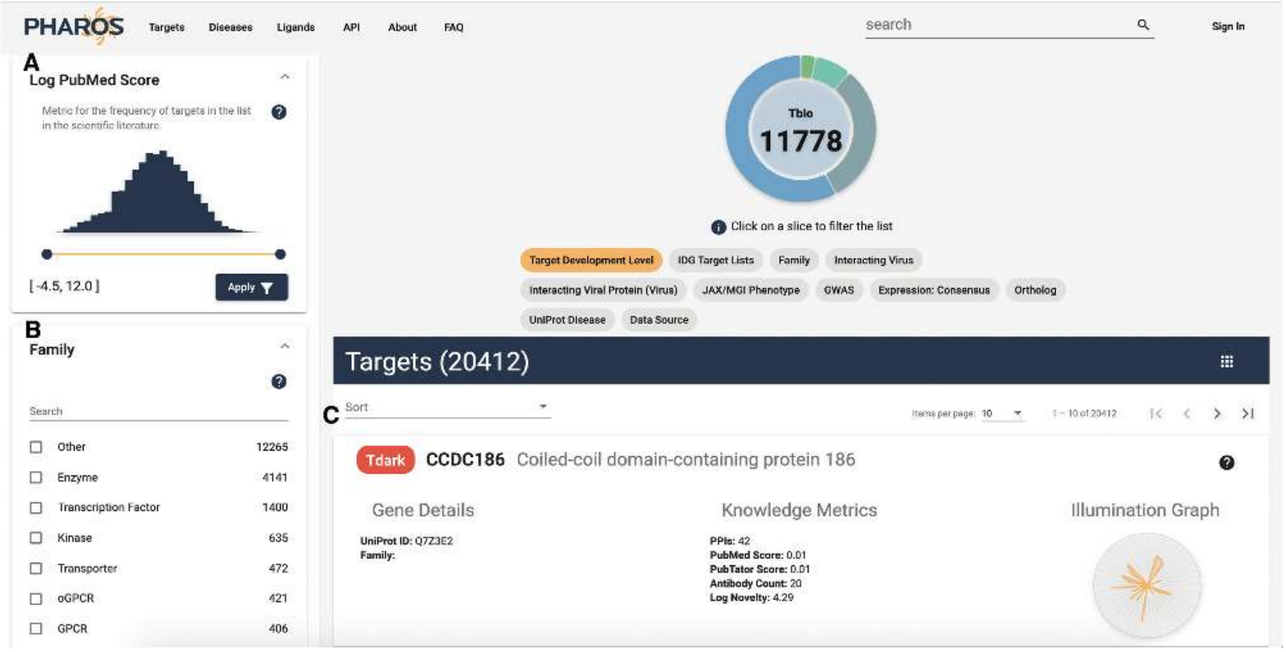


Figure 1. Chart of the TDL changes between TCRD v3.0 and v6.7. The decrease of Tdark and subsequent increase of other development levels shows an overall increase in target illumination.



Metric for the frequency of targets in the list in the scientific literature.

[-4.5, 12.0] Apply

Search

☐	Other	12265
☐	Enzyme	4141
☐	Transcription Factor	1400
☐	Kinase	635
☐	Transporter	472
☐	oGPCR	421
☐	GPCR	406

Click on a slice to filter the list

Targets (20412)

Items per page: 10 1 – 10 of 20412

CCDC186 Coiled-coil domain-containing protein 186

Gene Details

UniProt ID: Q7Z3E2
Family:

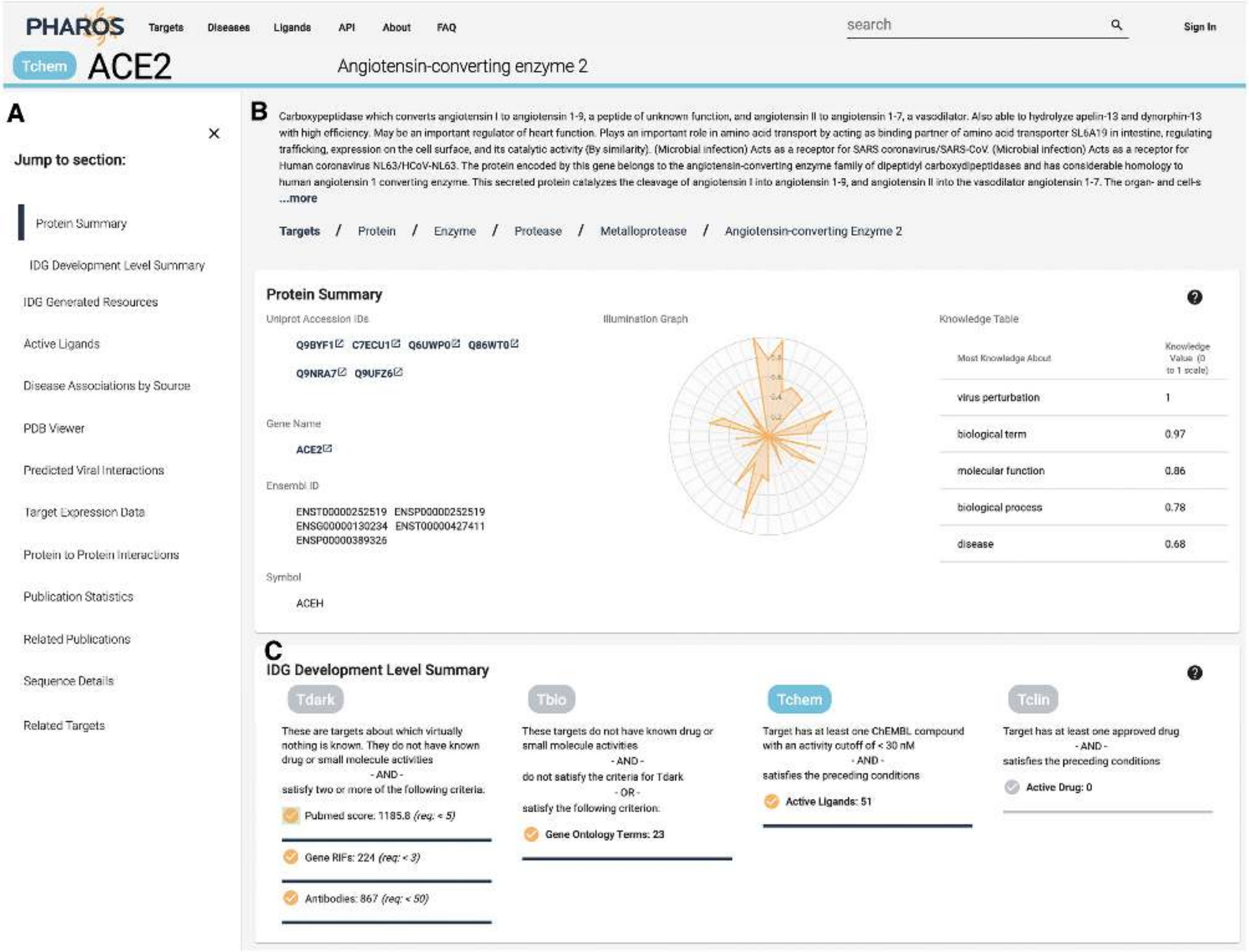
Knowledge Metrics

PPIs: 42
PubMed Score: 0.01
PubTator Score: 0.01
Antibody Count: 20
Log Novelty: 4.29

Figure 2. Browse targets page, showing new numeric slider facets for Log PubMed Score range (**A**), searchable protein Family filter panel (**B**) and improved target card view (**C**). The Log PubMed Score filter also shows the definition section displayed.

2020). TCRD has been enhanced by inclusion of new data types and data from emerging resources, which were prepared for machine-learning readiness. The scope of TCRD has also been expanded, moving past the initial area of focus of the druggable genome to aggregate data about the entire human proteome. The first published version of Pharos was based on TCRD version 3.0, the latest version uses TCRD version 6.7, which currently aggregates data from 78 data sources (see Supplementary Information). Furthermore, the search functionality on the Pharos web server has been upgraded to use the graph querying language (GraphQL; https://graphql.org/) API that facilitates faster data retrieval, directly from TCRD.

In the current paper, we describe changes implemented for the 2021 version, such as new data sources and how data from these sources have been integrated into TCRD and presented in Pharos. The latest architecture of the database and both new and improved features implemented in the Pharos platform are described in the following sections of the paper.



ACE2

Angiotensin-converting enzyme 2

Jump to section:

Protein Summary
IDG Development Level Summary
IDG Generated Resources
Active Ligands
Disease Associations by Source
PDB Viewer
Predicted Viral Interactions
Target Expression Data
Protein to Protein Interactions
Publication Statistics
Related Publications
Sequence Details
Related Targets

Carboxypeptidase which converts angiotensin I to angiotensin 1-9, a peptide of unknown function, and angiotensin II to angiotensin 1-7, a vasodilator. Also able to hydrolyze apelin-13 and dynorphin-13 with high efficiency. May be an important regulator of heart function. Plays an important role in amino acid transport by acting as binding partner of amino acid transporter SL6A19 in intestine, regulating trafficking, expression on the cell surface, and its catalytic activity (By similarity). (Microbial infection) Acts as a receptor for SARS coronavirus/SARS-CoV. (Microbial infection) Acts as a receptor for Human coronavirus NL63/HCoV-NL63. The protein encoded by this gene belongs to the angiotensin-converting enzyme family of dipeptidyl carboxydipeptidases and has considerable homology to human angiotensin 1 converting enzyme. This secreted protein catalyzes the cleavage of angiotensin I into angiotensin 1-9, and angiotensin II into the vasodilator angiotensin 1-7. The organ- and cell-s
...more

Targets / Protein / Enzyme / Protease / Metalloprotease / Angiotensin-converting Enzyme 2

Protein Summary

Uniprot Accession IDs

Q9BYF1 C7ECU1 Q6UWP0 Q86WT0
Q9NRA7 Q9UFZ6

Gene Name

ACE2

Ensembl ID

ENST00000252519 ENSP00000252519
ENSG00000130234 ENST00000427411
ENSP00000389326

Symbol

ACEH

Illumination Graph

Knowledge Table

Most Knowledge About	Knowledge Value (0 to 1 scale)
virus perturbation	1
biological term	0.97
molecular function	0.86
biological process	0.78
disease	0.68

IDG Development Level Summary

Tdark

These are targets about which virtually nothing is known. They do not have known drug or small molecule activities
- AND -
satisfy two or more of the following criteria.

✓ Pubmed score: 1185.8 *(req: < 5)*
✓ Gene RIFs: 224 *(req: < 3)*
✓ Antibodies: 867 *(req: < 50)*

Tbio

These targets do not have known drug or small molecule activities
- AND -
do not satisfy the criteria for Tdark
- OR -
satisfy the following criterion:

✓ **Gene Ontology Terms: 23**

Tchem

Target has at least one ChEMBL compound with an activity cutoff of < 30 nM
- AND -
satisfies the preceding conditions

✓ **Active Ligands: 51**

Tclin

Target has at least one approved drug
- AND -
satisfies the preceding conditions

✓ **Active Drug: 0**

Figure 3. Target details view for ACE2, showing the new table of contents section (**A**), Target overview section (**B**) and TDL section (**C**).

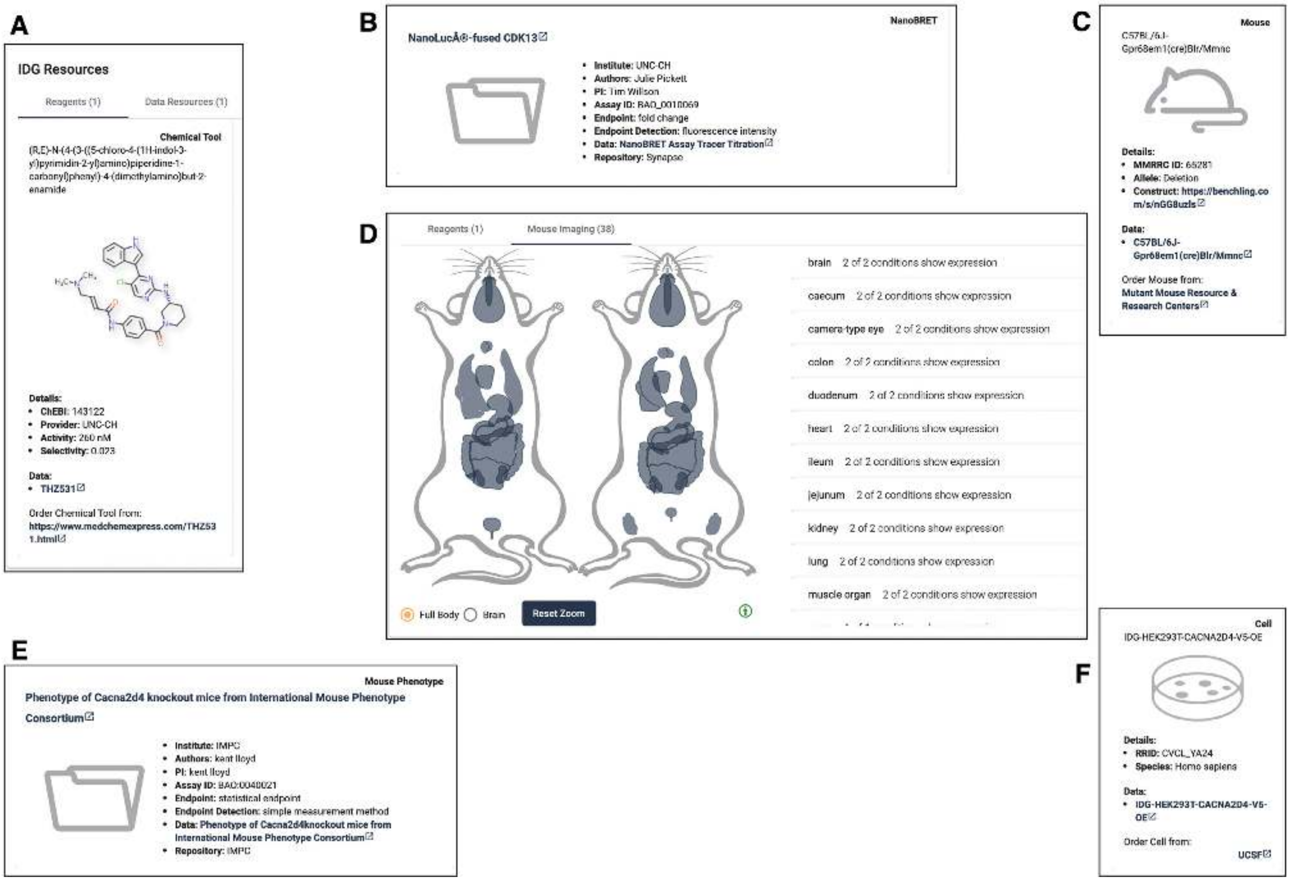
MATERIALS AND METHODS

The newly added data includes mouse and rat proteins from UniProt (5), with their associated phenotype data extracted from the International Mouse Phenotyping Consortium (6) and the Rat Genome Database (7), respectively. The Disease Ontology (DO) (8) data were further extended and newer ontologies such as Rat Disease Ontology (9) and Mammalian Phenotype Ontology (10) were added to facilitate comparison of target-disease and target-phenotype associations across multiple species. Similarly, additional data was included from GWAS (11) and OMIM (https://omim.org/) resources. The 'Disease details' page reflects these changes in addition to the other improvements made in displaying the list of associated targets. Throughout this manuscript, the term 'target' refers to '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.

Target expression data was primarily extracted from GTEx (12), the Human Protein Atlas (HPA) (13), UniProt and TISSUES (14). The GTEx dataset was further extended to include sex-specific expression values. In addition, cell line expression data was added from HPA and the Human Proteome Map (15) and expression data was collected for orthologous genes from 17 different species. Other expression datasets integrated into the latest version are the Cancer Cell Line Encyclopedia (16) and cell perturbation expression data from the Library of Integrated Network-Based Cellular Signatures (17). Furthermore, the target expression panel was visually upgraded with interactive anatomograms (18) (https://www.ebi.ac.uk/gxa) for both sexes which further provides systematic mappings to the source tissues.

Protein–protein interaction (PPI) is another data type that was included in the latest version by adding the most recent PPI data from STRING 11 (19). Given the COVID-19 pandemic, viral-human PPI data were added from P-HIPSTer (20) to help explore PPIs between viral pathogens and human proteins.

In addition to the new data from the aforementioned sources, TCRD has been continuously updated when newer versions of source databases (e.g. ChEMBL (21), DrugCentral (22), JensenLab PubMed scores (23) are released. As the



IDG Resources

Reagents (1) Data Resources (1)

Chemical Tool

(R,E)-N-(4-(3-((5-chloro-4-(1H-indol-3-yl)pyrimidin-2-yl)amino)piperidine-1-carbonyl)phenyl)-4-(dimethylamino)but-2-enamide

Details:

- ChEBI: 143122
- Provider: UNC-CH
- Activity: 260 nM
- Selectivity: 0.023

Data:

- THZ531

Order Chemical Tool from: https://www.medchemexpress.com/THZ531.html

NanoBRET

NanoLucÂ®-fused CDK13

- Institute: UNC-CH
- Authors: Julie Pickett
- PI: Tim Willson
- Assay ID: BAO_0010069
- Endpoint: fold change
- Endpoint Detection: fluorescence intensity
- Data: NanoBRET Assay Tracer Titration
- Repository: Synapse

Mouse

C57BL/6J-Gpr68em1(cre)Blr/Mmnc

Details:

- MMRRC ID: 65281
- Allele: Deletion
- Construct: https://benchling.com/s/nGG8uzls

Data:

- C57BL/6J-Gpr68em1(cre)Blr/Mmnc

Order Mouse from: Mutant Mouse Resource & Research Centers

Reagents (1) Mouse Imaging (38)

brain 2 of 2 conditions show expression

caecum 2 of 2 conditions show expression

camera-type eye 2 of 2 conditions show expression

colon 2 of 2 conditions show expression

duodenum 2 of 2 conditions show expression

heart 2 of 2 conditions show expression

ileum 2 of 2 conditions show expression

jejunum 2 of 2 conditions show expression

kidney 2 of 2 conditions show expression

lung 2 of 2 conditions show expression

muscle organ 2 of 2 conditions show expression

● Full Body ○ Brain Reset Zoom

Mouse Phenotype

Phenotype of Cacna2d4 knockout mice from International Mouse Phenotype Consortium

- Institute: IMPC
- Authors: kent lloyd
- PI: kent lloyd
- Assay ID: BAO:0040021
- Endpoint: statistical endpoint
- Endpoint Detection: simple measurement method
- Data: Phenotype of Cacna2d4knockout mice from International Mouse Phenotype Consortium
- Repository: IMPC

Cell

IDG-HEK293T-CACNA2D4-V5-OE

Details:

- RRID: CVCL_YA24
- Species: Homo sapiens

Data:

- IDG-HEK293T-CACNA2D4-V5-OE

Order Cell from: UCSF

Figure 4. IDG generated resources for ACE2, consisting of small molecule reagents (**A**), and data (**B**). Targets with mouse cell lines, such as GPR68, also have information about that resource (**C**) as well as a mouse tissue expression viewer (**D**). Other data can include, such as the case for CACNA2D4, mouse phenotype (**E**) and cell line (**F**) data. Supplementary Table S2 contains a full breakdown of data types and fields collected.

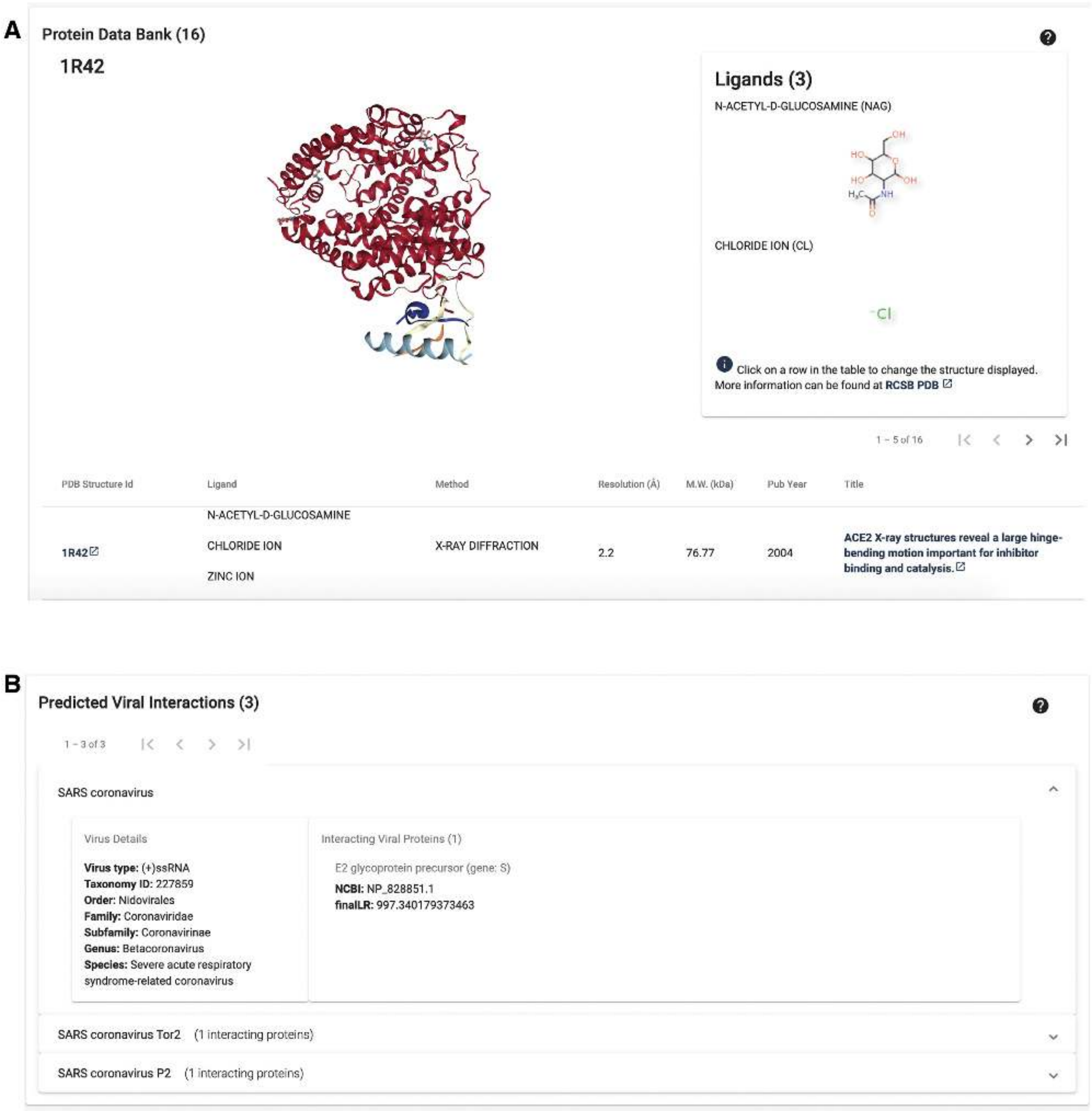
amount of data available for each target changed with each release of TCRD, the target development levels (TDLs) for respective targets were recalculated when the target criteria changed. This was done using automated scripts. Very briefly, the TDL is one of four potential values: Tclin, Tchem, Tbio or Tdark. **Tclin** are protein drug targets via which approved drugs act (24–26), which currently includes 659 human proteins; **Tchem** are proteins that are not **Tclin**, but are known to bind small molecules with high potency (currently $N = 1607$); **Tbio** includes proteins that have Gene Ontology (27) 'leaf' (lowest level) term annotations based on experimental evidence; or meet two of the following three conditions: A fractional publication count (28) above 5, three or more Gene RIF, 'Reference Into Function' annotations (https://www.ncbi.nlm.nih.gov/gene/about-generif), or 50 or more commercial antibodies, as counted in the Antibodypedia portal (29). 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.

For a further in depth exploration of each additional TCRD dataset and its database changes, we direct the reader to Supplementary Materials.

RESULTS

In the first 2 years following the initial release, the Pharos team frequently demoed the site, with a focus on obtaining user feedback, which provided the user-centered design information needed as Pharos underwent a ground-up rewrite in late 2018. By focusing on user needs, such as improving the target details page hierarchy and navigation, and adding more detailed explanations of the terms and ideas represented within Pharos, we were able to streamline target presentation, and simplify many pages. Many tables were swapped with sortable list elements, akin to a shopping site, allowing more data to be shown than a table would allow and also allowing us to display dynamic data as desired. Pharos was also redesigned with a focus on mobile usability, adjusting styles and visualizations depending on the screen used. Pharos is also installable as a Progressive Web App, which allows users to install Pharos as a mobile application on their device, increasing the ease of access for Pharos.

To increase speed and responsiveness, we switched from a fully server-side rendered application to a hybrid client and server-side application, allowing faster page rendering and data retrieval. We replaced the REST API with a GraphQL instance, which adds flexibility to data retrieval, as well as potentially reducing the amount of data being sent over the network. A discussion of GraphQL bench-



Protein Data Bank (16)

1R42

Ligands (3)

N-ACETYL-D-GLUCOSAMINE (NAG)

CHLORIDE ION (CL)

ⓘ Click on a row in the table to change the structure displayed. More information can be found at **RCSB PDB**

1 – 5 of 16

PDB Structure Id	Ligand	Method	Resolution (Å)	M.W. (kDa)	Pub Year	Title
1R42	N-ACETYL-D-GLUCOSAMINE CHLORIDE ION ZINC ION	X-RAY DIFFRACTION	2.2	76.77	2004	**ACE2 X-ray structures reveal a large hinge-bending motion important for inhibitor binding and catalysis.**

Predicted Viral Interactions (3)

1 – 3 of 3

SARS coronavirus

Virus Details

Virus type: (+)ssRNA
Taxonomy ID: 227859
Order: Nidovirales
Family: Coronaviridae
Subfamily: Coronavirinae
Genus: Betacoronavirus
Species: Severe acute respiratory syndrome-related coronavirus

Interacting Viral Proteins (1)

E2 glycoprotein precursor (gene: S)
NCBI: NP_828851.1
finalLR: 997.340179373463

SARS coronavirus Tor2 (1 interacting proteins)

SARS coronavirus P2 (1 interacting proteins)

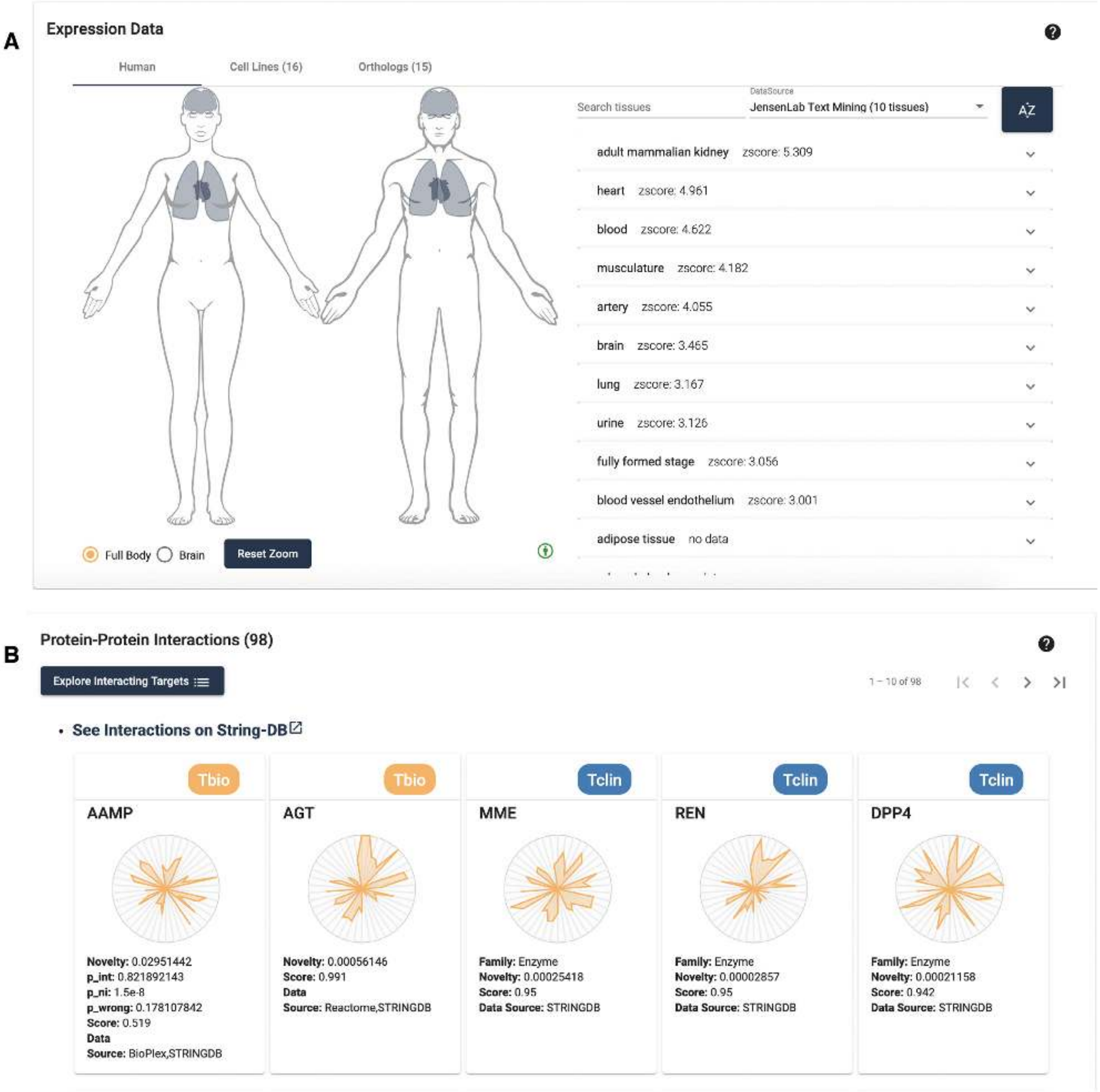
Figure 5. Target details view for ACE2, showing the improved Protein Data Bank data viewer (**A**) and predicted viral interactions (**B**).

marking is summarized in Supplementary Table S1. Documentation for the GraphQL format, as well as an interactive sandbox with several sample queries can be found at https://pharos.nih.gov/api.

Browse pages

Filter functionality has been expanded, and users are able to examine the entire list of values for each filter, as well as search for text within the filter values. In the case of numeric values, such as novelty and PubMed score, a range slider allows users to refine results. All filters also have a '?' help button that allows users to view a quick definition of the filter or to visit the original source, if desired. These features are displayed in Figure 2. In addition, sub-lists are generated on entity (target/disease/ligand) pages that can be viewed in their respective list browsers. It is therefore possible to browse the list of disease or ligands associated with a specific target, or targets associated with a disease or ligand. For associated diseases, additional numeric filters such as association score and interaction scores are available, as well as ligand affinity measurements for associated ligand lists.

By registering, users are able to select targets of interest and save them as custom lists, which are added to the main filter panel, and are always available. This allows users to



adult mammalian kidney zscore: 5.309
heart zscore: 4.961
blood zscore: 4.622
musculature zscore: 4.182
artery zscore: 4.055
brain zscore: 3.465
lung zscore: 3.167
urine zscore: 3.126
fully formed stage zscore: 3.056
blood vessel endothelium zscore: 3.001
adipose tissue no data

AAMP
Novelty: 0.02951442
p_int: 0.821892143
p_ni: 1.5e-8
p_wrong: 0.178107842
Score: 0.519
Data Source: BioPlex,STRINGDB

AGT
Novelty: 0.00056146
Score: 0.991
Data Source: Reactome,STRINGDB

MME
Family: Enzyme
Novelty: 0.00025418
Score: 0.95
Data Source: STRINGDB

REN
Family: Enzyme
Novelty: 0.00002857
Score: 0.95
Data Source: STRINGDB

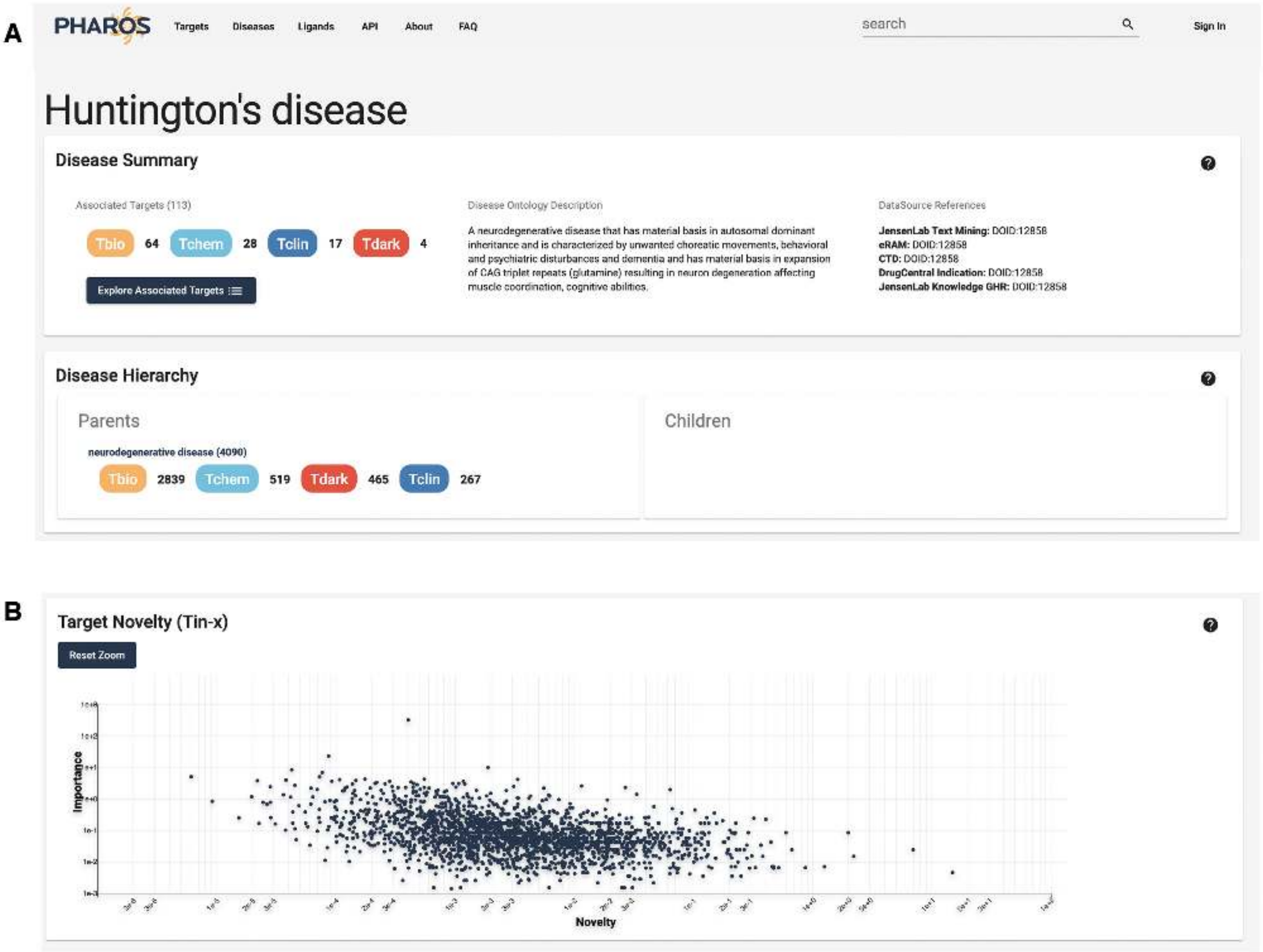
DPP4
Family: Enzyme
Novelty: 0.00021158
Score: 0.942
Data Source: STRINGDB

Figure 6. Target details view for ACE2, showing the tissue expression section, with highlighted tissues (**A**), and protein to protein interaction section (**B**). The frequency of updating for resources integrated in TCRD differs. Text-mined sources are updated more frequently, and changes in the scientific literature allow us to track certain associations sooner. Therefore, selecting one of these sources (as displayed) will show that ACE2 is expressed in the lungs.

further refine a filtered list of targets, as well as view how this new list may be further filtered, or examine the makeup of the list by the various filter categories.

Target details pages

In improving the target details pages, we relied on user interviews that we had conducted to determine the optimal layout and ordering of data on the page. In the newest version of Pharos, target detail pages start off with synonyms, identifiers and a broad overview of the knowledge about the target, as shown in Figure 3. Next, the user is able to examine the TDL criteria in relation to the specific target. Drug and ligand data comes next, followed by disease association data. Target expression and interaction data follows, with publication data, amino acid sequence information – including the ProtVista (30) graphical representation, and filters for related targets completing the page. Sections where no data are available, such as drug and ligand browsers for **Tdark** and **Tbio** targets, are automatically removed. The navigation panel is always visible on the left hand side, allowing users to easily jump to the section of interest. Most data panels have been improved by adding tooltips to further explain properties, and a help panel that contains addi-



A neurodegenerative disease that has material basis in autosomal dominant inheritance and is characterized by unwanted choreatic movements, behavioral and psychiatric disturbances and dementia and has material basis in expansion of CAG triplet repeats (glutamine) resulting in neuron degeneration affecting muscle coordination, cognitive abilities.

JensenLab Text Mining: DOID:12858
eRAM: DOID:12858
CTD: DOID:12858
DrugCentral Indication: DOID:12858
JensenLab Knowledge GHR: DOID:12858

Figure 7. Disease details page for Huntington's disease, showing Disease Ontology description and hierarchy (**A**), and TIN-X plot showing novelty targets and their importance, mapped with a log scale (**B**).

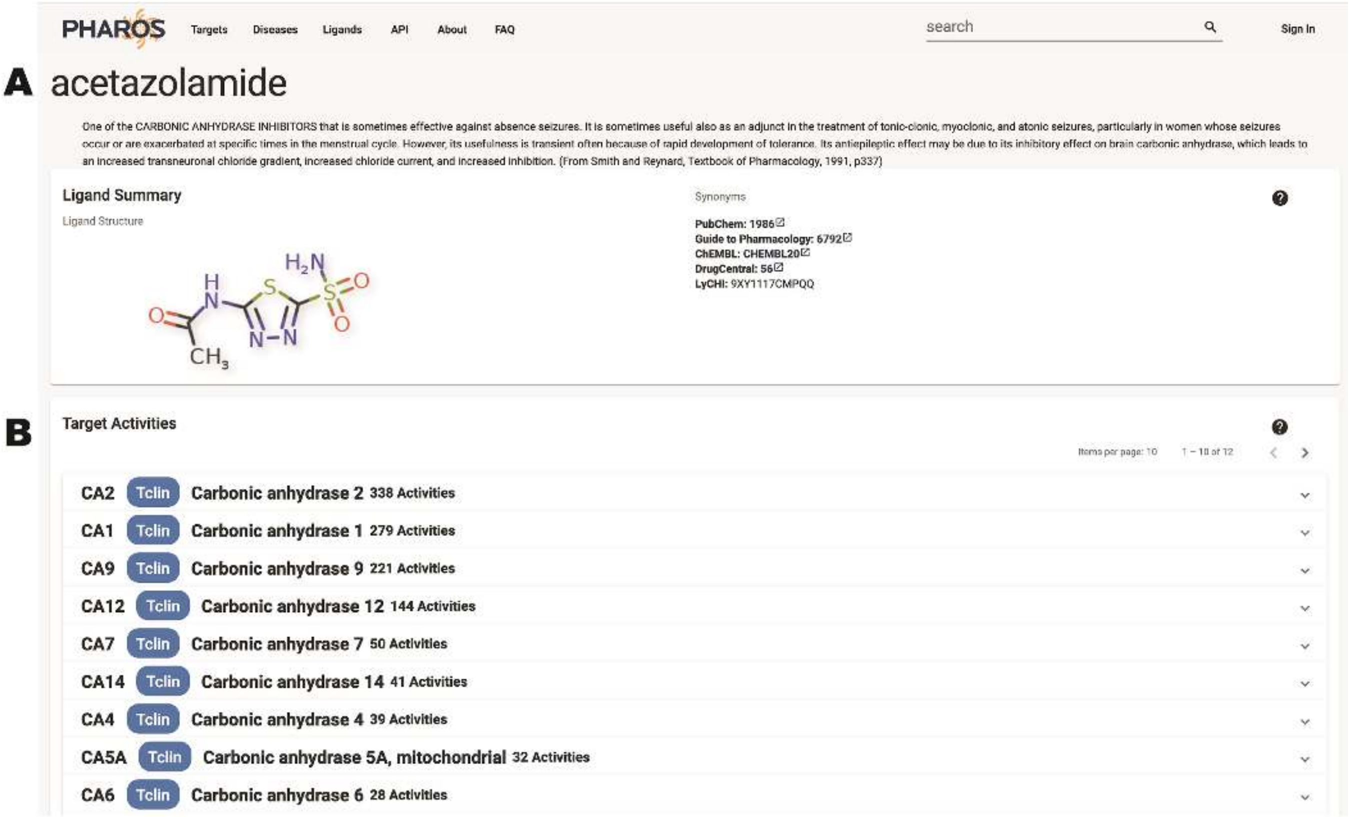
tional information, such as definitions, explanatory articles or raw data. Additional panels have been added to display additional data.

TDL descriptions. One addition to the previous Pharos UI is a description panel. While the various TDL criteria are well documented, it was not apparent to users how, and which TDL criteria are met by a given target. This section displays which criteria have been met, as well as the level or score. A highlighted checkbox indicates which TDL criteria have been met. This provides valuable evidence that illustrates to the user how TCRD generated this ranking. For **Tdark** and **Tbio**, this is also a useful indicator of what criteria are still deficient, and how close to the cutoff they are, potentially driving research in those areas. This panel is also featured in Figure 3.

IDG resources. A feature of the IDG program that has expanded since the initial publication is that of new data or reagents, which are generated by IDG Consortium members. Where data or reagents are available, a browseable section is displayed in Pharos, allowing users to navigate to the data set, and even order them in the case of physical resources, as shown in Figure 4. For targets with mouse expression data, we have incorporated our anatamogram image (discussed below) using male/female and brain mouse images. We refer the reader to Supplementary Materials for an in depth discussion of IDG Consortium resources available in Pharos.

Improved ligand/drug section. Drug and ligand browsing is now done via a pageable list. Users can now page through the entire list of approved drugs or active ligands, or open the entire list in the browse view, allowing for filtering of these lists, while specifying the number of targets each ligand is active on.

Disease associations. The disease associations panel has been greatly expanded, both in the amount of featured resources, now incorporating DisGeNet (31) and eRAM (32), as well as in size and placement. The source of each disease association is displayed as a collapsible panel that bundles multiple sources for the same disease. Upon expanding this panel, users are able to examine the evidence used to make this association. This allows users to discover provenance, an essential attribute of data aggregated in TCRD.



acetazolamide

One of the CARBONIC ANHYDRASE INHIBITORS that is sometimes effective against absence seizures. It is sometimes useful also as an adjunct in the treatment of tonic-clonic, myoclonic, and atonic seizures, particularly in women whose seizures occur or are exacerbated at specific times in the menstrual cycle. However, its usefulness is transient often because of rapid development of tolerance. Its antiepileptic effect may be due to its inhibitory effect on brain carbonic anhydrase, which leads to an increased transneuronal chloride gradient, increased chloride current, and increased inhibition. (From Smith and Reynard, Textbook of Pharmacology, 1991, p337)

PubChem: 1986
Guide to Pharmacology: 6792
ChEMBL: CHEMBL20
DrugCentral: 56
LyCHI: 9XY1117CMPQQ

Items per page: 10 1 – 10 of 12

CA2	Tclin	Carbonic anhydrase 2	338 Activities
CA1	Tclin	Carbonic anhydrase 1	279 Activities
CA9	Tclin	Carbonic anhydrase 9	221 Activities
CA12	Tclin	Carbonic anhydrase 12	144 Activities
CA7	Tclin	Carbonic anhydrase 7	50 Activities
CA14	Tclin	Carbonic anhydrase 14	41 Activities
CA4	Tclin	Carbonic anhydrase 4	39 Activities
CA5A	Tclin	Carbonic anhydrase 5A, mitochondrial	32 Activities
CA6	Tclin	Carbonic anhydrase 6	28 Activities

Figure 8. Ligand details page for acetazolamide, shows ligand description information as well as synonyms and identifiers (**A**). Target activity is shown with expanded activity for CA2 (**B**).

It is also possible to view the list of associated diseases as a browseable list, again, allowing for filtering and more refined data analysis.

PDB visualizations. While PDB identifiers have always been available in Pharos, they were previously displayed as a list of linkouts with no additional information (33). In Pharos 3.0, the PDB section was expanded and redesigned to display three-dimensional (3D) structures for proteins, including bound ligands where available. This 3D visualization, employing NGL Viewer (34,35), is highly interactive, allowing users to drag, zoom, pan and spin the structure to thoroughly examine it. This interface also enables users to browse the PDB identifiers and structures associated with a given target. Clicking on a table row loads the respective PDB entry, and displays the ligands shown in the structure. Other details include the reference source for the id, as well as the method used to generate the structure. This feature is highlighted in Figure 5.

Predicted viral interactions. Predicted viral-human protein interactions from the P-HIPSTer atlas are displayed in a separate panel that shows a list of viruses and their associated taxonomic data, specifically for the targeted human protein. Each virus also lists the viral proteins that the human protein is predicted to interact with, in addition to a likelihood ratio, which measures the strength of the sequence- and structure-based prediction. These features are shown in Figure 5, using ACE2, the angiotensin converting enzyme 2, as example. ACE2 is the functional entry receptor for both severe acute respiratory syndrome (SARS) coronaviruses, SARS-CoV (36) and SARS-CoV-2 (37).

Target expression data. The target expression data panel has also been expanded, as shown in Figure 6. We replaced the static human figure with an expanded, interactive anatomogram (18). This visualization contains both male and female figures, allowing sex-specific tissue information (e.g., from GTEx) to be highlighted. Users can toggle the brain image separately, for more fine-grained brain tissue expression visualization. All figures are zoomable and pannable, to improve visibility. When possible, all source tissues have been mapped to standardized Uberon (38) identifiers, enabling us to merge repetitive tissues. Expression data can be viewed by source, and tissue values are searchable. Clicking on a tissue panel, like the disease associations section, allows users to view detailed provenance of the data, both showing the source and relevant confidence values for the tissue expression. Gradient coloration allows us to display the tissue expression level, either a qualitative value (low medium, high), or a numeric value scale, depending on the data source, giving users a quick visualization of this data. We note that several sources have an overlap in data and content; however, they may also present unique data. Pharos displays the variety between sources, which can help users decide what sources to focus on. It also illustrates that the text-mining-based resources tend to be updated more

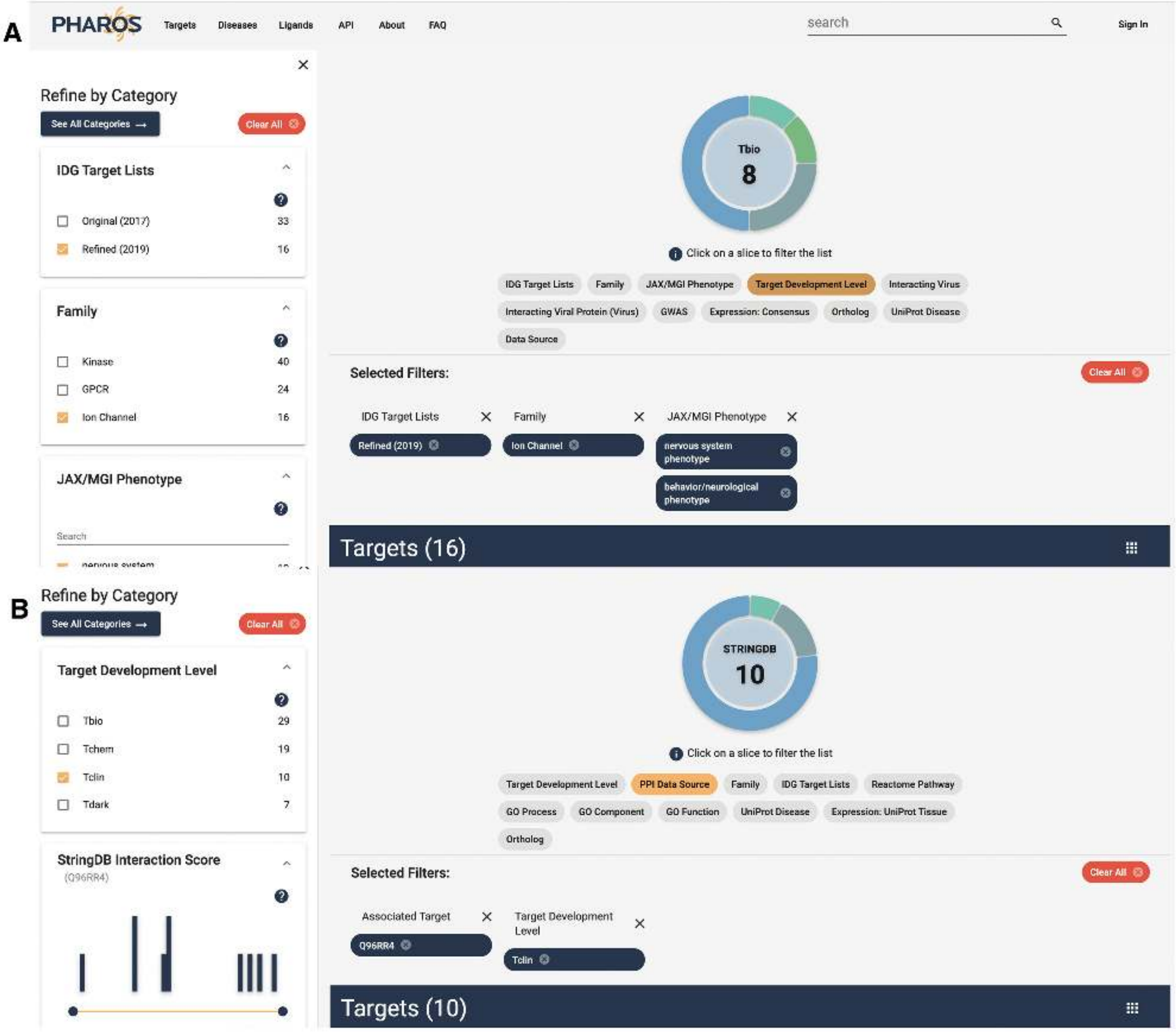


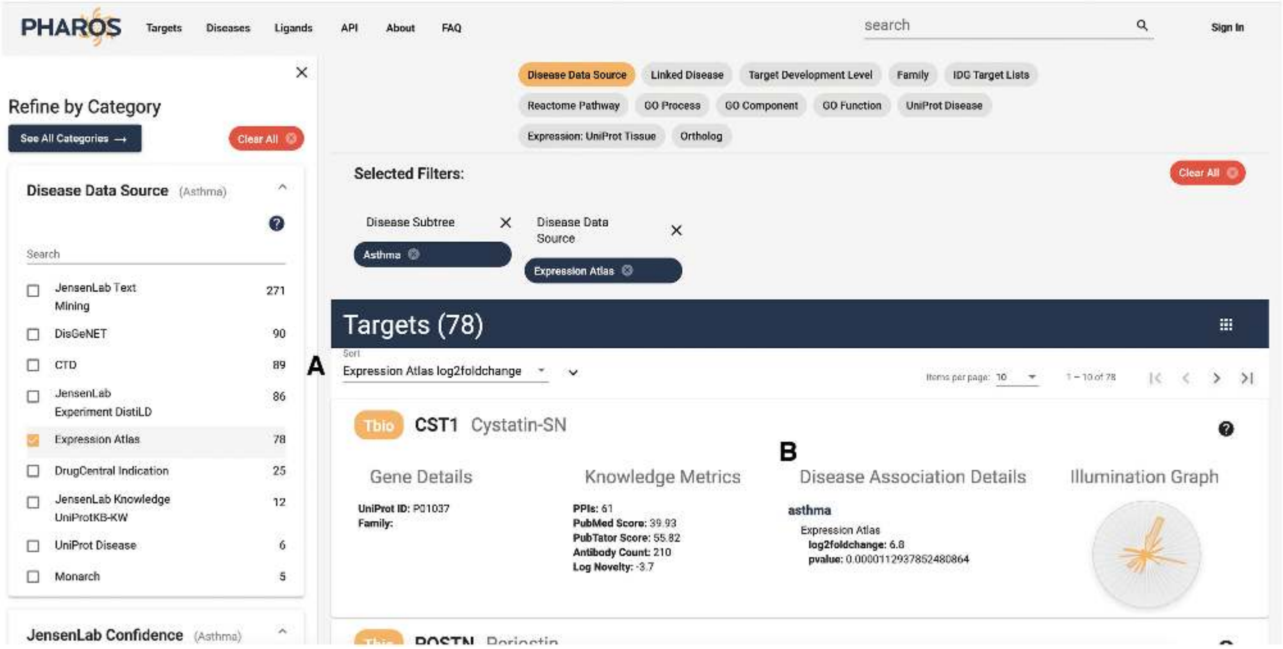
Figure 9. Target list described in *Target List Search* use case (**A**), and target list described in *Binding partner search* use case (**B**). Both panels show selected filters, including the ability to filter target lists by IDG featured sub-lists.

often than frequently accessed scientific databases that aggregate experimental data. When available, cell lines and orthologs are displayed with expandable evidence panels.

Protein–protein interactions. The introduction of data from Bioplex (39), STRING and Reactome (40) has allowed us to display rich PPI data, and allows users to examine how multiple targets may interact, as shown in Figure 6. This panel consists of a pageable list of targets known to interact with the current target. These targets display the illumination graph, which gives users a quick visual representation of the level and types of knowledge available for the target. The originating data source is also displayed, as well as the confidence of this interaction. This list of PPIs may be viewed either directly in STRING, or viewed in the browse page as a filterable list.

Disease details pages

While Pharos remains target-centric, it is equally important to be able to examine diseases in relation to their associated targets. As such, Pharos has always included diseases as a browseable list, as well as a disease details page. Pharos 3.0 expands on the data collected and displayed about diseases. When available, a description from the Disease Ontology is shown. The hierarchy of each specific disease is also shown, allowing users to view parent or child diseases or syndromes, as well as the siblings for each disease. In addition, a breakdown of the targets associated with each disease is shown, which can also be opened in the target browse page, allowing users to filter the list of associated targets. The target-disease Novelty score is shown in the Target-Importance Novelty eXplorer (TIN-X, (41)) panel. Similar to the target based TIN-X view, this is a scatterplot of novel targets related to the disease of interest. *Nov-*



Disease Data Source (Asthma)

Search

☐	JensenLab Text Mining	271
☐	DisGeNET	90
☐	CTD	89
☐	JensenLab Experiment DistiLD	86
☑	Expression Atlas	78
☐	DrugCentral Indication	25
☐	JensenLab Knowledge UniProtKB-KW	12
☐	UniProt Disease	6
☐	Monarch	5

JensenLab Confidence (Asthma)

Disease Subtree: Asthma

Disease Data Source: Expression Atlas

Targets (78)

Sort: Expression Atlas log2foldchange

Items per page: 10 1 – 10 of 78

Tbio **CST1** Cystatin-SN

Gene Details
UniProt ID: P01037
Family:

Knowledge Metrics
PPIs: 61
PubMed Score: 39.93
PubTator Score: 55.82
Antibody Count: 210
Log Novelty: -3.7

Disease Association Details
asthma
Expression Atlas
log2foldchange: 6.8
pvalue: 0.0000112937852480864

Illumination Graph

Tbio POSTN Periostin

Figure 10. List of targets associated with asthma, filtered by Data Source: Expression Atlas, and filtered by Expression Atlas log2foldchange value (**A**). Also shown is the Disease Association Details column, with additional association-specific details (**B**).

elty estimates the scarcity of publications about a target, whereas *Importance* estimates the strength of the association between that target and a specific disease (42). The *X*-axis shows the Novelty of the target-disease relationship, while the *Y*-axis shows the Importance of that target to the disease, both on a $\log_{10}$ scale. This feature is displayed in Figure 7. When available, help definition buttons are also shown for each section of data.

Ligand details pages

The ligand details page has been expanded to include a description of the ligand, where available. A 2D depiction of the ligand chemical structure is shown, together with synonyms and identifiers from PubChem (43), Guide to Pharmacology (44), ChEMBL and DrugCentral. Activity values are displayed in collapsible panels, and sorted by target. By clicking on a target header, users are able to examine activity values for that ligand on the selected target. Activity type, value and mechanism of action are shown, in addition to reference publication for each activity, allowing users to directly view the source. Figure 8 illustrates this page display.

Use cases

Target list search. A neuroscientist who studies ion channels is seeking understudied proteins for a student project. Filtering to targets listed by the IDG Consortium as targets of interest (https://druggablegenome.net/IDGProteinList), in the ion channel family, with a nervous system phenotype as described by the Mouse Phenotype Ontology from the Jackson Laboratory yields 16 potential targets of interest. The refined list view is shown in Figure 9.

Binding partner search. A scientist is studying the function of CAMKK2, and hypothesizing on its role in behavior. From the CAMKK2 target details page, she clicks 'Explore Interacting Targets' to open the potential binding partners in the Target List. Selecting **Tclin**, she filters the list to 10 potential binding partners that have approved drugs, with new therapeutic effects and drug-related side effects that can assist in her research. She is also able to save this list of targets as a custom list, allowing her to easily revisit and analyze these targets. This workflow is also shown in Figure 9.

Disease-based search. A researcher studying asthma searches for this top level disease term and enters the asthma disease details page. He wants to find proteins that are differentially expressed in asthma patients. Starting from the Asthma disease details page, he clicks 'Explore Associated Targets' to see 470 potential targets. Filtering those with data from the Expression Atlas (18), using the 'Expression Atlas' value from the 'Disease Data Source' filter, yields 78 targets, which can then be sorted by $\log_2$foldChange using the drop-down sort feature, as shown above the target list in Figure 10. He finds several with high $\log_2$foldChange, meaning the expression level changes drastically between disease and non-disease conditions, and low p-value, meaning the results are statistically significant. Also of note, for lists of associated targets, an additional column has been added to the target card, showing disease association details, such as the evidence associated with a given data source.

Finding experimental tools. A researcher has found that a Voltage-gated potassium channel, KCNS2, is expressed in the nucleus they are working on. It's role is unknown. The target details page for KCNS2 lists three approved drugs (e.g. amifampridine) which are known to block this channel. The IDG Resources show a cell line and a mouse genetic construct, which can be ordered from the IDG Consortium (UCSF, in this case). The researcher is now able to design a set of *in vitro* and *in vivo* experiments to help elucidate the role of the channel.

Machine learning based on PPI networks to prioritize targets for a disease. A novel target ranking approach could be developed that relies on deep network representation learning of a PPI network annotated with disease-specific knowledge attributes derived from TCRD/Pharos. This involves mapping the enriched PPI network into a feature space using the neural network framework, Gat2Vec (45). Gat2Vec employs a shallow neural network model to facilitate joint learning on the structural and attribute contexts of a given network. In this case, the PPIs provide the structural context and the disease-related knowledge provides the attribute contexts. The feature space generated can be used to develop machine learning models that can predict the ranking of protein targets in the context of a disease. Additional protocols for extracting available data for specific targets of interest, as well the differences in knowledge availability between understudied and more studied targets, are discussed elsewhere (46).

DISCUSSION

With the newly added data, interactive visualizations and enhanced search capabilities, TCRD and Pharos in their current forms serve as IDG resources that facilitate better exploration of the dark and understudied regions of the human genome. The central idea of the resource continues to be enriching knowledge around human targets and monitoring their therapeutic development levels. One area that received significant attention in the current update is the search mechanism which was upgraded with the implementation of GraphQL API. To improve user interaction with the new API, example queries have been made available on the website. The TDL descriptions panel and the target expression anatomograms are the significantly improved components of the Pharos web interface. Further, most data that have been recently integrated into TCRD facilitate application of artificial intelligence/machine learning (AI/ML) techniques for target evaluation and drug repositioning (47), and are accessible in ML ready format. Most of the above described developments are based on the feedback obtained directly from the website visitors, several demo sessions and presentations at scientific meetings.

Ongoing work focuses on utilizing the AI/ML-ready data such as target-disease, target-phenotype and PPI data to develop AI/ML models for prioritization of dark targets and better understanding of disease biology. Currently, we are evaluating ways to aggregate experimental data uncertainties, specifically data quality and reliability (48), similar to the *in silico* toxicology protocols (49). We continue to extend our efforts in enhancing the disease page to list associated targets, rank them by the consensus strength of association to a disease and further facilitate filtering to narrow down to targets of potential interest to researchers. The database can be expected to be further enriched with new data and data types that contribute to the ultimate goal of illuminating the dark targets.

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

Rajarshi Guha, for his immeasurable time spent not only working on and promoting Pharos 1.0, but also for his mentorship as we began the process of revamping Pharos. Oleg Ursu, for developing the first machine learning ready version of TCRD. Gorka Lasso, for supplying the P-HIPSTer virus–human protein interactions dataset.

FUNDING

National Institutes of Health (NIH) Common Fund [CA224370 to S.L.M., C.G.B., L.L.J., A.W., G.B., J.J.Y., T.I.O., U24TR002278 to D.V., A.K., J.H., A.W., S.C.S. and T.I.O.]; Novo Nordisk Foundation [NNF14CC0001 to L.J.J.]; Intramural Research Program, Division of Preclinical Innovation, NIH NCATS (to D.T.N., K.K., T.S., N.S., P.D., E.W., A.S.). Funding for open access charge: NIHCA224370.

Conflict of interest statement. L.J.J. is co-founder and scientific advisory board member of Intomics A/S. 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 is on the scientific advisory board of ChemDiv Inc. and InSilico Medicine.

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. 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.

3. Oprea,T.I., Bologa,C.G., Brunak,S., Campbell,A., Gan,G.N., Gaulton,A., Gomez,S.M., Guha,R., Hersey,A., Holmes,J. *et al.* (2018) Unexplored therapeutic opportunities in the human genome. *Nat. Rev. Drug Discov.*, **17**, 317–332.
4. Galperin,M.Y., Fernández-Suárez,X.M. and Rigden,D.J. (2017) The 24th annual Nucleic Acids Research database issue: a look back and upcoming changes. *Nucleic Acids Res.*, **45**, D1–D11.
5. Consortium,U.P. (2019) UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.*, **47**, D506–D515.
6. 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.
7. Smith,J.R., Hayman,G.T., Wang,S.-J., Laulederkind,S.J.F., Hoffman,M.J., Kaldunski,M.L., Tutaj,M., Thota,J., Nalabolu,H.S., Ellanki,S.L.R. *et al.* (2020) The Year of the Rat: The Rat Genome Database at 20: a multi-species knowledgebase and analysis platform. *Nucleic Acids Res.*, **48**, D731–D742.
8. 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–D946.
9. Hayman,G.T., Laulederkind,S.J.F., Smith,J.R., Wang,S.-J., Petri,V., Nigam,R., Tutaj,M., De Pons,J., Dwinell,M.R. and Shimoyama,M. (2016) The Disease Portals, disease-gene annotation and the RGD disease ontology at the Rat Genome Database. *Database*, **2016**, baw034.
10. Smith,C.L. and Eppig,J.T. (2009) The mammalian phenotype ontology: enabling robust annotation and comparative analysis. *Wiley Interdiscip. Rev. Syst. Biol. Med.*, **1**, 390–399.
11. 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.
12. Consortium, GTEx (2020) The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science*, **369**, 1318–1330.
13. Thul,P.J. and Lindskog,C. (2018) The human protein atlas: a spatial map of the human proteome. *Protein Sci.*, **27**, 233–244.
14. 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.
15. 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.
16. Barretina,J., Caponigro,G., Stransky,N., Venkatesan,K., Margolin,A.A., Kim,S., Wilson,C.J., Lehár,J., Kryukov,G.V., Sonkin,D. *et al.* (2012) The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, **483**, 603–607.
17. Stathias,V., Turner,J., Koleti,A., Vidovic,D., Cooper,D., Fazel-Najafabadi,M., Pilarczyk,M., Terryn,R., Chung,C., Umeano,A. *et al.* (2020) LINCS Data Portal 2.0: next generation access point for perturbation-response signatures. *Nucleic Acids Res.*, **48**, D431–D439.
18. Papatheodorou,I., Moreno,P., Manning,J., Fuentes,A.M.-P., George,N., Fexova,S., Fonseca,N.A., Füllgrabe,A., Green,M., Huang,N. *et al.* (2020) Expression Atlas update: from tissues to single cells. *Nucleic Acids Res.*, **48**, D77–D83.
19. 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.
20. Lasso,G., Mayer,S.V., Winkelmann,E.R., Chu,T., Elliot,O., Patino-Galindo,J.A., Park,K., Rabadan,R., Honig,B. and Shapira,S.D. (2019) A structure-informed atlas of human-virus interactions. *Cell*, **178**, 1526–1541.
21. 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.
22. 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.
23. Pletscher-Frankild,S., Pallejà,A., Tsafou,K., Binder,J.X. and Jensen,L.J. (2015) DISEASES: text mining and data integration of disease-gene associations. *Methods*, **74**, 83–89.
24. 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.
25. Ursu,O., Glick,M. and Oprea,T. (2019) Novel drug targets in 2018. *Nat. Rev. Drug Discov.*, **18**, 328.
26. Avram,S., Halip,L., Curpan,R. and Oprea,T.I. (2020) Novel drug targets in 2019. *Nat. Rev. Drug Discov.*, **19**, 300.
27. 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.
28. Pafilis,E., Frankild,S.P., Fanini,L., Faulwetter,S., Pavloudi,C., Vasileiadou,A., Arvanitidis,C. and Jensen,L.J. (2013) The SPECIES and ORGANISMS resources for fast and accurate identification of taxonomic names in text. *PLoS One*, **8**, e65390.
29. Björling,E. and Uhlén,M. (2008) Antibodypedia, a portal for sharing antibody and antigen validation data. *Mol. Cell. Proteomics*, **7**, 2028–2037.
30. Watkins,X., Garcia,L.J., Pundir,S., Martin,M.J. and Consortium,U.P. (2017) ProtVista: visualization of protein sequence annotations. *Bioinformatics*, **33**, 2040–2041.
31. Piñero,J., Ramírez-Anguita,J.M., Saüch-Pitarch,J., Ronzano,F., Centeno,E., Sanz,F. and Furlong,L.I. (2020) The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Res.*, **48**, D845–D855.
32. Jia,J., An,Z., Ming,Y., Guo,Y., Li,W., Liang,Y., Guo,D., Li,X., Tai,J., Chen,G. *et al.* (2018) eRAM: encyclopedia of rare disease annotations for precision medicine. *Nucleic Acids Res.*, **46**, D937–D943.
33. 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.
34. Rose,A.S., Bradley,A.R., Valasatava,Y., Duarte,J.M., Prlic,A. and Rose,P.W. (2018) NGL viewer: web-based molecular graphics for large complexes. *Bioinformatics*, **34**, 3755–3758.
35. Rose,A.S. and Hildebrand,P.W. (2015) NGL Viewer: a web application for molecular visualization. *Nucleic Acids Res.*, **43**, W576–W579.
36. Li,W., Moore,M.J., Vasilieva,N., Sui,J., Wong,S.K., Berne,M.A., Somasundaran,M., Sullivan,J.L., Luzuriaga,K., Greenough,T.C. *et al.* (2003) Angiotensin-converting enzyme 2 is a functional receptor for the SARS coronavirus. *Nature*, **426**, 450–454.
37. Hoffmann,M., Kleine-Weber,H., Schroeder,S., Krüger,N., Herrler,T., Erichsen,S., Schiergens,T.S., Herrler,G., Wu,N.-H., Nitsche,A. *et al.* (2020) SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor. *Cell*, **181**, 271–280.
38. 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.
39. Huttlin,E.L., Bruckner,R.J., Navarrete-Perea,J., Cannon,J.R., Baltier,K., Gebreab,F., Gygi,M.P., Thornock,A., Zarraga,G., Tam,S. *et al.* (2020) Dual proteome-scale networks reveal cell-specific remodeling of the human interactome. bioRxiv doi: https://doi.org/10.1101/2020.01.19.905109, 19 January 2020, preprint: not peer reviewed.
40. 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.
41. 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.
42. Oprea,T.I. (2019) Exploring the dark genome: implications for precision medicine. *Mamm. Genome*, **30**, 192–200.
43. Kim,S., Chen,J., Cheng,T., Gindulyte,A., He,J., He,S., Li,Q., Shoemaker,B.A., Thiessen,P.A., Yu,B. *et al.* (2019) PubChem 2019

update: improved access to chemical data. *Nucleic Acids Res.*, **47**, D1102–D1109.

44. Armstrong,J.F., Faccenda,E., Harding,S.D., Pawson,A.J., Southan,C., Sharman,J.L., Campo,B., Cavanagh,D.R., Alexander,S.P.H., Davenport,A.P. *et al.* (2020) The IUPHAR/BPS guide to Pharmacology in 2020: extending immunopharmacology content and introducing the IUPHAR/MMV guide to Malaria Pharmacology. *Nucleic Acids Res.*, **48**, D1006–D1021.
45. Sheikh,N., Kefato,Z. and Montresor,A. (2019) gat2vec: representation learning for attributed graphs. *Computing*, **101**, 187–209.
46. 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.
47. Levin,J.M., Oprea,T.I., Davidovich,S., Clozel,T., Overington,J.P., Vanhaelen,Q., Cantor,C.R., Bischof,E. and Zhavoronkov,A. (2020) Artificial intelligence, drug repurposing and peer review. *Nat. Biotechnol.* **38**, 1127–1131.
48. Klimisch,H.J., Andreae,M. and Tillmann,U. (1997) A systematic approach for evaluating the quality of experimental toxicological and ecotoxicological data. *Regul. Toxicol. Pharmacol.*, **25**, 1–5.
49. Myatt,G.J., Ahlberg,E., Akahori,Y., Allen,D., Amberg,A., Anger,L.T., Aptula,A., Auerbach,S., Beilke,L., Bellion,P. *et al.* (2018) In silico toxicology protocols. *Regul. Toxicol. Pharmacol.*, **96**, 1–17.