The main goal of the Knowledge Management Center (KMC) for the Illuminating the Druggable Genome (IDG) program is to aggregate, update and articulate protein-centric data, information and knowledge for the entire human proteome with emphasis on understudied proteins from the 3 families that are the focus of the IDG ("IDG List"). The long-term objective of the KMC is to encourage and support biomedical research aimed at understudied proteins by providing an extensive resource of data, information, knowledge, methods and reagents for the entire human proteome, and to support the growing online community focused on understudied proteins. With focus on the IDG List and human proteins, the KMC will enable support for expanded coverage for non-human proteins of therapeutic interest and other associated human health data, in order to catalyze novel biomedical discoveries. To support the overall IDG objective, and to maintain, update and improve these integrated resources, the KMC draws upon expertise from multiple knowledge domains, specifically biology, chemistry and medicine, as well as computer science, graphic design and web programming. Specifically, for the Phase 2 of the IDG KMC we propose 4 Aims:1. Create an automated workflow that captures relevant public data for the entire proteome and manual annotations for the IDG list. The KMC knowledge management system will be built around knowledge graphs, focused on five major branches of the target knowledge tree, tkt: *Genotype, Phenotype, Expression, Structure & Function*, and *Interactions & Pathways*, respectively. Aim 2: Design, develop and implement a protein knowledgebase with Data Analytics support. Our protein-centric biomedical knowledge base, TCKB (Target Central Knowledgebase) will be comprised of the data, knowledge and information container, together with its codebase and software pipelines. TCKB will be the repository for experimental, processed and computed data and reagents originating from the IDG DRGCs (Data and Resource Generation Centers). We will provide informatics and modeling support for DRGC activities. Aim 3: We will expand, improve and maintain Pharos. Particularly "knowledge packages," support automated data summaries for Protein Dossiers, and actively seek feedback from our community. Aim 4. Outreach to scientific community. We will support a series of activities that will leverage TCKB, Pharos and other IDG resources to increase adoption of IDG work, while observing FAIR (findable, accessible, interoperable, reusable) principles for our knowledgebase, portal and pipelines. The KMC will engage in community outreach by leading tutorials and feedback sessions and dissemination of the Pharos system. To meet its goals, the KMC will coordinate all core activities in close coordination with the IDG Steering Committee and IDG Project Scientists (PS), and include members of the IDG Consortium (IDG-C), other NIH Common Fund programs, NIH Commons, as well as other initiatives.

The Knowledge Management Center (KMC) for the Illuminating the Druggable Genome (IDG) program plans to aggregate, update and articulate protein-centric data, information and knowledge for the entire human proteome with emphasis on understudied proteins from the 3 families that are the focus of the IDG. The KMC long-term objective is to encourage and support biomedical research aimed at understudied proteins by providing an extensive resource of data, information, knowledge, methods and reagents for the entire human proteome, and to support the growing online community focused on understudied proteins.

FACILITIES AND OTHER RESOURCES
UNIVERSITY OF NEW MEXICO

Translational Informatics Division

The Translational Informatics Division (TID) is housed in the Innovation Discovery & Training Complex (IDTC) together with the UNM Center for Molecular Discovery (UNMCMD). TID is part of the UNM Health Sciences Center, which includes the School of Medicine, UNM Hospital, and the New Mexico Cancer Center, an NCI-designated Comprehensive Cancer Center.

Founded in 2012 by Tudor Oprea (as Division Chief) and by then Chair of Internal Medicine (DoIM) Pope Moseley, TID aims to provide integration of translational informatics services in support of clinical and basic research within DoIM. TID members are formally trained across multiple disciplines including medicine, chemistry, biochemistry, genetics, informatics, computer science and engineering. As the largest Department in the entire University (over 260 Faculty, across 17 Divisions and Centers), DoIM provides direct access to clinician scientists with a variety of medical specialties and research interests. TID has also maintained collaborations with UNM's departments of Computer Science, Mathematics and Statistics, and Chemistry and Chemical Biology, and with the two National Laboratories based in New Mexico, namely Sandia National Laboratories and Los Alamos National Laboratory. This setting combines world class biomedical research, clinical care, education and community service, and is ideally suited for translational research such as this project.

TID maintains a specially equipped server room of 190 sf hosting the clusters and enterprise servers of the Division, and a conference room of 300 sf. Our locally maintained and operated major cluster (Pinon) has 792 Intel Xeon compute cores and 33 nodes with 64 GB RAM each. Members of TID have access to UNM Center for Advanced Research Computing (CARC) which hosts a TID-dedicated computing system (Synergy) and provides access to campus-wide computing clusters. CARC currently has over 3000 CPU cores and 92k NVIDIA Tesla K40M CUDA cores spanning a variety of distributed and shared-memory architectures. Online working NAS and nearline storage is provided by the Research Storage Consortium (RSC) HP x9000/7400 system ~1.5 PB (raw) configured as RAID6, with integrated tape library for data archiving. TID's LAN is connected via 1 Gbps+ routers, with typical Internet bandwidth ~200Mbps. Network security features include a custom dual-DMZ architecture and industry standard VPN access for a variety of high performance privacy models. Each TID member is computationally well-equipped, typically with a 64-bit 4-core 8GB RAM workstation, a high-end laptop, and access to local, UNM-CARC, and cloud-based servers and clusters.

TID's software cyberinfrastructure includes a variety of free and commercial tools efficiently addressing a wide range of computational tasks. For any such tool, practical usability requires suitable hardware, prerequisite configurations, and expertise for effective use. A few of the enterprise server components readily available are: PostgreSql, MySql, Tomcat, Jena. For statistics, data analysis, visualization and machine learning: R, Tibco-Spotfire, Weka, Tableau, MKS-Simca, and Mesa Analytics. For cheminformatics: ChemAxon, OpenEye, RDKit, OpenBabel, Leadscope. For web development: Django, RShiny, Lift. OSs in current use include: CentOS, Ubuntu, SuSE, Mac OSX and Windows. Programming languages include: Perl, Python, R, Java, Scala, JavaScript, PHP and C++. Other (licensed, not open-access) resources include Truven MarketScan (health informatics database), Cerner HealthFacts (deidentified electronic medical records database) and Statista.com (statistics from a variety of areas, including consumer reports, pharmaceutical, etc.).

TID has access to a large variety of scientific enterprise hardware and software, as listed above. Highly relevant to this RFA are the digital assets we maintain, listed in Table 1 below. A variety of custom client tools provide access to online resources via programmable-web APIs (e.g. REST). The Department of Internal Medicine provides administrative and secretarial support.

Table 1. Digital Assets developed and maintained by TID

Resource	Category / Name	Description	Reference
Drugcentral	**Database:** Online drug compendium	DrugCentral provides information on active ingredients, pharmaceutical products, drug mode of action, indications, pharmacologic action.	Ursu et al., *DrugCentral: online drug compendium.* Nucleic Acids Res. 2016. PMID: 27789690 http://drugcentral.org
	Database: Chemical bioactivity database	CARLSBAD contains selected high confidence bioactivities with associated protein targets, compounds, and chemical patterns, scaffolds and MCSes.	Mathias et al., *The CARLSBAD database: a confederated database of chemical bioactivities.* Database (Oxford). 2013; 2013:bat044. PMID: 23794735. http://carlsbad.health.unm.edu
TCRD/Pharos	**Database:** target central research database, exposed via Pharos, the user interface portal	TCRD is the central resource behind the Illuminating the Druggable Genome Knowledge Management Center (IDG- KMC).	Nguyen et al., *Pharos: Collating protein information to shed light on the druggable genome*. Nucleic Acids Res. 2016. PMID: 27903890. http://pharos.nih.gov
Illuminating the Druggable Genome	**Website:** IDG Consortium	The IDG website, targetcentral.ws provides timely access to all activities and information of the IDG pilot phase.	http://targetcentral.ws
Badapple 	**Webapp:** scaffold promiscuity detection	Badapple is a method for rapidly identifying likely promiscuous compounds via associated scaffolds.	Yang et al., *Badapple: promiscuity patterns from noisy evidence.* J Cheminform. 2016:8(29):1-14. PMID: 27239230 http://pasilla.health.unm.edu/tomcat/badapple

Resource	Category / Name	Description	Reference
	Webapp: integrative navigation in pharmacological space	iPHACE explores the polypharmacology of drugs and cross-pharmacology of targets.	Garcia-Serna et al., *iPHACE: integrative navigation in pharmacological space*, Bioinformatics. 2010;26(7):985-6. PMID: 20156991 http://agave.health.unm.edu/iphace/
Drug-Likeness	**Webapp:** assess drug likeness	Drug-likeness filter based on molecular fragments.	Ursu O, Oprea TI. *Model-Free Drug-Likeness from Fragments*, J Chem Inf Model. 2010; 50(8):1387-94. PMID: 20726597 http://pasilla.health.unm.edu/tomcat/drug-likeness
TIN-X 	**Webapp:** Target Importance and Novelty eXplorer	Interactive visualization tool for discovering interesting associations between diseases and potential drug targets	http://newdrugtargets.org/

National Center for Advancing Translational Sciences (NCATS) FACILITIES AND ENVIRONMENT

The NCATS main laboratories are in Rockville, Maryland,10 minutes north of the main National Institutes of Health (NIH) Campus in Bethesda, MD; sharing our building are three biotechnology companies, Sanaria Biosciences, Biorelliance, and Canon Life Sciences, as well as the NIH Imaging Probe Development Center, with which the NCATS collaborates to develop imaging derivatives of its probes. We are at the heart of the Shady Grove Life Sciences Center, a biotechnology cluster that includes The Institute for Genome Research, the J. Craig Venter Institute, Shady Grove Adventist Hospital, the University of Maryland Montgomery County Campus, and the Johns Hopkins University (JHU) Biotechnology Campus, including the National Cancer Institute facilities at JHU Montgomery County, all of which are within walking distance of the NCATS. The NCATS occupies 35,000 square feet of laboratory and office space in this building, with all Cores located adjacent to each other to maximize interactions.

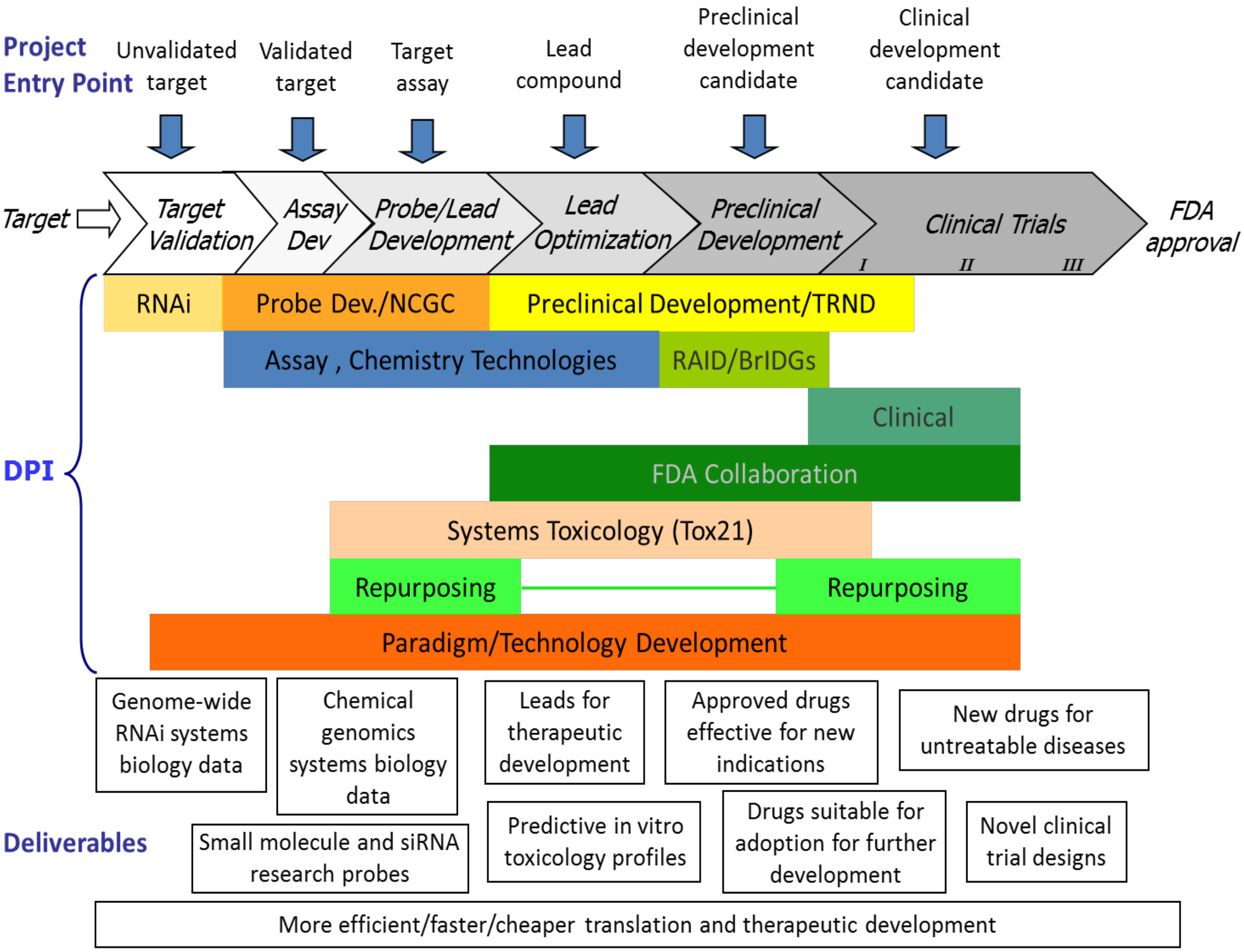
Institutional Environment

The NCATS Division of Pre-clinical Innovation (DPI) was previously known as the NIH Chemical Genomic Center (NCGC), which was originally part of the National Human Genome Research Institute (NHGRI) since its founding in 2003. At the end of 2011, NCATS was named as a stand alone Center at NIH and the NCGC was transplanted into this newly erected Center. This in turn gives it access to the enormous resources of the National Institutes of Health campus in and around Bethesda. One of the largest research campuses in the world, the NIH intramural program includes over 6,000 scientists working in every area of science, as well as the NIH Clinical Center, the world's largest clinical research hospital and an absolutely unique resource for translating basic biomedical discoveries into improved treatments for human disease.

The environment at NCATS is ideal for all aspects of this work: for stem cell work, for toxicological screening work, and for genomics and genomics data analysis work. NCATS now houses a Stem Cell Translation Laboratory under Dr. Ilyas Singec. Toxicological screening is supported by high-speed robots, acoustic and pintool dispensers, compound libraries, and readers. Even more impressive are the staff who support these activities, including dedicated qHTS screening experts e.g. Drs. James Inglese and Marc Ferrer, robotics engineers and programmers and Information Technology group led by Sam Michael, a Compound Management group led by Paul Shinn, and brilliant Data Analysis team led by Dr. Noel Southall. The support groups are capable, cooperative, and frequently innovative. NCATS central technology is quantitative high throughput screening (qHTS) but we also have researchers working in: toxicology, genomics, cancer genetics, chemoinformatics, modeling, and in all aspects of drug development. Most importantly, the atmosphere among the principle scientists is also highly collaborative and intellectually aggressive. NCATS has an amazing culture in which to innovate and to drive translational research.

For informatics support, the data center that NCATS resides in was originally designed by Human Genome Sciences, Inc. to support its massive genome sequencing operation. Thus, the data center has a large footprint with extensive cooling systems and power infrastructure. NCATS has developed a High Performance Computing infrastructure in the data center to support its public platforms such as Pharos and BARD, which will support large-scale public use of chemical biology data. The HPC infrastructure includes more than 200 CPU cores, along with a small GPU cluster, backed by XXXTB disk storage. The data center also supports NCATS internal computing needs for large-scale high throughput dose response generation across more than 50 bioassays per year. To date, the computational infrastructure generation is over 300 million dose response curves. Additionally, virtual screening using structure-based or ligand-based approaches are supported on the HPC, as well as RNAi data analysis. The servers host a number of public databases that span data types from pathways, genes, proteins to bioassays, siRNAs, and medicinal chemistry data. The data center is equipped with advanced data backup and a management system, which has an additional redundant backup in another NIH building. The HPC has been built using an extensive set of HP servers as nodes and utilizes open source software for its operating environment. The infrastructure is well suited to support the computational aims of the IDG KMC.

National Center for Advancing Translational Sciences: An Integrated Pipeline


Project Entry Point
Unvalidated target
Validated target
Target assay
Lead compound
Preclinical development candidate
Clinical development candidate
Target
Target Validation
Assay Dev
Probe/Lead Development
Lead Optimization
Preclinical Development
Clinical Trials
I
II
III
FDA approval
DPI
RNAi
Probe Dev./NCGC
Preclinical Development/TRND
Assay , Chemistry Technologies
RAID/BrIDGs
Clinical
FDA Collaboration
Systems Toxicology (Tox21)
Repurposing
Repurposing
Paradigm/Technology Development
Deliverables
Genome-wide RNAi systems biology data
Chemical genomics systems biology data
Leads for therapeutic development
Approved drugs effective for new indications
New drugs for untreatable diseases
Small molecule and siRNA research probes
Predictive in vitro toxicology profiles
Drugs suitable for adoption for further development
Novel clinical trial designs
More efficient/faster/cheaper translation and therapeutic development

Resources & Facilities:

The European Bioinformatics Institute (EMBL-EBI) is the largest bioinformatics centre in Europe, providing data and services to a large global user base. EBI's total disk resources include 50 peta bytes of storage and >17,000 CPU cores with ~100 Tb of RAM. Data served from remote data centres are externally replicated daily via 10 GB/s connection; dedicated database servers are provided for production databases and R and amazon cloud computing resources are available. Virtualized infrastructure is used as standard and supported by the Technical Services Cluster, which includes specialist DBAs, networking and storage experts. Systems support for ORACLE, postGRES and MySQL databases. EBI's infrastructure supports > 11 million web hits per day, includes Aspera for rapid file transfer and a local cloud infrastructure 'Embassy Cloud' which can be used to host third party user data securely while providing co-location with public data reducing data transfer costs and improving access.

Infrastructure and service activities are supported by approx. 50 people in the Technical Services Cluster delivering desktop and email support, systems infrastructure and applications, database services, web development and production. Access to 500+ Virtualised Machines (VM), Oracle and postGres databases, and our remote London Data Centre, relay and disaster recovery data centres, local data centres and Embassy cloud compute services are provided at no cost to this proposal. EBI has over 50PB of data. Internal analysis activities are supported by a 30,000 core compute cluster while a cloud environment with over 1,000 cores (and growing) supports collaborative analysis activities.

Dr. Leach has office space for 25 people. Video and teleconferencing is available. A visitor programme provides a means of hosting visiting collaborators, and institutional collaborator accounts are available for academic collaborations to allow sharing of code and documents.

Facilities and resources

The Novo Nordisk Foundation Center for Protein Research at the University of Copenhagen (UCPH) has excellent physical facilities, newly renovated offices and lab space comprising more than 3,000 m^2 in all. The Disease Systems Biology program has in the order of 10 large offices (room for 2-8 individuals in each), meeting space and a lecture room at its disposal. In addition to the dedicated database server, hosting databases used by IDG KMC, the UCPH partner has access to an excellent high-performance computing (HPC) infrastructure. This platform provides us with unique opportunities to perform advanced analyses and to integrate data across experimental domains. The installation consists of powerful compute units (shared memory machines and clusters) and a large professional multi-tier storage unit with tape backups. The focus is on the ability to perform extremely demanding computational analyses of very large amounts of biological data. The installation is also tailored to handling person sensitive data (such as electronic patient records) in an ultra-secure sub-environment, which is used also for commercially sensitive data.

Facilities and Resources:

Molecular Connections Pvt. Ltd. is a ‘for profit’ informatics company based in Bangalore, India. We were established in the year 2000 and have a staff of around 1800 today with a minimum qualification of Masters or engineering. The science team is supported by a strong IT support team in terms of technology. Over the years we have gained considerable experience, both by working through our clients and development of database products, in the area of manual annotation. We offer data mining services on a wide area of life science and chemical sciences disciplines. Our services have spanned abstraction of knowledge in genomics, compound/gene/protein annotation, signalling pathways, diseases, etc. from peer reviewed publications, patents and also conference proceedings.

The annotation capabilities include in-depth analysis of peer reviewed articles and patents for building pathway networks in normal, pathological or stress conditions, genome variation and diseases/drug response, adverse effect, compound-protein interactions and drug metabolism, etc. The pathways networks are generated taking into consideration the paracrine signaling between different cell types within a system [example: Factors from keratinocytes influencing melanocyte signalling in skin] or cross-talk between pathways within a given cell [Melanocortin receptor pathway cross-talk with Endothelin signaling]. All the knowledge mined are captured and delivered in formats specified by the clients. MC has extensive knowledge and expertise in mapping entities through Ids/nomenclature available in public domain to enable data uniformity and integration.

MC was part of IMEx and PSIMEx grants for molecular interaction curation and has contributed to IntAct (EBI hosted database). MC also actively contributed to substrates and their cleavage sites for proteases in Merops database, and currently to host-pathogen interaction for PHI-Base. Two “ontologies” developed by MC are available in BioPortal demonstrating understanding and conceptualization of ontology development capabilities. The list of publications and links are provided in the Annexure.

We will be using experienced life science masters degree holders for the IDG-KMC human curation work and the project will be headed by Dr. Arathi Raghunath with inputs from other senior colleagues. Our established IT support will enable delivering data in the defined format.

Molecular Connections has an established work flow which ensures high quality output as depicted in the next page.

SPECIFIC AIMS

The overall goal of this proposal is to aggregate, update and articulate protein-centric data, information and knowledge for the entire human proteome via the **IDG** (Illuminating the Druggable Genome) Knowledge Management Center (**KMC**), with emphasis on understudied proteins from the 3 families that are the focus of the IDG ("**IDG List**"). During the pilot phase, the KMC compiled, processed, structured and integrated data and information from 51 gene- and protein-focused datasets and developed and maintained seven websites, including the primary KMC portal, **Pharos** (pharos.nih.gov). To support the overall IDG objective, and to maintain, update and improve these integrated resources, the **KMC** draws upon expertise from multiple knowledge domains, specifically biology, chemistry and medicine, as well as computer science, graphic design and web programming. Specifically, for the Phase 2 of the IDG KMC we aim to**:**

Aim 1: Create an automated workflow that captures relevant public data for the entire proteome and manual annotations for the IDG list. Drawing on existing KMC resources and pipelines, we will improve the workflows currently implemented in **TCRD** (the Target Central Resource Database) and redesign the data around knowledge graphs focused on the five major branches of the target knowledge tree, **tkt**: *Genotype, Phenotype, Expression, Structure & Function*, and *Interactions & Pathways*, respectively. Data from the **tkt** branches include normal vs. disease, mutations, isoforms, age/gender differences and other alternatives as needed. We will develop automated pipelines in order to create an automated system, and will automate quality metrics, provenance and other meta-data, as well as knowledge graph mapping wherever possible. Community feedback will assist our human curation efforts for the **IDG List**, with focus on quality metrics, clarification of articulated information and accumulation of *specific* knowledge. Scientific justification for adding new (primary) sources to the automated workflow will be first to address *knowledge gaps in the 5* ***tkt*** *branches*, coupled with responding to specific requirements from the IDG steering committee and Program Scientists (**IDG-SC-PS**) and the IDG Consortium (**IDG-C**).

Aim 2: Design, develop and implement a protein knowledgebase with Data Analytics support. Given our 3-year experience with **TCRD** as primary database for **Pharos**, we will design, develop and implement a protein-centric biomedical knowledge base, **TCKB** (Target Central Knowledgebase). **TCKB** will capture, structure and integrate data elements from the five **tkt** branches (**Aim 1)**. **TCKB** will be supported by, designed and centered around protein-related ontologies and knowledge graphs, and will serve as primary repository for public domain data and information archived and updated through automated workflows, coupled with manual input for the **IDG List** (**Aim 1**). Nosology, linking phenotypes and diseases to genes, alleles and proteins, and other ontology approaches are central to the design and implementation of TCKB. **TCKB** will also serve as repository for experimental, processed and computed data originating from other IDG components such as the **DRGCs** (Data and Resource Generation Centers). IDG generated data will be stored in appropriate **tkt** branches, and made available via the **Pharos** portal, as needed. The Pharos API will support data analytics services such as **TIN-X** (Target Importance and Novelty eXplorer) and third-party applications. The KMC will provide analytics and machine learning support for IDG DGRCs, as needed.

Aim 3: Expand, improve and maintain Pharos. We will implement "knowledge packages," a mechanism to enrich target dossiers, with (semi-)automated summaries and automatic enrichment with related information (tool compounds, reagents, publications). We will support data summaries and integration with external tools and data generators, and expose data analytics methods from **TCKB**. We will implement network methods to enhance discoverability within TCKB and support sophisticated filter queries that include dependencies between data types. We will implement mechanisms to actively elicit feedback on portal features and performance and improve documentation of the **Pharos** portal and its API.

Aim 4: Outreach to scientific community. We propose a series of activities that will leverage **TCKB**, **Pharos** and other IDG resources to not only increase the adoption of our work, but also to tackle its sustainability beyond the life of this grant. We will use DockerStore to facilitate the deployment and routine execution of our framework; we will use a variety of approaches to implement the **FAIR** (findable, accessible, interoperable, reusable) principles for our knowledgebase, portal and pipelines. We will perform community outreach by leading tutorials and feedback sessions and dissemination of the Pharos system. Finally, we will coordinate with **IDG-SC-PS**, **RDOC**, **DRGCs**, and other stakeholders to enhance our outreach activities.

The KMC aims meet specific IDG challenges, with wide ranging applications in life sciences and health care. The KMC will 1) facilitate broad use of protein knowledge and software by making it **FAIR**; 2) enable and support a data ecosystem that accelerates basic and translational research for understudied proteins. To meet these aims, the KMC will coordinate its activities with all members of the **IDG Consortium** on a collaborative basis, supporting activities from **DRGCs** and **RDOC**, as well as any other future members of this program, in close coordination with the IDG Steering Committee and IDG Project Scientists (PS). The KMC will also support collaborative efforts to integrate the IDG project with other initiatives, such as the NIH Common Fund Program and NIH Commons, as well as international programs.

RESEARCH STRATEGY

A. LONG TERM GOAL OF THE KMC

The overall goal of the KMC is to encourage and support biomedical research aimed at the understudied proteins by providing an extensive resource of data, information, knowledge, methods and reagents for the *entire human proteome* (**HuP**), and to support the growing online community focused on understudied proteins. Focused on proteins that are on the **IDG List**, the KMC will aggregate, update and articulate protein-centric data, information and knowledge for the **HuP**, while enabling support for expanded coverage for non-human proteins of therapeutic interest and other associated human health data, in order to catalyze novel biomedical discoveries. To meet this goal, the KMC will coordinate its core activities with members of the IDG Consortium (**IDG-C**), other NIH Common Fund programs, NIH Commons and other initiatives

B. PILOT PHASE OF THE IDG KMC

"The reluctance to work on the unknown"[1] is inherent to the scientific endeavor, partly due to our tendency to choose research subjects more likely to confirm what we already know or believe.[2] In a deliberate attempt to focus on understudied proteins, NIH launched the IDG initiative[3] in 2014. Part of this initiative, the IDG KMC led by Oprea (PI) was tasked to systematize general and specific biomedical knowledge by processing a wide array of genomic, proteomic, chemical and disease-related resources. The KMC Pilot focused on conceptual development, designed and implemented approaches aimed at eventually enabling a sustainable system to support understudied proteins research (**Box 1**). *Pilot KMC activities are summarized below.*

BOX 1: Overview of KMC Digital Resources

- **DrugCentral:** Chemical, pharmacological and regulatory information for active pharmaceutical ingredients and products. Includes drug identification codes, MoA and pharmacologic action for targets, pharmaceutical product information, indications, contra-indications and off-label indications.[4] It maps 4,475 active ingredients to 89,942 pharmaceutical products (CC-BY-SA 4 license).
- **Drug Target Ontology (DTO):** Interactive visual framework for drug discovery data based on formalized and standardized classifications and annotations of druggable proteins[5].
- **Harmonizome:** Pre-processed collection[6] from 70 major online resources with ~72 million gene/protein attributes, the Harmonizome is integrated in TCRD/Pharos. *Maintained by Avi Ma'ayan at Mt. Sinai.*
- **Pharos:** Providing public access to TCRD,[7] Pharos is a web-based Java platform that supports efficient and intuitive queries and browsing of all TCRD data. Features: Search filters to reduce lists of targets; Query saving capability for sharing; Dossier functionality to collate data during search or browsing. Pharos supports an API for programmatic access and inclusion in pipelining tools.
- **TargetCentral Resource Database (**TCRD) is the KMC data repository and primary data source for Pharos[7]. TCRD integrates 51 heterogeneous datasets, with >85 million gene/protein attributes. Starting with 4 families of interest for the IDG Pilot, 3TCRD was extended to **HuP** *(Table 1)*. Implemented as a relational database using the open source MySQL.[8] TCRD features Python scripts and an internal adaptor API to process data from a variety of sources in formats including databases, spreadsheets, text files, XML, JSON, and Web APIs. Available for complete download[9] (CC-BY-SA 4 license), and via a REST API.[10]
- **TargetCentral.ws** is the website of the entire IDG Consortium.[11] It features information about IDG members and highlights IDG activities.
- **TIN-X** is a visualization platform[12] that allows users to explore the association between proteins and diseases based on text mining data processed from scientific literature. The user can navigate protein-disease relations, and examine a scatter plot using two newly introduced bibliometric scores, *Importance* and *Novelty*.[13]

Background: For the purposes of KMC, we define *knowledge as the consensus of structured information aggregated from different sources*, and *information as structured data* ('form' is central to in-form-ation[14]), *with a contextual layer that supports data analytics*. Data, facts and models have an "expiration date", so knowledge itself is subject to revision and change. However, current biomedical knowledge provides context for the evaluation, interpretation and integration of emergent data, information and models.

Knowledge management (**KM**) *implies the ability to structure data into information*[15] while combining low-volume, high-quality data types, e.g., analyses of experimental data like high-resolution X-ray crystallographic structures or systematic meta-analyses like the Cochrane[16] systematic reviews, with high-volume (perhaps lower quality) data such as GWAS (Genome-wide association studies) or HTS (high throughput screening). KMC *automated algorithmic processing of structured data* by extracting and processing expression and functional data related to proteins and genes, molecular probes such as small molecules and antibodies, GWAS, disease associations and launched drugs information into **TCRD**.[7]

Concepts developed by the KMC: 1. To articulate information and knowledge compiled and archived by the KMC (**Box 2**) we proposed "**Target Development Levels**" (TDL) to classify all human proteins (**Table 1**) from a clinical, chemical and biological standpoint, with respect to our depth of knowledge (**Box 2**). TDL provides an overview into the current illumination levels and opportunities for targets of "druggable" protein families. *Except for Tclin, TDLs are assigned with no human curation.* Bioactivity data from papers and patents currently processed in ChEMBL[17] may progress proteins to Tchem. TDLs are an easy to interpret scheme that monitors knowledge accumulation using multiple levels of evidence (**Box 1**). Proteins in Tdark and Tbio are understudied and need "illumination", whereas Tchem and Tclin are likely to be well studied.

Table 1. Distribution of TDL categories by protein family for the "druggable" genome. See also **Box 2**

Target Class	All	Tclin	Tchem	Tbio	Tdark
GPCRs (non-olfactory)	406	96	113	145	52
Olfactory GPCRs	421	0	0	8	413
Kinases	634	50	390	163	31
Ion Channels	341	125	45	140	31
Nuclear Receptors	48	18	19	11	0
Transporters	473	26	46	287	114
Transcription Factors (TF)	1,400	0	27	866	507
Epigenetic (*)	280	12	53	178	37
Enzymes (**)	4146	184	495	2,607	860
Other	11,971	87	217	6,681	4,986
Total	*20,120*	*598*	*1405*	*11,086*	*7,031*
Phase 2 Ion Channels	117	32	12	49	24
Phase 2 Kinases	134	0	76	39	19
Phase 2 GPCRs	143	5	23	67	48
Phase 2 Total	394	37	111	155	91

*) includes 40 proteins not counted with TFs **) excludes kinases

IDG investigators Roth and Kroeze[18] proposed the *number of citation* (below 100) and the *number of chemicals* associated with each protein (below 10) to decide which non-olfactory GPCRs are understudied. Citation counts extend our fractional literature count[19] (or **JS**, "Jensen PubMed score"[20]) by emphasizing publication relevance. Counting ChEMBL compounds is a criterion similar to Tchem in TDL. Roth-Kroeze criteria take into account knowledge about biological functions as well as chemical matter with which proteins are known to interact. However, counting ChEMBL compounds cannot be applied to the **HuP**, since *most human proteins are not annotated in ChEMBL*. **Table 1** suggests that a small fraction of the genome (3%) is already "druggable"[21]. TDL criteria were validated using 4 external sources: R01 grants; granted patents; the Harmonizome; and gene ontology (GO) leaf terms – as outlined elsewhere[22]. For these datasets, Tdark is significantly different compared to Tbio/Tchem/Tclin (Kruskal-Wallis post-hoc pairwise Dunn test[23]); all differences are significant except Tchem-Tclin. Clinical trials, required for the Tchem to Tclin progression, are not represented by these 4 data types.

BOX 2: Target Development Level overview

- **Tclin** ("clinical") are Drug Targets: entities with molecular mass present in living systems that, upon interaction with therapeutic agents or their byproducts, modify biologic responses leading to therapeutic outcomes. "Living systems" encompass humans or animals (e.g., for veterinary drugs) and foreign organisms such as viruses. The drug-target interaction leads, directly or indirectly, to observable clinical outcomes. Tclin proteins are linked to at least one approved drug (that is, an active pharmaceutical ingredient) by Mode of Action (MoA).
- **Tchem** ("chemistry") proteins are known to bind small molecules with high potency but lack MoA links to approved drugs. The interaction between proteins and small molecules (and sometimes approved drugs) are usually studied in the context of a disease, and are typically subject to medicinal chemistry efforts. Bioactivity of at least one small molecule is above a specific cutoff for any Tchem protein. Current thresholds are ≤ 30nM for kinases, ≤ 100nM for G-protein coupled and nuclear receptors, ≤ 10μM for ion channels and ≤ 1μM for other targets. Bioactivity values were extracted from ChEMBL[17] and DrugCentral[4].
- **Tbio** ("biology") refers to those proteins that have confirmed Mendelian disease phenotype in OMIM[28] (i.e., at least two publications); or have GO leaf term annotations[29] based on experimental evidence; or meet 2of the following 3 conditions: Fractional PubMed abstract count above 5;[19] 3 or more NCBI Gene RIF annotations[30], or 50 or more commercial antibodies, counted from data made available by the Antibodypedia database[31]. Assignment to Tbio implies that these proteins are not associated with drug MoA, and that bioactivity values for small molecules fall outside Tchem cut-off criteria.
- **Tdark** ("dark genome") refers to the remaining proteins that have been manually curated at the primary sequence level in UniProt[32], yet do not meet any of the criteria for Tclin, Tchem or Tbio. Evidence may be available concerning tissue location, dysregulation, inferred function via homology, etc. Tdark proteins have the least knowledge and molecular tools available, representing unexplored opportunities.

2. **Quantifying the knowledge gap:** All the data, information and knowledge aggregated and processed within TCRD confirms the existence of a *knowledge gap*, as documented by the statistically significant differences discussed above. The bias towards well-described proteins[1] is confirmed not only with respect to NIH funding patterns, patents and publications, but also GWAS and mouse phenotype data (*vide infra*), the availability of molecular probes such as antibodies and small molecules, and even queries in the STRING[24] database.[22] *Almost two out of five human proteins (38%) are dark.* NIH acknowledged that illumination should directly target understudied proteins: IDG Phase 2 funding opportunity announcements use, in part, metrics explored by the KMC (JS and R01 counts), which are criteria to quantify the knowledge and funding deficit.

$$N_i = 1/\sum_k \frac{1}{T_k} \quad \text{Eq. 1}$$

$$I_{ij} = \sum_k \frac{1}{T_k \cdot D_k} \quad \text{Eq. 2}$$

3*.* **Two bibliometric concepts** proposed by KMC, rooted in the **JS**, are used in the TIN-X platform **(Box 1)**: **Novelty**, which estimates the scarcity of publications about a protein target (see Eq. 1); and **Importance**, which estimates the strength of the association between that protein target and a specific disease (see also Eq. 2). Here, Tk and Dk are the numbers of targets and diseases in publication (k), respectively; summation over all publications includes target (i), and for importance, it also includes disease (j). Like JS, TIN-X uses fractional counts to reflect strength of association: When a paper mentions three targets, each protein receives a one-third fractional count. Similar counts are applied for diseases. Fractional counts were developed by IDG member Lars Jensen, **CPR** (the Novo Nordisk Foundation Center for Protein Research).

4. **Therapeutic Intent:** When analyzing drug indications,[25] Michel Dumontier (Maastricht University; also Letter of Support, LoS), Stuart Nelson (former Head, Medical Subject Headings, MeSH,[26] who will join KMC in Phase 2), and Oprea collaboratively developed the concept of **therapeutic intent**[27] to improve the quality and completeness of approved drug indications in DrugCentral for downstream use in drug repurposing. Therapeutic intent includes disease concepts and contextual aspects such as pre-existing conditions (e.g., drug A is indicated for complications of disease B), co-prescribed medications (e.g., drug A is indicated for patients having symptoms caused by drug B), particular genotypes (e.g., specific mutations), as well as a variety of other phenotypic, anatomic or temporal parameters. Therapeutic intent is relevant to precision medicine and drug repurposing. *An R01 to develop this further was submitted on 02/05/17.*

Novel approaches and technologies designed and implemented by the KMC: We developed several approaches specific for unifying and aggregating protein-centric knowledge. 1) *DTO* was built based on the need for a formal semantic model[33] for druggable targets, including related information, e.g, protein, gene, protein domain, protein structure, binding site, small molecule drug, MoA, protein tissue localization, disease association, and other types of information in TCRD. DTO is exposed in Pharos. 2) *Mode-of-action target annotations*, which were exhaustively mapped for 893 human and pathogenic biomolecular targets for 1,578 FDA-approved drugs,[34] formed the basis for Tclin; 3) *Financial analysis of MoA drug targets* based on global sales for approved drugs, which established for example that GPCRs are the most commercially successful drug target class (~900 billion USD in 5 years).[35] 4) *TCRD software:* The entire codebase used to build TCRD is available in github: https://github.com/stevemathias/TCRD/. For each of the datasets in TCRD v4, this repository includes a loading program, instructions on obtaining the needed source data file(s) (where applicable), and example output of running the programs. All preprocessing and analytical code used to produce KMC-generated data is included in the repository. 5) *Pharos software:* The Pharos platform source code is available at https://spotlite.nih.gov/ncats/pharos. This Git repository also includes instructions on building and running Pharos from sources. In addition to source code, a Docker image containing Pharos and TCRD v4 are available from https://hub.docker.com/r/ncats/pharos/; 6) *Patent analytics:* For KMC, the ChEMBL team identified 20,941 bioactivities from patents, covering 11,358 compounds and 1,134 assays covering 210 protein targets. These data, not present in biomedical literature, include 37 Tbio and 1 Tdark targets. **Data for seven Phase 2 targets were uncovered by patent data extraction.** *GPR6* (Patent US-6420563-B1) *and HCAR1* (Patent US-8507473-B2), *are GPCRs with bioactive chemicals,* some ≤ 100 nM, *which would progress them from Tbio* (current TDL in TCRD) *to Tchem*. The implication is that some of these proteins were investigated, but not reported (see **Aim 2**). 7) KMC developed 6 digital resources (**Box 1**).

Impact and Outreach of the KMC Pilot: With Lenat & Feigenbaum's observation in mind, *"If you don't know very much to begin with, don't expect to learn a lot quickly,"*[36] we acknowledge that the KMC pilot was initially in a growing, less visible phase. However, KMC websites gained visibility in 2016: According to Google Analytics data, DrugCentral had over 18,500 unique visitors (799 pageviews/day) and Pharos had 13,500 visitors (219 pageviews/day) in the past 11 months, while the IDG website, TargetCentral.ws had over 19,000 visitors (168 pageviews/day) in one year. Our MoA paper[34] reached 40,000 pageviews in 3 months, an Altmetric score of 450, and had an F1000 review. *The **TDL** classification has been adopted by KEGG[37] (human genes only) and ChEMBL, and is available through the Cytoscape STRING app.[38]* Among external collaborations, KMC helps the International Mouse Phenotype Consortium (**IMPC**) prioritize genes (LoS, from Steve Brown), and the GWAS Catalog visualize data (LoS from Helen Parkinson; also **Aim 2).**

During the Pilot phase, KMC members co-organized symposia at the following meetings: *American Medical Informatics Association*, Washington DC (11/14); *American Chemical Society*, Philadelphia PA (8/16); *EuroQSAR*, Verona, Italy (9/16); and *Society for Laboratory Automation and Screening*, Washington DC (2/17), respectively. KMC was presented at 25 international conferences, two webinars and three workshops, as well as three NIH events (NCATS Council; NHLBI internal, and Common Fund conference). Our *Nature Reviews Drug Discovery* poster[35] that provides an overview of KMC, was distributed to all NRDD subscribers; the online version had 4,456 visitors and 5,415 pageviews in under 3 months. The impact of the KMC is further substantiated by 25 Letters of Support: 17 from academia including large consortia (9 countries, 4 continents), 8 from large and small pharmaceutical companies, acknowledging the utility of Pharos and KMC.

C. OVERVIEW OF THE KMC: GOALS, AIMS AND STRUCTURE

The KMC has a unique objective, to automatically compile, process and evaluate – by human curation where possible - a highly diverse set of protein- and gene-centric data, with the explicit goal of advancing research for understudied proteins via target selection and prioritization. Since the amount of data being published exceeds the limits of human processing,[39] there is a critical unmet need to process biomedical data, structure and generalize it, and where possible create convergent, distilled views of specific protein knowledge. The emergence of new knowledge was the focus of a bibliometric evaluation by Edwards et al.[1] They examined how many of the newly sequenced proteins are the subject of new studies, 10 years after the completion of the Human Genome Project. They concluded that the process of druggable target selection is conservative, and that limited progress has been made regarding newly discovered proteins.

No single method for target selection and prioritization has been successful.[40] Target selection criteria vary from disease relevance to puzzling functional features, unexpected role of mutations, and other properties. Commercial aspects (e.g., access to reagents and funding) and personal bias can influence the selection process. To assist target selection, the KMC **quantified the knowledge gap** by combining knowledge about therapeutics (Tclin), chemogenomics (Tchem), disease and phenotypic traits together with functional & expression data (Tbio), noting the absence of such knowledge and the lack of molecular probes for Tdark. The TDL evaluation system offers an unbiased way to pinpoint understudied proteins.

The KMC team developed novel concepts, applications and visualization tools, as well as websites, with the ultimate goal of prioritizing research for understudied proteins. Some of these lines of research will be continued and enhanced (*vide infra*). We will deploy new analytics and predictive methods for processing data via Pharos. Human curation will play an equally important role in the abstraction process, particularly for **IDG List**. For Phase 2, *KMC will be functionally structured according to the four major aims*, as follows**:**

Aim 1: Create an automated workflow that captures relevant public data for the entire proteome and manual annotations for the IDG list. TCRD resources and pipelines developed at **UNM** will be adapted and expanded to provide systematic coverage for the five major **tkt** branches (see **Table 2**). Scientific justification for adding new sources and data to the automated **TCKB** workflow will be based on knowledge gaps in the 5 **tkt** branches. For the entire proteome, **UNM** will adapt the automated process, with support from **EBI** for patents and from **CPR** for text mining. However, for specific knowledge gaps, **UNM** will work closely with **Molecular Connections (MC),** an India-based subcontractor, to provide in-depth human curation for the proteins of interest to IDG. Community feedback will be derived from Pharos (maintained by **NCATS**), from the **RDOC** and **DRGCs**, and KMC will develop and apply quality metrics where possible.

Aim 2: Design, develop and implement a protein knowledgebase with Data Analytics support. TKCB will be designed and implemented at **UNM**. Although most elements of TCKB will be developed in-house, we will seek feedback from DRGCs, RDOC, and other IDG stakeholders, as well as from the scientific communit, particularly with respect to data analytics. Data analytics and machine learning (ML) support for DGRCs will be provided by **UNM**, **EBI** (patents), **CPR** and **NCATS**, based on specific expertise.

Aim 3: Expand, improve and maintain Pharos. The primary responsibility for Pharos will reside with the **NCATS** team, with feedback and support from **UNM**, the IDG Consortium and the community at large. Hardware and software infrastructure will be provided by **NCATS** *at no cost to this grant*. KMC will monitor NIH Commons policies and business models, and deploy Pharos (and TCKB if appropriate) into the cloud. Data summaries and integration will be developed at **NCATS** with support from **UNM**, **CPR** and **MC**.

Aim 4: Outreach to scientific community. TCKB, Pharos and other IDG resources will require sustainable development and concerted efforts to increase the visibility of IDG resources. DockerStore versions of TCKB and Pharos will be made available by **NCATS** with each new release, observing **FAIR**[41] principles. **KMC** will coordinate with **RDOC**, **DRGCs**, the **IDG-SC-PS**, NIH Common Fund and other stakeholders, as needed.

Limitations: There is an ongoing reproducibility crisis,[42] because companies like Bayer[43] and Amgen[44] have openly discussed the low reproducibility rates (33% and 11%, respectively) of high impact studies. A large number of biomedical publications are false,[45] and many findings in psychological research may or may not be true.[46] From a **KM** perspective, wrong data cannot be filtered out without a concerted community effort. Such an effort is beginning to take shape: the Reproducibility Project[47] is posting results in its dedicated *eLife* cancer biology collection.[48] ***Building a knowledge base starts by asserting what is true.*** That in itself does not attract significant funding: The Cancer Biology Reproducibility Project relies on a $2 million private donation,[49] which for 29 replication projects amounts to $70,000 per project. Despite interest from its top leadership[42], NIH does not have a mechanism to fund reproducibility. Illuminating other people's work is not without controversy[49], and not without limitations.[50] Thus, awareness of the *knowledge gap* for dark proteins underscores the need for basic science, its part in illuminating protein functions and roles in human disease.

Risk Mitigation: Acknowledging the knowledge gap is the first step towards deliberate interrogation of dark proteins, which is the core mission of IDG. However, this reproducibility crisis also suggests that the scientific community needs to lay down the foundations of verity, and carefully vet what is known. There is no *"we hold these truths to be self-evident"* basis in modern science. Biology does not appear to lend itself to an axiomatically build foundation. But managing knowledge implies knowledge maintenance *and change*, and we collectively share the responsibility of *sifting truthful science from questionable results.* In practical terms, the **IDG-C** will have to outline standards of practice, assist KMC in deciding what resources are trustworthy, and provide feedback on algorithms and models so that they maintain relevance to the IDG core mission.

Expertise of the Team: The KMC team has extensive **KM** and informatics experience, specifically in cheminformatics, translational informatics, medicine, pharmacology, text mining, knowledge management and representation. For 3 years, we collaboratively improved the quality and completeness of DrugCentral, TCRD, Pharos and all other KMC resources, while improving automated workflows. KMC Phase 2 will strengthen this relationship for continued collaboration in protein research. We will communicate via teleconferencing (we used GoToMeeting for 3 years) every 2 weeks, and in-person meetings such as at IDG events. Oprea is a Guest Professor at CPR, and will travel to Denmark (and UK) 2-4 times a year.

- *Tudor Oprea, MD, PhD,* is Professor of Medicine and Chief, Translational Informatics, at the **UNM** School of Medicine (SOM). He co-developed concepts for lead-likeness,[51,52] systems chemical biology[53] and mapped molecular drug targets,[34] before turning towards **KM** applied to drug discovery[54-56] and health-record analytics.[57,58] Oprea curated several drug databases: WOMBAT-PK,[59] DRUGSDB[60] and DrugCentral[61]. His work led to anti-cancer clinical trials for Raltegravir[62] and Ketorolac[63]. As PI for KMC, he coordinated knowledge management efforts across multiple levels.[64] His recent interest is to combine chemical, target and disease similarity for target[65] and drug repurposing.[66,67]
- *Stuart J. Nelson, MD, FACP, FACMI,* Research Professor at **UNM** SOM, is a board-certified internist and well-known medical informatician. He worked 16 years at NLM, as Head of MeSH, with the responsibility for editing of the Unified Medical Language System (UMLS).[68,69] Nelson developed RxNorm[70] and DailyMed. Nelson participated in the design[71] of the Medical Text Indexer (MTI). He has considerable experience in UMLS and RxNorm on lowering human effort while maintaining quality in indexing and curation.
- *Andrew Leach, D.Phil*. is Head of Chemistry Services at **EBI** since 2016, after decades of experience in drug discovery at GSK where he was involved in the development of platform capabilities and in leading therapeutic project portfolios based around target families.
- *Lars Juhl Jensen, PhD* has for the past eight years led a Systems Biology group at **CPR**, University of Copenhagen. His team has developed the DistiLD[72], DISEASES[73], COMPARTMENTS[74] and TISSUES[75] databases which integrate specific protein- and gene-centric properties. He made essential contributions to the STRING[24] and STITCH[76] databases of protein-protein and protein-small molecule associations. He also pioneered the use drug side-effects in drug repurposing[77].
- *Anton Simeonov, PhD* is the Scientific Director of the intramural division of Preclinical Innovation at **NCATS**. Author or inventor on more than 145 peer-reviewed scientific publications and patents, Simeonov has a truly diverse background, ranging from bioorganic chemistry and molecular biology to clinical diagnostic research and development. Simeonov trained as a postdoctoral fellow at Scripps under Richard Lerner and Kim Janda.

D. APPROACH

Aim 1: Create an automated workflow that captures relevant public data for the entire proteome and manual annotations for the IDG List. In the Pilot phase, KMC integrated 51 datasets (over 85 million protein attributes), constructing **TCRD**, a heterogeneous, multi-dimensional database with rigorous, automated entity resolution and mapping and a coherent, extensible schema designed for consumption by scientists and for downstream processing. *Managing protein data led us to identify five major* ***tkt*** *branches: Genotype, Phenotype, Expression, Structure & Function, and Interactions & Pathways, each with appropriate sub-branches (e.g., normal vs. disease, gender, etc.)*. Each dataset was evaluated for *i)* relevance, *ii)* quality and *iii)* added value prior to inclusion, which connotes in each case a custom, automated, data type level **ETL** (extract, transform, load) process. We will build upon this foundation by upgrading existing and incorporating new datasets, and we will extend into new areas such as phenomics, animal models, real world evidence, structural biology, pharmacology and cheminformatics. **Table 2** summarizes planned TCRD resources. Major updates of **TCRD** datasets represent a vital task with respect to quantity, quality and staying current. For **IDG List**, we will 1) identify **tkt** branches with the least coverage per protein, 2) run automated PubMed *and* 3) PatSeq (primary sequences via Lens.org) patent queries, which will help us fill knowledge gaps.

Table 2. Inventory of data sources (to be) processed by KMC. Those to be incorporated are in *italics*.

tkt Category	Data Types	Sources/Datasets
Genotype	Symbols, Names, IDs, Orthologs, Families, Genetic variants	HGNC[80], *GeneCards[81]*, UniProt[82], Panther[83], *eggNOG[84]*, *ClinVar[85]*, *RGD[86]*, *MedGen[87]*, *dbVar[88]*, *GTR[89], InParanoid[90]*
Phenotype	Disease associations, Knockout Mouse/Rat Phenotypes, GWAS	DISEASES[91], IMPC[92], *Monarch[78]*, GWAS Catalog[93], OMIM[94], PubTator[95], Disease Ontology[96], *MedGen[87]*, *RGD[86]*
Interaction / pathway	Protein-ligand, protein-protein interactions (transient and permanent PPIs)	ChEMBL[97], PubChem[98], DrugCentral[99], WikiPathways[100], Pathway Commons[101], Reactome[102], KEGG[37], Harmonizome[103], *STRING[104]*, *CTD[105]*, *GIANT[106]*, IntAct[107], *GuideToPharmacology[108]*,*PharmGKB[109]*, *LINCS[110]*
Structure / Function	Protein domains, 3D structure, and molecular function	GO[111], *PDB[112], PFam[113], InterPro[114], CATH[115], ECOD[116]*
Expression	Gene regulation and expression	GTEx[117], Expression Atlas[118], HPA[119], HPM[120], *TCGA[121]*

We will detect knowledge gaps, evaluate the need for integration into **TCKB** and prioritize human curation. We rely on **IDG-C** and community feedback (e.g., Gene Wiki) to prioritize **IDG List** human curation, while adding quality metrics, exposing *articulated* information and increasing *specific* knowledge. **TCKB** will be designed to link phenotypes and diseases to genes, alleles and proteins, using ontologies developed in-house (e.g., DTO), as well as those developed and maintained by the Monarch Initiative.[78] Stuart Nelson will supervise nosology aspects for **TCKB** design and implementation. This will ensure appropriate integration of disease classification, central to **TIN-X**, and compatibility of TCKB and Pharos to the *NCATS Biomedical Data Translator* project[79] (see also LoS from Melissa Haendel). Although we map "diseases" primarily under Phenotype, we note that disease information will be incorporated in all **tkt** branches, where relevant.

1. Principled, prioritized identification of datasources, databases, datasets and datatypes: Datasource, database, dataset and datatype are inconsistently used terms, which leads to confusion. One dataset (e.g., the Harmonizome) may aggregate data from multiple sources and combine it with primary data, which in turn may be experimental or computed. For KMC, *a database is an implementation of a dataset container.* Inferences from literature may be manually curated, or automated, or semi-automated. Some sources serve an indexing role (PubMed, BioCADDIE, SciCrunch) but are not primary sources, nor analyses. Some well-established organizations (NLM, EBI) manage many resources that provide high value metadata and organizational approaches. Data type may refer to scientific content, but also to methods and levels of processing. An assertion regarding binding could be extracted from free text or from direct experiment. An assertion may be raw, or processed by statistical, heuristic, or manual methods. Assertions may be empirical, computational, or both. E.g., protein structure data, based on X-ray crystallography requires computational interpretation; and GWAS, where genomic and phenomic variations need interpretation. Domain-informed criteria are necessary for effective quality and confidence metrics. Using **TCKB** data combined with **tkt** category, we will identify *non-redundant sources* so data is aggregated with accountability for quality evaluation. We will consider **Information Geometry** to compute distance measures between existing and novel data, to assist with incorporation of new data into TCKB (LoS from Evangelos Coutsias). Provenance metadata is a prerequisite for dataset integration. Effective **KM** also considers volume and velocity. Volume approaches must use powerful back-end technologies, but also data-reduction and visualization via UIs for scientific comprehensibility and use cases. Velocity demands timely updates, and responsiveness to data ecosystem evolution, which depends on community engagement and outreach.

2. One Phase 2 goal is to identify ***"proteins potentially able to bind drug-like molecules."*** Computational methods have extensively addressed this problem,[122-127] but none of these studies were validated on blind sets, i.e., on proteins not related to training/test sets. Use of computational methods at KMC led to the discovery of GPER agonists[128] and antagonists,[129] (NCI R01) inhibitors for the Cdc42 and Rac1 small GTP-ases[63] (NCATS R21) and GLUT5, the fructose-only transporter.[130] The GLUT5 project was just awarded an R01 (R01GM123103), effective 4/1/17. We will use the 10%-dedicated funds (if approved by **IDG-SC-PS**) to compute **IDG List** protein druggability and support **DGRC** experiments (e.g., LoS from Peter Sorger).

Case Study: Photoreactive fragment-based ligand discovery, combined with quantitative chemical proteomics identified over 2000 fragment-binding proteins.[131] *Of these, 7 kinases and 3 ion channels (ICs) are on the* **IDG List**. CLIC2, an intracellular chloride IC, and 194 other Tdark proteins were also identified. CLIC2 and 17 other IC genes are downregulated in dilated cardiomyopathy,[132,133] causing impaired contractility that leads to heart failure. Expression (dys)regulation in a disease state may provide insights into pathology and protein protein interactions (**PPIs**): Downregulation of CLIC2 triggers cytoplasmic Ca^{2+} release via ryanodine receptors (RYRs). Blockade of RYR/CLIC2 to regulate Ca^{2+} release could thus be beneficial.

3. This *Case Study* illustrates the need of **rapid human curation response** and meta-tagging of supplementary materials. Since such data would not readily be available through automated **TCKB** pipelines, KMC will use human curation directly (or via **MC**) to integrate such data into TCKB. **MC has agreed to provide human curation for 2,000 papers annually, for a total of 12,000 papers for the duration of this proposal, according to priorities provided by IDG-C (in coordination with KMC).** There is also a need to *expand beyond the human genome* for proteins of particular interest (e.g., non-human drug targets) as well as interfacing and capturing their associated human health data, as mandated by **IDG-C**.

4. Gene Wiki and Wikidata. Started in 2007, the Gene Wiki initiative[134,135] benefited from crowdsourcing of content after Wikidata added significant value via structured-data content (in 2012). Gene Wiki is a vibrant community, closely aligned to the WikiProject "Molecular and Cell Biology" (MCB). The popularity and reliability of Wikipedia are well established, even for scientific domains.[136] MCB articles are classified by importance and quality and may be regarded *de facto* as peer review. Wikidata is a publicly accessible RDF (Resource Description Framework) triple store and Sparql (the RDF query language) endpoint, thereby well suited for automated integration within TCKB. *We will integrate Gene Wiki via Wikidata with TCKB as an additional data sources, and as a platform for community curation and crowdsourcing.*

Aim 2: Design, develop and implement a protein knowledgebase with Data Analytics support. 1. *A major TCRD upgrade, TCKB is a reimagined container, no longer comprised of data, information and knowledge, but also incorporating its codebase and software tools.* The expertise was acquired by the KMC team while developing TCRD and its ETL-related processes is reflected in the current suite of protocols and in the approaches by which new or substantially revised datasets are considered for integration. Many datasets have been evaluated and rejected for one or more criteria related to relevance, quality or value.

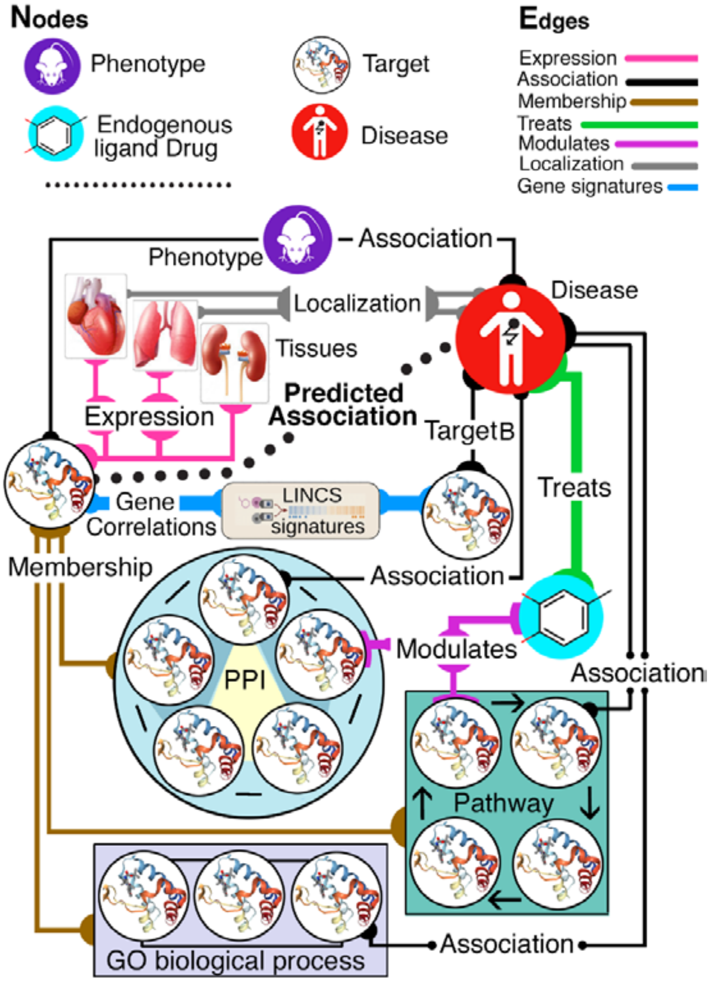


Figure 1. TCKB knowledge graph representation, showing meta-paths along the major **tkt** branches.

Co-development of Pharos with TCKB is synergistic: New functionality must be justified by use cases via Pharos, and needs to incorporate UI design principles, visualization and domain specific understanding of each feature for data summary, drill-down, and link-out. Should data be stored or referenced? *In the ideal world with an infinitely fast and programmable web, TCKB would require little local storage and consist mainly of processing pipelines.* However, KMC has specific stewardship and storage responsibilities for findings and methodology from DRGCs. KMC will collaborate with **DRGCs**, **RDOC** and other **IDG-C** members to integrate their data via TCKB, and maximize the benefit for the scientific community in addition to other channels, e.g., publications.

2. Knowledge Graph: Projection of TCKB data into a graph data structure (**tkt**) with typed nodes and edges will enable use of network based analytical algorithms. The **tkt** concept with its 5 major branches (**Fig 1**) is essential to our Analytics component, and will support the development of ML algorithms for imputing function and disease associations to a target. Using gene orthology relationships from the eggNOG[84] and InParanoid[90] databases, the model organism data from the five **tkt** branches, related to different organisms, will be fused into a single “pseudo-protein”, thus enabling *network-based inferences for function and phenotype across organisms*. This is *information articulation* (new relations articulate the data).

3. Network link predictions using meta-paths. Biological system networks (BSN) are heterogeneous with multiple node and edge types (**Fig 1**). Relationship prediction algorithms designed for homogeneous or social networks are not well suited for BSNs. Recent developments in heterogeneous[143,144] and BSN[145] relationship predictions introduced and formalized a new framework that takes into account BSN heterogeneity by defining type specific node-edge paths or meta-paths.[146] *A meta-path encodes type-specific network topology between the source node* (e.g., Protein target) *and the destination node* (e.g., Disease or Function). In **Fig. 1**, type specific meta-paths are: (Target — (member of) → PPI Network ← (member of) — Protein –– (associated with) → Disease) and (Target — (expressed in) → Tissue ← (localized in) — Disease). Type-specific meta-path counts can be combined using, e.g., degree-weighted paths to dampen the effect of highly connected nodes. We will apply Relationship Ontology[147] (RO) to formalize meta-paths types. RO formalizes and describes a rich set of relationships between biological entities, and is developed by **Monarch Initiative**[78] scientists, with whom we collaborate (LoS from Melissa Haendel and Tudor Groza).

Employing RO will ensure **FAIR** representations of meta-paths and enable computational re-use. The original BSN application used logistic regression and ridge logistic regression. The meta-path framework, however, can be adapted to a variety of classification algorithms (**Table 3**). *By applying ML approaches with typed meta-paths on BSNs, KMC will provide predictions for " function" or "role in disease."* Variable importance estimates will allow sorting of meta-paths in decreased order of importance, leading to insights into the different biological processes and interactions that will contribute to predicted "function" or "role in disease." The process of applying meta-path methodology to generate these predictions is depicted in **Fig 2**.

Machine learning and artificial intelligence (AI) are subject to rapid developments. Acknowledging that improved and better ML/AI algorithms will become available, we will allow the user community to apply new algorithms to TCKB data, in particular meta-path derived predictors and observations. The REST API will be language (ex. Python, Java, etc.) agnostic and support access via a Bioconductor[148] package. The Bioconductor platform provides support for biological modeling and analyses of complex biological data, and supports advanced analytical pipelines and ML algorithms. Google released its TensorFlow ML library[149] as an R package[150]. Experimental **DRGC** data stored in **TCKB** will be projected onto the **tkt** *knowledge graph*, to enable new meta-paths and enrich existing ones. New meta-paths or enrichments will effectively lead to new predictions. *When focused on the* **IDG List**, *this process will suggest new functions and disease associations, which DRGCs can follow up for confirmation or falsification.* Some of the 10% funds for collaborations with other IDG awardees will be used to validate these assertions, if approved by **IDG-SC-PS**.

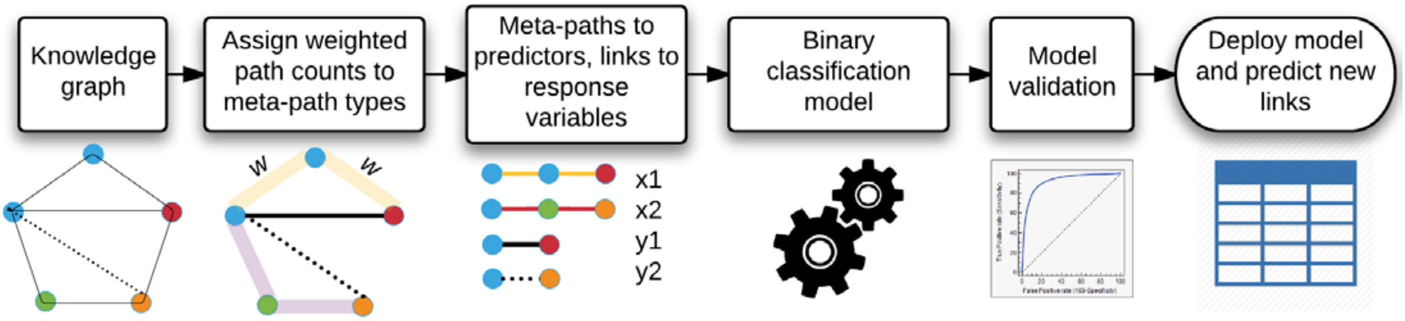

Knowledge graph → Assign weighted path counts to meta-path types → Meta-paths to predictors, links to response variables → Binary classification model → Model validation → Deploy model and predict new links

Figure 2. Analytical algorithms workflow for information *articulation* in KMC.

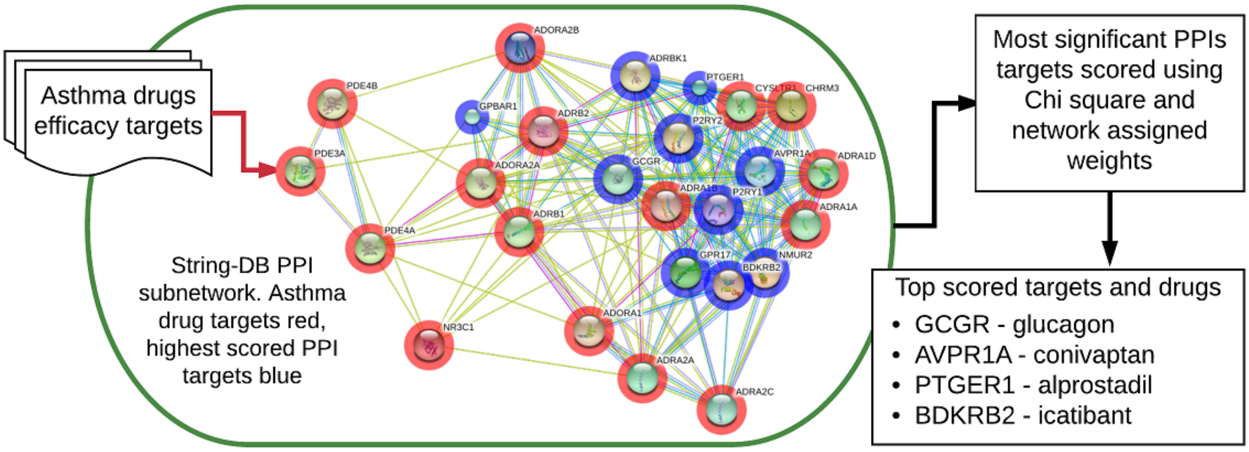
Table 3. Algorithms for consideration in the **Aim 2** *Data Analytics* component

Algorithm/Method	Ref
Regression models & derivatives (logistic, ridge, GLM, etc.)	137
Support vector machines	138
Naive bayesian classifier	139
PLS Discriminant analysis	140
Random Forests	141
Neural Networks	142

Moreover, Bioconductor integration will expose TCKB, **tkt**, meta-path predictors, known/missing links between proteins and functions/diseases, together with high confidence predictions, to the scientific community. Bioconductor is supported and used by a large community of computational biologists, featuring over 1,200 packages for computational biology, genomic data analysis and visualization. In addition to Bioconductor, users have access to the Comprehensive R Archive Network (CRAN)[151] that features over 10,000 packages for statistical and ML applications. Numerous learning resources are available for Bioconductor[152] including Coursera[153] classes. *This is part of our outreach program* **(Aim 4).**

Case Study: A **tkt** meta-path example is Target — (member of) → String[154] PPI network ← (member of) — Target — (modulated by) — Drug — (treats) — Disease (*TmPmPdGtD meta-path*). When using drug indications, MoA targets and bioactivities from DrugCentral,[99] we found ~135,000 unique paths along the *TmPmPdGtD meta-path*. Scoring Target — Disease associations using weighted Chi-squared statistics[155] and mapping these Target-Drug-Disease tuples along the TmPmPdGtD meta-path relveals the top scoring drug that could be potentially repurposed in Asthma; see **Fig 3**. Several of these top associations were confirmed by PubMed searches[156–158], illustrating the potential value of this methodology. We expect to improve the predictive power of this approach by including additional meta-paths, as depicted in **Fig 1**.

3. Cheminformatics-based target space calibration: Knowledge of approved drugs and associated Tclin (MoA) targets, extracted from DrugCentral, will be integrated into the KMC predictive methodology by associating chemical substructures and scaffolds with therapeutic areas and target families, an approach we explored in CARLSBAD.[159] This will serve as gold standard for BSN meta-paths, and accessible via Pharos.


Asthma drugs efficacy targets

String-DB PPI subnetwork. Asthma drug targets red, highest scored PPI targets blue

Most significant PPIs targets scored using Chi square and network assigned weights

Top scored targets and drugs
- GCGR - glucagon
- AVPR1A - conivaptan
- PTGER1 - alprostadil
- BDKRB2 - icatibant

Figure 3. Drug and Target repurposing using meta-paths of PPIs for MoA targets, applied to asthma.

4. Predictive model exchange: Algorithm development is a key KMC goal. Equally important is their wide dissemination via Pharos. This approach may not always be suitable (e.g., for algorithms with long run times). Model reuse will be enabled using the *Predictive Model Markup Language* (PMML) as a platform-independent way to specify and serialize predictive model to XML documents. An industry accepted standard,[160] PMML is supported by commercial (e.g., IBM SPSS) and non-commercial (e.g., Knime, R) applications. PMML supports most algorithms listed in **Table 3**. This approach will not bind KMC to specific platforms such as Bioconductor/Knime. PMML documents support model specification (meta-data, provenance, etc.) as well as normalization and other data transformations. *Limitations and risk mitigation:* This approach applies only to a fixed class of predictive models. However, the supported model types cover the majority of types used in computational biology & chemistry, for example cheminformatics ML repositories[161]. A broader limitation is that PMML does not support arbitrary algorithms. PMML specification supports feature (descriptor) descriptions, but such descriptions may not be recognized by arbitrary tools that support PMML documents. *To mitigate this limitation, we will include scripts to compute features based on their PMML specifications.*

Aim 3: Expand, improve and maintain Pharos. Developed at **NCATS**, the Pharos platform[7] is publicly deployed and actively used by the scientific community (**Box 1**). Initially designed for ~1,800 targets and four protein families (IDG Pilot), Pharos now serves information from 51 diverse datasets on more than 20,000 human proteins. It is comprised of a web-based graphical user interface (GUI) and an industry standard API (HATEOS[162], JSON-LD), with over 1.2 million lines of source code. Based on the Play framework and Akka reactive platform, Pharos is a scalable and high performance enterprise system with average response times under 1s. Designed with input from multiple IDG-funded groups, Pharos is a modern, responsive interface to TCRD, providing seamless search, browsing and filtering activities. Full auto-suggest text search provides flexibility in how and what a user can search for across text, numeric, chemical structure and biological sequence. Sophisticated indexing mechanisms ensure sub-second search times across 3 million entities. Modern UI features such as shopping carts and faceted search enable "serendipitous discovery."[163] Pharos implements multiple interactive visualizations to summarize search results and compare proteins, diseases and ligands (**Fig. 4**). Pharos supports bookmarkable downloads of all KMC data in standard formats (CSV, JSON, SDF) via the GUI and API. The latter allows easy inclusion into custom workflow tools. In summary Pharos represents a robustly engineered foundation that will support the development of the IDG Portal and serve the IDG community by enabling intuitive and effective access to TCKB. By enhancing "serendipitous discovery," Pharos enables researchers to more effectively explore the dark genome. **We will create new functionalities utilizing TCKB to provide actionable support for and around user queries**. These features will enhance search and filtering capabilities and enable integration of community-developed tools and workflows. Pharos supports analytics but will not replace large scale platforms (e.g., Spotfire). To ensure responsiveness, analytics will be precomputed if possible, following the principle of "analysis via browsing."

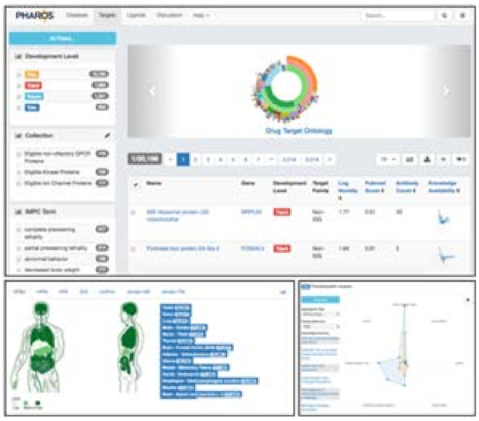


Figure 4. Views of the Pharos UI, highlighting the browsing interface (top), visualization of tissue expression (left) and summary of knowledge availability (right).

1. Knowledge Packages: Key for the KMC effort is to enable investigators within and outside the IDG to process preliminary evidence in an accessible manner, and to design experiments that would clarify function or role of understudied proteins in biology or disease. To support this effort, we will extend the Dossier functionality to enable the creation of "knowledge packages." The Dossier is currently a list of entities supporting various operations (visualize, download, merge, copy, etc.). The "knowledge package" will go beyond the list of entities by enriching them with summaries of multiple targets and automatic suggestions of tools, reagents and relevant publications. We will address this in the following manner:

1.1. *Expand entity coverage.* The Dossier functionality will include all entities supported by Pharos, not just protein and diseases. Dossiers will also support **tkt** categories, allowing users to draw inferences based on the Dossier contents. We will support the inclusion of items associated with a top level **tkt** category. Flagged as associated with a top-level entity, these data would visually be presented as a data "snippet."

1.2. Automatic Dossier enrichment. We will implement a recommendation architecture based on "knowledge similarity," so a *Dossier will be automatically enriched by suggesting related targets and diseases.* These recommendations will be based on similarity matrices, introducing a Target Knowledge Vector (**TKV**) to represents **tkt** category elements (e.g., primary sequence, expression data, etc.). Discrete properties such as phenotype may be represented as TKV elements. Continuous properties (e.g., **JS**, number of PDB entries) may be represented by probability mass functions. TKVs range in dimension from 80 (Tdark) to over 50,000

(Tclin). A generalized Tanimoto coefficient will be used to compute similarity. For **tkt** categories, we will subset TKVs to compute category specific similarities. Using similarity to a protein of interest, Pharos will recommend other proteins, according to computed similarity type (e.g., Genotype similarity vs Structure/Function similarity). Via the Pharos UI we will track usage of the recommendations to further prioritize and inform development. The TKV approach, derived for proteins, is readily applicable to diseases. **NCATS** will consult with Stuart Nelson at **UNM**, and with the **Monarch Initiative**, on how to best compute *disease similarity*. Along with computed recommendations, we will automatically add relevant, linked reagents such as *chemical probes and antibodies for proteins and animal models for diseases*. Automated links to **IDG-C** researchers working on any of the proteins in a dossier will be provided. Users will have the option to ignore such recommendations and links, which will be available both via the GUI and the API.

1.3. Automatically summarize a collection of entities in a Dossier. Enriched Dossiers can join relevant, related pieces of data; however, users need to derive a coherent description of Dossier contents. By summarizing a Dossier using semi-structured text transforms the dossier from a “bag of entities” into a narrative about them. We will develop approaches to summarizing the contents of a Dossier. This effort will proceed in two stages. First, we will summarize the proteins in a dossier using **tkt**-based dimensions; these will also include **TDL** categories, tissues where the proteins are most highly expressed (normal vs disease), etc. We will seek community feedback from **IDG-C** members and external Pharos users to select and expose specific dimensions. The results of this summary will be both tabular and visual, using charts that are appropriate for each **tkt** type being summarized. Visual summaries will employ the HighCharts library[164] to support download and reuse, e.g. inclusion in reports and grant proposals. Second, we will reuse methodology developed in **Aim 3.2.2.** to generate textual summaries of protein collections, initially for proteins within the same target family, subsequently for arbitrary collections, employing a data fusion[165] strategy. Text sources across the set of proteins, as well as numeric and other data types will be merged; this merged set of data will be treated as a single “pseudo-protein” (similar to **Aim 2.2**), and will serve as input for methodology described in **4.1.**

2. Generate data summaries and integrate with external tools and data generators. Pharos presents KMC data linked out to original sources. We aim to present new knowledge derived from **TKCB**:

2.1. Integration hooks to support external tools. Pharos supports several visualizations, from target collection summaries (e.g., sunburst visualization of hierarchical **DTO** terms) to information around a single target (see **Fig. 4**), and including third-party visualizations (e.g., the Harmonogram[6]). To support community contributions of new visualizations and other such tools, the Pharos API will provide “integration hooks,” API endpoints for user-app registration with the Pharos platform. On registration, apps will be executable by Pharos, similar to BARD plugins.[166] This will allow us to reuse community- developed summarization tools.

2.2. Automatically generate a text summary of all data available for a protein. The current protein page lists many independent data types. Our goal is to automatically generate a concise and (close to-) natural language summary of the data: First, we will explore text summarization techniques[167,168]. Since many data types are not plain text, we will develop a rubric to characterize the relevance, importance and reliability of a dataset (or datatype). Second, using “integration hooks” (above) we will plug in existing tools (e.g., Classifier4J[169] and MEAD[170]). We will develop a benchmark to judge the quality of summarization techniques by manually constructing a summary for 100 **IDG List** targets, and will utilize human raters from **MC** and, as we improve, from **IDG-C** members, to score manually generated vs. automated summaries. This approach to crowdsourcing performance metrics[171,172] is appropriate in this setting. Using these benchmarks, we will identify the best target summarization method(s) and generate summaries for each protein webpage.

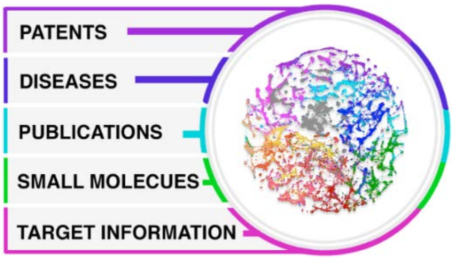


Figure 5. tkt network visualization, highlighting regions according to dominant data types

3. Network integration to enhance discoverability and support complex queries. TKCB contains a variety of data types that relate to different entities (targets, ligands) and different scales (molecular to organismal). While Pharos supports effective, intuitive searching, the user is still restricted to exploring individual dimensions separately. Though filters can be thought of as a multi-dimensional “slice” through **TKCB**, the data types are not exposable in an integrated manner. We address these limitations by:

3.1. Quantifying and visualizing tkt domains. Pharos filters mandate one to explicitly specify the characteristic of interest. Since there are >140 filters, users may not be aware of a relevant filter. We will extract **tkt** subnets corresponding to different knowledge branches. These branches already group knowledge types (see **Table 2**). We will present **tkt** subsets through the lens of a protein family or a disease class. Interactive network visualization and navigation will enable users to zoom into specific **tkt** regions to identify relevant information (**Fig. 5**). **This represents a transformative step as Pharos will assist users in exploring tkt /TCKB,** rather than requiring users to make arbitrary choices of filters. Skupin et al[173] and Boyack et al[174] developed network visualizations of scientific literature and computed clusters characterizing specific scientific domains.

We will implement a clustering approach based on a combination of topic models,[175] enhanced using numeric data collected by the TCKB via the Harmonizome and self-organizing maps (SOM).[176] This clustering process will identify Targets and Diseases (using Named Entity Extraction technology developed at **CPR**) and use these as labels for validation of an unsupervised topic modeling process based on LDA, Latent Dirichlet Allocation[177]. Extracted Proteins and Diseases will be used to identify clusters enriched in specific targets and/or diseases, allowing us to correlate topics as extracted by the LDA model to those targets and diseases. In parallel, we will generate an unsupervised clustering using the SOM and map the resultant clusters to those obtained from LDA. We will characterize each approach using cluster quality metrics.[178] We will generate consensus clusters by combining clusters from both methods. Since all clusters will be pre-computed, we will present each clustering, ordered by cluster quality metrics. We will also explore approaches to include (with **EBI**), as they fit into the LDA workflow. Protein-cluster associations will be stored in a dynamic relational database for rapid retrieval of clusters (**tkt** domains) associated with any given protein. **Tkt** domains will be displayed using an interactive network visualization. In this network, nodes represent **tkt** categories - phenotype, genotype, etc. - with data across the five branches as edges. When visualized in its entirety, this will represent a TCKB-wide clustering. We will implement interactive UI elements, based on the Cytoscape.js infrastructure[179] that will allow the user to drill down into specific knowledge domains. Using visual cues such as color and translucence, users will be able to highlight subnets based on specific features (e.g., a disease term). We will link the current set of filters to this visualization so that 140 pre-existing filters will also be available to select subnets. The proposed network visualization will explicitly bring such connections to the users' attention. The **tkt** networks will be searchable and filterable based on data type and prediction confidence (e.g., from meta-path models, **Aim 2**).

3.2. Integrating search and inference using Bayesian networks. Pharos does not have an integrated view of the associations between entities, which prevents us from supporting sophisticated filters that include entity dependencies. As a result, queries such as "*Find all enzymes annotated with 'farnesyltransferase activity' (molecular function) have 'increased bone mineral density' (associated phenotype), and are modulated by 'antineoplastic'* (drugs)" are not feasible. For understudied targets where such annotations are missing, predicted associations (**Aim 2.3**) will enable similar queries across the **HuP**. We will use Bayesian Networks (BNs) as a probabilistic framework to capture the interactions between ligands, genes, tissues, phenotypes, and diseases. We will leverage Pharos search facilities (i.e., facet, text, structure, and sequence) to seamlessly integrate BN's for various inference tasks. A BN is a directed acyclic graph with nodes represent domain variables and edges denote causal relationships between variables. A BN query amounts to efficient evaluation of the joint probability functions subject to new evidence. BNs allow us to combine exploratory (i.e., explicit searching and filtering) and inference queries that lead to an integrated view of the druggable genome. We will implement BNs following the methodology described in Cooper & Herskovits.[180] Given that network architecture is a key challenge (a 10-node network will have 10^{18} possible structures) in the construction of BNs, we propose to confine the possible network configurations to a set of predefined templates, inspired from the 5 **tkt** branches. Nodes will be organized into layers according to their entity type, and the size of the network will thus depend on the resolution of the data in each layer. By merging **tkt** and BN networks at different data resolution and network configurations, we enable the interrogation of the druggable genome across scales and modalities, resulting in cross-cutting queries that may afford drug repurposing and novel discovery opportunities.

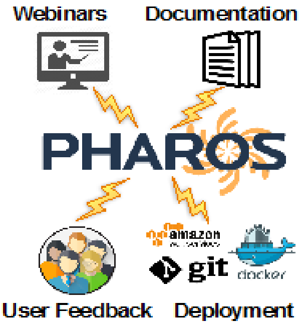


Figure 6. Examples of outreach activities.

Aim 4: Outreach to scientific community. KMC members will engage in a series of activities to leverage **TCKB**, **Pharos** and other **IDG** resources, in order to increase adoption of IDG output (e.g., TDLs, focus on knowledge gaps and understudied proteins), in close coordination with **RDOC**, **DRGCs** and **IDG-SC-PS**. Key to this Aim is sustainability of KMC activities beyond the life of this award. User feedback is essential in identifying limitations and ways to improve Pharos and TCKB. Thus**,** *gathering feedback with the aim of improving content, usability and features desired by the community* are important tasks (see **Fig. 6**).

4.1. User feedback. The KMC will provide several ways to enable user feedback via the Pharos UI. 1) Explicit feedback links will be made available (e.g., a failed search provides a link for users to notify us whether it should have been successful). 2) Each protein detail page supports comments via Disqus.[181] This allows users to post (moderated) comments about a given protein and to start a discussion with other users. 3) We will instrument the Pharos UI to record which interface features (e.g., which filters or which visualizations) receive the most attention in terms of clicks. One use, noted in **Aim 3.1.2**, is to track which classes of recommendation are most frequently used in a Dossier. This data will enable us to prioritize development and drop unused features from the interface, leading to a tailored user experience.

4.2. Educate users. KMC will provide tutorial or feedback sessions, either at Bethesda, MD or when feasible, at user sites. These sessions will serve to introduce users to Pharos and TCKB, highlight their utility to scientific problems, and allow users to provide feedback about the platform. Several sessions have run during the pilot phase. In Phase 2, we will focus on groups from domains not initially considered in the pilot phase (e.g., cardiovascular community). *Work in this aim will be coordinated with the* ***RDOC***.

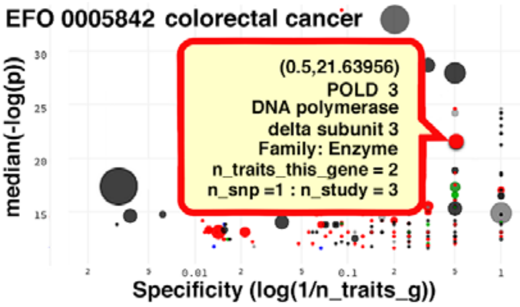


Fig. 7. GWA-X allows visualization of traits per gene on a specificity vs. probability plot.

4.3. External collaborations: 1) IDG helped **IMPC** production centers prioritize production of knockout mouse strains. To date, *130 new knockout strains have been produced as a result*: We will continue to collaborate with IMPC and support IMPC "landing pages" in Pharos. An IDG landing page is already available at the IMPC website, mousephenotye.org. 2) We will adapt the TIN-X platform to visualize the GWAS Catalog by introducing *Specificity* (1/mapped traits per gene) on the X axis, and using the median negative $\log_{10}$(probability) (from the Catalog) on the Y axis. Dot size is proportional to the number of publications linking that particular trait and gene. GWA-X, the GWAS-eXplorer (**Fig. 7**), can be incorporated into the TIN-X framework: instead of Disease Ontology, we will use EBI's **EFO** (experimental factor ontology) as left-side navigation bar. This allows filtering by EFO hierarchy. GWA-X has been prototyped for several IDG List genes (see also LoS from Parkinson).

4.4. Enhance integration with NIH Commons and dissemination. First, we will prepare a new Docker image for every new release of Pharos and TCKB, available at https://hub.docker.com/r/ncats/pharos/, our public repository of images, to encourage usage by industrial scientists (see LoS from BMS, Merck, Genentech, and Lilly). Second, we will monitor **NIH Commons policies and business models**, and deploy our TCKB pipeline and Pharos into the cloud. *Source code developed in this proposal will be available via this repository under a permissive BSD license.* Third, we will implement FAIR principles[41] and expose new annotations in **DrugCentral**, new protein annotations in **TCKB,** with new assertions as **nanopublications**[182] or as the complete DrugCentral/TCKB dump. Fourth, we will work with BD2K colleagues to establish a **Commons Framework Working Group on Benchmark Datasets**, where our focus will be for understudied proteins. Finally, we have investigated deployment of the Pharos platform to Amazon Web Services (AWS) and a Pharos instance is available at http://pharos.ncats.io. Deployments will further enable users with minimal local hardware resources to deploy the system in a cloud setting.

4.5. Enhance documentation. Extensively documented, Pharos features *About* and *Help* pages, frequently asked questions (FAQs) and tooltips on individual UI elements. We will expand the FAQs and add additional help text as suggested by users. We will develop documentation for KMC models and support and third-party model documentation. We will update the REST API documentation using the Swagger toolset[183] and bring it in line with other APIs used in the life sciences (e.g., SmartAPI[184]). We will keep up to date with community-developed standards, e.g., data formats and provenance, coordinating with the **IDG-SC-PS** and **RDOC**.

E. TIMELINE

Aim	Task	Y1		Y2		Y3		Y4		Y5		Y6	
1,2,3	Maintain pilot phase resources	■	■										
1,2,3,4	Initiate research plan organizationally and technically	■	■										
1	Target Knowledge Tree (tkt), nosology, adapt pipelines	■	■										
1	Integrate Monarch phenotypes, GeneWiki, other new datasets		■	■	■	■	■						
1	Monitor, evaluate and integrate other new datasets			■	■	■	■	■	■	■	■	■	
1	Manual curation for prioritized genes and related entities		■		■		■		■		■		■
1	Develop DRGC integration & community feedback pipelines		■	■	■	■	■	■					
2	Develop TCKB 1.0, API w tkt based schema			■	■	■	■	■					
2	Develop IDG Bioconductor Pkg & Knime toolkit				■	■	■	■	■	■	■		
2,3	Develop Knowledge Graph, meta-path & TKV analytics				■	■	■	■	■				
3	Develop Pharos 1.0, API with Knowledge Packages			■	■	■	■	■					
3	Develop Pharos networks, clustering & visualization analytics			■	■	■	■	■	■	■			
3	Develop Pharos community-tool integration system						■	■	■	■			
4	Develop and publicize TargetCentral websites			■	■	■	■	■	■	■	■	■	■
4	Collaborative outreach w/ RDOC via meetings, communities, media	■	■	■	■	■	■	■	■	■	■	■	■
4	Sustainability & data/software sharing efforts prioritized										■	■	■

REFERENCES

1. Edwards, A. M. *et al.* Too many roads not taken. *Nature* **470,** 163–165 (2011).
2. Nickerson, R. S. Confirmation Bias: A Ubiquitous Phenomenon in Many Guises. *Rev. Gen. Psychol.* **2,** 175–220 (1998).
3. Illuminating the Druggable Genome | NIH Common Fund. Available at: https://commonfund.nih.gov/idg/index. (Accessed: 31st January 2017)
4. Ursu, O. *et al.* DrugCentral: online drug compendium. *Nucleic Acids Res.* **45,** D932–D939 (2017).
5. Drug Target Ontology. Available at: http://drugtargetontology.org/. (Accessed: 24th March 2017)
6. Rouillard, A. D. *et al.* The harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins. *Database* **2016,** (2016).
7. Nguyen, D.-T. *et al.* Pharos: Collating protein information to shed light on the druggable genome. *Nucleic Acids Res.* **45,** D995–D1002 (2017).
8. Wikipedia contributors. MySQL. *Wikipedia, The Free Encyclopedia* (2017). Available at: https://en.wikipedia.org/w/index.php?title=MySQL&oldid=764140600. (Accessed: 13th February 2017)
9. Index of /tcrd/download. Available at: http://juniper.health.unm.edu/tcrd/download. (Accessed: 24th March 2017)
10. Target Central REST API. Available at: http://juniper.health.unm.edu/tcrd/api.html. (Accessed: 13th February 2017)
11. Tid, I. IDG. Available at: http://targetcentral.ws/. (Accessed: 24th March 2017)
12. Target Importance and Novelty Explorer (TIN-X). *TIN-X* (2014). Available at: http://newdrugtargets.org/. (Accessed: 13th December 2016)
13. Cannon, D. C. *et al.* TIN-X: Target Importance and Novelty Explorer. *Bioinformatics* (2017).
14. Capurro, R. & Hjørland, B. The concept of information. *Ann. Rev. Info. Sci. Tech.* **37,** 343–411 (2003).
15. Agarwal, P. & Searls, D. B. Can literature analysis identify innovation drivers in drug discovery? *Nat. Rev. Drug Discov.* **8,** 865–878 (2009).
16. The Cochrane Collaboration. Available at: http://www.cochrane.org/. (Accessed: 16th March 2017)
17. Gaulton, A. *et al.* The ChEMBL database in 2017. *Nucleic Acids Res.* **45,** D945–D954 (2017).
18. Roth, B. L. & Kroeze, W. K. Integrated Approaches for Genome-wide Interrogation of the Druggable Non-olfactory G Protein-coupled Receptor Superfamily. *J. Biol. Chem.* **290,** 19471–19477 (2015).
19. Pafilis, E. *et al.* The SPECIES and ORGANISMS Resources for Fast and Accurate Identification of Taxonomic Names in Text. *PLoS One* **8,** e65390 (2013).
20. RFA-RM-16-024: Knowledge Management Center for Illuminating the Druggable Genome (U24). Available at: https://grants.nih.gov/grants/guide/rfa-files/RFA-RM-16-024.html. (Accessed: 24th March 2017)
21. Hopkins, A. L. & Groom, C. R. The druggable genome. *Nat. Rev. Drug Discov.* **1,** 727–730 (2002).
22. Oprea, T. I. *et al.* Unexplored Therapeutic Opportunities in the Human Genome. *Nat. Rev. Drug Discov.* (2017). *Invited as Analysis paper by the editor, Peter Kirkpatrick.*
23. Dunn, O. J. Multiple Comparisons Using Rank Sums. *Technometrics* **6,** 241–252 (1964).
24. Szklarczyk, D. *et al.* STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.* **43,** D447–52 (2015).
25. Wikipedia contributors. Indication (medicine). *Wikipedia, The Free Encyclopedia* (2016). Available at: https://en.wikipedia.org/w/index.php?title=Indication_(medicine)&oldid=736851214. (Accessed: 10th January 2017)
26. Nelson, S. J. Medical Terminologies That Work: The Example of MeSH. in *2009 10th International Symposium on Pervasive Systems, Algorithms, and Networks* (2009). doi:10.1109/i-span.2009.84
27. Nelson, S. J. *et al.* The need for formalization of therapeutic intent. *J. Am. Med. Inform. Assoc.*
28. Amberger, J., Bocchini, C. A., Scott, A. F. & Hamosh, A. McKusick's Online Mendelian Inheritance in Man (OMIM). *Nucleic Acids Res.* **37,** D793–6 (2009).
29. Ashburner, M. *et al.* Gene Ontology: tool for the unification of biology. *Nat. Genet.* **25,** 25–29 (2000).
30. About Gene RIF - Gene - NCBI. Available at: https://www.ncbi.nlm.nih.gov/gene/about-generif. (Accessed: 13th February 2017)
31. Kiermer, V. Antibodypedia. *Nat. Methods* **5,** 860–861 (2008).
32. UniProt Consortium. UniProt: a hub for protein information. *Nucleic Acids Res.* **43,** D204–12 (2015).
33. Lin, Y. *et al.* Drug Target Ontology to Classify and Integrate Drug Discovery Data. *bioRxiv* 117564 (2017). doi:10.1101/117564
34. Santos, R. *et al.* A comprehensive map of molecular drug targets. *Nat. Rev. Drug Discov.* **16,** 19–34 (2017).
35. Oprea, T. I. et al. Unexplored opportunities in the druggable human genome. (2016). Poster available

at http://www.nature.com/nrd/posters/druggablegenome/index.html (Accessed: 16th March 2017)
36. Lenat, D. B. & Feigenbaum, E. A. On the thresholds of knowledge. *Artif. Intell.* **47,** 185–250 (1991).
37. Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y. & Morishima, K. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* **45,** D353–D361 (2017).
38. Cytoscape App Store - stringApp. Available at: http://apps.cytoscape.org/apps/stringapp. (Accessed: 25th March 2017)
39. Hunter, L. & Cohen, K. B. Biomedical language processing: what's beyond PubMed? *Mol. Cell* **21,** 589–594 (2006).
40. Knowles, J. & Gromo, G. Target selection in drug discovery. *Nat. Rev. Drug Discov.* **2,** 63–69 (2003).
41. Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* **3,** 160018 (2016).
42. Collins, F. S. & Tabak, L. A. Policy: NIH plans to enhance reproducibility. *Nature* **505,** 612–613 (2014).
43. Prinz, F., Schlange, T. & Asadullah, K. Believe it or not: how much can we rely on published data on potential drug targets? *Nat. Rev. Drug Discov.* **10,** 712 (2011).
44. Begley, C. G. & Ellis, L. M. Drug development: Raise standards for preclinical cancer research. *Nature* **483,** 531–533 (2012).
45. Ioannidis, J. P. A. Why most published research findings are false. *PLoS Med.* **2,** e124 (2005).
46. Open Science Collaboration. PSYCHOLOGY. Estimating the reproducibility of psychological science. *Science* **349,** aac4716 (2015).
47. Reproducibility Project: Cancer Biology Wiki. Available at: https://osf.io/e81xl/wiki/home/. (Accessed: 17th March 2017)
48. eLife. *eLife* Available at: https://elifesciences.org/collections/reproducibility-project-cancer-biology. (Accessed: 17th March 2017)
49. Baker, M. & Dolgin, E. Cancer reproducibility project releases first results. *Nature* **541,** 269–270 (2017).
50. Replication studies offer much more than technical details. *Nature* **541,** 259–260 (2017).
51. Teague, S. J., Davis, A. M., Leeson, P. D. & Oprea, T. The Design of Leadlike Combinatorial Libraries. *Angew. Chem. Int. Ed Engl.* **38,** 3743–3748 (1999).
52. Oprea, T. I., Davis, A. M., Teague, S. J. & Leeson, P. D. Is there a difference between leads and drugs? A historical perspective. *J. Chem. Inf. Comput. Sci.* **41,** 1308–1315 (2001).
53. Oprea, T. I., Tropsha, A., Faulon, J.-L. & Rintoul, M. D. Systems chemical biology. *Nat. Chem. Biol.* **3,** 447–450 (2007).
54. Garcia-Serna, R., Ursu, O., Oprea, T. I. & Mestres, J. iPHACE: integrative navigation in pharmacological space. *Bioinformatics* **26,** 985–986 (2010).
55. Manallack, D. T., Prankerd, R. J., Yuriev, E., Oprea, T. I. & Chalmers, D. K. The significance of acid/base properties in drug discovery. *Chem. Soc. Rev.* **42,** 485–496 (2013).
56. Liu, T., Oprea, T., Ursu, O., Hasselgren, C. & Altman, R. B. Estimation of Maximum Recommended Therapeutic Dose Using Predicted Promiscuity and Potency. *Clin. Transl. Sci.* **9,** 311–320 (2016).
57. Jensen, A. B. *et al.* Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients. *Nat. Commun.* **5,** 4022 (2014).
58. Vazquez Guillamet, R., Ursu, O., Iwamoto, G., Moseley, P. L. & Oprea, T. Chronic obstructive pulmonary disease phenotypes using cluster analysis of electronic medical records. *Health Informatics J.* (2016). doi:10.1177/1460458216675661
59. Olah, M. *et al.* in *Chemical Biology* 760–786 (Wiley-VCH Verlag GmbH, 2007).
60. Oprea, T. I. *et al.* Associating Drugs, Targets and Clinical Outcomes into an Integrated Network Affords a New Platform for Computer-Aided Drug Repurposing. *Mol. Inform.* **30,** 100–111 (2011).
61. Ursu, O. *et al.* DrugCentral: online drug compendium. *Nucleic Acids Res.* **45,** D932–D939 (2017).
62. Williamson, E. A. *et al.* Targeting the transposase domain of the DNA repair component Metnase to enhance chemotherapy. *Cancer Res.* **72,** 6200–6208 (2012).
63. Oprea, T. I. *et al.* Novel Activities of Select NSAID R-Enantiomers against Rac1 and Cdc42 GTPases. *PLoS One* **10,** e0142182 (2015).
64. Tid, I. IDG. (mar 22, 2016). Available at: http://targetcentral.ws/. (Accessed: 12th December 2016)
65. Oprea, T. I. & Mestres, J. Drug repurposing: far beyond new targets for old drugs. *AAPS J.* **14,** 759–763 (2012).
66. Oprea, T. I. *et al.* Drug Repurposing from an Academic Perspective. *Drug Discov. Today Ther. Strateg.* **8,** 61–69 (2011).
67. Oprea, T. I. & Overington, J. P. Computational and Practical Aspects of Drug Repositioning. *Assay Drug Dev. Technol.* **13,** 299–306 (2015).
68. Nelson, S. J., Powell, T., Srinivasan, S. & Humphreys, B. L. Unified Medical Language

System®(UMLS®) Project. in *Encyclopedia of library and information sciences* 5320–5327 (2010).

69. McCray, A. T. & Nelson, S. J. The representation of meaning in the UMLS. *Methods Inf. Med.* **34,** 193–201 (1995).
70. Nelson, S. J., Zeng, K., Kilbourne, J., Powell, T. & Moore, R. Normalized names for clinical drugs: RxNorm at 6 years. *J. Am. Med. Inform. Assoc.* **18,** 441–448 (2011).
71. Aronson, A. R. *et al.* The NLM Indexing Initiative. *Proc. AMIA Symp.* 17–21 (2000).
72. DistiLD database. Available at: http://distild.jensenlab.org/about.html. (Accessed: 24th March 2017)
73. Pletscher-Frankild, S., Pallejà, A., Tsafou, K., Binder, J. X. & Jensen, L. J. DISEASES: text mining and data integration of disease-gene associations. *Methods* **74,** 83–89 (2015).
74. Binder, J. X. *et al.* COMPARTMENTS: unification and visualization of protein subcellular localization evidence. *Database* **2014,** bau012 (2014).
75. Santos, A. *et al.* Comprehensive comparison of large-scale tissue expression datasets. *PeerJ* **3,** e1054 (2015).
76. Szklarczyk, D. *et al.* STITCH 5: augmenting protein-chemical interaction networks with tissue and affinity data. *Nucleic Acids Res.* **44,** D380–4 (2016).
77. Campillos, M., Kuhn, M., Gavin, A.-C., Jensen, L. J. & Bork, P. Drug target identification using side-effect similarity. *Science* **321,** 263–266 (2008).
78. Monarch Initiative. Available at: https://monarchinitiative.org/. (Accessed: 22nd March 2017)
79. Biomedical Data Translator Program | National Center for Advancing Translational Sciences. *National Center for Advancing Translational Sciences* (2016). Available at: https://ncats.nih.gov/translator. (Accessed: 25th March 2017)
80. Gray, K. A., Yates, B., Seal, R. L., Wright, M. W. & Bruford, E. A. Genenames.org: the HGNC resources in 2015. *Nucleic Acids Res.* **43,** D1079–85 (2015).
81. Stelzer, G. *et al.* The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analyses. *Curr. Protoc. Bioinformatics* **54,** 1.30.1–1.30.33 (2016).
82. The UniProt Consortium. UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* **45,** D158–D169 (2017).
83. Mi, H. *et al.* PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements. *Nucleic Acids Res.* **45,** D183–D189 (2017).
84. Huerta-Cepas, J. *et al.* eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Res.* **44,** D286–93 (2016).
85. Landrum, M. J. *et al.* ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res.* **44,** D862–8 (2016).
86. Shimoyama, M. *et al.* The Rat Genome Database 2015: genomic, phenotypic and environmental variations and disease. *Nucleic Acids Res.* **43,** D743–50 (2015).
87. Halavi, M., Maglott, D., Gorelenkov, V. & Rubinstein, W. MedGen. (2013).
88. dbVar. Available at: https://www.ncbi.nlm.nih.gov/dbvar. (Accessed: 21st March 2017)
89. Rubinstein, W. S. *et al.* The NIH genetic testing registry: a new, centralized database of genetic tests to enable access to comprehensive information and improve transparency. *Nucleic Acids Res.* **41,** D925–35 (2013).
90. Sonnhammer, E. L. L. & Östlund, G. InParanoid 8: orthology analysis between 273 proteomes, mostly eukaryotic. *Nucleic Acids Res.* **43,** D234–9 (2015).
91. Pletscher-Frankild, S., Pallejà, A., Tsafou, K., Binder, J. X. & Jensen, L. J. DISEASES: Text mining and data integration of disease–gene associations. (2014). doi:10.1101/008425
92. Bradley, A. *et al.* The mammalian gene function resource: the International Knockout Mouse Consortium. *Mamm. Genome* **23,** 580–586 (2012).
93. Welter, D. *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* **42,** D1001–6 (2014).
94. OMIM - Online Mendelian Inheritance in Man. Available at: https://omim.org/. (Accessed: 22nd March 2017)
95. Wei, C.-H., Kao, H.-Y. & Lu, Z. PubTator: a web-based text mining tool for assisting biocuration. *Nucleic Acids Res.* **41,** W518–22 (2013).
96. Kibbe, W. A. *et al.* Disease Ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data. *Nucleic Acids Res.* **43,** D1071–8 (2015).
97. Gaulton, A. *et al.* The ChEMBL database in 2017. *Nucleic Acids Res.* **45,** D945–D954 (2017).
98. Wang, Y. *et al.* PubChem: a public information system for analyzing bioactivities of small molecules. *Nucleic Acids Res.* **37,** W623–33 (2009).
99. Ursu, O. *et al.* DrugCentral: online drug compendium. *Nucleic Acids Res.* **45,** D932–D939 (2017).

100. Kutmon, M. *et al.* WikiPathways: capturing the full diversity of pathway knowledge. *Nucleic Acids Res.* **44,** D488–94 (2016).
101. Cerami, E. G. *et al.* Pathway Commons, a web resource for biological pathway data. *Nucleic Acids Res.* **39,** D685–90 (2011).
102. Fabregat, A. *et al.* The Reactome pathway Knowledgebase. *Nucleic Acids Res.* **44,** D481–7 (2016).
103. Rouillard, A. D. *et al.* The harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins. *Database* **2016,** (2016).
104. Szklarczyk, D. *et al.* The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Res.* **45,** D362–D368 (2017).
105. Davis, A. P. *et al.* The Comparative Toxicogenomics Database: update 2017. *Nucleic Acids Res.* **45,** D972–D978 (2017).
106. Greene, C. S. *et al.* Understanding multicellular function and disease with human tissue-specific networks. *Nat. Genet.* **47,** 569–576 (2015).
107. Orchard, S. *et al.* The MIntAct project--IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res.* **42,** D358–63 (2014).
108. Southan, C. *et al.* The IUPHAR/BPS Guide to PHARMACOLOGY in 2016: towards curated quantitative interactions between 1300 protein targets and 6000 ligands. *Nucleic Acids Res.* **44,** D1054–68 (2016).
109. Hernandez-Boussard, T. *et al.* The pharmacogenetics and pharmacogenomics knowledge base: accentuating the knowledge. *Nucleic Acids Res.* **36,** D913–8 (2008).
110. NIH LINCS Program. Available at: http://lincsproject.org/. (Accessed: 22nd March 2017)
111. The Gene Ontology Consortium. Expansion of the Gene Ontology knowledgebase and resources. *Nucleic Acids Res.* **45,** D331–D338 (2017).
112. Rose, P. W. *et al.* The RCSB protein data bank: integrative view of protein, gene and 3D structural information. *Nucleic Acids Res.* **45,** D271–D281 (2017).
113. Finn, R. D. *et al.* The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.* **44,** D279–85 (2016).
114. Finn, R. D. *et al.* InterPro in 2017-beyond protein family and domain annotations. *Nucleic Acids Res.* **45,** D190–D199 (2017).
115. Dawson, N. L. *et al.* CATH: an expanded resource to predict protein function through structure and sequence. *Nucleic Acids Res.* **45,** D289–D295 (2017).
116. Schaeffer, R. D., Liao, Y., Cheng, H. & Grishin, N. V. ECOD: new developments in the evolutionary classification of domains. *Nucleic Acids Res.* **45,** D296–D302 (2017).
117. Wheeler, H. E. *et al.* Survey of the Heritability and Sparse Architecture of Gene Expression Traits across Human Tissues. *PLoS Genet.* **12,** e1006423 (2016).
118. Petryszak, R. *et al.* Expression Atlas update--an integrated database of gene and protein expression in humans, animals and plants. *Nucleic Acids Res.* **44,** D746–52 (2016).
119. Uhlén, M. *et al.* Proteomics. Tissue-based map of the human proteome. *Science* **347,** 1260419 (2015).
120. Kim, M.-S. *et al.* A draft map of the human proteome. *Nature* **509,** 575–581 (2014).
121. The Cancer Genome Atlas Home Page. *The Cancer Genome Atlas - National Cancer Institute* Available at: https://cancergenome.nih.gov/. (Accessed: 25th March 2017)
122. Borrel, A., Regad, L., Xhaard, H., Petitjean, M. & Camproux, A.-C. PockDrug: A Model for Predicting Pocket Druggability That Overcomes Pocket Estimation Uncertainties. *J. Chem. Inf. Model.* **55,** 882–895 (2015).
123. Wong, G. Y., Leung, F. H. F. & Ling, S. H. Predicting protein-ligand binding site using support vector machine with protein properties. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **10,** 1517–1529 (2013).
124. Lee, H. S. & Im, W. Ligand binding site detection by local structure alignment and its performance complementarity. *J. Chem. Inf. Model.* **53,** 2462–2470 (2013).
125. Schmidtke, P., Bidon-Chanal, A., Luque, F. J. & Barril, X. MDpocket: open-source cavity detection and characterization on molecular dynamics trajectories. *Bioinformatics* **27,** 3276–3285 (2011).
126. Zhang, Z., Li, Y., Lin, B., Schroeder, M. & Huang, B. Identification of cavities on protein surface using multiple computational approaches for drug binding site prediction. *Bioinformatics* **27,** 2083–2088 (2011).
127. Le Guilloux, V., Schmidtke, P. & Tuffery, P. Fpocket: an open source platform for ligand pocket detection. *BMC Bioinformatics* **10,** 168 (2009).
128. Bologa, C. G. *et al.* Virtual and biomolecular screening converge on a selective agonist for GPR30. *Nat. Chem. Biol.* **2,** 207–212 (2006).
129. Dennis, M. K. *et al.* In vivo effects of a GPR30 antagonist. *Nat. Chem. Biol.* **5,** 421–427 (2009).
130. George Thompson, A. M. *et al.* Discovery of a specific inhibitor of human GLUT5 by virtual screening

and in vitro transport evaluation. *Sci. Rep.* **6,** 24240 (2016).
131. Parker, C. G. *et al.* Ligand and Target Discovery by Fragment-Based Screening in Human Cells. *Cell* **168,** 527–541.e29 (2017).
132. Molina-Navarro, M. M. *et al.* Differential gene expression of cardiac ion channels in human dilated cardiomyopathy. *PLoS One* **8,** e79792 (2013).
133. Ortega, A. *et al.* Patients with Dilated Cardiomyopathy and Sustained Monomorphic Ventricular Tachycardia Show Up-Regulation of KCNN3 and KCNJ2 Genes and CACNG8-Linked Left Ventricular Dysfunction. *PLoS One* **10,** e0145518 (2015).
134. Su, A. I., Good, B. M. & van Wijnen, A. J. Gene Wiki Reviews: Marrying crowdsourcing with traditional peer review. *Gene* **531,** 125 (2013).
135. Burgstaller-Muehlbacher, S. *et al.* Wikidata as a semantic framework for the Gene Wiki initiative. *Database* **2016,** (2016).
136. Giles, J. Internet encyclopaedias go head to head. *Nature* **438,** 900–901 (2005).
137. Freedman, D. *Statistical Models: Theory and Practice*. (2005).
138. Cortes, C. & Vapnik, V. Support-vector networks. *Mach. Learn.* **20,** 273–297 (1995).
139. Russell, S. & Norvig, P. *Artificial Intelligence: A Modern Approach, Global Edition*. (2016).
140. Wold, S., Sjöström, M. & Eriksson, L. PLS-regression: a basic tool of chemometrics. *Chemometrics Intellig. Lab. Syst.* **58,** 109–130 (2001).
141. Tin Kam Ho & Ho, T. K. The random subspace method for constructing decision forests. *IEEE Trans. Pattern Anal. Mach. Intell.* **20,** 832–844 (1998).
142. Haykin, S. *Neural Networks: A Comprehensive Foundation*. (IEEE, 1999).
143. Sun, Y., Barber, R., Gupta, M., Aggarwal, C. C. & Han, J. Co-author Relationship Prediction in Heterogeneous Bibliographic Networks. in *2011 International Conference on Advances in Social Networks Analysis and Mining* (2011). doi:10.1109/asonam.2011.112
144. Liang, W., He, X., Tang, D. & Zhang, X. in *Lecture Notes in Computer Science* 305–319 (2016).
145. Himmelstein, D. S. & Baranzini, S. E. Heterogeneous Network Edge Prediction: A Data Integration Approach to Prioritize Disease-Associated Genes. *PLoS Comput. Biol.* **11,** e1004259 (2015).
146. Fu, G. *et al.* Predicting drug target interactions using meta-path-based semantic network analysis. *BMC Bioinformatics* **17,** 160 (2016).
147. Mungall, C. *et al.* Relationship Ontology. *zenodo* Available at: https://zenodo.org/record/437891.
148. Bioconductor. Available at: https://www.bioconductor.org/. (Accessed: 19th March 2017)
149. TensorFlow: An open-source software library for Machine Intelligence. *TensorFlow* Available at: https://www.tensorflow.org/. (Accessed: 19th March 2017)
150. TensorFlow R package. Available at: https://rstudio.github.io/tensorflow/. (Accessed: 19th March 2017)
151. The Comprehensive R Archive Network. Available at: https://cran.r-project.org/. (Accessed: 19th March 2017)
152. Bioconductor - Courses and Conferences. Available at: https://www.bioconductor.org/help/course-materials/. (Accessed: 19th March 2017)
153. Bioconductor for Genomic Data Science - Johns Hopkins University | Coursera. *Coursera* Available at: https://www.coursera.org/learn/bioconductor. (Accessed: 20th March 2017)
154. Szklarczyk, D. *et al.* STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.* **43,** D447–52 (2015).
155. Pasek, J., from Alex Tahk, W. S. A., by Gene Culter, S. C. M. F. R.-C. A. C. & Schwemmle., M. weights: Weighting and Weighted Statistics. (2016).
156. Cavallari, J. M., Jawaro, T. S., Awad, N. I. & Bridgeman, P. J. Glucagon for refractory asthma exacerbation. *Am. J. Emerg. Med.* **35,** 144–145 (2017).
157. Akbary, A. M., Wirth, K. J. & Schölkens, B. A. Efficacy and tolerability of Icatibant (Hoe 140) in patients with moderately severe chronic bronchial asthma. *Immunopharmacology* **33,** 238–242 (1996).
158. Machado-Carvalho, L., Roca-Ferrer, J. & Picado, C. Prostaglandin E2 receptors in asthma and in chronic rhinosinusitis/nasal polyps with and without aspirin hypersensitivity. *Respir. Res.* **15,** 100 (2014).
159. Mathias, S. L. *et al.* The CARLSBAD database: a confederated database of chemical bioactivities. *Database* **2013,** bat044 (2013).
160. PMML 4.1 - General Structure. Available at: http://dmg.org/pmml/v4-1/GeneralStructure.html. (Accessed: 25th March 2017)
161. Ruusmann, V., Sild, S. & Maran, U. QSAR DataBank repository: open and linked qualitative and quantitative structure-activity relationship models. *J. Cheminform.* **7,** 32 (2015).
162. Fielding, R. T. & Taylor, R. N. Principled design of the modern Web architecture. in *Proceedings of the 22nd international conference on Software engineering* 407–416 (ACM, 2000).

163. Sakai, T. & Nogami, K. Serendipitous search via wikipedia: a query log analysis. in *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval* 780–781 (ACM, 2009).
164. Highsoft, A. S. Highcharts-Interactive JavaScript charts for your webpage. (2013).
165. Ritchie, M. D., Holzinger, E. R., Li, R., Pendergrass, S. A. & Kim, D. Methods of integrating data to uncover genotype-phenotype interactions. *Nat. Rev. Genet.* **16,** 85–97 (2015).
166. Howe, E. A. *et al.* BioAssay Research Database (BARD): chemical biology and probe-development enabled by structured metadata and result types. *Nucleic Acids Res.* **43,** D1163–70 (2015).
167. Mishra, R. *et al.* Text summarization in the biomedical domain: a systematic review of recent research. *J. Biomed. Inform.* **52,** 457–467 (2014).
168. Gonzalez, G. H., Tahsin, T., Goodale, B. C., Greene, A. C. & Greene, C. S. Recent Advances and Emerging Applications in Text and Data Mining for Biomedical Discovery. *Brief. Bioinform.* **17,** 33–42 (2016).
169. Lothian, N. Classifier4j. (2005).
170. [No title]. Available at: http://www.summarization.com/mead/. (Accessed: 26th March 2017)
171. Good, B. M. & Su, A. I. Crowdsourcing for bioinformatics. *Bioinformatics* **29,** 1925–1933 (2013).
172. Khare, R., Good, B. M., Leaman, R., Su, A. I. & Lu, Z. Crowdsourcing in biomedicine: challenges and opportunities. *Brief. Bioinform.* **17,** 23–32 (2016).
173. Skupin, A., Biberstine, J. R. & Börner, K. Visualizing the topical structure of the medical sciences: a self-organizing map approach. *PLoS One* **8,** e58779 (2013).
174. Boyack, K. W. *et al.* Clustering more than two million biomedical publications: comparing the accuracies of nine text-based similarity approaches. *PLoS One* **6,** e18029 (2011).
175. Liu, L., Tang, L., Dong, W., Yao, S. & Zhou, W. An overview of topic modeling and its current applications in bioinformatics. *Springerplus* **5,** 1608 (2016).
176. Kohonen, T. Self-organized formation of topologically correct feature maps. *Biol. Cybern.* **43,** 59–69 (1982).
177. Blei, D. M., Ng, A. Y. & Jordan, M. I. Latent Dirichlet Allocation. *J. Mach. Learn. Res.* **3,** 993–1022 (2003).
178. Kaufman, L. & Rousseeuw, P. J. in *Finding Groups in Data* 126–163 (John Wiley & Sons, Inc., 1990).
179. Franz, M. *et al.* Cytoscape.js: a graph theory library for visualisation and analysis. *Bioinformatics* **32,** 309–311 (2016).
180. Cooper, G.F. & Herskovits, E. A Bayesian method for the induction of probabilistic networks from data. *Mach Learn* 9: 309-347 (1992).
181. Disqus – The #1 way to build your audience. *Disqus* Available at: https://disqus.com/. (Accessed: 26th March 2017)
182. Groth, P., Gibson, A. & Velterop, J. The anatomy of a nanopublication. *Inf. Serv. Use* **30,** 51–56 (2010).
183. Swagger – The World’s Most Popular Framework for APIs. Available at: http://swagger.io/. (Accessed: 26th March 2017)
184. smartAPI Specification. Available at: https://websmartapi.github.io/smartapi_specification/. (Accessed: 26th March 2017)

Consortium/ Contractual Arrangements

The University of New Mexico Health Sciences Center (UNM HSC) is prepared to enter into subcontract arrangements with the European Bioinformatics Institute (EBI), the National Center for Advancing Translational Sciences (NCATS), the University of Copenhagen, and Molecular Connections.

Authorized officials of the partnering institutions have agreed to collaborate on the proposed NIH project, *Knowledge Management Center for Illuminating the Druggable Genome*. The project period is November 1, 2017 through October 31, 2023.

The effort will be led by Dr. Tudor Oprea, in UNM HSC's Department of Internal Medicine, Division of Translational Informatics. Project Leads at the partnering institutions are Dr. Andrew Leach (EBI), Dr. Anton Simeonov (NCATS), Dr. Lars Jensen (U Copenhagen), and Dr. Arathi Raghunath (Molecular Connections).

Upon receipt of a Grant Notice of Award from the National Institutes of Health, UNM HSC will issue a subaward agreement approved by the Regents of the University of New Mexico, and as outlined in the subawardee's budget and justification as provided for this application.

Each institution will each make significant scientific contributions to the study, and will comply with all pertinent NIH regulations and policies. The Project Lead at each institution will take responsibility for the scientific conduct of their respective component of the study.

List of Letters of Support for KMC Phase II:

1. Ola Spjuth - Uppsala University
2. Michel Dumontier - Maastricht University
3. Renxiao Wang - Shanghai Institute of Organic chemistry, Chinese Academy of Sciences
4. Olivier Taboureau - University of Paris Diderot
5. David Wild - Data2Discovery
6. Jeffrey M. Blaney - Genentech
7. Jordi Mestres - Chemotargets
8. Carleton R. Sage - Beacon Discovery
9. Timothy J. Mitchison - Harvard University
10. Meir Glick - Merck
11. Tudor Groza - Garvan Institute
12. Andrea Cavalli - University of Bologna
13. Peter Sorger - Harvard University
14. Garland Marshall - Washington University in St. Louis
15. Helen Parkinson - EMBL-EBI
16. Michal Vieth - Eli Lilly
17. Melissa Haendel - Monarch Initiative
18. Ying Ding - Indiana University
19. Steve Brown and Terrence Meehan - IMPC
20. Stephen Johnson - BMS
21. Evangelos A. Coutsias - Stony Brook University
22. Scott E. Martin – Genentech
23. Finkbeiner – Gladstone Institutes
24. Susumu Tomita – Yale University
25. Anton Simeonov – NCATS

UPPSALA
UNIVERSITET

2017-03-16

1 (1)

Ola Spjuth, Docent
Associate Professor

Department of
Pharmaceutical Biosciences
and Science for Life
Laboratory

Visiting address:
Husargatan 3

Postal address:
Box 591, Biomedicum
SE-751 24 Uppsala
Sweden

Telephone:
+46 18–471 4281

Web page:
http://www.farmbio.uu.se

To whom it may concern,

I would hereby like to add my support to the *Illuminating the Druggable Genome* (IDG) project. My research group at Uppsala University is focused on developing methods for improving predictive pharmacology, toxicology and metabolism using machine learning methods. Having access to relevant integrated data sources is of uttermost importance for us, and the Target Central Resource Database with the Pharos interface constitute important and useful components. Further, we believe the Target Development Level classification scheme to be useful for target prioritization in our research.

My group is currently setting up automated systems for continuous modeling where predictive models are re-trained as new data becomes available. We deploy these models on cloud computing resources, or more specifically using containers orchestrated by the Kubernetes framework developed by Google. The analysis pipelines we automate using scientific workflow software, and we refine existing solutions including Luigi and Pachyderm to support the features necessary to sustain continuous modeling.

We are aiming to work with IDG to develop e.g. target druggability models that can be contributed back to IDG KMC, at no cost for the consortium.

Sincerely,

Ola Spjuth, Associate Professor
Department of Pharmaceutical Biosciences and
Science for Life Laboratory
PO Box 591, Uppsala University
Uppsala, Sweden

Email: ola.spjuth@farmbio.uu.se
Ph: +46 (0)70 425 06 28

Maastricht University

March 28, 2017

Dear Dr. Oprea,

It is my pleasure to write this letter in support of your U24 application to develop and implement a Knowledge Management Center (KMC) for the Illuminating the Druggable Genome (IDG) Program. The aim to create a research consortium focused on the elucidation of functionally undercharacterized proteins that make part of the druggable genome is of clear and undeniable importance. As a biomedical informatics researcher focused on developing computational methods for drug discovery, access to the highest quality data and tools is critical to my research.

I am most familiar with two tools of your doing: DrugCentral and Pharos. Our group uses DrugCentral as a source of high quality drug information that we use in our machine learning methods predict drug indications. In particular, our preliminary analysis determined that the drug indications available in DrugCentral were of much higher quality than those available in other sources, owing largely to the semi-automated pipeline that you have put in place. Moreover, this higher quality translated into better predictions as compared to when we used other non-curated sources such as SIDER or improved coverage over sources such as NDF-RT. Importantly, the effort that you have put into curating these indications to the UMLS means that we can reliably use ontologies as part of our learning process, which consistently outperforms using other sources of indications that are described in natural language, such as DrugBank. Moreover, the breadth of coverage also means that we have better predictions across a broad set of therapeutic areas.

I only recently began to use Pharos, but I must say that it is impressive. Indeed, the ability to rapidly drill down through diseases, targets, and ligands across a dozen or so sources enables our group to ensure that our discoveries are truly novel, and not just a symptom of poor data collection. It’s become a go-to reference tool, and we certainly look forward to seeing more of these kinds of tools and technologies being developed. Should there ever be an opportunity for us to contribute use cases, provide feedback, or “test drive” any new software or features arising from the consortium, do let me know. Best of luck on your application!

Sincerely,

Michel

Michel Dumontier

Distinguished Professor of Data Science
Maastricht University
michel.dumontier@maastrichtuniversity.nl
http://dumontierlab.com

SHANGHAI INSTITUTE OF ORGANIC CHEMISTRY
CHINESE ACADEMY OF SCIENCES
345 Lingling Road, Shanghai 200032, P. R. China

Mar 5th, 2017

To Whom It May Concern:

I am writing this letter to indicate my strong support to the **PHAROS** and the **DrugCentral** web site.

I am leading a group at the Shanghai Institute of Organic Chemistry, Chinese Academy of Science. My group is engaging on molecular-targeted drug design by combining bioinformatics, molecular modeling, organic synthesis and biological studies. My group members have been using the PHAROS and the DrugCentral web site regularly for information related to drug targets and small-molecule marketed drugs or drug candidates. In my opinion, these two web sites are valuable public resources for researchers working in this field. Although the scale of both web sites is relatively smaller as compared to some data resources (such as ChEMBL@EBI or PubChem@NIH), they provide unique information or annotations that are not available from those more comprehensive ones. For example, the knowledge of the "druggable genome" provided by PHAROS is exceptionally useful. Besides, all information on both web sites are well-organized, eye-catching, and convenient to be delivered to the fingertips of the user.

In conclusion, we love to use those two web sites. We certainly hope that they will be supported continuously by governmental funding.

Yours sincerely,

Renxiao Wang

Renxiao Wang, Professor
State Key Laboratory of Bioorganic Chemistry, Shanghai Institute of Organic chemistry
Chinese Academy of Sciences
Shanghai 200032, China

DATA2DISCOVERY INC

P.O. Box 5456
Bloomington, IN 47407
812.202.6190
www.d2discovery.com

March 10th, 2017

Professor Tudor I Oprea, MD PhD
Translational Informatics Division
Department of Internal Medicine
University of New Mexico Health Sciences Center

Dear Professor Oprea,

Re: “Illuminating the Druggable Genome Knowledge Management Center” U24 Application

I write to express strong support for your proposal for continuation of the Knowledge Management Center of the Illuminating the Druggable Genome project, as it enters Phase 2 (RFA-RM-16-024). Your curation efforts from DrugCentral and TargetCentral are critical in that they provide accurate information regarding drugs and their mode of action. This is just one aspect that informs the “druggable” genome in Pharos, as one of the many trascriptomic, phenotypic and genomic information sources that you combined in that online resource.

I am Founder and President of Data2Discovery, a corporation that is partnering with pharmaceutical companies to help them build the next generation of translational data and computational infrastructure for drug discovery. For those companies, access to high quality, curated information from the public domain that is provided in a sustainable fashion is critical, and of very high value. We believe pharmaceutical companies are destined to become “data companies” in that all discovery will be data-driven. The ongoing work of the Knowledge Management Center will directly facilitate new discoveries in pharmaceutical companies, as well as fuelling small companies like ours that are building infrastructure on top of high quality data sources. It is worth noting that the innovation in the Knowledge Management Center together with our innovation in our company has already resulted in submission of an NIH SBIR proposal to feasibility test a highly novel approach to drug repurposing that we think could have a dramatic impact on pharmaceutical research.

It is my opinion that your proposal has potential of game changing impact for both academic and industrial research groups with respect to target selection and prioritization, molecular probe identification and disease tailoring. The international multidisciplinary team you have assembled is well suited to continue developing automated informatics pipelines for the entire proteome and to bring forward new interfaces and data analytics in an aggressive timeline.

The tools that you are developing are well suited for the analysis and prioritization of dark genes, such as the 395 genes of interest to the Phase 2 of this project.

Yours truly,

David Wild
President & Founder

université PARIS DIDEROT PARIS 7

MTi Molécules Thérapeutiques in silico

Instituts thématiques Inserm
Institut national de la santé et de la recherche médicale

Letter of support for Pr. TI Oprea NIH grant renewal

March 06, 2017

Recommendation letter for Pr. TI Oprea

It is with utmost interest that we, at the University of Paris Diderot, have learned about the efforts made by Pr. TI Oprea at UNM, in the process of applying for a continuation of the IDG initiative, through NIH grant.

In our group at the unit of "Molecules Therapeutiques in SIlico", we are highly interested in polypharmacology and the integration of chemogenomics data with other biological sources of information, i.e. transcriptomics, protein-protein interactions, pathways, clinicals, among others. Therefore, we were extremely thankful, when Pharos was released in 2016. It allowed us to integrate one of the most accurate chemogenomics database to our in house chemical-diseases platform and start to use it in our daily research study.

We planned also to educate our students to the different implementation in the Pharos database during our master in "In silico drug design" at Paris Diderot. The performance of such tool is usually highly appreciated by our students, as the tool is very flexible and allows exploiting outcomes rapidly.

Such tool will be definitively of interest in industry and academia for the annotation and prediction of bioactive compounds in drug discovery.

Therefore, I am convinced that Pharos will become one of the top databases in chemogenomics to consult in drug discovery and I would like to give to TI. Oprea and his application my strongest possible recommendation, so they will be able to pursue their active work on keeping their tool a MUST to have.

Yours sincerely,

Olivier Taboureau, Professor, PhD

Genentech
A Member of the Roche Group

March 14, 2017

Prof. Tudor I Oprea, MD PhD
Translational Informatics Division
Department of Internal Medicine
University of New Mexico Health Sciences Center
Via email to TOprea@salud.unm.edu

Re: "Illuminating the Druggable Genome Knowledge Management Center" U24 Application

I am excited about your proposal to continue the "Knowledge Management Center of Illuminating the Druggable Genome" project, as it enters Phase 2 (RFA-RM-16-024). Your curation efforts from DrugCentral and TargetCentral provide accurate information regarding drugs and their mode of action. This is just one aspect that informs the "druggable" genome in Pharos, as one of the many transcriptomic, phenotypic and genomic information sources that you combine in that online resource.

I have led Genentech's Small Molecule Drug Discovery Computational Chemistry and Cheminformatics Group since starting at Genentech in Oct 2007. I received my Ph.D. in Pharmaceutical Chemistry at UCSF in 1982 and have worked in industrial drug discovery research ever since, in both large and small pharma/biotech. My main areas of focus have been QSAR (Quantitative Structure-Activity Relationships), ligand-based design approaches, structure-based design, fragment-based discovery, and industrial chemical informatics.

My group collaborates with all ~20 ongoing small molecule discovery project teams. We are responsible for all aspects of modelling and informatics, including understanding proteins which are biologically and/or structurally related to our primary project targets. We focus primarily on our in-house data, which is a full-time effort. We have little resource available to tackle curating, integrating, and analysing the many valuable publicly available databases. Your software and database tools will complement our in-house efforts very well. Your approaches are particularly creative and insightful.

Your proposal will have high impact for both academic and industrial research groups with respect to target selection and prioritization, molecular probe identification, and identifying potential off-targets. Your international, multidisciplinary team is well suited to continue developing automated informatics pipelines for the entire proteome with new interfaces and data analytics.

The tools that you are developing are well suited for the analysis and prioritization of “dark” genes, such as the 395 genes of interest to the Phase 2 of this project. I find this area to be particularly exciting, as this is an underexplored area with great potential for both industry and academia.

Sincerely,

Jeffrey M Blaney

Jeffrey M. Blaney, Ph.D.
Director, Computational Chemistry & Cheminformatics
Small Molecule Drug Discovery, Genentech
1 DNA Way, South San Francisco, CA 94080
blaney.jeff@gene.com

Chemotargets

Barcelona, March 14th, 2017

To: Prof. Tudor I Oprea, MD PhD
Translational Informatics Division
Department of Internal Medicine
University of New Mexico Health Sciences Center
Via email to TOprea@salud.unm.edu

Re: "Illuminating the Druggable Genome Knowledge Management Center" U24 Application

As a regular user of the valuable resources developed in your Translational Informatics Division, I enthusiastically support your proposal for continuation of the Knowledge Management Center of the Illuminating the Druggable Genome (IDG) project, as it enters Phase 2 (RFA-RM-16-024). Your curation and integrative efforts in DrugCentral and TargetCentral provide accurate information regarding the connection between drugs, their targets, and their therapeutic use. This is just one aspect that illuminates the "druggable" genome in Pharos, a long-awaited on-line resource combining transcriptomic, phenotypic and genomic information.

Our company develops novel computational methodologies and graphical tools to help designing small molecule pharmaceuticals in the context of presonalized medicine. To do that, we rely on carefully annotated data on drugs, targets, and diseases from both public and commercial sources that we then integrate and use internally to generate predictive models.

Therefore, I strongly believe that your proposal has the potential of having a game-changing impact for precision modelling in both academia and industry. The international multidisciplinary team you have assembled is well suited to continue developing automated informatics pipelines for the entire proteome and to bring forward new interfaces and data analytics in an aggressive timeline. Last but not least, the tools that you are developing are well suited for the analysis and prioritization of the 395 dark genes of interest to the Phase 2 of this project.

Yours sincerely,

Dr. Jordi Mestres, CEO
Chemotargets S.L.
Barcelona, Spain
Email: jordi.mestres@chemotargets.com

Harvard Medical School
DEPARTMENT OF SYSTEMS BIOLOGY

200 Longwood Avenue
Warren Alpert 536
Boston, Massachusetts 02115-5730
(617) 432-3805

Timothy J. Mitchison , Ph.D., F.R.S.
Hasib Sabbagh Professor of Systems Biology
timothy_mitchison@hms.harvard.edu

March 13, 2017

To: Prof. Tudor I Oprea, MD PhD
Translational Informatics Division
Department of Internal Medicine
University of New Mexico Health Sciences Center
Via email to TOprea@salud.unm.edu

Re: "Illuminating the Druggable Genome Knowledge Management Center" U24 Application

Dear Dr. Oprea

This letter is to state my enthusiastic support regarding your proposal for continuation of the Knowledge Management Center of the Illuminating the Druggable Genome project, as it enters Phase 2 (RFA-RM-16-024). You may recall that in October 2015, we both attended a conference in Berlin, Germany. Upon learning about the IDG initiative, I was intrigued to learn that NIH is encouraging scientists to migrate towards the study of what you call the "ignorome" or "dark targets". I'm extremely enthusiastic re your efforts to generate tools to prompt experimental study of less-studied and un-studied candidate targets.

When I teach pharmacology to PhD students, and in the research work in my group and local labs, I am continually struck by how much money and effort is spent on a rather small number of well know drug targets. We make a big point of this in the pharmacology class I direct, and I encourage students to start thinking more "out of the box". However, it is challenging to find novel candidate targets by browsing pubmed, which is heavily biased towards known targets, or raw genomic resources, which are very difficult for biochemists to use effectively. In this light your efforts to curate medicines in DrugCentral, and proteins via TargetCentral, are very useful resources regarding drugs and targets. I believe that this type of information will spur novel approaches to drug discovery in academia and pahrma/biotech. I also look forward to using them to develop teaching units for my class.

The work of my own group relates particularly to microtubules, cancer and neurodegeneration. We also have an active interest in systems biology, which relies on integrating functional, structural and genomics aspects of drugs and targets.

It is my opinion that your proposal can impact the process of target prioritization as it helps bring focus on dark proteins, specifically with respect to phenotype and disease associations. Since 2015, when I became aware of your international team effort, I have followed with interest your tool development for mining the “dark genome”. Your team is well suited for the continued deployment of automated pipelines and human curation concerning the human proteome, as well as for generating analytics and visualization apps for the dark genes of interest to the IDG Phase 2.

Sincerely,

T. J.

Timothy J. Mitchison
Professor of Systems Biology

BEACON
DISCOVERY
6118 Nancy Ridge Drive, San Diego, CA 92121

March 10, 2017

To: Prof. Tudor I Oprea, MD PhD
Translational Informatics Division
Department of Internal Medicine
University of New Mexico Health Sciences Center
Via email to TOprea@salud.unm.edu

Re: "Illuminating the Druggable Genome Knowledge Management Center" U24 Application

Dear Dr. Oprea:

I would like to express my strong support for the continuation of the Knowledge Management Center of the Illuminating the Druggable Genome project, as it enters Phase 2 (RFA-RM-16-024). The curation efforts from DrugCentral and TargetCentral provide accurate information regarding drugs and their mode of action. The collation of these data informs the "druggable" genome in Pharos, as one of the many transcriptomic, phenotypic and genomic information sources that you combined in that online resource.

I have been working in the field of drug discovery and target enablement for almost 20 years. I have a Ph.D. in Biochemistry and Molecular Biology from the university of California, Santa Barbara and did my postdoctoral work in X-ray Crystallography in Robert Stroud's lab at the University of California, San Francisco. My work relates to discovery of novel targets to treat CNS and liver diseases and we have an active interest in integrating structural, functional and disease aspects of G-Protein Coupled Receptors (GPCRs).

Since target enablement increasingly depends on the combination of multiple lines of informatics data sources, I strongly believe that your proposal has the potential for fundamental impact for both academic and industrial research groups. Rational target selection, prioritization and molecular probe identification are the cornerstones of the discovery of new medicines. Therefore, it is my hope that the international multidisciplinary team you have assembled continues to develop automated informatics tools and brings forward new approaches to genomic and proteomic integration in the future.

The tools that you are developing are well suited for the analysis and prioritization of dark genes. We are especially interested in GPCRs, which make up a significant fraction of the 395 genes of interest to the Phase 2 of this project.

Sincerely,

Carleton R. Sage, Ph.D.
Vice President, Computational Sciences
Beacon Discovery
csage@beacondiscovery.com

MERCK

Meir Glick, Ph.D.
Director
Informatics

Merck Research Laboratories
33 Avenue Louis Pasteur
Boston, MA 02115-5727
USA
Office 617-992-3221
Cell 781-856-2268

Date: March 10th, 2017

To: Prof. Tudor I Oprea, MD PhD
Translational Informatics Division
Department of Internal Medicine
University of New Mexico Health Sciences Center
Via email to TOprea@salud.unm.edu

Re: "Illuminating the Druggable Genome Knowledge Management Center" U24 Application

Dear Prof. Oprea,

This is to state my enthusiastic support regarding your proposal for continuation of the Knowledge Management Center of the Illuminating the Druggable Genome project, as it enters Phase 2 (RFA-RM-16-024). Your curation efforts from DrugCentral and TargetCentral provide accurate information regarding drugs and their mode of action. This is just one aspect that informs the "druggable" genome in Pharos, as one of the many transcriptomic, phenotypic and genomic information sources that you combined in that online resource.

As the Director of Informatics at Merck where I am responsible for the vision, strategy and delivery for a world class group that influences all of Research. Before joining Merck in 2015 I was the head of the in silico Lead Discovery at the Novartis Institutes for BioMedical Research where I worked since 2003. I was a trained as a postdoc in the area of computational chemistry in the University of Oxford and received my PhD in computational chemistry from the Hebrew University of Jerusalem. My work relates to Alzheimer's disease and we have an active interest in integrating structural, functional and disease aspects of proteases.

It is my opinion that your proposal has potential of game changing impact for both academic and industrial research groups with respect to target selection and prioritization, molecular probe identification and disease tailoring. The international multidisciplinary team you have assembled is well suited to continue developing automated informatics pipelines for the entire proteome and to bring forward new interfaces and data analytics in an aggressive timeline.

The tools that you are developing are well suited for the analysis and prioritization of dark genes, such as the 395 genes of interest to the Phase 2 of this project.

Sincerely,
Meir Glick

384 Victoria Street
Darlinghurst NSW 2010
Sydney, Australia

T 61 2 9355 5717
E t.groza@garvan.org.au
www.garvan.org.au

GARVAN
INSTITUTE
Breakthrough Medical Research

Dr Tudor Groza
Phenomics Team Leader

17 March 2017

To: Prof. Tudor I Oprea, MD PhD
Translational Informatics Division
Department of Internal Medicine
University of New Mexico Health Sciences Center
Via email to TOprea@salud.unm.edu

Re: "Illuminating the Druggable Genome Knowledge Management Center" U24 Application

I've been monitoring with interest the resources and data types integrated in the Pharos portal, as part of the IDG KMC (Illuminating the Druggable Genome Knowledge Management Center). Your protein-focused view of biomedical knowledge articulates data from a variety of sources, such as experimental data, phenotypes, diseases, funding and patents, to name a few. Your Target Development Level classification is particularly intriguing, since it draws knowledge not only from publications, but also drug labels, chemogenomic resources and OMIM. These data, available via Pharos and its API, can be interfaced with decision-support methods, and could be used for automated concept recognition.

I am the Phenomics Team Leader at the Garvan Institute, Chief Technology Officer of Garvan's spin-off Genome.One and Principle Investigator on the Monarch Initiative. The Garvan Institute of Medical Research is a leading biomedical and genomics research institute, which focuses on understanding the role of genes and molecular and cellular processes in health and disease, as the basis for developing future preventions, treatments and cures. As part of this ecosystem, my team develops algorithms and platforms covering various aspects of the phenotype-genotype analytics stack – including pharmaco-genomics, where Pharos is of a critical relevance.

In several aspects, the IDG KMC knowledge base is complementary to currently on-going efforts within the Monarch Initiative, particularly with respect to genomic, trascriptomic, and therapeutic information. Among other aspects that inform the "druggable" genome we believe that our efforts with human phenotype ontology are likely to be of benefit to the IDG KMC.

Therefore, I wish to convey my support with respect to your competitive renewal proposal (RFA-RM-16-024) for continuation of the IDG KMC, as this project enters Phase 2. Your automated pipelines and human curation are likely to be of benefit to the scientific community at large.

Sincerely,
Tudor Groza

Hi-+-S
Harvard Program in Therapeutic Science

200 Longwood Avenue
Warren Alpert 440
Boston, MA 02115-5730

Peter K. Sorger, Ph.D.
Otto Krayer Professor of Systems Pharmacology
Head of the Harvard Program in Therapeutic Science

phone: (617) 432-6901
fax: (617) 432-6990

peter_sorger@hms.harvard.edu
Administrative Coordinator: Christopher Bird
christopher_bird@hms.harvard.edu

Prof. Tudor I Oprea, MD PhD
Translational Informatics Division
Department of Internal Medicine
University of New Mexico Health Sciences Center
Via email to TOprea@salud.unm.edu

March 20, 2017

Re: “Illuminating the Druggable Genome Knowledge Management Center” U24 Application

Dear Tudor,

I am writing to express enthusiastic support for your proposal to continue the Knowledge Management Center (KMC) for Illuminating the Druggable Genome project as it enters Phase 2 (RFA-RM-16-024). My group has found that the curation you have done for DrugCentral and TargetCentral provides useful and accurate information on the drugs (and their mode of action) that we are actively analyzing in our NIH LINCS Center (http://lincs.hms.harvard.edu/).

As you know, Gary Johnson (who runs a Phase 1 IDG Center) and I are collaborating on a data generation Center submission to IDG Phase 2 focusing on dark kinases. Our computational and informatics plan for this Center (Aim 1 in the proposal we shared with you) envisions very close coordination with you on the creation of a “dark kinase” information resource that combines data from IDG and LINCS. In preparing our proposal we made extensive use of data you have gathered together in Pharos, including the many trascriptomic, phenotypic and genomic data sources that you have combined into an integrated information resource. Our plan (if we are funded) is to work with you to build APIs and metadata standards that allow rapid integration of data across multiple NIH projects as a means to rapidly accumulate and disseminate information about the druggable genome. We have found relevant data in TCGA, LINCS, the SGC and other large-scale efforts.

I firmly believe that your proposal has the potential to shift the paradigm for target selection and prioritization toward understudied proteins, particularly the subset that is associated with human disease. Your team is extremely well suited to continued development of automated pipelines for integration of -omic data and effective prioritization of dark genes. I know that our work in this area will be greatly aided by the expertise of the impressive team you have assembled for the Phase 2 IDG KMC. For example, my LINCS team, working in collaboration with the Johnson group, recently deposited preliminary cheminformatics analysis on the dark kinome at https://github.com/sorgerlab/DarkKinome, and I believe such data could ideally be distributed as part of TargetCentral (TCRD) via the IDG KMC portal, Pharos.

I also believe that effective reuse of ideas and resources developed in LINCS for IDG Phase 2 (and vice-versa) is essential, particularly with respect to knowledge management. Funds for such activities are scarce, and we must make optimal use of the progress we have made (with support from the IDG Program Scientists appointed by NIH). Since both IDG KMC and LINCS are open-access, open science initiatives, this type of coordination would clearly benefit the scientific community at large.

I look forward to continuing to work with your team in IDG Phase 2.

Sincerely yours

Peter Sorger

DIPARTIMENTO DI FARMACIA E BIOTECNOLOGIE

To: Prof. Tudor I Oprea, MD PhD
Translational Informatics Division
Department of Internal Medicine
University of New Mexico Health Sciences Center
Via email to TOprea@salud.unm.edu

Re: "Illuminating the Druggable Genome Knowledge Management Center" U24 Application

Dear Tudor,

Following your poster at the EuroQSAR conference in Verona, Italy, in September 2016, I've become interested in the information that the team you lead has integrated on behalf of the IDG KMC (Illuminating the Druggable Genome Knowledge Management Center). The protein view in Pharos aggregates a diverse set of transcriptomic, proteomic, therapeutic and phenotypic elements, in addition to text mining data from diseases, patents and NIH funding. The TDL concept combines orthogonal sources of information from drug labels, ChEMBL and PubMed, in addition to OMIM and GO terms, which I believe will be very useful for drug discovery. As a Professor in Medicinal Chemistry at the University of Bologna and Research Director at the Italian Institute of Technology, I often look for innovative tools and databases for both my classes and my research activities. An initiative, such as Pharos may represent a great asset for my daily activities, and I will indeed be proposing it to students, postdocs, and collaborators.

In several aspects, the IDG KMC knowledge base a useful resource for searching proteins associated with tauopathies such as Alzheimer's disease, for which I have a long-term interest. Actually, while the amyloid hypothesis has been dominant over the last 15 years, the recent failure of the Merck's BACE inhibitor, verubecestat, raises important question around this hypothesis bringing the tau mechanism to forefront of the scene. Identifying novel targets within this hypothesis represents the first steps for a new drug discovery project. In this respect, your "TIN-X" resource appears to provide a useful, interactive tool for evaluating novel proteins, and their association with this disease based on PubMed abstracts.

Hereby, I wish to express my enthusiasm for supporting your IDG Phase 2 proposal (RFA-RM-16-024) for continuation of the IDG KMC. In conclusion, I personally believe this initiative may represent a very important asset for academic drug discovery, for classes in pharmacy, medicinal chemistry, and related fields, and for basic research in chemical biology, target identification and validation. I'm therefore very glad to support this initiative now and in the future.

Sincerely,

Andrea Cavalli, PhD
Professor of Medicinal Chemistry
Research Director

Washington University in St. Louis

SCHOOL OF MEDICINE

Department of Biochemistry
and Molecular Biophysics

March 1, 2017

To whom it may concern:

This letter supports the development of Pharos as a useful tool for those interested in drug discovery and development. By compiling the correlations between drug targets and compound activity, unknown interactions can be revealed. The breadth of literature is such that a database such as Pharos is essential for efficient review of interactions.

Our industrial colleagues have access to database tools developed in house that provide them competitive advantages over academic research. Pharos will help level the playing field and deserves NIH support.

Sincerely.

Garland R. Marshall

Garland R. Marshall
Professor

EMBL-EBI

Helen Parkinson
EMBL-EBI
Wellcome Genome Campus
Hinxton, Cambridgeshire
CB10 1SD, UK

20 March 2017

To: **Prof. Tudor I Oprea, MD PhD**
Translational Informatics Division
Department of Internal Medicine
University of New Mexico Health Sciences Center
Via email to TOprea@salud.unm.edu

Re: Competitive NIH renewal for the "Illuminating the Druggable Genome Knowledge Management Center" U24 Grant Application RFA-RM-16-024.

Dear Tudor,

It is my pleasure to write in support of your IDG renewal. Since 2014, when the IDG KMC started, our group has been following your progress with interest. The amounts of data aggregated and integrated by IDG KMC, made available via the Pharos portal, are unique in the manner of data type presentation and visual summary. By highlighting vignettes related to grant funding, patents and publications, disease, mouse phenotype, functional and structural information, Pharos summarizes a highly diverse set of data types and we have used it as a user experience paradigm for complex data integration.
As you know, our group maintains the GWAS Catalog (www.ebi.ac.uk/gwas) which provides curated GWAS SNP-trait relationships for genome wide association studies. We have recently discussed your proposal to visualize GWAS data via GWAX, the GWAS eXplorer, which allows users to evaluate specificity on a per-gene and per trait basis. This method is complementary to our current GWAS Catalog visualization methods and in line with the plans we have in our project renewal. We will therefore be delighted to work with you to determine the utility of your visualisation tools for our users, how we could use these with complete summary statistics and to work with you as a user of the GWAS Catalog in determining our new API's utility for your project and users.
We have also worked in collaboration on the IMPC project in delivering IDG specific pages, in evaluating analysis methods and would like also to continue this fruitful collaboration. I am pleased to offer my support to your proposal and look forward to a continued and productive collaboration,

Yours faithfully,

Helen Parkinson

Head of Molecular Archives

Eli Lilly and Company

Lilly Corporate Center

Indianapolis, Indiana 46285

U.S.A.

March 22, 2017

Requested by:

Tudor I. Oprea, M.D., Ph.D.
Professor, Chief, Translational Informatics Division
Department of Internal Medicine
The University of New Mexico
Albuquerque, NM 87106

To whom it may concern,

I am providing this Letter of Support for the project titled, " Illuminating the Druggable Genome Knowledge Management Center" proposed by prof. Tudor Oprea from The University of New Mexico. From my understanding the purpose of this research is to prioritize disease relevant targetable proteins. The work described in the proposal is important to initiate the development of chemical probes, which are a critical target validation tool in the pharmaceutical industry.

In several aspects, the IDG KMC has the potential to become a unique knowledge repository for target selection in drug discovery. For example, the target development level is a provides information for scientists interested in understanding the target selection process, as it combines multiple data elements from literature, chemogenomic and disease databases, and drug labels.

I have served as the chair of the Kinase Panel of the Expert Target Panel committee for the Illuminating the Druggable Genome Knowledge Management Center, IDG KMC in 2015 and 2016, and I along with other Industrial and Academic Scientific Experts provided general scientific advice on the committee.

As described, this proposed research IDG Phase 2 proposal (RFA-RM-16-024) will continue to contribute to the wider scientific community.

With warmest regards,

Michal Vieth, PhD,
Senior Research Advisor
Global Computational Chemistry

Mail code LIB
3181 SW Sam Jackson
Park Rd
Portland, OR 97239-3098
Phone: 503-407-5970
Fax: 503-494-3227
www.ohsu.edu/library

Melissa Haendel, Ph.D.
Associate Professor
OHSU Library and Dept. of Medical Informatics and Clinical Epidemiology
haendel@ohsu.edu

March 23, 2017

Prof. Tudor I Oprea, MD PhD
Translational Informatics Division
Department of Internal Medicine
University of New Mexico Health Sciences Center

RE: "Illuminating the Druggable Genome Knowledge Management Center" U24 Application."

Dear Tudor,

As you know, I co-lead the NIH Monarch Initiative project, (https://monarchinitiative.org/), which focuses on the ontology-based integration of cross-species gene, genotype, variant, disease, and phenotype data and the development of the Human Phenotype Ontology (HPO). This work aims to allow scientists and clinicians to facilitate the understanding of diseases by providing means to compare phenotypes and syndromes and thereby enable the improved interpretation of genetic alterations that may be causal for human disease. We have used our curated knowledge base to identify human disease genes and are facilitating the diagnosis of patients that have evaded correct diagnosis for decades in collaboration with the NIH Undiagnosed Disease Program (UDP). We have recently spent considerable effort creating openly available APIs over our large-scale integrated data to ensure that the data is usable and actionable for the scientific community.

In this context, I would be the first to acknowledge that managing data, information and knowledge in a digital and human friendly manner is no trivial task. The way your Knowledge Management Center on behalf of the Illuminating the Druggable Genome NIH Common Fund project has managed to develop a protein-centric repository is remarkable. On behalf of the Monarch Initiative, I want to express my **enthusiastic support for the continuation of the IDG KMC, and to support your proposal for competitive renewal as IDG begins Phase 2 (RFA-RM-16-024).**

Target prioritization for drug discovery is an essential activity in both academic and industrial research, and IDG KMC is uniquely poised to streamline this effort with seamless data aggregation, knowledge integration and articulation. The automated informatics pipelines and human curation efforts that Pharos makes available to the scientific community enables novel avenues of research; your efforts to mine the NIH R01 awards are poised to foster communication and collaboration, and bring forward new methods for target discovery.

The IDG has already been able to utilize Monarch resources to extract relevant gene-disease associations and integrate them with the rich target classification data you have developed for the IDG project. As the IDG program moves into the implementation phase, we are happy to continue to collaborate with you to further incorporate Monarch resources and tools into the IDG. Specifically, our algorithmic work on integrating nosologies and their associations with phenotypes, alleles, and proteins, and other ontology-based integration approaches are synergistic with your proposed design and implementation of the TargetCentral Knowledgebase you are developing.

We will also explore how we can best leverage IDG resources to advance the integration of IDG target prioritization and classification with the Monarch model system phenotype associations, as well as in the context of our role in the NCATS Data Translator project that aims to create a mechanistic classification of disease involving drug targets.

I wish you the best of luck with your application and am looking forward to our collaboration with you and to contribute to the success of the IDG program.

Sincerely,
Dr. Melissa Haendel

Associate Professor, OHSU library and Dept. of Medical Informatics and Epidemiology
Oregon Health & Science University

SCHOOL OF INFORMATICS
AND COMPUTING

INDIANA UNIVERSITY
Bloomington

March 22, 2017

Professor Tudor I Oprea, MD PhD

Translational Informatics Division

Department of Internal Medicine

University of New Mexico Health Sciences Center

Dear Professor Oprea,

Re: “Illuminating the Druggable Genome Knowledge Management Center” U24 Application

Gladly I provide this letter of strong support for your proposal for continuation of the Knowledge Management Center of the Illuminating the Druggable Genome project, as it enters Phase 2 (RFA-RM-16-024). Your curation efforts from DrugCentral and TargetCentral are critical in that they provide accurate information regarding drugs and their mode of action. This is just one aspect that informs the “druggable” genome in Pharos, as one of the many trascriptomic, phenotypic and genomic information sources that you combined in that online resource.

I am Associate Professor of Informatics at Indiana University and Director of the Web Science Lab, also Associate Director of the Data Science Online Program. My research interests include Semantic Web Technologies and Data-driven Knowledge Discovery, particularly to biomedical networks. Your plan to apply meta-path knowledge graph analytics to the target druggability domain is of great interest and potential. As you know, at IU we have developed Chem2Bio2RDF, and SLAP, and recently completed and published a study involving meta-path analysis with the PubChem team (Fu et al., BMC Bioinfo, 2016).

This proposal has great promise for illuminating the druggability of understudied targets by collecting relevant high value data and applying analysis methods including meta-path based analytics. Your team is well qualified and equipped to continue developing automated informatics pipelines for the entire proteome with new interfaces and data analytics. I look forward with great interest to employing your resources in my work, and expect and wish for great success for you, your colleagues and project.

Yours sincerely

Ying Ding, Ph.D.

Associate Professor of Informatics
Associate Director of Data Science Online Program
Director of Web Science Lab
Faculty of Center for Complex Networks and Systems Research
Faculty of Chemical Informatics Center
Core Faculty of Cognitive Science
School of Informatics and Computing
Indiana University

Stony Brook University

Department of Applied Mathematics and Statistics

March 25, 2017

To: Prof. Tudor I Oprea, MD PhD
Translational Informatics Division
Department of Internal Medicine
University of New Mexico Health Sciences Center
Via email to TOprea@salud.unm.edu

Dear Tudor:

I am pleased to write and confirm my enthusiastic support for the project you describe in your application to the NIH, entitled "Illuminating the Druggable Genome Knowledge Management Center".

For many years, we have shared an interest in the interface between the diversity of the largely unexplored therapeutically-relevant chemical space and the correlation between therapeutic activity and molecular diversity. As you know, in our previous collaboration we produced a new molecular structure fingerprint, the topological index, which uniquely characterizes the ring structure of a multicyclic system of up to 8 rings, offering a means to rapidly compare highly complex structures.

Building on this experience, I am now very excited to collaborate with you on the new ideas in your proposal that would develop new information theory tools to optimize collation of diverse datasets into your TargetCentral Knowledgebase. I am happy to offer assistance in applying ideas from **Information Geometry** to develop inference based models of mutual information for combining these databases and deriving a meaningful distance measure in the combined space.

I am eager to collaborate with you and your group, and advise you on the mathematics. Our collaboration is facilitated through regular video exchanges, to discuss progress and strategies to advance model design.

I am very much looking forward to our cross-disciplinary interactions that have already been very fruitful for both our endeavors.

Best wishes,

E. A. Coutsias

Evangelos A. Coutsias
Professor,
Applied Mathematics and Statistics

Prof. Tudor I Oprea, MD PhD
Translational Informatics Division
Department of Internal Medicine
University of New Mexico Health Sciences Center
TOprea@salud.unm.edu

Dear Tudor

We are writing with enthusiastic support for your NIH Common Fund application to continue your activities with Illuminating the Druggable Genome (IDG) Knowledge Management Center (KMC). As Principle Investigators with the International Mouse Phenotyping Consortium (IMPC), we have benefited from the tools and informatics services that the IDG-KMC provides. The IMPC is a G7-recognised global research infrastructure whose goal is to build a catalogue of gene function by generating and characterising a knockout mouse strain for every protein-coding gene. The IMPC is funded in large part by the $100 Million KOMP2 NIH Common fund including funding the entirety of the IMPC data coordination center based at MRC Harwell and EMBL-EBI, as well mouse production and phenotyping in the USA, Canada and UK.

A critical part of the IMPC informatics infrastructure is to coordinate and help prioiritse production of mouse strains across a dozen global production centers. Our first interactions with the IDG-KMC were several years ago when your group started providing lists of “druggable genes” that had little known functional information. The KOMP centers eagerly incorporated these into the mouse knockout production plans and am happy to report that over 130 knockout strains have been produced as a result (summary here: https://www.mousephenotype.org/data/secondaryproject/idg). We extended our collaboration by inviting you to our annual KOMP Meeting in 2015 where we saw a prototype of the Pharos portal and the powerful Target Development Levels classification that provides a summary of all known knowledge for all genes. We had been looking for a way to demonstrate how the IMPC is providing new knowledge and tools for poorly understood genes and are now incorporating these classifications into our infrastructure and to a “gene knowledge summary” page we will be providing on the IMPC portal that links back to Pharos. In turn, we were very excited to work with you to facilitate incorporating IMPC data into the IDG-KMC where it is now available in context with the other rich sources of information.

As the IMPC moves into the second phase, we are increasing our efforts to disseminate the knowledge generated by our colleagues as widely as possible. The IDG-KMC is an important part of these outreach efforts to the pharmaceutical research community. Therefore, we hope to see the NIH support your competitive KMC renewal as the IDG enters a second phase.

Sincerely,

SDM Brown

Steve Brown
Chair, IMPC Steering Committee
Director, MRC Harwell Institute, Harwell, UK

Terrence Meehan

Terrence Meehan
PI- KOMP2 MPI2 grant
Mouse Informatics Coordinator
European Molecular Biology Laboratory-
European Bioinformatics Institute

Bristol-Myers Squibb

P.O. Box 4000 Princeton, NJ 08543-4000 609-252-4000
www.bms.com

March 26, 2017

Prof. Tudor I Oprea, MD PhD
Translational Informatics Division
Department of Internal Medicine
University of New Mexico Health Sciences Center
Via email to TOprea@salud.unm.edu

Dear Tudor,

I am writing to express my enthusiastic support for the continuation of the Knowledge Management Center of the Illuminating the Druggable Genome project, as it enters Phase 2 (RFA-RM-16-024). The rich target and mode-of-action information that you have curated from DrugCentral and TargetCentral are just one aspect of the important information that makes up Pharos. The many trascriptomic, phenotypic and genomic information sources that you combined in that online resource provide an exceptional resource for searching and understanding the “druggable” genome.

My role at Bristol-Myers Squibb includes the leadership of two computationally focused groups. The first is the Molecular Analytics group, a small team dedicated to cheminformatics research including target identification efforts, screening collection design, and data mining and machine learning. My other key responsibility includes the leadership of computational chemistry activities related to Immunology and Oncology drug discovery efforts at BMS. As my research combines structural, functional, and chemical information to drive the identification of potential therapeutics, a resource like Pharos holds great potential for significantly impacting the efforts of my team.

I strongly believe that your proposal has the potential to become a key resource for both academic and industrial research groups with respect to target selection and molecular probe identification. The international multidisciplinary team you have assembled is well suited to continue developing automated informatics pipelines for the entire proteome and to bring forward new interfaces and data analytics in an aggressive timeline. I look forward to me and my team making effective use of these tools, and the data they provide, in our research efforts to bring new medicines to help the lives of patients.

Regards,

Stephen R. Johnson

Stephen R. Johnson, Ph.D.
Sr. Principal Scientist
Bristol-Myers Squibb

Scott Martin
Senior Scientific Manager
Discovery Oncology
Genentech, Inc.
Phone 650 225 6724
Fax 650 742 5179
Email martin.scott@gene.com

Genentech
A Member of the Roche Group

March 27, 2017

Dear Review Committee:

I am writing this letter in support of continued development of the Pharos user interface to explore existing information on druggable targets in the human genome. This platform is very relevant in our current age of big data and many disconnected data sources. Scientists need centralized tools to research information relating to their targets of interest. This is especially relevant in my own work in the area of functional genomics. My group conducts large-scale screens using CRISPR, RNAi, and small molecule platforms. Accordingly, we are continuously searching for information on screen hits, and relationships between actives. Pharos provides useful tools in this regard, and is worth continued efforts to further develop its utility.

I've used the Pharos interface to explore information on targets from genome-scale screens. I find a number of the features intuitive and useful. For example, not only are standard citations for targets presented with corresponding PMIDs, but NIH grant applications are also linked. These can be extremely insightful, as they are written by experts in respective target areas, and can provide information beyond that found in published manuscripts. Protein-protein interactions are also easy to explore, as are GO term enrichments. Both can be contrasted with other screen hits to look for relationships and develop hypotheses. The listing of known chemical inhibitors is yet another useful resource that can enable follow-up experiments. All of this information can be collected in a very clever tool termed the "Dossier".

Overall, the existing Pharos portal is a promising resource for scientists exploring novel targets. There are a number of improvements that can make the tool of even greater use. I look forward to expanded function in the years to come.

Sincerely,

Scott E. Martin

Scott E. Martin, Ph.D.
Senior Scientific Manager
Group Leader, Functional Genomics
Department of Discovery Oncology
Genentech
martin.scott@gene.com

GLADSTONE
INSTITUTES

STEVEN FINKBEINER, M.D., Ph.D.
Direct Line: (415) 734-2508
Fax: (415) 355-0824
Steve.finkbeiner@gladstone.ucsf.edu
http://gladstoneinstitutes.org/

SCIENCE
OVERCOMING
DISEASE

March 13, 2017

Dear Dr. Anton Simeonov:

We have been using Pharos for the last three years and have watched its development over that time period. We began using Pharos, to obtain information on orphan genes that we are working on through the IDG program. Pharos is the first site we visit when looking up any orphan gene. The site now also includes data from all genes within the genome, so we have begun to use the site to search for genes for other non-IDG projects.

The platform is visually appealing and user friendly. It is easier to navigate in comparison to its competitors (e.g., Open Targets). The front page has the simplicity of the Google search home page, and it has incorporated features such as the 'Amazon cart' idea to store search information. It also has some slick visualization tools, namely the 'Explore' feature. During its three years in development it has added help and training tutorial videos that also publicize new site features.

In addition to cataloguing information from publically available databases (e.g., GTEX, PubMed) it also incorporates information on the targets investigated by other Tech Dev groups enabling more efficient collaboration across the IDG Consortium. The filter option enables an unbiased screening of the targets. In addition to providing phenotypic filtering options it also provides some unique options including grant activity (e.g., number of RO1s funded on the target).

The next stage would be to use Pharos to provide 'novel' knowledge about the orphan targets, by using bioinformatics interrogation and predictive modeling methods. Based on our experience with the Pharos platform, we find it to be a very useful tool in its current state and one that has great promise for future development. As such, we recommend further investment to help develop Pharos so that it can achieve its full potential.

With kind regards,

Steven Finkbeiner, M.D., Ph.D.
Director, Taube/Koret Center for Neurodegenerative Disease and Gladstone Institutes
Professor, Departments of Neurology and Physiology,
University of California, San Francisco

SMF/ksn

1650 Owens Street
San Francisco, CA 94158
Phone 415.734.2000
www.gladstoneinstitutes.org

Yale University

Susumu Tomita, Ph.D.
Professor

Department of Cellular and Molecular Physiology
Department of Neuroscience
CNNR program
Yale University School of Medicine
295 Congress Avenue BCMM454B/441
P.O. Box 9812
New Haven, Connecticut 06536-0812

Telephone: 203 785-7201
Fax: 203 785-4951
Email: Susumu.Tomita@yale.edu

Dr. Anton Simeonov
Scientific Director
Division of Pre-Clinical Innovation
National Center for Advancing Translational Sciences
National Institutes of Health

March 11, 2017

Dear Anton,

I am pleased to support your Pharos platform. As you know, we are working on ion channel biology and have revealed ion channel constituents, physiology and phenotypes in vivo. While our identification of constituents or functional modulators of ion channels, we have identified new molecules for each ion channels. At that time, we actively use the Pharos website for examining gene functions and characteristics. Furthermore, when we are selecting new targets, we are checking the Pharos to identify the understudied Dark genes. Thus, we would like to see the Pharos for future.

In addition, we would like to keep providing our feedback to the Pharos website. Dr. Rajarshi Guha from your lab visited my laboratory last year, and he provided us instructions, and his instructions were so helpful, and my lab members were very impressed about it. In the meantime, we gave him feedback, and our feedback seems reflected to organization of the Pharos website. We are glad to see it. In the future, as a customer and data generator, we would like to further contribute to the Pharos.

Hope we can continue to work together.

Sincerely,

Susumu Tomita

Susumu Tomita

Susumu Tomita, Ph.D.
Professor of Neuroscience
Professor of Cellular and Molecular Physiology

DEPARTMENT OF HEALTH & HUMAN SERVICES Public Health Service

National Institutes of Health
National Center for Advancing
Translational Sciences
9800 Medical Center Drive, MSC 3370
Bethesda, MD 20892-3370
PH (301) 217-5721

Tuesday, March 14, 2017

Dr. Tudor Oprea, Translational Informatics Division

MSC09-5025; 1 University of New Mexico, Albuquerque, New Mexico 87131, USA

RE: Knowledge Management Center for Illuminating the Druggable Genome (U24)

Dear Prof. Oprea,

I am writing to express my enthusiasm and support for your collaboration with Dr. Rajarshi Guha to develop the next phase of the IDG Knowledge Management Center, comprising the Target Central Resource Databases (TCRD) and Pharos platform. The proposed IDG Portal will be based on the pre-existing Pharos platform, jointly developed by your team and us, which has successfully been serving the IDG community for the past year. The proposed work will implement new features in Pharos that will lead to significant advances in the characterization of knowledge about the dark druggable proteome. The proposed network approaches to defining knowledge domains and the use of Bayesian networks to capture dependencies between targets, diseases, ligands and other entities will provide users with powerful tools to search for information as well as obtain recommendations and pointers to other relevant and related data. Together with improvements in the user interface, including novel visualizations and integration with external tools, Pharos will play a key role in serving the IDG community as a centralized hub for data and methods.

The mission of NCATS is to catalyze the generation of innovative methods and technologies that will enhance the development, testing, and implementation of diagnostics and therapeutics across a wide range of human diseases and conditions; this includes the development of tools and platforms that generate, collate, present and disseminate information about targets that can be used by the scientific community to advance research on these targets for therapeutic purposes. As such, one of our focuses is industrial-scale small molecule assay development, screening, informatics, and medicinal chemistry to discover and develop chemical probes for use in the study of gene, protein, and cell functions, and to identify new targets and drugs for medically underserved diseases. From its inception, NCATS (formerly the NIH Chemical Genomics Center, NCGC) has also focused on technology and paradigm development, resulting in innovations such as our quantitative high throughput screening (qHTS) and a wide range of informatics platforms. The resources of NCATS include new state-of-the-art assay development, robotic screening, informatics, synthetic and analytical chemistry laboratories, and drug ADME study capabilities. NCATS' infrastructure will be available to you to support this important research. You will have my full organizational support to use the appropriate tools and technologies to successfully fulfill all the aims as outlined in your grant application. For our work on this project, we will not be requesting support for federal employee salaries, per NIH policy, as we are part of the NIH Intramural Research Program.

We look forward to further developing Pharos as the platform underlying the IDG KMC and are very interested in this exciting joint research.

With very best regards,

Anton Simeonov, Ph.D.
Scientific Director, National Center for Advancing Translational Sciences, NIH

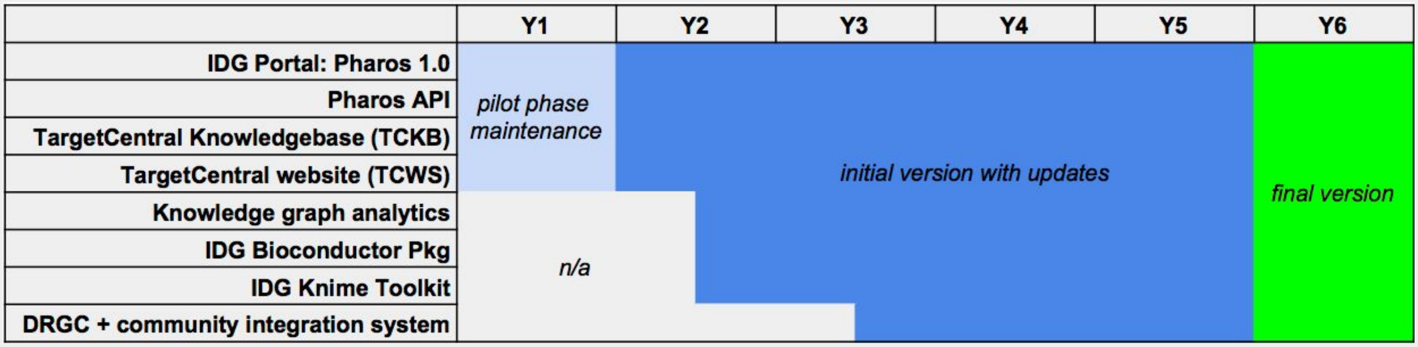
Resource Sharing Plan

Scientific progress and public health benefit from the free, open and effective communication of research data and knowledge. This basic truth is formalized in the mission and goals of NIH, and of research universities and organizations worldwide. We are fully committed to this mission both in principle and specifically as required by NIH resource sharing instructions including modifications specified for this RFA. We recognize that effective resource sharing means going well beyond mere data and software downloads, and have previously in the IDG pilot phase and other projects implemented effective resource sharing practices. Details of this resource sharing plan follow.

- **All NIH-approved resource sharing policies developed by the IDG SC** we will abide by.
- **Licensing provenance metadata** will be managed in a consistent regime to facilitate maximal sharing of pass-through data in compliance with source licensing.
- **The NIH Genomic Data Sharing Policy** we will abide by as applicable.
- **Industry standard code sharing systems** (e.g. Github, Bitbucket, Docker) will be employed for sharing of originated data, metadata, installers and computational protocols.
- **Reagents, tools or protocols** generated by the IDG Consortium are a particular focus, whether from KMC, DRGCs, or RDOC. Such resources will be promoted, in collaboration with IDG RDOC, through outreach and devoted pages of appropriate IDG websites, as with Pilot Phase http://targetcentral.ws TechDev pages.
- **Data sharing** is a central and ongoing goal of IDG, and KMC has many specific and general plans and deliverables in this effort, solely and in collaboration with RDOC and DRGCs, detailed in the research plan. The portal Pharos is the primary instrument of data sharing, via GUI, via downloaded subsets, and via the Pharos API.
- **Software sharing and dissemination** is of particular interest to our team. We have extensive experience with open-source community projects as users and developers (with contributions to R [10.18637/jss.v018.i05], the CDK [PMID: 16796559] and the Blue Obelisk [PMID: 21999342]). We are committed to a robust software development methodology, employing modern programming languages, accepted software design standards, profession software engineering practice and wise use of 3rd party libraries. This approach will facilitate reuse, extension, and independent continuation of our efforts. Specific instance of effective software sharing by team members include the IDG Pilot Phase repositories for: Pharos code[1], Pharos installer[2], TCRD build code[3], Drug Target Ontology (DTO)[4], and the TIN-X code[5].
- **Computational tools, protocols, and methods** and sharing approaches vary widely, from publication (more or less replicable) to source code (more or less comprehensible), to APIs (more or less reliable). Accordingly, our plan mobilizes specific industry standard and community driven platforms designed and proven to facilitate scientific dissemination. Three such projects are: **R/Bioconductor, and Knime**, all free and open source. The breadth, vitality and value of R is quite remarkable, with 1000s of developers and scientists contributing, benefiting and sharing code, methods and ideas. Bioconductor is a bioinformatics community project built upon R. Knime is a rapidly growing workflow development, deployment and sharing system. These platforms effectively provide an environment for independent development and extensibility.

[1] https://spotlite.nih.gov/ncats/pharos
[2] https://hub.docker.com/r/ncats/pharos
[3] https://github.com/stevemathias/TCRD
[4] https://github.com/schurerlab/DTO
[5] https://bitbucket.org/dccannon/tin-x

- **Improvements or customizations by others** to IDG analysis codes are facilitated and encouraged by the adoption of community based development platforms such as Git, R, Bioconductor and Knime, as an intrinsic aspect of these approaches and this plan. However, KMC remains responsible for producing and updating official versions of core components, including example code designed to promote community development.
- **Timelines and timeliness** are essential, since delays in sharing have negative impact to science and public health. The timeline chart provided indicates specific measurable and meaningful commitments, but key aspects of our plan go further. **Software and analytical pipelines sharing** via participation in online community platforms such as R and Knime are continual efforts.
- **Communication and coordination within the IDG Consortium** is essential for external effective resource sharing to the broader scientific community. This KMC plan explicitly involves coordination with RDOC, the DRGCs, IDG SC and NIH staff. The Pilot Phase TargetCentral.ws website is illustrative, by organizing and publicizing IDG-wide technologies, resources, progress, news and events, including TechDev-supplied metadata. Our plan is to expand on this functionality, with use-case driven data packages, and further integrated DRGC, consortium-wide, and community input.
- **Complete data and tool sets for the studied proteins** is a specific focus in our plan. Data about the IDG initial focus 395 Kinase, IC and GPCR proteins will be featured and available as "knowledge packages" designed to address the needs of domain scientists in target prioritization and selection.



	Y1	Y2	Y3	Y4	Y5	Y6
IDG Portal: Pharos 1.0	*pilot phase maintenance*	*initial version with updates*				*final version*
Pharos API						
TargetCentral Knowledgebase (TCKB)						
TargetCentral website (TCWS)						
Knowledge graph analytics	*n/a*					
IDG Bioconductor Pkg						
IDG Knime Toolkit						
DRGC + community integration system						

Table 1: Software dissemination timeline